

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/28-1-1480>

Lexikos 28

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/28-1-1480>

Lexikos 28

Redakteur

Editor

Elsabé Taljard

Resensieredakteur

Review Editor

T. Harteveld



African Association for Lexicography

AFRILEX-REEKS 28:2018

AFRILEX SERIES 28:2018



BURO VAN DIE WAT

STELLENBOSCH

Uitgewer Publisher

BURO VAN DIE WAT
Posbus 245
7599 STELLENBOSCH

Kopiereg © 2018 deur die uitgewer
Alle regte streng voorbehou
Eerste uitgawe 2018

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

ISBN 978-0-9946528-4-3
ISSN 2224-0039

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

Licensed under Creative Commons License CC BY 4.0.

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

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

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

Lexikos is elektronies beskikbaar by Sabinet, AJOL, Ebsco en Proquest
Lexikos is available online from Sabinet, AJOL, Ebsco and Proquest

Indekse Indexes

Arts and Humanities Citation Index®, Current Contents®/Arts & Humanities, Current Contents®/Social and Behavioral Sciences, ERIH Plus, 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

Span van Roterende Redakteurs / Team of Rotating Editors

Dr. H.S. Ndinga-Koumba-Binza (RSA en Gaboen/RSA and Gabon)
Prof. D.J. Prinsloo (RSA)
Prof. Elsabé Taljard (RSA)

Adviesraad / Advisory Board

Prof. A. Adamska-Salaciak (Pole/Poland)
Prof. H. Béjoint (Frankryk/France)
Prof. H. Chimhundu (Zimbabwe)
Prof. F. Dolezal (VSA/USA)
Prof. R.H. Gouws (RSA)
Prof. R.R.K. Hartmann (Groot-Brittanje/Great Britain)
Prof. W. Martin (België en Nederland/Belgium and The Netherlands)
Prof. I.A. Mel'čuk (Kanada/Canada)
Prof. A.M.F.J. Moerdijk (Nederland/The Netherlands)
Dr. J. Tent (Australië/Australia)
Prof. J. Van Keymeulen (België/Belgium)
Prof. P.G.J. van Sterkenburg (Nederland/The Netherlands)
Prof. L.S. Vikør (Noorweë/Norway)
†Prof. H.E. Wiegand (Duitsland/Germany)

Redaksiekomitee / Editorial Committee

Dr. M.M. Bagwasi (Botswana)
Prof. H.L. Beyer (Namibië/Namibia)
Prof. W.A.M. Carstens (RSA)
Prof. E. Chabata (Zimbabwe)
Prof. C.J. Conradie (RSA)
Prof. A.E. Feinauer (RSA)
Prof. R. Finlayson (RSA)
Dr. S. Hadebe (Zimbabwe)
Prof. I.M. Kosch (RSA)
Dr. P.A. Louw (RSA)
Mnr./Mr K.J. Mashamaite (RSA)
Prof. P.A. Mavoungou (Gaboen/Gabon)
Dr. V.M. Mojela (RSA)
Mnr./Mr M.C. Mphahlele (RSA)
Prof. D. Nkomo (RSA en Zimbabwe/RSA and Zimbabwe)
Prof. T.J. Otlogetswe (Botswana)
Prof. A.N. Otto (RSA)
Prof. P.H. Swanepoel (RSA)

Inhoud / Contents

Voorwoord	x
Foreword	xi
<i>Elsabé Taljard</i>	
'n Woord van AFRILEX	xii
A Few Words from AFRILEX	xiii
<i>Herman L. Beyer</i>	
Redaksionele doelstellings	xiv
Editorial Objectives	xv

Artikels / Articles

On Recent Proposals to Abolish Polysemy and Homonymy in Lexicography	1
<i>Herman L. Beyer</i>	
Corpus-driven Bantu Lexicography. Part 1: Organic Corpus Building for Lusoga	32
<i>Gilles-Maurice de Schryver and Minah Nabirye</i>	
Corpus-driven Bantu Lexicography. Part 2: Lemmatisation and Rulers for Lusoga	79
<i>Gilles-Maurice de Schryver and Minah Nabirye</i>	
Corpus-driven Bantu Lexicography. Part 3: Mapping Meaning onto Use in Lusoga	112
<i>Gilles-Maurice de Schryver and Minah Nabirye</i>	
New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the Diccionarios Valladolid-UVa	152
<i>Pedro A. Fuertes-Olivera, Sven Tarp and Peter Sepstrup</i>	

'n Leksikografiese datatrekkingstruktuur vir aanlyn woordeboeke <i>Rufus H. Gouws</i>	177
A Lexicographic Approach to Teaching the English Article System: Help or Hindrance? <i>Sugene Kim</i>	196
An Empirical Study of EFL Learners' Dictionary Use in Chinese– English Translation <i>Pengcheng Liang and Dan Xu</i>	221
Once Again Why Lexicography Is Science <i>Tinatin Margalitadze</i>	245
The Effectiveness of Using Dictionaries as an Aid for Teaching Standardization of English-based Sports Terms in Serbian <i>Mira Milić, Tatjana Glušac and Aleksandra Kardoš</i>	262
Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources <i>Carolin Müller-Spitzer, María José Domínguez Vázquez, Martina Nied Curcio, Idalete Maria Silva Dias and Sascha Wolfer</i>	287
Polish Americans in the History of Bilingual Lexicography: The State of the Art <i>Mirosława Podhajecka</i>	316
Semi-automating the Reading Programme for a Historical Dictionary Project <i>Tim van Niekerk, Johannes Schäfer and Ulrich Heid</i>	343
Objectivity, Prescription, Harmlessness, and Drudgery: Reflections of Lexicographers in Slovenia <i>Alenka Vrbinc, Donna M.T.Cr. Farina and Marjeta Vrbinc</i>	361
Towards Chinese Learner's Dictionaries for Foreigners Living in China: Some Problems Related to Lemma Selection <i>Mei Xue and Sven Tarp</i>	384

- Enhancing the Learnability of Chinese–English Dictionaries for Chinese as a Foreign Language Learners: The Neglected Legacy of Robert Morrison in His Compilation of *Wuche Yunfu* (1819)
Ying Ye, Xiangqing Wei and Wenlong Sun 405

Projekte / Projects

- Corpus-Based Research on Terminology of Turkish Lexicography (CBRT-TURKLEX)
Erdoğan Boz, Ferdi Bozkurt and Fatih Dođru 428
- Web-based Exploration of Results From a Large European Survey on Dictionary Use and Culture: ESDexplorer
Sascha Wolfer, Iztok Kosem, Robert Lew, Carolin Müller-Spitzer and Maria Ribeiro Silveira 440

Leksiko-opname / Lexicosurvey

- Which Learning Tools Accompanying the Paid Online Version of LDOCE Do Advanced Learners of English Find Useful?
Bartosz Ptasznik and Robert Lew 448

Leksikohuldeblyk / Lexicotribute

- Herbert Ernst Wiegand (08 Januarie 1936–03 Januarie 2018) 461
Herbert Ernst Wiegand (08 January 1936–03 January 2018) 463
Rufus H. Gouws

Resensieartikel / Review Article

- Das Rumäniendeutsche in der Neuauflage (2016) des *Variantenwörterbuchs des Deutschen*. Ioan Lăzărescu zum 65. Geburtstag gewidmet
Doris Sava 465

Resensies / Reviews

- Pedro A. Fuertes-Olivera. *The Routledge Handbook of Lexicography*
Dai Lingzhen 486

María José Domínguez Vázquez, Fabio Mollica and Martina Nied
Curcio (Eds.). *Zweitsprachige Lexikographie zwischen Translation und
Didaktik* 494
Maria Smit

Publikasieaankondigings / Publication Announcements 504

Voorskrifte aan Skrywers 505
Instructions to Authors 506

Voorwoord

Die 28ste uitgawe van *Lexikos* bevestig weer eens sy status as ware internasionale tydskrif met sy wortels in Afrika. Hierdie jaar se uitgawe bevat bydraes uit België, Serwië, Slowenië, China, Spanje en Pole. Die navorsing waaroor daar gerapporteer word, strek van historiese leksikografie, oor hedendaagse vraagstukke in die leksikografie, tot vooruitskouings oor wat die toekoms vir dié dissipline inhou. Leksikograwe is duidelik bewus van die bedreigings, maar ook van die geleenthede wat die elektroniese media vir die leksikografie inhou. As redakteur is ek egter besorg oor die gebrek aan artikels oor die Afrikataal-leksikografie in Suid-Afrika. Die feit dat ons elf amptelike tale het waarvan tien Afrikatale is, bied 'n rykdom navorsingsgeleenthede aan beide praktiese en teoretiese leksikograwe, en ek wil ons plaaslike leksikograwe aanmoedig om hul kennis en kundigheid met die res van die leksikografiegemeenskap te deel.

In die loop van die jaar het ons met hartseer verneem van die afsterwe van prof. Herbert Ernst Wiegand, een van die reuse in die metaleksikografie. Ons het gedink dat dit gepas is om 'n huldeblyk oor hom te publiseer en ons dank aan prof. Rufus Gouws vir hierdie bydrae.

Die uitgee van *Lexikos* is 'n spanpoging. In dié verband wil ek graag me. Tanja Hartevelt en me. Hermien van der Westhuizen van die WAT bedank vir hulle toewyding om seker te maak dat *Lexikos* aan hulle hoë tegniese standaarde voldoen. Ek wil ook graag vir prof. Danie Prinsloo en dr. Steve Ndinga-Koumba-Binza bedank — ek het groot waardering vir hulle bydrae en ondersteuning. 'n Spesiale woord van dank gaan aan die keurders. Keuring van artikels is 'n ondankbare en dikwels tydrawende taak, maar die toewyding van ons keurders verseker dat die hoë standaard waaraan ons oor die jare heen gewoon geraak het, gehandhaaf word. Laastens, 'n woord van dank aan ons outeurs sonder wie se bydraes ons nie 'n tydskrif sal hê nie. Ek is dankbaar vir die positiewe gees waarin outeurs op keurders se kommentaar reageer. Dit dra alles by tot 'n stimulerende leksikografiese gesprek.

Die redakteurs van *Lexikos* 29 is profs. Danie Prinsloo en Dion Nkomo. Die ervaring van die ou garde en die entoesiasme van die jong bloed sal ongetwyfeld 'n onvergeetlike uitgawe van *Lexikos* tot gevolg hê!

Elsabé Taljard
Redakteur

Foreword

The 28th edition of *Lexikos* once again confirms its status as a true international journal with its roots in Africa. This year's edition contains contributions from Belgium, China, Slovenia, Serbia, Spain and Poland. The research reported on range from historical lexicography, through current issues in lexicography, to predictions on what the future holds for this discipline. Lexicographers are clearly very aware of both the challenges and the opportunities offered by the electronic media. Of some concern to me as editor though, is the dearth of articles dealing specifically with African language lexicography in South Africa. Having eleven official languages, of which ten are African languages, offers a wealth of research opportunities to both practical and theoretical lexicographers, and I would like to encourage our local lexicographers to share their knowledge and expertise with the rest of the lexicographic community.

During the course of the year, we learned with sadness of the passing away of Prof. Herbert Ernst Wiegand, one of the giants in metalexigraphy. We have therefore deemed it fitting to publish a tribute to him, and thank Prof. Rufus Gouws for this contribution.

The publication of *Lexikos* is a team effort. In this regard, I would like to thank Ms Tanja Harteveld and Ms Hermien van der Westhuizen of the WAT for their commitment to make sure that *Lexikos* meets their exacting technical standards. I would also like to extend my gratitude to Prof. Danie Prinsloo and Dr Steve Ndinga-Koumba-Binza, whose input and support I value greatly. A special word of thanks goes to the reviewers. Reviewing articles is a thankless and often time-consuming task, but the commitment of our reviewers ensures that the high standard to which we have gotten used over the years, is maintained. Finally, I would like to thank our authors without whose contributions we would not have a journal. I am grateful for the positive spirit with which authors respond to reviewers' comments. It all contributes to a stimulating lexicographic discourse.

The editors of *Lexikos* 29 will be Profs Danie Prinsloo and Dion Nkomo. The experience of the old guard and the enthusiasm of youth will most certainly result in a memorable edition of *Lexikos*!

Elsabé Taljard
Editor

'n Woord van AFRILEX

Die African Association for Lexicography (AFRILEX) bly dankbaar en trots daarop om 'n internasionaal gevestigde en hoog aangeskrewe Goue-Oop-Toegang-vaktydskrif soos *Lexikos* as sy mondstuk te hê. Sonder hierdie waardevolle bate wat so kundig bestuur word deur die Buro van die WAT as uitgewer, sou die Vereniging veel armer wees. Daarom moet die Hoofredakteur en personeel van die Buro van die WAT geloof word vir hulle toewyding tot die metaleksikografiese gesprek en die uitbou van die wetenskap, bo en behalwe hulle dagtaak as praktiese leksikograwe. Dit is onder andere hierdie omvattende benadering tot die leksikografie wat die Buro 'n onbetwiste leier in Afrika-leksikografie maak.

Die redaksie van hierdie nommer was in die besonder vaardige hande van prof. Elsabé Taljard, 'n jarelange Raadslid van AFRILEX van die Universiteit van Pretoria. Sy is in die Buro van die WAT se kenmerkende tradisie van professionaliteit en leksikografiese noukeurigheid bygestaan deur me. Tanja Harteveld as resensieredakteur, met uitstekende tegniese ondersteuning deur me. Hermien van der Westhuizen.

Dit is my voorreg om namens die Raad en lede van AFRILEX die redaksionele span, die Buro van die WAT en bydraende outeurs van harte te bedank vir nommer 28 van *Lexikos*.

Herman L. Beyer
President: AFRILEX

A Few Words from AFRILEX

The African Association for Lexicography (AFRILEX) remains grateful and proud to have an internationally established and highly regarded Gold Open Access journal like *Lexikos* as its mouthpiece. Without this valuable asset, so expertly managed by the Bureau of the WAT as publisher, the Association would have been much poorer. For this reason, the Editor-in-Chief and staff of the Bureau of the WAT should be praised for their dedication to metalexigraphic discourse and the development of the discipline, above and beyond their core business of practical lexicography. It is, among other things, this comprehensive approach to lexicography that makes the Bureau an undisputed leader in lexicography in Africa.

The editorship of this volume was in the very capable hands of Prof. Elsabé Taljard, a long-standing Board member of AFRILEX from the University of Pretoria. She was assisted in the Bureau of the WAT's fine tradition of professionalism and lexicographic thoroughness by Ms Tanja Harteveld as review editor, with excellent technical support by Ms Hermien van der Westhuizen.

It is my privilege to, on behalf of the Board and members of AFRILEX, sincerely thank the editorial team, the Bureau of the WAT and contributing authors for volume 28 of *Lexikos*.

Herman L. Beyer
President: AFRILEX

Redaksionele doelstellings

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

Die strewe van die AFRILEX-reeks is:

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

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

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

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

Editorial Objectives

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

The objectives of the AFRILEX Series are:

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

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

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

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

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/28-1-1480>

On Recent Proposals to Abolish Polysemy and Homonymy in Lexicography

Herman L. Beyer, *Department of Language and Literature Studies, University of Namibia, Windhoek, Namibia, and Department of Afrikaans and Dutch, Stellenbosch University, Stellenbosch, South Africa (hbeyer@unam.na)*

Abstract: Two articles appeared recently in *Lexikos* that propose the abolishment of homonymy and polysemy in lexicography, particularly in dictionaries with a text reception function only. This contribution identifies two main theoretical premises of the proposal in these articles and challenges them. They are: (i) a theory of the lemma as linguistic sign; and (ii) the results of dictionary criticism. Under examination, it is found that both premises fail to support the proposal with regard to polysemy. With regard to homonymy, the first premise is proven invalid, and the second is found to be valid. This implies that the theoretical basis for the proposal should either be reviewed (for which the lexicographical communication theory is offered), or the proposal should rely on the sole practical and unproven argument of data accessibility. The contribution simultaneously develops a potential broad framework for the lexicographical communication theory. The framework constitutes a lexicographical text grammar, which is presented as a parallel communication code to elements of the lexicographic text theory and linguistic grammars. It is argued that dictionary articles constitute texts in which these two grammars overlap to varying degrees, representing a hybrid form of textual communication.

Keywords: LEXICOGRAPHICAL COMMUNICATION THEORY, GRAMMAR, HOMONYMY, LEXICOGRAPHICAL COMMUNICATION, LEXICOGRAPHICAL GRAMMAR, LINGUISTIC SIGN, LINGUISTICS, POLYSEMY, SEMIOTICS, LEXICOGRAPHICAL TEXT THEORY

Opsomming: Oor onlangse voorstelle vir die wegdoen van polisemie en homonimie in leksikografie. Twee artikels het onlangs in *Lexikos* verskyn wat voorstel dat weggedoen word met homonimie en polisemie in die leksikografie, spesifiek in woordeboeke met slegs 'n teksresepsiefunksie. Hierdie bydrae identifiseer twee teoretiese hoofpremisses vir die voorstel en bevraagteken hulle. Die premisse is: (i) 'n teorie van die lemma as taalteken; en (ii) die resultate van woordeboekkritiek. By nadere ondersoek word bevind dat beide die premisse faal met betrekking tot polisemie. Met betrekking tot homonimie word die eerste premis as ongeldig bewys, en die tweede een word geldig bevind. Die bevindinge hou in dat die teoretiese basis vir die voorstel óf hersien moet word (waarvoor die teorie van leksikografiese kommunikasie aangebied word), óf op die enkele praktiese en onbewese argument van datatoeganklikheid moet steun. Terselfdertyd ontwikkel die bydrae 'n potensiele breë raamwerk vir die teorie van leksikografiese kommunikasie. Die raamwerk verteenwoordig 'n leksikografiese teksgrammatika, wat as 'n kommunikasiekode parallel tot elemente van die teorie van leksikografiese tekste en taalkundige grammatikas aangebied word. Daar word aangevoer dat woordeboekartikels uit tekste bestaan waarin hierdie twee

grammatikas in wisselende mates oorvleuel en as sodanig 'n hibriediese vorm van tekstuele kommunikasie verteenwoordig.

Sleutelwoorde: GRAMMATIKA, HOMONIMIE, LEKSIKOGRAFIESE GRAMMATIKA, LEKSIKOGRAFIESE KOMMUNIKASIE, POLISEMIE, SEMIOTIEK, TAALKUNDE, TAALTEKEN, TEORIE VAN LEKSIKOGRAFIESE KOMMUNIKASIE, TEORIE VAN LEKSIKOGRAFIESE TEKSTE

1. Introduction

Two articles appeared recently in *Lexikos* that propose the abolishment of homonymy and polysemy in lexicography. The first article claims that "polysemy and homonymy do not exist" and that "in lexicography we can do well without these terms" (Bergenholtz and Agerbo 2014: 31). The apparent overall rejection of these concepts is also clear from the title of the article: "There is No Need for the Terms Polysemy and Homonymy in Lexicography". The second article builds on the work presented in the first, but it displays a more moderate attitude towards the relevant concepts, stating that "the existence of homonymy and polysemy as concepts in the field of linguistics is acknowledged," that arguments can be advanced for the abolishment of the "traditional distinction between homonymy and polysemy", and that the proposal to abolish polysemy and homonymy is limited to "the communicative situation where a mother-tongue speaker or a foreign language speaker encounters text reception problems" (Bergenholtz and Gouws 2017: 110, 112, 125).

The first article (Bergenholtz and Agerbo 2014) describes three models according to which homonymy and polysemy can be dealt with in dictionaries:

- Model I: the "traditional" model, where homonyms are linguistically distinguished as formally identical but separate lexemes on the grounds of semantic non-relatedness and/or different etymologies, each represented by a separate lemma sign and dictionary article, and polysemy on the grounds of the relatedness of semantic values that can be assigned to one lexeme, i.e. polysemic values presented in one article.
- Model II: a model that rejects the notions of homonymy and polysemy, and assigns only one semantic value to a given lemma: In model I, a set of two homonyms, each with three polysemic values, would be presented as two formally identical lemma signs representing each of the homonyms, each lemma sign with its own article containing three polysemic values. Given model II, the same set of lexical items would be presented as six formally identical lemma signs, each with its own article representing one semantic value only; no polysemic or homonymic relations would be signalled.
- Model III: "words that are orthographically similar but have different inflectional paradigms (also within the same part of speech) are defined as

homonyms, whereas orthographically similar words belonging to the same part of speech and with the same inflectional paradigm are defined as polysems [sic]" (Bergenholtz and Agerbo 2014: 29).

In the first article, model III is favoured because it is "closer to the solution that dictionary users are familiar with" (Bergenholtz and Agerbo 2014: 34).

The second article (Bergenholtz and Gouws 2017) attempts to build a case for the model II solution on the basis of two main theoretical premises:

- a lexicographic theory of the lemma as linguistic sign by Bergenholtz and Agerbo (2014);
- criticism of a selection of Danish and English dictionary articles.

The first aim of this contribution is to challenge these premises and therefore the validity of model II on the following points, which will be elaborated in the indicated sections to construct the argument:

- Bergenholtz and Agerbo's (2014) lexicographic theory of the lemma as linguistic sign is flawed as well as irrelevant: section 2.
- The model II solution does not address Bergenholtz and Gouws's (2017) criticism of existing dictionary articles, but merely transfers a number of perceived metalexigraphic problems from one lexicographic text structure type to another, potentially adding unnecessary complications for lexicographical communication in the process: section 3.

In the course of arguing the above points, a potential broad framework for the theory of lexicographical communication (or: lexicographical communication theory), as introduced by Beyer (2014) and Beyer and Augart (2017), is developed in subsection 2.3 on the basis of linguistic grammar. This is the second aim of this contribution. The basic tenets of the lexicographical communication theory are that (i) at its core, lexicography is an exercise in communication, and (ii) this communication is indirect communication mediated by text (Beyer and Augart 2017: 8). The description of dictionary article text structures in the theory of lexicographic texts (or: lexicographic text theory), developed primarily by H.E. Wiegand within a general theory of lexicography, is "completely taken over from formal syntax" (Wiegand 1996: 136), which can be observed in that theory's presentation of (abstract) microstructures in the form of hierarchical tree structures similar to the presentation of sentence constituents in context-free (i.e. phrase structure) grammars (cf. Gouws, Heid, Schweickard and Wiegand 2013: articles 3–10). This method has inspired the grammar framework that will be presented for the lexicographical communication theory. Consequently, similarities between the framework presented and the relevant elements of the lexicographic text theory will be evident, and will be accounted for where necessary for the purposes of the discussion.

2. Bergenholtz and Agerbo's lexicographic theory of the lemma as linguistic sign

Bergenholtz and Agerbo (2014) employ De Saussure's (2013) model of the *linguistic sign* to evaluate the status of a set of word types. This evaluation forms the main premise of their proposal to abolish the concepts *polysemy* and *homonymy* in lexicography. It will be shown in this section that this premise is conceptually flawed and that therefore the conclusion based on it is logically false. First, however, it is necessary to clarify the relevant terms within the Saussurean model.

2.1 (Linguistic) sign, code and sign system

The term *sign* is defined as follows by Bock (2014: 57):

def₁ A **sign** is something that represents or stands for something else, where the 'something else' may refer to an idea, object, value or phenomenon. The sign is not 'the something' itself, but rather a representation of that thing.

While signs in themselves have *values*, they can only assume *meaning* in relation to other signs (De Saussure 2013: 134ff). This requires signs to possess paradigmatic and syntagmatic properties which allow them to function in various relations with other signs (cf. De Saussure 2013: 144-148). The sum of the paradigmatic and syntagmatic properties of all signs that belong to the same sign system can be referred to as that sign system's *code*. A *sign system*, then, consists of two primary components: (i) a set of signs, and (ii) a set of rules, known as a code, which describes the paradigmatic and syntagmatic properties of the signs that allow them to be combined to signal meanings (cf. Bock 2014: 57-58). In linguistic terms, *sign system* is equated to a particular language (e.g. English), *set of signs* is equated to that language's lexicon, and *code* is equated to the language's grammar (Bock 2014: 57-58).

A *linguistic sign* is a sign (<def₁) that functions within a linguistic code: English words are linguistic signs inasmuch as they function within the linguistic code of the English grammar. De Saussure (2013: 77) defines a linguistic sign as a combination of two "intimately linked" elements, namely a "concept and a sound pattern"¹. Chandler (2007: 14ff) uses the equivalent terms *signified* and *signifier*, and Bergenholtz and Agerbo (2014) use the equivalent *content* and *expression*. Although this article is a response to Bergenholtz and Agerbo (2014) and Bergenholtz and Gouws (2017), Chandler's terms will be used in the following discussion, because they bear the closest resemblance to the original terms proposed by De Saussure (i.e. French *signifiant* and *signifié*). A (linguistic) sign, then, is "the whole that results from the association of the signifier [expression] with the signified [content]" (Chandler 2007: 15), which can, in the

style of De Saussure (2013: 77), be presented in the following diagram:



Figure 1: The constitution of the *sign*, according to De Saussure (2013), in the terms of Chandler (2007)

An alternative presentation of the same concept in table format, which will be used in this article, looks as follows:

Table 1: An alternative representation of the concept *sign* according to De Saussure (2013), in the terms of Chandler (2007)

Sign	
Signifier	Signified

2.2 Bergenholtz and Agerbo's application of the term *linguistic sign*

Bergenholtz and Agerbo (2014: 31) claim that "we cannot speak about polysemy and homonymy if we relate these terms to the linguistic sign. However, in lexicography we can do well without these terms." This claim is based on the following argument (Bergenholtz and Agerbo 2014: 31):

quote₁ In the lexicographical tradition [...] a lemma is not a linguistic sign because a lemma can represent different lexical words (sometimes it represents only one lexeme, in other cases it represents several lexemes). Hence, there is no solidarity between one expression [signifier] and one content [signified].

The argument is followed by the model II proposal as a "radical solution [...] where we discard polysemy and homonymy and instead connect each lexical word to its own lemma," because only then "the lemma could be defined as a linguistic sign" (Bergenholtz and Agerbo 2014: 31).

In the following subsections different aspects of Bergenholtz and Agerbo's application of the term *linguistic sign* will be scrutinised.

2.2.1 All (types of) words are linguistic signs

The model II solution depends on Bergenholtz and Agerbo's evaluation of the lemma as a linguistic sign in certain uses and not a linguistic sign in other uses.

This evaluation is conducted within the context of a broader evaluation of the status of a set of word types vis-à-vis the concept *linguistic sign*, namely so-called orthographic words, text words, grammatical words, lexical words (lexemes) and dictionary words (lemmata) (Bergenholtz and Agerbo 2014: 30-31). The broader evaluation can be summarised in the following table:

Table 2: Summary of Bergenholtz and Agerbo's (2014) evaluation of a set of word types

Word type	Description	Linguistic sign?
orthographic word	A sequence of letters between blanks and sentence signs (like commas), also search strings in e-dictionaries.	No
text word	A concrete word in a text with a specific spelling, meaning, grammar, etc.	Yes
grammatical word	An expression with at least one nucleus morpheme and for adverbs, verbs and nouns also at least one grammatical morpheme. A grammatical word belongs to a certain inflection paradigm.	No
lexical word (lexeme)	An abstraction for an amount of grammatical words belonging to the same stem and the same inflection paradigm.	Yes
lemma	An abstraction for an amount of grammatical words, but it is not the same as a lexical word, because, contrary to lexical words, different stem meanings do not result in different lemmata.	No

In every case in table 2, a word type is judged to be a linguistic sign or not on the basis of the perceived presence or absence of a combination of signifier and signified to form a sign. In fact, each judgement is based on the prerequisite for the existence of a sign per se (cf. def₁; Chandler 2007: 15), and not necessarily of a linguistic sign, because the requirement of functioning specifically in a linguistic code is not tested (except perhaps with the type *text word*).

Table 2 clearly shows that every word type *represents or stands for* some concept as summarised under the heading "Description" (<def₁; Chandler 2007: 15), which presupposes signification, i.e. a combination of signifier and signified, *in every case*. This is an obvious refutation of every "No"-judgement, i.e. of every judgement that a particular word type is not a linguistic sign. Moreover, Bergenholtz and Agerbo's (2014: 31) argument in quote₁ above that "a lemma is not a linguistic sign because a lemma can represent different lexical words" is self-contradictory: If a lemma (or any other word type) *represents or stands for* x, y and/or z, it follows that it is a sign. This can be illustrated by listing an exem-

plar of each word type and indicating how that exemplar is a sign by aligning its signifier and a representation of its signified, as in table 3:

Table 3: Examples of word types and their sign values

Ref.	Word type	Sign value	
		Signifier	Representation of the signified
1	orthographic word	<i>flush</i>	'the grapheme sequence ⟨f, l, u, s, h⟩'
2	text word	<i>flushes</i>	' <i>flushes</i> in "Tom has played two flushes so far"'
3	grammatical word	<i>flushes</i>	'the grammatical word paradigm { <i>flushes</i> (n., pl.: 'reddening'), <i>flushes</i> (n., pl.: 'hand of cards'), <i>flushes</i> (n., pl.: 'piece of wet ground')}'
4	lexical word (lexeme)	<i>flush</i>	'the inflection paradigm { <i>flush</i> , <i>flushes</i> '}
5	lemma	<i>flush</i>	'the lexeme <i>flush</i> '

Table 3 shows the various signs' values. Additionally, each of the signs can be proven to be a *linguistic sign*, because each can function in terms of its word type and assume *meaning* in paradigmatic and syntagmatic relations to other signs in the code of the English grammar. More directly, the mere fact that each category could be designated a type of *word* indicates the linguistic sign status of every category member. Compare their respective occurrence in the following grammatical English sentences (numbered in correspondence to "Ref." in table 3) (cf. also Murphy 2010: 11f and Cruse 2011: 47):

- (1) [The orthographic word] *flush* consists of five graphemes.
- (2) [The text word] *flushes* in "Tom has played two flushes so far" means 'more than one hand of cards all of the same suit'.
- (3) [The grammatical word] *flushes* represents a grammatical word paradigm.
- (4) [The lexeme] *flush* represents an inflection paradigm.
- (5) [The lemma] *flush* represents a lexeme.

Sentences (1) to (5) demonstrate that each word functions not only as a sign, but also as a linguistic sign.

The conclusion is therefore that, in the first place, and contrary to Bergenholtz and Agerbo's (2014) evaluation, all word types in table 2 are signs because signification is proven in all cases. In the second place, they are specifically linguistic signs because they function within a linguistic code, in this case that of English.

There are, however, more obvious and general problems with Bergenholtz and Agerbo's (2014) lexicographic theory of the lemma as linguistic sign. These are dealt with in the following subsections.

2.2.2 Representation of the signified is not the signified

Compare the following dictionary article from the *Oxford South African Concise Dictionary* (Van Niekerk and Wolvaardt 2010: 449):

da₁ **flush**³ ■ n. (in poker or brag) a hand of cards all of the same suit.

Leaving the homonymy indicator |³| and the register item |(in poker or brag)| aside for the moment, Bergenholtz and Agerbo (2014) would argue that the lemma in da₁ is a linguistic sign because there is solidarity between one expression (signifier: the lemma sign form) and one concept (signified: |a hand of cards all of the same suit|). Semiotically speaking, however, there is a fundamental problem with this argument.

The signifier is the "sensory part" of the sign which "implies reference to the whole [i.e. the sign itself — HLB]" (De Saussure 2013: 77). It is "the *material (or physical) form* of the sign — it is something which can be seen, heard, touched, smelled or tasted" (Chandler 2007: 15). The signified is "generally of a more abstract kind" (De Saussure 2013: 76). Chandler (2007: 16) explains that De Saussure's "*signified* is not to be identified directly with [...] a referent but is a *concept* in the mind — not a thing but a notion of a thing." (Cf. also Peirce 1985, Sebeok 2001: 5-6, Danesi 2004: 4-6, Hébert 2018.)

The point being made is that whereas the signifier has a physical form, the signified is abstract: It is physically imperceptible. A lexicographic paraphrase of meaning — ostensibly referred to as a *meaning* by Bergenholtz and Agerbo (2014) and Bergenholtz and Gouws (2017)⁴ — is a physically perceptible signal; therefore, it is impossible to equate it to a signified (or, in Bergenholtz and Agerbo's (2014) terms, a *content*). Rather, the lexicographic definition |a hand of cards all of the same suit| in da₁ constitutes a complex sign (in the form of a syntagma) associated with the signified 'flush' in the very same way that the lemmatically represented word form *flush* constitutes a simple sign associated with the same signified.⁵ The logical conclusion is that the lemmatically represented form and the lexicographic definition are two *equivalent* signs. This fact becomes clearer when the lexicographic definition is replaced by a word synonym in a monolingual dictionary and by a translation equivalent in a bilingual dictionary. (Bergenholtz and Agerbo (2014: 34) assert that their theory applies to "both monolingual and bilingual dictionaries; there are no significant differences".) As wholes, then, the lemma sign and lexicographic definition in da₁ are indirectly equivalent signs: the lemma in the form of a sign representing a simple linguistic sign with the value 'flush'_i and the lexicographic definition in the

form of a syntagma as signifier of a complex sign with the meaning 'flush'. The relevant relations can be represented in figure 2:

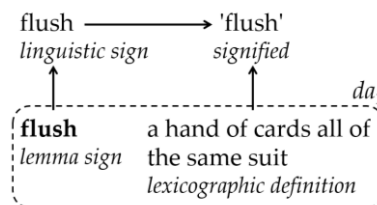


Figure 2: A simplified representation of the semiotic relations involving the lemma sign and lexicographic definition in da_1 , and the signified ("x → y" = x refers to y)

It follows that a dictionary article, or any text for that matter, cannot contain a signified/content. A monolingual dictionary article simply coordinates signs in one and the same sign system that share the same signified, in exactly the same way that a bilingual dictionary article coordinates signs in a source sign system with signs in a target sign system that share signifieds, explained in linguistic terms by Zgusta (1971: 294) as the semantic coordination of a set of lexical items in one language with that of another. With regard to the purposes of a specific dictionary, the lexicographic definitions, word synonyms and/or translation equivalents function as representations of (or *comments* on) the signifieds associated with the lemmatically represented signs; they are not — and cannot possibly be — the signifieds in themselves. In the case of a dictionary article of a polysemic lemma, the lemma sign represents a set of linguistic signs with identical signifiers (which, in model I, normally constitute a lexeme), while the semantic and pragmatic comments on the various identified senses represent the set of signifieds co-constituting the respective signs. From the number of senses so distinguished, together with data on inflection, the number of signs that are (partially) represented in the dictionary article can be inferred, if necessary, although this would hardly fulfil one of the purposes of a dictionary with only a text reception function. This, in short, is the semiotic nature of the typical dictionary article as text.

The above exposition clearly shows that the semiotic requirement that a dictionary article should represent "solidarity between one expression [signifier] and one content [signified]" (Bergenholtz and Agerbo 2014: 31) is untenable, regardless of the dictionary's purposes. In semiotic terms, a monosemic dictionary article in effect coordinates at least two signifiers that can signify the same signified. This represents one of the core problems in lexicography: how to represent the signified of a particular signifier in terms of another signifier or signifiers.

A further problem with the semiotic requirement pertains to the question of inflected word forms as linguistic signs, which is the focus of the next subsection.

2.2.3 Inflected words are (also) linguistic signs

Gallmann (1991) assigns all formal (i.e. physical) features of the linguistic sign to the signifier, while all grammatical and semantic features are assigned to the signified, in line with the concept of the sign (cf. again Peirce 1985, Sebeok 2001: 5-6, Danesi 2004: 4-6, Chandler 2007: 15-16, De Saussure 2013: 77, Hébert 2018). Therefore, inflected and non-inflected word forms constitute separate linguistic signs, since an inflected word form as sign differs both in terms of signifier (i.e. formal features) and signified (i.e. grammatical features) from its non-inflected form. Bergenholtz and Agerbo (2014: 30) also evaluate so-called text words, which include inflected forms, as linguistic signs (cf. table 3 and sentence (2) in 2.2.1). This can be illustrated with a simple example in table 4:

Table 4: Inflected and non-inflected word forms as separate linguistic signs

Sign	
Signifier	Representation of the signified
<i>ampersand</i>	'&'
<i>ampersands</i>	'& & ...'

Bergenholtz and Gouws (2017: 125) regard inflected forms as "different variant forms of the expression [signifier] with the same contents [signified]." From the above it is clear that this is an untenable position. It also contradicts Bergenholtz and Agerbo's (2014: 30) evaluation of text words as linguistic signs. Even orthographic variants, like *realise* and *realize*, are separate signs: Although they share the same signified, they have distinctive signifiers. After all, a (linguistic) sign exists only as "solidarity between *one* expression [signifier] and *one* content [signified]" (Bergenholtz and Agerbo 2014: 31; my emphasis — HLB). Bergenholtz and Gouws's mistaken semiotic definition of inflected forms seems to originate from Bergenholtz and Agerbo's (2014: 30) evaluation of a lexeme as a linguistic sign (cf. table 2), which is of course correct in itself; however, a lexeme's signified constitutes an entire inflection paradigm and not only the stem of such a paradigm (cf. table 3). It would seem that properties of the concept *lexeme* (a linguistic notion) have been confused with that of the concept *sign* (a semiotic notion).

If Bergenholtz and Agerbo's (2014) semiotic requirement that a lemma should be a linguistic sign with one signifier and one signified is to be met, then it follows that every inflected word form should also be lemmatised instead

of merely indicating inflection possibilities in the article of a stem. This is obviously not Bergenholtz and Agerbo's (2014) and Bergenholtz and Gouws's (2017) positions, from which it would appear that they contradict their own requirements. Therefore, Bergenholtz and Agerbo's (2014: 34) claim that model II is not "connected to any theoretical contradictions" does not hold water.

Besides the foregoing, it will be argued in the following subsection that typical lexicographical communication, especially via the medium of the typical dictionary article, is conducted within a sign system that is different from the natural language that is the object of the lexicographical communication in a particular instance. This implies that in lexicographical communication the lemma is in fact *not* a *linguistic sign*, but a sign in a different code, namely a lexicographical code, and is therefore a *lexicographic sign*.

2.3 The lemma as non-linguistic sign (in a linguistically-based theory of lexicography)

The lexicographical communication theory takes a global view of the potential of linguistic theory for meta-lexicography, i.e. linguistic theory not merely to explain the representation of lexical data in dictionaries, but also to form a basis for explaining how lexicographical communication functions (cf. Beyer 2014: 40). An attempt to construct such a basis will be outlined in this subsection as part of the discussion of the lemma as sign. Although the linguistic perspective is inspired by the lexicographic text theory, there are important areas of divergence between the lexicographic text theory and the lexicographical communication theory, as will be indicated where relevant.

2.3.1 A lexicographic sign system

The fact that dictionary articles typically comment on the lexical features of a particular natural language obscures the fact that such comments are typically not encoded in that language, but in a hybrid sign system that merely partially resembles and overlaps with the relevant language, yet is significantly distinct from it. Compare the following two texts (*text₂* being a slightly adapted version of a dictionary article from the *South African Oxford Secondary School Dictionary* (Reynolds 2006: 57)):

text₁ This is a paragraph about the word *bigwig*. The word *bigwig* is a word in English, and it is spelt as b, i, g, w, i, g. It is a noun. It is also an informal word, so be careful not to use it in a formal context; if you hear it or read it in a text, you will know that the speaker or author is using informal language in that instance. The word *bigwig* has only one semantic value, namely 'an important person'.

text₂ **bigwig** *n.* (*informal*) an important person

Text₁ is a text in natural language which adheres to the grammar of English. Text₂ obviously does not adhere to the grammar of English, yet it successfully communicates the same contents than text₁ does — but only for someone who knows how to interpret it. A literate mother-tongue speaker of English would easily interpret text₁ fully and correctly, but this does not imply that they would be able to fully and correctly interpret text₂. Conversely, it is possible for someone who does not know English at all to at least partially interpret text₂ correctly and even to answer a limited set of user questions (e.g. that the form *bigwig* is a lexeme in English and that it has only one sense), provided that they are "text₂-literate", in spite of the fact that they would not be able to interpret text₁ at all. Since humans make meanings through the creation and interpretation of signs (Sebeok 2001, Chandler 2007: 14), human communication requires sign systems. Because text₂, which seems to be an English text, successfully communicates only between parties with some type of competence in addition to their competence in English, it follows that text₂ adheres to a sign system that is at least partially different from English.

The lexicographic text theory would argue that text₁ has been subjected to textual condensation in a process of lexicographic textualization in order to produce text₂, which means that text₂ is some condensed version of text₁ (cf. Wiegand 1996a). Textual condensation would involve operations identified as shortening, abbreviating, omitting, shifting, substituting, summarising and embedding (Wiegand 1996a: 139). Some of these operations correspond to a greater or lesser degree to some of the operations identified and described in text linguistics, particularly abbreviation, substitution and ellipsis. However, the critical distinction is that text linguistics explains the relevant operations within the framework of the grammar of the relevant language, for example De Beaugrande and Dressler (1981) with regard to English, and Carstens (1997) with regard to Afrikaans. In contrast, the operations of textual condensation that would render text₂ as a condensed version of text₁ cannot be explained within the framework of the grammar of English. It follows then that text₁ and text₂ are created within the frameworks of different codes: text₁ within the framework of the grammar of English, and text₂ within the framework of some other code. This fact has required the lexicographic text theory to develop elaborate sub-theories of textual condensation (cf. Wiegand 1996a) and addressing structure (cf. Wiegand and Gouws 2013) to construct an inter-code bridge between text₁ and text₂. These sub-theories in fact amount to the description of an alternative code to the grammar of English in order to make the rendering of text₂ possible. For this reason, the lexicographical communication theory does not recognise text₂ as any *version* of text₁, but rather views text₁ and text₂ as distinctly separate texts that happen to encode the same set of lexicographic messages by means of distinctly separate sign systems: text₁ by means of the English language, and text₂ by means of a lexicographic sign system (which, in this case, overlaps with English in some ways), effectively making text₁ and text₂ *textual translation equivalents* of each other.

Although text₂ does not adhere to the grammar of English but ostensibly contains English words and even an English syntagma, it might be argued that it constitutes a version of text₁ because the reader can successfully interpret text₂ through processes of inference such as described by for example the theory of conversational implicature (cf. Grice 1991) and relevance theory (cf. Sperber and Wilson 1995, Clark 2013), to arrive at the propositions in text₁. In this regard Sperber and Wilson (1995: 12-13) note the following:

Inferential and decoding processes are quite different. An *inferential process* starts from a set of premises and results in a set of conclusions which follow logically from, or are at least warranted by, the premises. A *decoding process* starts from a signal and results in the recovery of a message which is associated to the signal by an underlying code, and signals do not warrant the messages they convey.

It is clear that the highly sophisticated and intricate lexicographic text theory has developed a general code for lexicographic texts, because every functional text segment identified and described by the theory is assigned a specific unit of lexicographic data that it transmits. This means that there is a fixed association between signal and message, and that the receiver of such a text *decodes* the signal to recover the lexicographic message. Therefore, during optimal lexicographical communication, encoding and decoding takes place rather than implicature and inferencing. This implies "an underlying code", which, as has been seen, is not the grammar of English, but a distinct lexicographical code.

When text₁ and text₂ are evaluated against the foregoing argument, the conclusion is that text₁ is an English text, but that text₂ is not an English text, although it is a text about English. It is clear that there is an overlap of codes (and sign systems) in text₂, but this in itself is not an unusual phenomenon. Although it is not equally evident, there is also an overlap of codes in text₁. Chandler (2007: 149) points out that "various kinds of codes overlap, and the semiotic analysis of any text or practice involves considering several codes and the relationships between them." Based on a range of code typologies found in the literature of semiotics, Chandler (2007: 149-150) distinguishes between three main classes, of which two are relevant for the current discussion, namely:

- **social codes**, including natural/verbal language (with phonological, syntactic, lexical, prosodic and paralinguistic subcodes), bodily codes, commodity codes and behavioural codes;
- **textual codes**, including scientific codes, aesthetic codes, genre codes, rhetorical codes, stylistic codes and mass media codes.

A language like English obviously belongs to the class of social codes, but text₁ is created through an overlap between the social code and a particular textual code in order to produce a paragraph. Arguably, the social code is the primary code and the textual code is the secondary code (cf. also De Saussure 2013 on

the spoken vs. written modes of natural language). Given that lexicographical communication almost exclusively takes place through the medium of specialised types of text (and not in sound form as in the case of natural language), it can be argued that a particular textual code (which is significantly different from that of text₁, even to the extent that it in fact constitutes a different sign system) is the primary code of text₂, which is overlapped to a certain degree by a social code, in this case English. Therefore, lexicographical communication like in text₂ takes place by means of a distinct lexicographic sign system. The sign systems that have been studied the most extensively and scientifically are natural languages because they are the "primary and most pervasive" codes in any society (Chandler 2007: 149). This has given rise to the extensive discipline of modern linguistics. It therefore makes sense to consider the potential value of linguistic theory in attempting to describe a lexicographic sign system. Such a specific text-based sign system could be referred to as a *lexicographic language*, or *l-language* (as opposed to a natural language, or "n-language"). It should be noted that, because of its text-based nature, an *l-language* is not a type of natural language and is not represented by an element of Chandler's class of social codes or described by linguistics; rather, it is represented by a type of textual code. The sign |■| in da₁ (cf. 2.2.2), for example, is not a linguistic sign, but it belongs to the lexicon of the relevant *l-language*. The partial term *language* is merely used for lack of a better alternative.

With regard to an *l-language* as sign system, *set of signs* is equated to *lexicographic lexicon* (or: *l-lexicon*), and *code* is equated to *lexicographical grammar* (or: *l-grammar*). The sign |■| in da₁, for example, would be an element of the *l-lexicon* of the *l-language* used in the dictionary involved. In the following section natural language grammars will be highlighted briefly to provide a background for the introduction of an *l-grammar* in section 2.3.3.

2.3.2 Natural language grammars

Traditionally, a natural language grammar consists of the following components:

- *phonetics* and *phonology*, describing the sound system of the language;
- *morphology*, describing word formation;
- *syntax*, describing sentence formation;
- *semantics*, describing the meaning of words and sentences;
- *pragmatics*, describing the use of the language in context.

In a traditional grammar, the largest unit of study is any of the various types of sentence. Consider the following simple English sentence:

s₁ A lemma represents a lexeme.

An English phonetics and phonology would study the speech sounds and phonological processes involved in pronouncing the sentence, for example that *a* is pronounced [ə], and that [ə] does not assimilate with the following sound [l] because it is a lateral.

Morphology would for example note that the verb *represents* is an inflected form of *represent*, and that *represent* is a diachronic derivative of the order [present]_V.

Syntax would identify and describe the order of the various sentence constituents, for example in the following linear representation of the constituent syntax of *s*₁:

[_S[_{NP}[_{DET}[_{ART} A]] [_N lemma]]_{NP} [_{VP}[_V represents] [_{NP}[_{DET}[_{ART} a]] [_N lexeme]]]]

From the above description the following set of syntactic rules could be derived: S → NP VP; NP → DET N; DET → ART; VP → V NP

Semantics would describe the semantic values of respective words and the propositions that are encoded in the sentence, and the relations between them, for example:

Lexical semantics: *lemma* → [- animate], [+ abstract], [+ countable], etc.

Sentence semantics: REPRESENT(a lemma, a lexeme)

Pragmatics would describe the meaning of the sentence as an utterance in context, for example that it constitutes an assertion, that its interpretation can be described in terms of a cooperative principle of communication, how the subject relates to interlocutors' common ground through reference by means of the indefinite article *a*, etc.

In addition to traditional sentence-based grammars, the discipline of text linguistics expands the basic object of linguistic enquiry to the text or discourse as a whole (cf. De Beaugrande and Dressler 1981, Carstens 1997). According to Carstens (1997: 53-59), Van Dijk (1972) had a tremendous influence on the development of text research, particularly with his notion of a *text grammar*, which proposes that, like sentences, texts can be described in terms of a type of formal grammar, facilitated by a distinction between textual surface and deep structures. The following tasks are assigned to a text grammar by Van Dijk (1972: 11):

- to formally enumerate all and only grammatical texts of a language;
- to assign structural descriptions to each of these generated texts;
- to formulate rules in terms of which the textual deep structure can be derived from the textual surface structure; and
- to investigate textual surface structures.

The potential of a text grammar for lexicographic theory development is par-

ticularly attractive to the lexicographical communication theory, especially because of the generally highly conventionalised nature of lexicographic texts as it relates to the second basic tenet of the theory. Within the broader discipline of text linguistics, the seven elements of textuality, i.e. cohesion, coherence, intentionality, acceptability, informativity, situationality and intertextuality (cf. De Beaugrande and Dressler 1981, and Carstens 1997), are also of central relevance.

2.3.3 A text grammar as a lexicographical code

In line with the object of study in text linguistics, the largest unit of study in an *l*-grammar is any of the various types of *lexicographic text*, which entails that an *l*-grammar is essentially a type of text grammar. The lexicographic text theory, having empirically identified and meticulously described a range of lexicographic text types, provides a solid foundation in this regard.

Adopting and adapting concepts from linguistic theory, it is proposed that an *l*-grammar consists at least of the following components:

- an *l-syntax*, describing the order of the various text elements in a lexicographic text and the textual surface structure relations among them;
- an *l-morphology*, describing the formation of lexicographic items contained in a lexicographic text;
- an *l-semantics*, describing the lexicographic propositions encoded in lexicographic items and the textual deep structure relations among them;
- an *l-pragmatics*, describing the communicative functions of the various text elements and the textual deep structure relations among them.

An *l-phonology* could be added in cases where lexicographical communication takes place via the audio channel, for example the representation of pronunciation data relating to the target language by means of audio(-visual) signals in an e-dictionary.

The above *l*-grammar components can be illustrated by applying them to da_1 (repeated below):

da_1 **flush**³ ■ n. (in poker or brag) a hand of cards all of the same suit.

An *l-syntax* would identify and describe the order of the various text constituents in da_1 , for example in the hierarchical structure in figure 3:

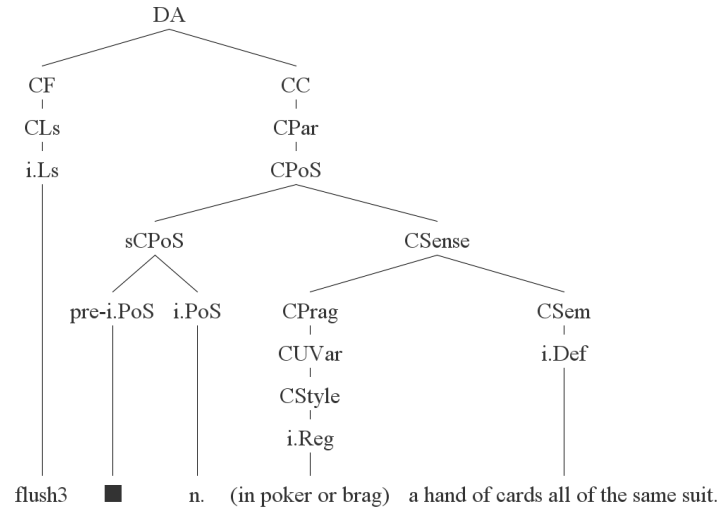


Figure 3: A constituent *l*-syntax of da_1

(Key: DA = dictionary article; CF = comment: form; CC = comment: concept; CLs = comment: lemma sign; i.LS = item: lemma sign; CPar = comment: paradigmatic properties; CPoS = comment: part of speech; sCPoS = sub-comment: part of speech; pre-i.PoS = pre-item: part of speech; i.PoS = item: part of speech; CSense = comment: sense; CPrag = comment: pragmatic value; CUVar = comment: usage variation; CStyle = comment: style; i.Reg = item: register; CSem = comment: semantic value; i.Def = item: *l*-definition)²

The following set of *l*-syntactic rules could be derived: DA → CF CC; CF → CLs; CLs → i.LS; CC → CPar; CPar → CPoS; CPoS → sCPoS CSense; sCPoS → pre-i.PoS i.PoS; CSense → CPrag CSem; CPrag → CUVar; CUVar → CStyle; CStyle → i.Reg; CSem → i.Def

An *l*-morphology would describe the formation of the *l*-items involved, e.g. the lemma sign | **flush**³ | consists of the lemma sign form | **flush** |, printed in roman and bold, and a suffix |³| in superscript; the pre-item to the part-of-speech item is a dark square | ■ |; the part-of-speech item | n. | is an abbreviation and printed in roman; the register item |(in poker or brag)| is a PP, circumfixed by parentheses and printed in roman; the lexicographic definition |a hand of cards all of the same suit| is a NP and printed in roman. With regard to the part-of-speech item | n. |, there is an overlap between the morphology of the *l*-grammar and the morphology of the target language's grammar, and with regard to the lexicographic definition |a hand of cards all of the same suit|, there is an overlap between the morphology of the *l*-grammar and the syntax of the target language's grammar. These overlaps accentuate the hybrid nature of the *l*-language.

The lexicographic text theory regards typographical features like parentheses as non-typographical structural markers, and bold print and italic print as typographical structural markers, all of which are elements of a set of non-functional text elements (cf. Wiegand 1990). The lexicographical communication theory, however, regards these features as *l*-morphemes and therefore as inherent component structures of *l*-items.

An *l*-semantics would describe the semantic value of each *l*-item as a union of form and *l*-proposition(s), for example in the table below:

Table 5: *L*-items and *l*-propositions in *da*₁

<i>L</i> -items	<i>L</i> -propositions
flush	<i>lp</i> ₁ : This is the dictionary article about the word <i>flush</i> . <i>lp</i> ₂ : The word <i>flush</i> is a word in SA English. <i>lp</i> ₃ : The word <i>flush</i> has the orthographic form ⟨f, l, u, s, h⟩.
³	<i>lp</i> ₄ : The word <i>flush</i> is a member of a homonym paradigm.
n.	<i>lp</i> ₅ : The word <i>flush</i> is a noun.
(in poker or brag)	<i>lp</i> ₆ : (As a noun) the word <i>flush</i> is a word in the register of poker or brag.
a hand of cards all of the same suit	<i>lp</i> ₇ : (As a noun) the word <i>flush</i> has the semantic value 'a hand of cards all of the same suit'.

An *l*-pragmatics would describe, among other things, the illocutionary force that accompanies every *l*-proposition to form the *l*-message encoded in the *l*-utterance. In terms of *da*₁, the illocutionary force STATEMENT would for example accompany *l*-propositions *lp*₁ to *lp*₅ and *lp*₇ in table 5, and the illocutionary force ADVICE could accompany *l*-proposition *lp*₆, depending on the dictionary's purposes and target user sociology.

The *l*-semantic information in table 5, coupled with the relevant *l*-pragmatic variables (specifically speech acts), explain how *text*₂ above communicates the same messages than *text*₁, but by means of a sign system that is distinct from English, namely an *l*-language.

2.3.4 The lemma (sign) as sign

From table 5 in the previous section it is clear that the lemma sign form | **flush** |, as it functions in *da*₁, is not a linguistic sign like in sentence *s*₁ (cf. 2.3.2), because in *da*₁ it does not display the paradigmatic and syntagmatic properties required to function in the English grammar. Whereas the lemma *flush* functions as a linguistic sign in sentence (5) in section 2.2.1, it functions as an *l*-sign

in the *l*-language of da_1 , representing a complete, multi-propositional *l*-utterance, as *l*-propositions lp_1 to lp_3 in table 5 demonstrate.

Furthermore, the *l*-status (as opposed to the linguistic status) of the lemma sign form |flush| can be illustrated by contrasting its salient paradigmatic and syntagmatic properties to those of the lemma as linguistic sign, as in table 6 below:

Table 6: Paradigmatic and syntagmatic properties of the lemma *flush* as linguistic sign and as *l*-sign

	Lemma <i>flush</i> as linguistic sign in (5)	Lemma <i>flush</i> as <i>l</i> -sign in da_1
Paradigmatic properties	— Can be replaced by any countable noun	— Can be replaced by any lemma sign form
Syntagmatic properties	<ul style="list-style-type: none"> — Forms the compulsory head of a NP — Functions as stem of inflected forms — Can be inflected by the plural-forming suffix <i>-es</i> — Can take AP, NP, NUM, etc. as pre-modifiers — Can take ADV, PP, S, etc. as post-modifiers 	<ul style="list-style-type: none"> — Forms the compulsory head of a CF — Functions as stem of i.LS — Takes the superfix³ [...] — Can take the suffix [^{<sup>}x_i_{</sup>] to indicate that it is an element (number x_i) of a homonym paradigm}

Consider the variation of da_1 in da_2 below:

da_2 *³ n. (in poker or brag) *flush* a hand of cards all of the same suit.

Dictionary article da_2 is preceded by an asterisk in the linguistic tradition of marking an ungrammatical construction, in this case an *l-ungrammatical* variation of da_1 because the lemma sign form does not conform to its *l*-syntactic and *l*-morphological properties within *l*-grammar $_{da_1}$, which can be expressed in the following rules:

l-syntax $_{da_1}$: DA → CF CC; CF → CLs; CLs → i.Ls

l-morphology $_{da_1}$: [x]_{i.Ls} = [** x **]_{i.Ls}; [x]_{i.Ls[+HOM, 3]} → [^{x _{}]_{i.Ls}}

(Key: x = superfix: print x in bold; ^{x} = superfix: print x in superscript. Compare Booij (2012: 119) for an interpretation of the morphological rule.)

The foregoing illustrates that, at least in principle, a lemma can function as both linguistic sign and *l*-sign. It functions as linguistic sign in a natural language sentence, and as *l*-sign in a dictionary article. Obviously, its primary function is

that of an *l*-sign. Therefore, again, any requirement that a lemma should be a linguistic sign in order to function in an *l*-grammar cannot be valid. This distinction would of course not affect the basic general norm that in order for a lemma to be considered for inclusion in the lemma list of a dictionary, such lemma (as an *l*-sign) should represent a linguistic sign in the treated lexicon.

2.4 Perspective

The discussion in the foregoing subsections (especially 2.2) demonstrate that Bergenholtz and Agerbo (2014) seemingly confuse aspects of semiotic theory with aspects of linguistic theory by attempting to disprove the existence of the linguistic phenomena of polysemy and homonymy through arguments of semiotics relating to the concept of the sign. The apparent confusion results in a misapplication of the Saussurean model of the linguistic sign, which invalidates their lexicographic theory of the lemma as linguistic sign. Furthermore, it is shown that the theory of the lemma as linguistic sign is irrelevant, because the lemma does not function as linguistic sign in lexicographical communication. Consequently, the first premise for the model II solution fails.

The validity of the second premise is the focus of the next section.

3. Criticism and model II implementation

In this section the criticism on existing dictionary articles by Bergenholtz and Gouws (2017) is examined. The model will also be implemented hypothetically with regard to one actual dictionary article series in the *Oxford South African Concise Dictionary* in order to identify and evaluate salient implications.

3.1 Criticism on existing dictionary articles dealing with homonymy and polysemy

Bergenholtz and Gouws (2017) offer a comparative criticism of the treatment of polysemy in three Danish and six English dictionaries to motivate the model II proposal. The criticism can be summarised in the following points:

- crit₁ The numbering of polysemic values are sometimes done in a non-transparent way and therefore polysemic values are distinguished unsystematically.
- crit₂ Just as many "meaning gaps" can be detected in the dictionaries as lemma gaps.
- crit₃ Different dictionaries that have the same lemma have different (numbers of) polysemic values for that lemma.

- crit₄ The same polysemic values in different dictionaries are ordered differently.
- crit₅ It is often unclear how polysemic values are distinguished in the same and in different dictionaries.

The general conclusion is that there is often greater consistency in lemma selection but a "lack of consistency in polyseme selection" among the dictionaries (Bergenholtz and Gouws 2017: 124). The criticism acknowledges that different dictionaries have different purposes and serve different user sociologies, which would account for some discrepancies, but not for all.

With regard to homonymy, it is argued that the distinction of homonyms does not serve the user sociology of a dictionary with only a text reception function (Bergenholtz and Gouws 2017: 125).

In the following subsection an existing series of dictionary articles will be adapted to show how the implementation of the model II solution would impact presentation and lexicographical communication. This will be followed by combined comments in subsection 3.3 on both the hypothetical model II implementation and the above criticism.

3.2 Hypothetical implementation of the model II solution

Dictionary article series *das*₁ below, extracted from the *Oxford South African Concise Dictionary* (Van Niekerk and Wolvaardt 2010: 449), will be adapted to the model II solution and presented as dictionary article series *das*₂.

Oxford South African Concise Dictionary article series *das*₁ = ⟨[flush¹]_{da} ... [flush⁴]_{da}⟩:

*das*₁ **flush**¹ ■ v. 1 (of a person's skin or face) become red and hot, typically through illness or emotion. 2 cleanse (something, especially a toilet) by passing large quantities of water through it. ► remove or dispose in such a way. 3 drive (a bird or animal, especially a game bird) from cover. 4 (of a plant) send out fresh shoots. ■ n. 1 a reddening of the face or skin. ► an area of warm colour or light. 2 a sudden rush of intense emotion. ► a period of freshness and vigour: *the first flush of youth*. 3 an act of flushing. 4 a fresh growth of leaves, flowers or fruit.

–DERIVATIVES **flusher** n.

flush² ■ adj. 1 completely level or even with another surface. 2 *informal* having plenty of money. ■ v. fill in (a joint) level with a surface.

–DERIVATIVES **flushness** n.

flush³ ■ n. (in poker or brag) a hand of cards all of the same suit.

flush⁴ ■ n. *Ecology* a piece of wet ground over which water flows without being confined to a definite channel.

Model II dictionary article series $das_2 = \langle [\text{flush}^1]_{da} \dots [\text{flushness}]_{da} \rangle$:

- das_2 **flush**¹ v. (of a person's skin or face) become red and hot, typically through illness or emotion.
flush² v. cleanse (something, especially a toilet) by passing large quantities of water through it.
flush³ v. remove or dispose by flushing (>**flush**²).
flush⁴ v. drive (a bird or animal, especially a game bird) from cover.
flush⁵ v. (of a plant) send out fresh shoots.
flush⁶ n. a reddening of the face or skin.
flush⁷ n. an area of warm colour or light.
flush⁸ n. a sudden rush of intense emotion.
flush⁹ n. a period of freshness and vigour: *the first flush of youth*.
flush¹⁰ n. (of a person's skin or face) an occurrence of becoming red and hot, typically through illness or emotion.
flush¹¹ n. an act of cleansing (something, especially a toilet) by passing large quantities of water through it.
flush¹² n. an act of removing or disposing by flushing (>**flush**²).
flush¹³ n. an act driving (a bird or animal, especially a game bird) from cover.
flush¹⁴ n. a fresh growth of leaves, flowers or fruit.
flush¹⁵ adj. completely level or even with another surface.
flush¹⁶ adj. *informal* having plenty of money.
flush¹⁷ v. fill in (a joint) level with a surface.
flush¹⁸ n. (in poker or brag) a hand of cards all of the same suit.
flush¹⁹ n. *Ecology* a piece of wet ground over which water flows without being confined to a definite channel.
flusher¹ n. *informal* someone who easily becomes red in the face through emotion.
flusher² n. someone who drives a bird or animal (especially a game bird) from cover.
flusher³ n. something that is used to drive a bird or animal (especially a game bird) from cover.
flushness n. the state of being completely level or even with another surface.

3.3 Comments on Bergenholtz and Gouws's (2017) criticism and the model II implementation

Comments are presented in numbered paragraphs.

3.3.1. A total of 16 senses (including the subsenses introduced by | ► |) are presented in four dictionary articles in das_1 . (Bergenholtz and Gouws (2017) treat subsenses as separate polysemic values.) The number of dictionary articles

have increased to 23 in das_2 , representing an increase of 575%. This seems to contradict Bergenholtz and Gouws's (2017: 128) estimations that the number of dictionary articles would rise, "but not too much". It should be noted that the estimations are based on calculations involving the number of dictionary articles and polysemic values they represent in samples of the studied dictionaries (cf. Bergenholtz and Gouws 2017: 126-128). Therefore, it could be argued that either das_2 represents a statistical exception, or that the samples are not representative of the populations involved. Nevertheless, if the variables used in the calculations are applied in adapting das_1 to das_2 , then no more than 16 dictionary articles should have resulted: one dictionary article for every sense in das_1 . How, then, can the substantial surplus of seven dictionary articles (44%) be explained? To begin with, cognisance should be taken of the fact that the dictionary's target user group are mother tongue speakers of English. Firstly, derivatives are not lemmatised in das_1 ; rather, they are listed as such without further treatment at the end of the articles representing their stems (cf. $[\text{flush}^1]_{da}$ and $[\text{flush}^2]_{da}$). This presentation is sufficient for target users engaged in text reception tasks. In das_2 every derivative has to be lemmatised and treated in a separate article with regard to every relevant polysemic value of its stem. This accounts for the last four dictionary articles in das_2 . Secondly, the remaining three surplus dictionary articles, i.e. $[\text{flush}^{11}]_{da}$ to $[\text{flush}^{13}]_{da}$, are the result of the necessary deconstruction of the lexicographic definition |remove or dispose in such a way| of the subsense of polysemic value 2 in the dictionary article $[\text{flush}^1]_{da}$ (das_1). The reference of the phrase "in such a way" and textual cohesion is lost when each polysemic value is presented in a separate dictionary article, which necessitates the addition of an article and full lexicographic definition for every polysemic value which may be a referent of "such a way". The extent to which the loss of these two lexicographic strategies may cause an increase in dictionary articles are not accounted for by Bergenholtz and Gouws (2017), and they are possibly not the only potential causes, subject to the type of dictionary involved. This implies that the offered estimates of expected increases are not reliable.

3.3.2. In relation to the previous point, there are at least two ways of dealing with lexicographic definitions in das_2 that might have been briefer in articles of polysemic lemmata thanks to the relatively easy establishment of textual cohesion, like in $[\text{flush}^1]_{da}$ (das_1). The first method is to employ cross-references, like in $[\text{flush}^3]_{da}$ and $[\text{flush}^{12}]_{da}$ (das_2). This would require the numbering of lemma signs, for example as it is done in das_2 , in order to disambiguate reference addresses. The clear disadvantage of this method is that the target user would not obtain instant access to all data relating to the lemma. The second method is to write full definitions, like in $[\text{flusher}^1]_{da}$ to $[\text{flusher}^3]_{da}$. With regard to $[\text{flusher}^2]_{da}$ and $[\text{flusher}^3]_{da}$ the question might arise as to whether instead only one lemma sign could be listed with a lexicographic definition like |someone or something that drives a bird or animal (...) from cover| in order to avoid redundancy in the lexicographic definitions of two articles. The semiotic argu-

ment advanced by Bergenholtz and Agerbo (2014) would certainly oppose such a confluence, because clearly the linguistic sign represented by the lemma sign |**flusher**²| relates to a different signified (i.e. a person) than that represented by the lemma sign |**flusher**³| (i.e. something), requiring two linguistic signs which should each be represented by a separate lemma. Also compare the treatment of subsenses in the criticism, mentioned in paragraph 3.3.1. In this regard, Lyons (1977: 554) points out and demonstrates that "distinctions of sense [and therefore of separate linguistic signs and hence lemmata] can be multiplied indefinitely" and also result in "considerable redundancy in the dictionary", apparently contradicting the "not too much"-estimate in 3.3.1. If, on the other hand, the distinction between signifieds is regarded as not significant enough to warrant two dictionary articles and the semiotic requirement is consequently somewhat relaxed, the question soon arises as to when such types of distinction are to be regarded as significant, and when not. Different editorial teams would likely draw different conclusions, and the result would be that it is not always clear how different lemmata/articles are distinguished in the same and in different dictionaries. This state of affairs would attract the same type of criticism that is expressed in crit₅, the only difference being that it would relate to a different lexicographic text structure. Once the semiotic requirement is relaxed, it is not a great cognitive step to ultimately reach a point where it is argued that all different senses of a lexeme could be grouped together in one article with a single lemma sign as guiding element, like in [**flush**¹]_{da} (das₁).

3.3.3. In relation to the previous point, it is not axiomatic that the model II solution would offer easier access to sought data, and no proof to the contrary is provided by either Bergenholtz and Agerbo (2014) or Bergenholtz and Gouws (2017). Instead of having to navigate through a series of dictionary articles in order to find the (precise) relevant sense of a lexeme, it could very well be argued that the target user would find it more convenient to have to look up only one lemma sign and find all senses of the represented lexeme(s) in a single consolidated text. Access to data in single, multi-sense dictionary articles could be enhanced with a clearly differentiating *l*-morphology and smart microarchitectural design without having to resort to the model II solution. With regard to the favouring of model III by Bergenholtz and Agerbo (2014) on the grounds of user familiarity, Bergenholtz and Gouws (2017: 110) are doubtful: "Whether such an approach is convincing remains to be seen." Given the foregoing, the same can be said of the model II proposal.

3.3.4. As alluded to in paragraph 3.3.2, the implementation of the model II solution across dictionaries would not guarantee more uniform decision-making by different editorial teams or even members of the same editorial team than if model I were maintained. Therefore, much of Bergenholtz and Gouws's (2017) criticism of the treatment of polysemy in existing dictionaries would apply in equal measure to model II dictionaries, the only distinction

being that it would target different text structures: (i) It is clear that the dictionary articles in das_2 are not ordered systematically. Which criteria of article ordering should be applied, and how would they differ from the criteria employed to order polysemic values in dictionary articles? If different dictionaries order polysemic values differently ($\langle crit_4 \rangle$), they will most likely also order articles differently in model II. (ii) Similarly, if different dictionaries display different (numbers of) polysemic values in articles of the same lemma ($\langle crit_3 \rangle$), they will most likely display different (numbers of) articles with identical lemma signs in model II. (iii) Similarly, "meaning gaps" in model I dictionaries ($\langle crit_2 \rangle$) will be manifested as article gaps in model II dictionaries. (iv) Only the strictest instance of the model II solution would fully address $crit_1$, and that would result in a presently unpredictable inflation of dictionary articles (cf. 3.3.2). Therefore, it is highly unlikely that the model II solution could be implemented without eventually some relaxation of the semiotic requirement. The risk of non-transparent and unsystematic distinctions between articles would be directly proportional to the extent to which the semiotic requirement would be relaxed, and it would be even greater across dictionaries.

3.3.5. Bergenholtz and Gouws's (2017: 125) argument that the distinction of homonyms does not serve the user sociology of a dictionary with only a text reception function is clearly valid. The model II solution successfully accommodates this issue.

3.4 Perspective

In this section it was shown that Bergenholtz and Gouws's (2017) criticism of the treatment of polysemy in existing model I dictionaries is hardly addressed by the model II solution, although it deals successfully with the question of homonymy. There are also potential quantitative consequences of the implementation of model II that have not been accounted for. Furthermore, it is highly unlikely that model II could be implemented without some eventual relaxation of the semiotic requirement, which would similarly have potential consequences that have not been considered and may be difficult to estimate. These undescribed and unidentified variables would be costly to the integrity of the model II theory, if it was otherwise in order. The conclusion is that the final premise for the model II solution is questionable at best.

In the following section the potential for an alternative to the model II solution is outlined. It is based on the practical treatment of homonymy and polysemy in Van Dale dictionaries.

4. A potential alternative to model II: *l*-polysemy and *l*-homonymy

Instead of arguing for the disposal of polysemy and homonymy in lexicography, the concepts could be adapted to lexicography so that they are not

limited to linguistic interpretation. This calls for the introduction of *l-polysemy* and *l-homonymy*. All senses that are allocated to one dictionary article and whose treatments are addressed at one lemma sign constitute *l-polysemy*, regardless of whether such senses represent linguistic polysemy. Similarly, when more than one formally identical lemma sign form, each with its separate dictionary article, is presented, those lemma sign forms are *l-homonyms* and constitute an instance of *l-homonymy*, regardless of whether they represent linguistic homonymy. Whereas linguistic polysemy and homonymy pertain to lexemes, *l-polysemy* and *l-homonymy* pertain to lemma sign forms. Lemma signs |flush¹| to |flush¹⁹| in das₂ above (cf. 3.2), for example, constitute a paradigm of *l-homonyms*.

The application of *l-polysemy* and *l-homonymy* can be briefly illustrated by means of a set of articles from *Van Dale Online Gratis Woordenboek*⁶. In the interest of brevity, details and requirements of user sociology and dictionary purposes will not be accommodated here; the objective is to demonstrate the potential of the concepts and not to fully develop an alternative model to model II.

Consider the following dictionary article series, das₃:

- | |
|---|
| <p>¹as (<i>de; v(m)</i>); meervoud: <i>assen</i>) zie <i>x-as, y-as</i></p> <ol style="list-style-type: none">1. voorwerp waarom of waarmee iets ronddraait; = spil2. denkbeeldige lijn door het middel van een voorwerp, ruimte of vlak: <i>de as van de aarde; de as Berlijn-Rome</i> het bondgenootschap tussen Duitsland en Italië van 1936 tot 19433. lijn die een lichaam in twee symmetrische helften verdeelt <p>²as (<i>de; v(m)</i>); meervoud: <i>assen</i>)</p> <ol style="list-style-type: none">1. overblijfsel bij verbranding: <i>een huis in de as leggen</i> verbranden |
|---|

Dictionary article series das₃ = ⟨[¹as]_{da}, [²as]_{da}⟩ from *Van Dale Online Gratis Woordenboek* NL-NL

In das₃, two linguistic homonyms are distinguished and presented as separate lemma signs, i.e. |¹as| and |²as|. The first lemma is allocated three polysemic values, all relating to the semantic value 'axis'. The second lemma represents a monosemic lexeme with a lexicographic definition and cotext item signalling the semantic value 'ash'. In das₃ Van Dale applies a linguistic distinction between homonyms, i.e. two lexemes with identical form but unrelated semantic values. Here, *l-homonymy* corresponds to linguistic homonymy, and *l-polysemy* corresponds to linguistic polysemy. This is a typical application of model I.

In contrast, compare [as]_{da} below:

- | |
|--|
| <p>as</p> <ol style="list-style-type: none">1. (<i>verbrande resten</i>) ashes, ash (<i>van sigaret</i>): <i>gloeiende as</i> (glowing) embers; <i>een stad in de as leggen</i> reduce a city to ashes2. axle, (<i>drijf</i><i>as</i>) shaft3. (<i>meetkunde</i>) axis: <i>om zijn as draaien</i> revolve on its axis4. (<i>muziek</i>) A-flat |
|--|

Dictionary article [as]_{da} in *Van Dale Online Gratis Woordenboek NL-EN*

In dictionary article [as]_{da}, four senses are distinguished: The first sense is related to the homonym represented by the lemma sign |²as| in das₃, senses 2 and 3 are polysemic values related to the homonym represented by the lemma sign |¹as|, and sense 4 is related to a homonym not represented in das₃. In this article, obviously, homonyms are not represented by separate lemma signs. Therefore, *l*-polysemy does not correspond to linguistic polysemy, although there is some overlap. Although linguistic homonymy could be said to be involved, it is not represented (by *l*-homonymy). In linguistic terms, lemma sign |as| represents three lexemes. In semiotic terms, it represents four linguistic signs (cf. 3.2.2).

Finally, compare the following dictionary article series, das₄:

- | |
|---|
| <p>¹dwaas (<i>bijvoeglijk naamwoord, bijwoord</i>; vergrotende trap: <i>dwazer</i>, overtreffende trap: <i>dwaast</i>)</p> <ol style="list-style-type: none">1. zot, gek <p>²dwaas (<i>de; m,v</i>; meervoud: <i>dwazen</i>)</p> <ol style="list-style-type: none">2. gek, dwaas mens |
|---|

Dictionary article series das₄ = ⟨[¹dwaas]_{da}, [²dwaas]_{da}⟩ in *Van Dale Online Gratis Woordenboek NL-NL*

In das₄, two homonyms are distinguished and presented as separate lemma signs. From the paraphrases of meaning it is clear that both lemma signs represent lexemes with very closely related semantic values: [¹dwaas]_{da} (adj., adv.) the semantic value 'foolish', and [²dwaas]_{da} (n.) the semantic value 'fool'. Here, *l*-homonymy is distinguished on the basis of lemma signs that represent formally identical lexemes from different parts of speech. If these lexemes are considered to be grammatical homonyms (cf. Carstens 2018: 116-117), then *l*-homonymy corresponds to a form of linguistic homonymy. If, instead, they are considered to represent an instance of part-of-speech multifunctionality (cf. Gouws 1989: 126-129), then *l*-homonymy does not correspond to linguistic homonymy.

In paragraphs 3.3.2 and 3.3.3 above it was argued that target users might

prefer senses to be grouped under one lemma sign for ease of access to the relevant data on the represented lexeme(s), instead of each sense being presented in a separate dictionary article to satisfy some extra-metalexigraphic requirement. The concepts of *l*-polysemy and *l*-homonymy provide the theoretical space to address the target user sociology without the obligation to conform to unduly restrictive elements of linguistic or semiotic theory. The terms have the added advantage that their denotations can vary according to the *l*-grammar in which they are applied, as demonstrated above. This does not imply, however, that they do not need to be applied systematically and be based in lexicographic theory.

The use of *l*-homonymy and *l*-polysemy in [as]_{da} and das₄ yield similar results to model III. Yet, *l*-homonymy and *l*-polysemy represent a different model because it has a different theoretical base: Model III is predicated on the notion of polysemic and homonymic signifiers as defined by Bergenholtz and Agerbo (2014: 32) (although the notion of polysemic and homonymic relations between signifieds in fact defines linguistic polysemy and homonymy; cf. Hébert 2018), while *l*-homonymy and *l*-polysemy has the construct of an *l*-grammar as foundation. In lexicographic application, the flaws of the premises underlying model II also apply to model III (cf. 2).

5. Conclusion

This article has identified two main theoretical premises for Bergenholtz and Agerbo's (2014) and Bergenholtz and Gouws's (2017) model II solution to the treatment of polysemy and homonymy in dictionaries that have only a text reception function. Under examination, as reported in the foregoing sections, one of the premises have been proven invalid, and the second is only partially valid, inasmuch as it addresses homonymy. Both premises fail to support the proposed solution with regard to the question of polysemy in the dictionary type involved. This leaves only one argument cited in favour of the model II solution, namely that of data accessibility. However, the argument can equally well support a counter-model II conclusion, as shown in paragraphs 3.3.2 and 3.3.3, which can be theoretically defended by employing the notions of *l*-polysemy and *l*-homonymy in an *l*-grammar. Whether the model II solution or a solution involving *l*-polysemy and *l*-homonymy is the (more) valid one from a standpoint of practice, can only be proven by (independent) experimental user research based on a robust methodology. Even then, the general conclusion might entail that different target user groups prefer different solutions to the treatment of polysemy. Still, it is highly unlikely that a "pure" model II solution would be practicable.

During the course of the exposition in this article, a potential broad conceptual framework for the lexicographical communication theory was developed. In the same way that the well-established term *lemma* is used in meta-lexicography to distinguish a guiding element of a dictionary article from the

lexical item which it represents, the lexicographical communication theory introduces the notion of *l*-grammar (including *l*-polysemy and *l*-homonymy) parallel to linguistic grammar to distinguish lexicographic theory from linguistic theory, even while the former benefits from the latter.

Endnotes

1. Although De Saussure (2013: 77) uses the term *sound pattern*, signifiers are "now commonly interpreted as the *material (or physical) form of the sign*" (Chandler 2007: 15); cf. 2.2.2.
2. Due to space considerations the principles of this constituent *l*-syntax (and the *l*-grammar) are not elaborated here. They will be explained in future work. However, it should be noted that the terms *comment* and *item* have different denotations from the formally identical terms in the lexicographic text theory.
3. The term *superfix* is introduced to refer to an *l*-affix that is superimposed onto another form instead of prefixed, suffixed, circumfixed or suprafixed to it. It is an affix because it is a dependent *l*-morpheme and it contributes to the construction of *l*-meaning.
4. The term *meaning* is not defined in either article despite evidently not sharing the denotation De Saussure assigns to it (cf. 2.1). If it is used as a synonym for *signified/content*, the problem is even more acute.
5. Morphological simplexes can be regarded as *simple linguistic signs*, and morphological complexes and syntagmata as *complex linguistic signs* (cf. Cruse 2011: 12-13).
6. The representation of the Van Dale dictionary articles in this section do not fully correspond to the actual articles' *l*-morphology and microarchitecture.

References

A. Primary literature (dictionary data)

- Reynolds, M. (Ed.). 2006. *South African Oxford Secondary School Dictionary*. Cape Town: Oxford University Press Southern Africa.
- Van Dale Online Gratis Woordenboek. Accessed at: <https://www.vandale.nl/opzoeken> [10 August 2018].
- Van Niekerk, T. and J. Wolvaardt (Eds.). 2010. *Oxford South African Concise Dictionary*. Second edition. Cape Town: Oxford University Press Southern Africa.

B. Secondary literature

- Bergenholtz, H. and H. Agerbo. 2014. There is No Need for the Terms Polysemy and Homonymy in Lexicography. *Lexikos* 24: 27-35. DOI: <http://dx.doi.org/10.5788/24-1-1251>.
- Bergenholtz, H. and R.H. Gouws. 2017. Polyseme Selection, Lemma Selection and Article Selection. *Lexikos* 27: 107-131. DOI: <http://dx.doi.org/10.5788/27-1-1396>.
- Beyer, H.L. 2014. Explaining Dysfunctional Effects of Lexicographical Communication. *Lexikos* 24: 36-74. DOI: <http://dx.doi.org/10.5788/24-1-1252>.

- Beyer, H.L. and J. Augart.** 2017. From User Questions to a Basic Microstructure: Developing a Generative Communication Theory for a Namibian German Dictionary. *Journal for Studies in Humanities and Social Sciences* 6(2): 1-31.
- Bock, Z.** 2014. Introduction to Semiotics. Bock, Z. and G. Mheta (Eds.). 2014. *Language, Society and Communication: An Introduction*: 55-77. Pretoria: Van Schaik.
- Booij, G.** 2012. *The Grammar of Words. An Introduction to Linguistic Morphology*. Third edition. Oxford: Oxford University Press.
- Carstens, W.A.M.** 1997. *Afrikaanse tekslinguistiek: 'n inleiding*. Pretoria: J.L. van Schaik.
- Carstens, W.A.M.** 2018. *Norme vir Afrikaans. Moderne Standaardafrikaans*. Sixth edition. Pretoria: Van Schaik.
- Chandler, D.** 2007. *Semiotics. The Basics*. Second edition. New York: Routledge.
- Clark, B.** 2013. *Relevance Theory*. Cambridge: Cambridge University Press.
- Cruse, A.C.** 2011. *Meaning in Language. An Introduction to Semantics and Pragmatics*. Third edition. Oxford: Oxford University Press.
- Danesi, M.** 2004. *Messages, Signs and Meanings. A Basic Textbook in Semiotics and Communication Theory*. Toronto: Canadian Scholars' Press.
- De Beaugrande, R. and W. Dressler.** 1981. *Introduction to Text Linguistics*. London/New York: Longman.
- De Saussure, F.** 2013. *Course in General Linguistics*. Translated and annotated by Roy Harris. London/New York: Bloomsbury Academic.
- Gallmann, P.** 1991. Wort, Lexem und Lemma. Augst, G. and B. Schaeder (Eds.). 1991. *Rechtschreibwörterbücher in der Diskussion. Geschichte — Analyse — Perspektiven*. Theorie und Vermittlung der Sprache 13: 261-280. Frankfurt am Main/Bern/New York/Paris: Peter Lang.
- Gouws, R.H.** 1989. *Leksikografie*. Pretoria/Cape Town: Academica.
- Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.)**. 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/New York: De Gruyter Mouton.
- Grice, P.** 1991. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Hébert, L.** 2018. Elements of Semiotics. *Signo. Theoretical Semiotics on the Web*. Accessed at: <http://www.signosemio.com/elements-of-semiotics.asp> [13 August 2018].
- Lyons, J.** 1977. *Semantics. Volume 2*. Cambridge: Cambridge University Press.
- Murphy, M.L.** 2010. *Lexical Meaning*. Cambridge: Cambridge University Press.
- Peirce, C.S.** 1985. Logic as Semiotic: The Theory of Signs. Innis, E. (Ed.). 1985. *Semiotics. An Introductory Anthology*: 4-23. London/Johannesburg: Hutchinson.
- Sebeok, T.A.** 2001. *Signs: An Introduction to Semiotics*. Second edition. Toronto/Buffalo/London: University of Toronto Press.
- Sperber, D. and D. Wilson.** 1995. *Relevance, Communication and Cognition*. Second edition. Oxford: Blackwell.
- Van Dijk, T.A.** 1972. *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics*. Janua Linguarum: Studia Memoriae Nicolai Van Wijk Dedicata. Series Maior 63. The Hague: Mouton.
- Wiegand, H.E.** 1990. Printed Dictionaries and Their Parts as Texts. An Overview of More Recent Research as an Introduction. *Lexicographica* 6: 1-126.

- Wiegand, H.E.** 1996. A Theory of Lexicographic Texts: An Overview. *South African Journal of Linguistics* 14(4): 134-149.
- Wiegand, H.E.** 1996a. Textual Condensation in Printed Dictionaries. A Theoretical Draft. *Lexikos* 6: 133-158. DOI: <http://dx.doi.org/10.5788/6-1-1029>.
- Wiegand, H.E. and R.H. Gouws.** 2013. Addressing and Addressing Structures in Printed Dictionaries. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013: 273-314.
- Zgusta, L.** 1971. *Manual of lexicography*. The Hague: Mouton.

Corpus-driven Bantu Lexicography Part 1: Organic Corpus Building for Lusoga

Gilles-Maurice de Schryver, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of African Languages, University of Pretoria, Pretoria, South Africa (gillesmaurice.deschryver@UGent.be)*

and

Minah Nabirye, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of Teacher Education and Development Studies, Kyambogo University, Kampala, Uganda (minah.nabirye@UGent.be)*

Abstract: This article is the first in a trilogy that deals with corpus-driven Bantu lexicography, which is illustrated for Lusoga. The focus here is on the building of a so-called 'organic corpus' from scratch, while the next two instalments will deal with the use of that corpus on the macro-structural and microstructural levels, respectively. Not many detailed descriptions of corpus-building efforts exist for Bantu languages, so each and every step is discussed in detail, paying particular attention to the parameters that have to be taken into account, while not losing sight of the need to log the metadata either.

Keywords: BANTU, LUSOGA, CORPUS BUILDING, ORGANIC CORPUS, ORAL, WRITTEN, SOURCE, PERIOD, GENRE, TOPIC, METADATA

Obufunze: Omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu. Ekitundu 1: Okuzimba namukyukilo w'eitu ly'Olusoga. Olupapula luno n'olusooka ku isatu edhinaayogela ku musomo gw'omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu nga gulaga omulimu ogw'akolebwa ku Lusoga. Mu lupapula luno, eisila lili ku nzimba ya itu namukyukilo okuva ku ntandiiko. Ebitundu ebinaaba mu lupapula olw'okubili n'olw'okusatu biidha kugema ku nkozesa ya itu lino ku isa ly'omutindiigo ogw'ebizimbibwa mu mutegeko n'eisa elilaga eitu lino mu mwoleko ogw'azimbibwa mu mutindiigo n'engeli omusingi ogulimu bwe gulagibwa mu iwanika. Mu nnimi dha Bantu, emilimu egilaga omusingi guno tigitela kuwandiikibwaku mu butongole okusobola okumanhisa abo abayinza okuba nga bagasibwa. Kale buli kitundu ekiteesebwaku mu nnambika eli mu mpapula eisatu dhino kitoolayo buli kanhomelo ka bukodyo n'emitendela egy'agobelebwa ela gy'akozesebwa mu kusenvula omulimu gw'okuzimba omutimbo gw'ekyebungo ky'olulimi Olusoga gwonagwona.

Ebigambo ebikulu: BANTU, LUSOGA, OKUZIMBA EITU, EITU NAMUKYUKILO, ENDHOGELA, EMPANDIIKA, OBUVO, EKISEELA, ENNAMBIKA, EKINHUMYO, OMUTIMBO GW'EKYEBUNGO

1. Goal of the present study

In this article we wish to show how an electronic corpus for a Bantu language, especially an under-resourced Bantu language, may be assembled from scratch. We have lexicographic applications in mind, but such corpora may also be used (and *have* successfully been used) for Bantu corpus linguistics studies more generally. While Bantu corpora have been built for about two decades now, explicit descriptions of their composition are rare in the literature. For instance, in his MA dissertation de Schryver (1999: 103-117) devotes about 14 pages to the design, structure, contents and text collection of a 300 000-word Cilubà corpus, but to this date that study remains unpublished. When it comes to the descriptions of the corpora that have been assembled for the South African Bantu languages, these are typically less than a page long (de Schryver and Prinsloo 2000). On the other hand, corpus stability tests have been carried out for the South African Bantu languages (Prinsloo and de Schryver 2001, Prinsloo 2015), as well as attempts at multilingual corpus building and multilingual data extraction (de Schryver 2002, Prinsloo and de Schryver 2005). Scientific articles on the Zimbabwean corpora built under the umbrella of ALLEX/ALRI tend to focus on specific topics, such as tagging issues for a Shona corpus (Chabata 2000) or the sociolinguistic, political and economic considerations that influence the contents of a corpus of Zimbabwean Ndebele (Hadebe 2002). Even the latest version of the widely-used *Helsinki Corpus of Swahili* is not accompanied by a proper description (Hurskainen 2016).

The only exceptions to this pattern seem to be the corpora built to carry out corpus linguistics studies at BantUGent (i.e., the UGent Centre for Bantu Studies) where, for instance, the PhDs of Mberamihigo (2014), Nshemezimana (2016) and Misago (2018) describe the various Kirundi corpora built, or where the PhD of Kawalya (2017) describes the Luganda corpus that he used for his study. The building of a Lingála corpus may be found in the PhD of Sene-Mongaba (2013), reworked and expanded as Sene-Mongaba (2015). Our effort (Nabirye 2016), on which the Lusoga case study presented below is based, is also the result of PhD research undertaken at BantUGent.

With regard to corpus-building efforts for Lusoga, only one exploratory study has appeared so far (Nabirye and de Schryver 2011). In that study, the main focus was on the writing problems that the corpus builder encounters during the transcription of oral material and the implications for the corpus lexicographer when data is extracted from such a corpus. In contrast, of particular interest in the present study will be the parameters/axes that can be used to characterise the composition of a Bantu-language corpus, these being, in addition to oral vs. written, also the distribution of the sources, the periods, the genres and the topics. Orthographic issues will only briefly be recapped here. Furthermore, the value of detailed corpus documentation will be exemplified; this will be done by means of the inclusion of and reference to a comprehensive addendum. Corpus-query software will be mentioned in passing.

2. The Lusoga language and publications in Lusoga

Lusoga is a largely undocumented Great Lakes Bantu language classified as JE16 (Guthrie 1948, Maho 2009). According to the Uganda Bureau of Statistics, 2 062 920 people identified themselves as Basoga in 2002 (UBOS 2006: 12), a figure that grew by nearly half to a respectable 2 960 890 by 2014 (UBOS 2016: 71). While immediately acknowledging that not all people who claim to be Basoga also necessarily speak 'Lusoga', however defined,¹ one should still realise that several million people currently speak Lusoga, of which about two million are monolingual. While it might surprise that a language with up to three million speakers may be largely undocumented, it is fitting to recall that there are even endangered languages with millions of speakers (Adelaar 2014).

Lusoga was first reduced to writing near the end of the 19th century, as pointed out by Condon a century ago:

The Basoga Batamba had no written characters. Nor do any writings on rocks or pictorial characters exist. According to native report — and I mean natives of a ripe old age — there never was, as far as they remember, any means whatever of placing down their verbal utterances. All messages from one chief to another were committed to a trustworthy man, who learned the communication by heart, and so delivered the message by word of mouth. It is only within the last 15 years that the language of this people has been put in book form. (Condon 1911: 368)

The very first language data for Lusoga may be found in the 'vocabularies' included in Johnston (1902: 980-991) as well as in Condon (1911). However, we have found no evidence to suggest that Lusoga was documented in earnest prior to the 1960s. The earliest reference uncovered so far with an exclusive focus on Lusoga is the orthography of Byandala (1963). That booklet was followed by the documentation of Lusoga proverbs and riddles in Lyavala-Lwanga (1967, 1969). There is no record of Lusoga materials produced during the 1970s or the 1980s. Writing on and in Lusoga was again picked up in the 1990s. The first Lusoga publication in this period was the second version of the Lusoga orthography: Kajolya (1990). It was followed by two attempts at publishing a newspaper, which faltered shortly after: *Kodh'eyo* (1997–98) and *Ndimugezi* (1998–99). From the late 1990s and early 2000s onwards, the main output in Lusoga has come from the *Cultural Research Centre* (CRC), a religious body based in Jinja (e.g., CRC 1998a, 1999a, b, c, d, e, f, g, h, 2000a, b, 2002, 2005a, Kaluuba et al. 2010, CRC 2011).² Also, one very prolific writer is Gulere who, amongst others, self-published ten children's story books, which he placed online in various locations at various times and in various formats (Gulere 2011a, b, c, d, e, f, g, h, i, j). Gulere moreover self-published two translations, one of *Antigone*, a tragedy by the ancient Greek playwright Sophocles from 441 BC (Gulere 2007a), another of *The Bride*, a play in English by the Ugandan Austin L. Bukonya from 1987 (Gulere 2007b).³ Lastly, a first novel has now been published in Lusoga, written by Kuunya (2011a).

3. Building a corpus for Lusoga

3.1 Towards an organic (but structured), general-language, synchronic Lusoga corpus

The basics of corpus building for the Bantu languages have been described by de Schryver and Prinsloo (2000). The two important concepts that also applied to the building of our Lusoga corpus are that of an 'organic corpus' and that of a 'structured corpus'. An 'organic corpus' has been defined by Atkins, Clear and Ostler as follows:

[...] a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing, living language. [...] In order to approach a 'balanced' corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weaknesses identified and reported. In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually. [...] 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. (Atkins et al. 1992: 1, 4, 6)

De Schryver and Prinsloo link this to what they call a 'structured corpus' as follows:

Formulated differently, it is any corpus compiler's task to attempt to assemble a representative corpus for his/her specific need(s). Subsequent additions and deletions of sections should be seen as a balancing activity to rectify initial weaknesses, but more importantly, also to take account of and track a growing, living language. As such, there is no such thing as 'the' corpus of a certain language (variety). Rather, at any point in time one selects a certain number of texts from the range of available electronic texts (which might or might not be grouped together into sub-corpora), and uses 'a' corpus for the specific research one wishes to pursue. The minimum requirement for any organic corpus is thus that the corpus compiler(s) will have attempted to put some structure in assembling the range of electronic texts. Within this framework, any first attempt at compiling an organic corpus will at least result in a structured corpus. (de Schryver and Prinsloo 2000: 92)

Our Lusoga corpus is both structured and organic. On the whole, the organicity means that the overall size has increased and decreased over the years.

Corpus building for the Bantu languages is always slightly opportunistic, in that one adds the little existing written material one can get hold of, except when a serious imbalance results. In other words, to get going, one often makes do with an 'imperfect corpus', which is then modified later on, when 'better' data becomes available. Over and above this balancing act, the corpus used

should always attempt to be representative of the population that is the subject of the planned description or research. For a general-language corpus, the goal is consequently to acquire as many different genres as possible, that deal with as wide a topic range as possible. Existing written material for all but a few Bantu languages is unfortunately biased in this respect. Most are the result of (modern) missionary activities, so the genre *Biblical documents* tends to be overrepresented in many Bantu corpora. Conversely, for Bantu languages with a varied, vibrant and ongoing online media presence, the genre *Journalism* may be overrepresented, and within that, topics such as *Sports* and *Politics*. Of course, when the aim is to describe features of biblical works or journalistic texts, then such types of corpora may indeed be 'representative', and when multiple sources have been equally sampled, these corpora may also be 'balanced'. But if the goal is to describe the general language, then an effort needs to be made to achieve both representativeness and balance in another way. It is here that the material found in the oral component of a corpus may bring a solution, as it did for our Lusoga corpus (cf. *infra*, §3.5.1).

Another important point concerns the time period covered by a Bantu corpus. In all but a few cases, this will be 'the present', with that present optionally stretching back to a number of decades, maximum half a century. Although attempts are being made to build Bantu corpora with time-depths of at least half a century down to a century — such as for Zulu (de Schryver and Gauton 2002), Kirundi (Mberamihigo et al. 2016) and Luganda (Kawalya et al. 2018) — the only Bantu corpus containing substantial amounts of diachronic data that has been built (and used)⁴ is the set of corpora for the *Kikongo Language Cluster*, where some parts are up to four centuries old, while others go back to around 250 years ago (Bostoen and de Schryver 2015). For Lusoga, the aim has always been to build a synchronic corpus covering the general language. Material older than a few decades is in any case extremely rare for Lusoga (cf. *supra*, §2). When available, it was nonetheless included in an attempt to widen the genre/topic range.

3.2 The 0.5m Lusoga corpus

A first Lusoga corpus, of about half a million words, was built as part of the research leading to an MA dissertation. Its composition is as shown in Table 1 (adapted from Nabirye (2008: 70)).

Table 1: Genre distribution in the 0.5m Lusoga corpus

Genre	Tokens	%
Journalism (<i>Kodh'eyo</i> , <i>Ndimugezi</i>)	187 393	34.84%
Biblical documents (<i>New Testament</i> and others)	199 853	37.16%
Short stories and idioms (<i>Kintu</i> , <i>Ababita Ababiri</i> , etc.)	150 560	28.00%
SUM	537 806	100.00%

3.3 The 0.9m Lusoga corpus

For a corpus-based study of the Lusoga noun (de Schryver and Nabirye 2010) the Lusoga 'MA corpus' was supplemented with the full text of the *Eiwanika ly'Olusoga* (Nabirye 2009), being a monolingual Lusoga dictionary compiled without the use of a corpus. The reasoning at the time was that because the example sentences from that dictionary were the result of original fieldwork, they could as well form part of a Lusoga corpus. A number of reports written in Lusoga (from the Busoga clan leaders, the private sector, academia, etc.) were also added, as was the initial impetus for a true oral part of the Lusoga corpus (i.e., the first few transcriptions of conversations, interviews and songs). The make-up of this Lusoga 'noun corpus' is as shown in Table 2 (taken from de Schryver and Nabirye (2010: 100)).

Table 2: Genre distribution in the 0.9m Lusoga corpus

Genre	Tokens	%
Reference work (<i>Eiwanika ly'Olusoga</i>)	305 660	35.00%
Journalism (<i>Kodh'eyo, Ndimugezi</i>)	187 393	21.46%
Biblical documents (<i>New Testament</i> and others)	199 853	22.88%
Reports (from the Busoga clan leaders, private sector, academia, etc.)	24 166	2.77%
Short stories and idioms (<i>Kintu, Ababita Ababiri</i> , etc.)	150 560	17.24%
Transcriptions of conversations, interviews and songs	5 716	0.65%
SUM	873 348	100.00%

This version of the Lusoga corpus contained about 870 000 running words (tokens), and about 150 000 orthographically different words (types). Not only the transcriptions of conversations, interviews and songs but also the dictionary examples (together close to a third of the total) could be considered reductions of spoken data to text; the other genres being written texts from the start. From Table 3 (also taken from de Schryver and Nabirye (2010: 100)) one may further deduce that most sources are recent to very recent, with over 98% produced during the past two decades.

Table 3: Period distribution in the 0.9m Lusoga corpus

Period	Tokens	%
1960s	16 822	1.93%
1970s	—	—
1980s	—	—
1990s	457 978	52.44%
2000s	398 548	45.63%
SUM	873 348	100.00%

3.4 The 1.1m Lusoga corpus

Following the Lusoga noun study, and with the acquisition of more data to compensate for it, the dictionary data was again dropped from the Lusoga corpus. Although based on natural language production, the dictionary examples lacked the original context, and had in any case been 'selected' for their pedagogical value. As such, they did not have their place in a proper text corpus, that is, one that consists of large sections of free-flowing, running text. Instead, the symbolic oral section of about 6 000 tokens in the Lusoga 'noun corpus' was enlarged to well over 400 000 tokens. Furthermore, various texts translated from English, as well as digital-born Lusoga material, were also added, to obtain the corpus that was used for the study of the writing problems in a Lusoga corpus (Nabirye and de Schryver 2011). The composition of that new corpus is as shown in Table 4 (adapted from Nabirye and de Schryver (2011: 123)).

Table 4: Genre distribution in the 1.1m Lusoga corpus

Genre	Tokens	%
Journalism (<i>Kodh'eyo, Ndimugezi</i>)	187 393	17.07%
Biblical documents (<i>New Testament</i> and others)	199 853	18.20%
Reports (from the Busoga clan leaders, private sector, academia, etc.)	24 166	2.20%
Short stories and idioms (<i>Kintu, Ababita Ababiri</i> , etc.)	150 560	13.71%
Transcriptions of conversations, interviews and songs, as well as traditional ceremonies, speeches, sermons, radio broadcasts, etc.	413 827	37.69%
Translations from English (<i>PEAP (Poverty Eradication Action Plan), ICEE (International Centre for Eye Education), FIDA/PLAN (inheritance laws)</i> , etc.)	19 814	1.80%
Electronic texts (e-mails, mailing lists, <i>Facebook</i> , etc.)	102 365	9.32%
SUM	1 097 978	100.00%

This 1.1m Lusoga 'writing-problems corpus' — just as the earlier 0.9m Lusoga 'noun corpus' and the even earlier 0.5m Lusoga 'MA corpus' — was not annotated for any linguistic features. As such, these corpora were not tagged for parts of speech, nor lemmatised. They are known as 'raw corpora'.

3.5 The 1.7m Lusoga corpus

The latest iteration of the Lusoga corpus stands at over 1 700 000 tokens and about 200 000 types. The various text files of the 1.1m Lusoga 'writing-problems corpus' were cleaned up, re-assembled and renamed. New material was added for each genre except *Journalism*. For the latter, however, all the newspa-

per clippings were reprocessed with better software (cf. *infra*, §3.5.2). It is this version of the Lusoga corpus that we will now study in more detail.

3.5.1 Oral vs. written distribution

In contrast to the 0.5m Lusoga corpus, which had no transcribed text, and the 0.9m one with just 5 716 such tokens, a major effort in building the 1.7m Lusoga corpus went to expanding the oral component even further compared to the 1.1m Lusoga corpus. While the model of all modern corpora, the 100m *British National Corpus* (BNC 1994–2018), has set the standard for general-language corpora to contain 10% spoken material vs. 90% written material (Rundell and Stock 1992: 46), we managed to triple this conventional allocation of the spoken part in the total. In all, 216 audio files were transcribed, amounting to well over half a million tokens, as may be seen from Table 5, which corresponds to 31% of the total corpus, illustrated graphically in Figure 1.

Table 5: Statistics for the oral vs. written distribution in the 1.7m Lusoga corpus

Medium	No. of files	%	Tokens	%
Oral	216	55.24%	541 129	31.39%
Written	175	44.76%	1 182 562	68.61%
SUM	391	100.00%	1 723 691	100.00%

Oral vs. Written in the 1.7m Lusoga corpus

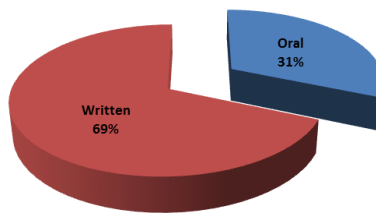


Figure 1: Pie chart showing the oral vs. written distribution in the 1.7m Lusoga corpus

There is nothing magic about attaining over half a million words of spoken data,⁵ nor about reaching a division of a third for oral vs. two-thirds for written data, but for a language which to this date is chiefly an oral language, it simply looked like a necessity in order to ensure that any explanations drawn from this corpus would also reflect real language usage. The oral component is sizeable enough so as to feature in every screenful of concordance lines, where oral and written material is instantly juxtaposed and may be cross-compared to

make sure there are no differences between oral vs. written language use that would need to be reported.

What is true is that there is an addictive aspect to corpus building, so a goal was set to reach about '100 hours of audio'. Indeed, the 541 129 tokens of transcribed material correspond to exactly 98 hours, 42 minutes, and 38 seconds of audio files. Transcribing half an hour of audio took on average two hours, which means that 400 hours were required for all the transcriptions (not counting the fieldwork and hours spent recording in the first place, nor the many hours to collect and log all the metadata and consent forms). The types of audio recorded and transcribed are varied, and include modern and traditional songs, radio talk shows, traditional ceremonies (as currently being performed), business meetings, interviews and dialogues.

3.5.2 Source distribution

The bulk of the written part of the 1.7m Lusoga corpus was assembled through the digitization of more or less every work, down to every snippet, ever written and published in Lusoga, whether commercially or produced as grey literature. A total of 85 sources were scanned in high resolution, after which the optical character recognition (OCR) tool of OmniPage (1995–2018) was utilised to turn the images into machine-readable texts.⁶ These 85 sources were good for about 670 000 tokens. OCR was also used to re-digitise large parts of the two short-lived Lusoga newspapers: *Kodh'eyo: Busoga etebenkere* (Kodh'eyo 1997–98) and *Ndimugezi n'omukobere: The factfinder* (Ndimugezi 1998–99). Due to the poor quality of the printing of these newspapers, the OCR output required substantial clean-up. The result was about 200 000 tokens of newspaper articles. A further 62 files were obtained electronically. These included self-published works found on the Internet, unpublished material from friends, private e-mail and mailing list communications, translations into Lusoga taken from government, NGO and commercial websites, as well as some religious material found online. All these texts together came to about 260 000 tokens. The translations we ourselves had made over the years, 15 of them, were also added, which contributed a further 25 000 tokens, as well as some of our own writings, six texts with just 2 500 tokens. The remainder consisted of low-resolution images of texts found online, as well as a single hand-written document, which were all retyped, adding another 25 000 tokens.

An overview of these various sources may be seen in Table 6. For a mostly undocumented and oral language like Lusoga, we must admit that we never expected to be able to reach nearly 1.2m tokens of material that had been written in one way or another. Extending the corpus building effort beyond the more obvious transcriptions and OCR, as seen in the last five bullets of Table 6, clearly helped in this regard (and in effect resulted in about a quarter of the written data).

Table 6: Statistics for the source distribution in the 1.7m Lusoga corpus

Source	No. of files	%	Tokens	%
ORAL				
• Transcriptions of audio	216	55.24%	541 129	31.39%
WRITTEN				
• OCR (optical character recognition)	85	21.74%	669 320	38.83%
• OCR + Retyping	2	0.51%	201 664	11.70%
• Electronic transfers	62	15.86%	258 990	15.03%
• Translations	15	3.84%	25 365	1.47%
• Own writings	6	1.53%	2 568	0.15%
• Retyping of images	4	1.02%	24 436	1.42%
• Retyping of hand-written document	1	0.26%	219	0.01%
SUM	391	100.00%	1 723 691	100.00%

3.5.3 Period distribution

As may be seen from the data presented in Table 7 and the bar chart shown in Figure 2, the 1.7m Lusoga corpus is essentially a synchronic corpus with a time-depth of just over 20 years.

Table 7: Statistics for the period distribution in the 1.7m Lusoga corpus

Period	No. of files	%	Tokens	%
1940s	1	0.26%	1 325	0.08%
1950s	—	—	—	—
1960s	2	0.51%	36 065	2.09%
1970s	—	—	—	—
1980s	1	0.26%	16 657	0.97%
1990s	44	11.25%	417 837	24.24%
2000s	139	35.55%	398 153	23.10%
2010s (to 2013)	204	52.17%	853 654	49.52%
SUM	391	100.00%	1 723 691	100.00%

Only four files represent the 1940s, 1960s and 1980s.⁷ The 1990s and 2000s are equally represented, with about 400 000 tokens each, while the 2010s (and only up to August 2013 at that) is represented by as many as 850 000 tokens. While each of the past two periods and the present one cover both oral and written material, up to 70% of the transcriptions concern spoken data from the 2010s, which is the main reason why the 2010s contain more material than any other period. Another is the flurry of primers that were produced in the 2010s, in the

wake of the recognition of Lusoga as a medium of instruction in 2005 (NCDC 2006: 5).

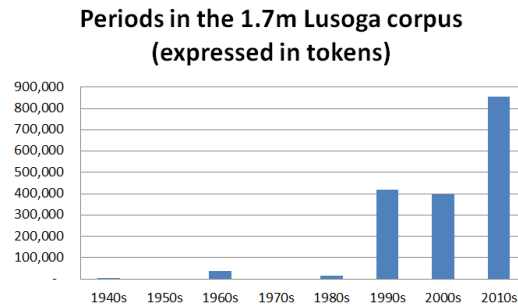


Figure 2: Bar chart showing the period distribution in the 1.7m Lusoga corpus

3.5.4 Genre distribution

The 391 files of the 1.7m Lusoga corpus were also grouped into 12 broadly-defined genres, as summarised in Table 8 and shown graphically in Figure 3. Three genres dominate, making up more than half the corpus: *Biblical documents* (23% of the tokens),⁸ *Literature* (16%) and *Radio talk shows* (15%). Also sizable are *Journalism* (12%) and *E-mails* (9%). Each of the next five genres contains about a twentieth (5%) of the total corpus: *Policy documents*, *Interviews*, *Songs*, *Celebrations*, and *Academic documents*. *Newsletters* and *Advertisements* each represent less than 1% of the total.

Table 8: Statistics for the genre distribution in the 1.7m Lusoga corpus

Genre	No. of files	%	Tokens	%
Biblical documents	44	11.25%	388 026	22.51%
Literature	64	16.37%	271 701	15.76%
Radio talk shows	41	10.49%	265 726	15.42%
Journalism	2	0.51%	201 664	11.70%
E-mails	18	4.60%	153 563	8.91%
Policy documents	19	4.86%	101 029	5.86%
Interviews	11	2.81%	94 693	5.49%
Songs	155	39.64%	86 028	4.99%
Celebrations	10	2.56%	82 138	4.77%
Academic documents	19	4.86%	68 662	3.98%
Newsletters	6	1.53%	10 027	0.58%
Advertisements	2	0.51%	434	0.03%
SUM	391	100.00%	1 723 691	100.00%

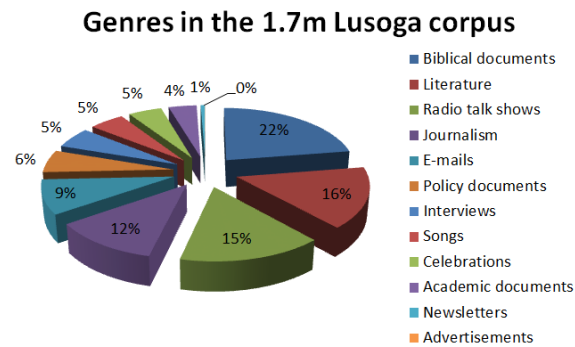


Figure 3: Pie chart showing the genre distribution in the 1.7m Lusoga corpus

3.5.5 Topic distribution

The different files in the Lusoga corpus were also grouped into 18 broadly-defined topics. To do so, related subjects were brought together, such as:

Health:

- Health planning
- Ill-health & death
- Rural health management
- Traditional healing
- AIDS scourge
- Eye-care education ...

Inspirational:

- Self-appreciation
- Jubilation
- Honouring activity
- Hope message
- Graduation ceremony ...

Even though a strict division between genre and topic is not always possible, and even though some files actually deal with various topics, the data shown in Table 9 may be considered to be a good approximation of the actual topics covered in the corpus.

While a quarter of the Lusoga corpus deals with *Religion*, the inverse also means that three-quarters does not, which is fine given the usual bias in Bantu-language corpora. The topic *Networking* actually covers such varied items as newspaper texts, mailing-list messages, songs about networking, and even advertisements. The other topic labels are self-explanatory. The data is shown graphically in Figure 4.

Table 9: Statistics for the topic distribution in the 1.7m Lusoga corpus

Topic	No. of files	%	Tokens	%
Religion	55	14.07%	439 915	25.52%
Networking	22	5.63%	355 761	20.64%
Health	41	10.49%	153 588	8.91%
Language ⁹	26	6.65%	126 449	7.34%
Sensitization	35	8.95%	102 061	5.92%
Politics	16	4.09%	97 785	5.67%
Fables	41	10.49%	92 470	5.36%
Marriage	36	9.21%	79 839	4.63%
Life	19	4.86%	75 237	4.36%
History	12	3.07%	59 739	3.47%
Proverbs	5	1.28%	45 556	2.64%
Inspirational	13	3.32%	31 443	1.82%
Science	5	1.28%	19 784	1.15%
Riddles	3	0.77%	15 694	0.91%
Relationships	31	7.93%	12 203	0.71%
Rehabilitation	13	3.32%	7 400	0.43%
Money	13	3.32%	6 469	0.38%
Gratitude	5	1.28%	2 298	0.13%
SUM	391	100.00%	1 723 691	100.00%

While the percentages for each of the broadly-defined topics as seen in Figure 4 may or may not reflect the actual allocation to each of these topics in the way Lusoga is used by millions of speakers on a daily basis in Busoga, what is relatively certain is that the coverage of the range and variation is rather wide in the 1.7m Lusoga corpus.

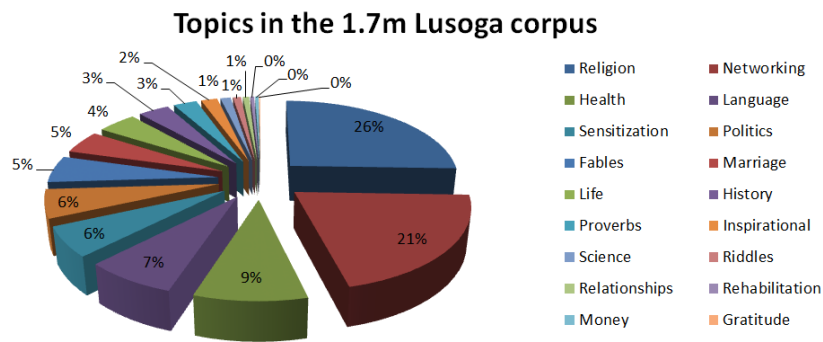


Figure 4: Pie chart showing the topic distribution in the 1.7m Lusoga corpus

3.5.6 The orthography in the corpus

Important to observe at this point is that the various orthographies as seen in the original written sources were left intact. Bar a few exceptions, there are no tone markings in the corpus.

This implies that the stated number of types (i.e. the orthographically unique words) is always slightly inflated compared to a corpus in which the spelling would have been homogenised. Working with a corpus that contains various spellings for some of the same words is not an insurmountable hurdle; it only means that one is dealing with some (evenly spread) noise as far as the type counts are concerned; the token counts, however, are (mostly) correct.

Although a number of Lusoga orthography guides exist, one must conclude that they did not have much impact on helping the different authors streamline their writing in Lusoga. But then, the majority of the texts which are now in the corpus were not necessarily meant for formal usage, so their authors did not adhere to a strict application of any orthographic rules. For example, biblical prayer books are in-house documents that are only employed for the purposes of religious teaching. The different short stories and the novel in Lusoga have all been produced informally and are written in a style that the authors feel is most appropriate at the time of writing. E-mails and website texts in Lusoga display a severely unregulated use of written Lusoga. Also, the type of written Lusoga found in this category of sources is often mixed with English. In addition, Lusoga is borrowing sounds from neighbouring languages, such as the palatal nasal [ɲ] which is not an indigenous Lusoga sound. One also notices a switch between the voiced labio-velar approximant [w] and the velar fricative [ʁ]; and the fact that the Lusoga dental sounds are being relegated to neighbouring alveolar sounds (which are easier to pronounce for non-native speakers). Most prominent is an ongoing discussion on whether Lusoga really has a trill [r], only a flap [ɾ], or neither of the two — which results in inconsistent uses of /r/ and /l/ in the orthography.¹⁰

Instances of orthography-based problems in writing Lusoga are shown in examples (1) & (2). (For the abbreviations in the glosses, see the explanations at the end.)

- (1) *enhyandhula* instead of *ennhandhula* 'introduction'
okuhwunga instead of *okuwunga* 'to catch an object mid-air'
cyatulirwa instead of *kyatulilwa* 'it is pronounced/spoken'
 [File ID: KiyinKbi | W • Literature • Language • 1969]

- (2) a. *Me Enterprise development oyinza okufuna bakakensa [...]*
me ... **o-yinza** **oku-fun-a** **ba-kakensa**
 CON ... SM_{2SG}-can 15-get-FV 2-expert
 But for Enterprise development you can get experts [...]
 [File ID: Mail1306 | W • E-mails • Networking • 2013]

- b. *What did I ng'omuntu do?*
... **nga** **o-mu-ntu** ...
... adv AUG-1-person ...
What did I as a person do?
[File ID: Mail1306 | W • E-mails • Networking • 2013]

In the examples in (1) the author decided to write the dental nasal as /nhy/, the voiced labio-velar approximant as /hw/, and the voiceless palatal plosive as /c/, as well as making distinctions in writing the trill after /i, e/ and the lateral flap after /u/ and /a/. The orthographic problems seen in examples of this nature seem to arise out of a need to use a phonetic-inspired orthography. Such orthographic interpretations may simply be idiosyncratic improvisations made in the absence of a proper (and popular) phonetic description of the sounds of Lusoga.

On the other hand, the examples in (2) reflect a user who is continuously code switching, and missing out on a few basic grammatical forms in the writing system. This is probably due to ignorance or the lack of a proper grounding in writing Lusoga.

The type of issues seen in the two examples can be generalised as occurring rather often in the informal written texts included in the corpus. While the spelling of the original texts was left intact, recognition errors might have been introduced during the OCR process, with some of the letters being machine unreadable and interpreted differently, even though we did our utmost to read through the OCRed material.

It is also probable that some 'errors' were introduced during the transcription process: while we tried to steer away from it, there was a tendency to 'over-correct' misspoken sections and hesitations, as the goal of our corpus-building efforts is not to use the material for, say, sociolinguistic studies of detailed turn-taking, but to use the material to uncover language as it was meant to be (Hanks 2012: 416).¹¹

We do trust that these 'inconsistencies' and 'errors' have not obscured the proper usages of Lusoga.

3.5.7 Querying the corpus

The 391 files of the 1.7m Lusoga corpus are stored as *plain text files*, and as such this 1.7m Lusoga corpus is also a 'raw corpus'. Raw corpora may successfully be searched using off-the-shelf corpus-query software like WordSmith Tools (Scott 1996–2018). WST was indeed used in this way to present the various corpus counts above, and will also be used for the macrostructural and microstructural illustrations in the next two parts of this set of three articles.

However, and as we will explain in Part 2, the 1.7m Lusoga corpus was also part-of-speech tagged and lemmatised for lexicographic purposes. Either or even both of these levels (i.e., the part-of-speech labels and/or the lemmas of

each orthographic word) may also be added as tags to all (or part of the) 1.7m tokens of the Lusoga corpus. Software such as WST is able to handle such *marked-up text files* as well.

3.5.8 Corpus file IDs, corpus filename bibliography and corpus metadata database

As could be seen in examples (1) and (2), for material excerpted from the corpus, it is good practice to mention the source from which it was taken. In (1) this information was presented following all the examples, and in (2) this was done on the line following the interlinear glossing and translation. In all cases, the corpus details are presented between square brackets.

In actual fact, for all material that is quoted from a corpus, whether for lexicographic purposes or more generally in corpus linguistics, three distinct levels of supplementary information may be provided for each source. At the quoted material itself a *File ID* may be provided, together with 'minimal information', here on whether the treated example is either taken from the written or the oral section of the corpus, and further information on the genre and topic, as well as the year or period, in the following format:

at examples

[File ID: Filename W(ritten) or O(ral) • Genre • Topic • Year or Period]
--

The *Filename* also serves as the entry point to Addendum 1, where further details on each source may be found. The author (or for audio, performer) as well as the title of the work (either as published or as given by us), the number of types and tokens for the work, the source of the work, the place of publication and publisher, as well as the number of pages of the work (or for audio, length of the recording) are all provided in that addendum. The format used for the twelve slots of information in Addendum 1 is always as follows:

in Addendum 1

Filename Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
--

For instance, the *Filename* for (1) above reveals the following in Addendum 1:

KiyinKbi Lyavala-Lwanga, E.J. 1969 (Kiyini Kibi) • Literature / Language • 19,256 / 7,737 • W ~ Retyping of image • Kampala: Milton Obote Foundation • 123 pp.

This type of information includes what one would find in a traditional bibliography (before the first bullet, after the penultimate and last bullets), but adds corpus-specific information to that (all the rest in-between).

in the corpus metadata database

Addendum 1 is an extract from a larger database, which, for the written sources and when relevant, also includes the translator and date of translation, as well as the edition number and year of original publication. For the oral material, that database additionally includes the date of the recording, and the names of the recorders and transcribers. Lastly, for each source the standardised type-token ratio (with a base of 1 000) and the standard deviation thereof are also given.¹²

A notes field is used for any additional information that needs to be mentioned. This *corpus metadata database*, which brings together all the metadata of the corpus in a structured format, is available electronically and may be consulted at BantUGent together with the corpus itself.

3.5.9 Original data database

While it is feasible to store all of the 391 files in one single folder, much more intelligent is to arrange the files into various folders and sub-folders, for instance reflecting the different genres (12 sub-folders) or topics (18 sub-folders). How this is organised for a particular corpus depends on the use that will be made of that corpus. Another division could be oral vs. written, or the use of sub-folders that reflect the different time periods in the corpus, or even combinations of all of the above using tiered sub-folders. What has furthermore proven to be very useful is to keep several copies of the corpus at hand: in each, one finds the same data, but structured differently.

What is of paramount importance, however, is to keep a parallel version of one of these corpus structures in a different (off-site) location, where all the *original files* are kept. There the original audio (.wav, .mp3, ...) and at times even videos (.mp4, .webm, ...) are stored, as well as the original web pages (.htm, .html, ...), documents (.doc, .pdf, ...) and images (.jpeg, .png, ...). Temporary files such as those used to turn scanned material ('image pdfs') with OCR software (e.g., .opd) into machine-readable images ('searchable pdfs') should also be kept there. This parallel version of the corpus, or *original data database*, not only functions as a backup from which the corpus files could be regenerated whenever this would prove to be necessary, but it is also the first place to go to whenever in doubt about a certain transcription (audio) or the orthography in an automatically-recognised (written) work. Published texts, with their formatting, and multimedia files furthermore contain more information than the text (.txt) versions in the corpus, which may at times and for certain purposes be useful to consult.

4. Discussion

In this article we have given a detailed description of the building of a general-language corpus for Lusoga, an under-resourced Bantu language. We showed that it is indeed possible to reach a substantial size, in this case 1.7 million tokens, a third of which consists of oral data, even though the building of this corpus has basically been a one- to two-person effort. This stands in sharp contrast to for instance the ALLEX/ALRI corpora, for which scores of students were sent into the field and as many were enlisted to transcribe the recordings.

Our corpus is an 'organic corpus', as material has not only been added over the years, but some of it has also been taken away, while still other parts were replaced after being reworked. Merely having more data does not necessarily mean one has better data, as one should keep an eye on balance as well. In the overview presented in the present article, the 1.7m Lusoga corpus is a 'raw corpus', in that it has not been annotated; but it was pointed out that with the results from Part 2, part-of-speech tags and/or lemma tags could enrich this corpus linguistically.

We also illustrated the importance of *knowing* one's corpus, not only in terms of the oral vs. written distribution, but similarly with regard to the distribution of the sources, periods, genres, and topics. Variations on our presentation are of course possible, and indeed in the PhDs of Mberamihigo (2014), Nshemezimana (2016) and Misago (2018) for Kirundi, as well as the PhD of Kawalya (2017) for Luganda, three-dimensional graphs are shown in addition, the third dimension representing the diachronic aspects of their corpora. The point, however, is that a detailed description of a corpus is needed if one is to make intelligent use of it.

As the details in the addendum indicate, we further place particular importance on the metadata of a corpus. Metadata may evidently be put to good use when actually *using* a corpus: for lexicographic ends, but also far beyond in the wider discipline of linguistics. There are no doubt differences between the spoken and the written forms of a language, and certain phenomena may be realised slightly differently depending on the genre or topic, just as word use differs with register. Likewise, for differences in word use depending on the author or performer, or even the publishing house of a certain work (each with their own style guide and own approach to copy-editing), and so on. Sub-corpora may indeed be assembled along such lines.

Reformulated, depending on how one intends to use a corpus, all the categorisations given so far may play an important role. But they do not inform each study in the same way. Within the field of lexicography, the first two and main uses of a corpus have to do with the creation of the macrostructure of a dictionary on the one hand, and the compilation of the articles in the microstructure on the other. These two topics will now be looked into, and illustrated for Lusoga, in two follow-up studies.

Abbreviations

#	noun class number	FV	final vowel
ADV	adverb	SG	singular
AUG	augment	SM _x	subject marker (of cl. or person x)
cl.	class		
CON	connective		

Acknowledgements

The research for this article was funded by the Special Research Fund of Ghent University. Thanks are due to the two anonymous referees.

Endnotes

1. In our work Lusoga, as in all subsequent mentions of 'Lusoga corpus', narrowly refers to the Lutenga variety only (Nabirye et al. 2016).
2. At the CRC library in Jinja, a substantial amount of grey literature may also be found, either written by the CRC staff itself, or facilitated by them. These works are mostly for internal use, of a religious nature and typically do not have a stated publisher, but may be 'assigned' to the CRC (e.g., CRC 1998b, Kasozi 2000, CRC 2003a, b, c, 2005b, 2008, Wabugoyera et al. 2008, CRC 2010, 2012a, b, c, d, e, f, g). Other religious works often do not have publication years, such as Mwesigwa (s.d.), except for those published by The Bible Society of Uganda, for which, see Endnote 8. Lately, the CRC has begun re-jacketing earlier works, including CRC (2009) and Kaluuba and Korse (2010). The CRC also played a pioneering role in producing the first grammars for Lusoga (Korse 1999, CRC 2004, Wambi et al. 2005, Kuunya 2011b), the first bilingual Lusoga–English dictionaries (Korse 2000, Gonza 2007), new orthographies (LULANDA and CRC 2001, 2004), as well as readers (e.g., Gulere and Wambi 2011).
3. Gulere also compiled a bilingual Lusoga–English dictionary (Gulere 2009).
4. At BantUGent a diachronic corpus for Swahili with a time-depth of up to two centuries is under construction. Research articles have not yet been published, however, although preliminary results have been presented at conferences (Devos and de Schryver 2013, 2016).
5. While not magic, Rundell and Stock (1992: 46) refer to this part of a corpus as the 'Holy Grail': 'Truly spontaneous speech, however — the everyday conversation of ordinary members of the public — has so far been available only in very small quantities and for lexicographers this remains the "Holy Grail".'
6. In earlier descriptions of corpus building for the Bantu languages, some attention was paid to the type of OCR errors one needs to attend to (de Schryver 1999: 116). Today's OCR software is however so performant that all one needs to remember is that the letter combination read as 'm' should often be corrected to the single letter 'm'.
7. Observe that material for the 1980s was found after all, in an academic publication (Cohen 1986), following a memorable search (Nabirye 2016: 25-27). Although eventually published in 1986, this edited material is based on recordings made two decades earlier, in 1966–1967.

8. A late entrant — in the sense that it came too late to be added to the 1.7m Lusoga corpus (apart from the fact that it may not have been desirable for reasons of representativeness and balance) — is the full Bible in Lusoga, which became available in 2014 (BSU 2014). As is normally the case with biblical works, the full Bible (BSU 2014) incorporates the New Testament (BSU 1998) — published earlier and included in the 1.7m Lusoga corpus. The New Testament itself incorporated the even earlier Gospel of Mark (BSU 1996), which in turn incorporated the still earlier Chapters 4 and 5 of the same gospel (BSU 1994). After the New Testament was released, at least one other edition appeared, with the addition of the Psalms from the Old Testament (BSU 2011).
9. The topic *Language* mainly includes material about teaching the language of Lusoga and instructional material for Lusoga (written in Lusoga), as well as website texts and journal abstracts on Lusoga (written in Lusoga).
10. See Nabirye et al. (2016) for more on these phonetic issues.
11. Or, as Kennedy (1998: 82) writes: 'A transcription is an imperfect written approximation of a speech event which exists initially as a dance of air molecules. The level of delicacy or amount of detail in a transcription is [...] related to the use to which the transcription will be put'.
12. As defined by Scott (1996–2018) 'the *standardised type/token ratio* (STTR) is computed every *n* words as Wordlist goes through each text file. By default, *n* = 1,000. In other words the ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of your text or corpus. A running average is computed, which means that you get an average type/token ratio based on consecutive 1,000-word chunks of text. (Texts with less than 1,000 words (or whatever *n* is set to) will get a standardised type/token ratio of 0.)'.

References

- Adelaar, W.F.H. 2014. Endangered Languages with Millions of Speakers: Focus on Quechua in Peru. *JournalLIPP* 3: 1-12.
- Atkins, B.T.S., J. Clear and N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.
- BNC. 1994–2018. British National Corpus. Available online at: <http://www.natcorp.ox.ac.uk/>.
- Bostoen, K. and G.-M. de Schryver. 2015. Linguistic Innovation, Political Centralization and Economic Integration in the Kongo Kingdom: Reconstructing the Spread of Prefix Reduction. *Diachronica* 32(2): 139-185 + 13 pages of supplementary material online.
- BSU. 1994. *Mariko Omutwe Ogwokuna n'Ogwokutaanu mu Lusoga [A Selection from St. Mark's Gospel Chapters 4 and 5 in Lusoga]*. Kampala: The Bible Society of Uganda.
- BSU. 1996. *Mariko. Amawulire Amalungi mu Lusoga [The Gospel of Mark in Lusoga]*. Kampala: The Bible Society of Uganda.
- BSU. 1998. *Endagaano Empyaka [New Testament]*. Kampala: The Bible Society of Uganda.
- BSU. 2011. *Endagaano Empyaka ni Zabbuli [New Testament and Psalms]*. Kampala: The Bible Society of Uganda.
- BSU. 2014. *Baibuli. Ekibono kya Katonda. Omuli n'ebitabo ebyetebwa deuterokanoniko/apokurifa [Bible. The Word of God, Which Also has the Books Known as Deuteronomy/Apocrypha]*. Kampala: The Bible Society of Uganda.

- Byandala, G.I.** 1963. *The Lusoga Orthography*. Iganga.
- Chabata, E.** 2000. The Shona Corpus and the Problem of Tagging. *Lexikos* 10: 75-85.
- Cohen, D.W.** 1986. *Towards a Reconstructed Past: Historical Texts from Busoga, Uganda* (Union Académique Internationale, Fontes Historiae Africanae Series Varia III). Oxford: Oxford University Press (for the British Academy).
- Condon, M.A.** 1911. Contribution to the Ethnography of the Basoga-Batamba Uganda Protectorate, Br. E. Africa. Part 2. *Anthropos: International Review of Ethnology and Linguistics* 6(2): 366-384.
- CRC.** 1998a. *Kintu*. Jinja: Cultural Research Center.
- CRC.** 1998b. *Priestly Ordination of Rev. Richard Kayaga Gonza, Rev. Silvester Makwali*. Bugembe: Diocese of Jinja.
- CRC.** 1999a. *Ababita Ababiri*. Jinja: Cultural Research Centre.
- CRC.** 1999b. *Akatabo Akasooka ak'Enfumo edh'Abasoga*. Jinja: Cultural Research Center.
- CRC.** 1999c. *Amagezi Tigamalwayo*. Jinja: Cultural Research Center.
- CRC.** 1999d. *Ensambo edh'Abasoga*. Jinja: Cultural Research Center.
- CRC.** 1999e. *Mwidhe Tufume*. Jinja: Cultural Research Center.
- CRC.** 1999f. *Obufunvu Magezi*. Jinja: Cultural Research Centre.
- CRC.** 1999g. *Omuvangano mu Busoga*. Jinja: Cultural Research Centre.
- CRC.** 1999h. *Twire ku Butaka*. Jinja: Cultural Research Centre.
- CRC.** 2000a. *Enkabi Ekifiini mu Busoga*. Jinja: Cultural Research Centre.
- CRC.** 2000b. *Lwaki Abakazi Tibabeeda Mulambo*. Jinja: Cultural Research Centre.
- CRC.** 2002. *Ebikoiko eby'Abasoga*. Jinja: Cultural Research Centre.
- CRC.** 2003a. *Diocesan Family Day*. Iganga: Diocesan Printery.
- CRC.** 2003b. *The Priestly Ordination for Rev. Deacon Mbaziira Henry Jude, Rev. Deacon Musana Paul*. Bugembe: Diocesan Printery.
- CRC.** 2003c. *Priestly Ordination of Rev. Serapio Kasuura Wamara Araali*. Bugembe: Diocese of Jinja.
- CRC.** 2004. *A Lusoga Grammar* (Revision of Korse 1999). Jinja: Cultural Research Centre.
- CRC.** 2005a. *Ebindi kw'Idembe ery'Obw'omuntu mu Nsi Yoonayoona*. Kisubi: Marianum Publishing Company.
- CRC.** 2005b. *Priestly Ordination of Deacon Mwangi Simon Gitua, Deacon Mugabe Paschal Atwooki, Deacon Jenga Fred*. Jinja: Little Sisters of St. Francis.
- CRC.** 2008. *Enhembo mu Mikolo Emitukuvu Egyo Busaserdooti, Obudyakoni ne Miruka e Kyebando Parish nga 02-08-2008*. Kyebando.
- CRC.** 2009. *Ensambo dh'Abasoga (Kisoga Proverbs)* (Revision of CRC 1999). Kisubi: Marianum Publishing Company.
- CRC.** 2010. *Installation of Rt. Rev. Bishop Charles Martin Wamika as Bishop of Diocese of Jinja*. Kisubi: Marianum Publishing Company.
- CRC.** 2011. *Ensengeka y'Omusomo gw'Ekikulu (Curriculum for Functional Adult Learners in Lusoga)*. Kisubi: Marianum Publishing Company.
- CRC.** 2012a. *Basoga Catholics in and Around Kampala*. Nsambya: Diocese of Jinja.
- CRC.** 2012b. *Missa mu Lusoga Ebiseera eby'Amatuuka n'Amazaalibwa*. Jinja: Diocese of Jinja.
- CRC.** 2012c. *Missa mu Lusoga Ebiseera eby'Amazuukira*. Jinja: Diocese of Jinja.
- CRC.** 2012d. *Missa mu Lusoga Ebiseera eby'Ekisiibo*. Jinja: Diocese of Jinja.
- CRC.** 2012e. *Missa mu Lusoga Ebiseera eby'Omwaka n'Enaku edh'Abatuukirivuu*. Jinja: Diocese of Jinja.
- CRC.** 2012f. *Missa mu Lusoga Ensengeka y'Emikolo gya Wiiki Entukuvu*. Jinja: Diocese of Jinja.

- CRC.** 2012g. *Thanks Giving Mass. for Rev. Sr. Restetuta Wangoye*. Jinja: Diocese of Jinja.
- de Schryver, G.-M.** 1999. *Bantu Lexicography and the Concept of Simultaneous Feedback, Some Preliminary Observations on the Introduction of a New Methodology for the Compilation of Dictionaries with Special Reference to a Bilingual Learner's Dictionary Cilubà-Dutch*. Unpublished M.A. dissertation. Ghent: Ghent University.
- de Schryver, G.-M.** 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies* 11(2): 266-282.
- de Schryver, G.-M. and R. Gauton.** 2002. The Zulu Locative Prefix ku- Revisited: A Corpus-based Approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201-220.
- de Schryver, G.-M. and M. Nabirye.** 2010. A Quantitative Analysis of the Morphology, Morphophonology and Semantic Import of the Lusoga Noun. *Africana Linguistica* 16: 97-153.
- de Schryver, G.-M. and D.J. Prinsloo.** 2000. The Compilation of Electronic Corpora, With Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1-4): 89-106.
- Devos, M. and G.-M. de Schryver.** 2013. From 'habitually going' to 'maybe': Grammaticalization and Lexicalization of an Epistemic Sentence Adverb in Swahili. Anon. (Ed.). 2013. *Abstracts of The 21st International Conference on Historical Linguistics*: 29. Oslo: University of Oslo.
- Devos, M. and G.-M. de Schryver.** 2016. From Usually Going to Epistemic Possibility. Origin and Development of an Epistemic Sentence Adverb in Swahili. Anon. (Ed.). 2016. *6th International Conference on Bantu Languages, Workshop on the Expression of Mood and Modality in Bantu Languages*: 11. Helsinki: University of Helsinki.
- Facebook.** 2004–18. Facebook Online Social Media and Social Networking Service. Available online at: <https://www.facebook.com>.
- Gonza, R.K.** 2007. *Lusoga–English Dictionary and English–Lusoga Dictionary* (revision of Korse 2000). Kampala: MK Publishers.
- Gulere, C.W.** 2007a. *Nantamegwa* (translation into Lusoga of *Antigone*, a play by Sophocles from 441 BC). Busembatya: Lusoga Language Academic Board.
- Gulere, C.W.** 2007b. *Omugole* (translation into Lusoga of *The Bride*, a play by Bukenya from 1987). Busembatya: Lusoga Language Academic Board.
- Gulere, C.W.** 2009. *Lusoga–English Dictionary / Eibwanio*. Kampala: Fountain Publishers.
- Gulere, C.W.** 2011a. *Abasikawutu*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011b. *Amagelo mu Nsiko*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011c. *Ebikete bya Busoga*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011d. *Ekidhuubo kya Giligoori*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011e. *Engabo ya Busoga*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011f. *Lusoga Nguli Namanha*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011g. *Ndi ni Mukazi Wange*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.

- Gulere, C.W.** 2011h. *Nsobola Nsobola*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011i. *Ogusolo ni Ekikaadh*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W.** 2011j. *Okusanhusa Tikwesanhusa*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Gulere, C.W. and M. Wambi.** 2011. *Lusoga Olusonhe*. Busembatia: Association of Lusoga Language Educationists, Researchers and Translators.
- Guthrie, M.** 1948. *The Classification of the Bantu Languages* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).
- Hadebe, S.** 2002. The Ndebele Language Corpus: A Review of Some Factors Influencing the Content of the Corpus. *Lexikos* 12: 159-170.
- Hanks, P.** 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hurskainen, A.** 2016. Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version. Available online at: <http://urn.fi/urn:nbn:fi:lb-2016011301>.
- Johnston, H.H.** 1902. *The Uganda Protectorate. An Attempt to Give Some Description of the Physical Geography, Botany, Zoology, Anthropology, Languages and History of the Territories under British Protection in East Central Africa, between the Congo Free State and the Rift Valley and between the First Degree of South Latitude and the Fifth Degree of North Latitude*. London: Hutchinson & Co.
- Kajolya, J.B.N.** 1990. *The Lusoga Orthography* (revision of Byandala 1963). Jinja: Lusoga Ecumenical Committee.
- Kaluuba, J.P., J. Kivuunike, J.S. Dhizaala and C. Nabirye.** 2010. *Gw'Olekera Abato: Okusoma kuleeta obusobozi*. Jinja: Cultural Research Centre, Marianum Publishing Company and Literacy and Adult Basic Education.
- Kaluuba, J.P. and P. Korse.** 2010. *Kintu* (revision of CRC 1998). Jinja: Cultural Research Centre.
- Kasozi, J.** 2000. *Mwidhe Tugye Tusenge*. Jinja: Diocese of Jinja.
- Kawalya, D.** 2017. *A Corpus-driven Study of the Expression of Modality in Luganda (Bantu, JE15)*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Kawalya, D., G.-M. de Schryver and K. Bostoën.** 2018. From Conditionality to Modality in Luganda (Bantu, JE15): A Synchronic and Diachronic Corpus Analysis of the Verbal Prefix *-andi-*. *Journal of Pragmatics* 127: 84-106.
- Kennedy, G.** 1998. *An Introduction to Corpus Linguistics* (Studies in Language and Linguistics). Harlow, Essex: Addison Wesley Longman.
- Kodh'eyo.** 1997-98. *Kodh'eyo: Busoga etebenkere* (a short-lived newspaper in Lusoga). Kampala: Kodh'eyo Publications.
- Korse, P.** 1999. *A Lusoga Grammar*. Jinja: Cultural Research Centre.
- Korse, P.** 2000. *Dictionary Lusoga-English / English-Lusoga*. Jinja: Cultural Research Centre.
- Kuunya, C.** 2011a. *Agakuba Omughafu*. Kisubi: Marianum Publishing Company.
- Kuunya, C.** 2011b. *Gulaama ey'Olusoga [Grammar of Lusoga]*. Kisubi: Marianum Publishing Company.
- LULANDA and CRC.** 2001. *Empandiika ey'Olulimi Olusoga Enkalamu [Standard Lusoga Orthography]*. Jinja: Lusoga Language Development Academy & Cultural Research Centre.

- LULANDA and CRC.** 2004. *Empandiika ey'Olulimi Olusoga Enkalamu [Standard Lusoga Orthography]* (2nd edition). Jinja: Lusoga Language Development Academy & Cultural Research Centre.
- Lyavala-Lwanga, E.J.** 1967. *Endheso dh'Abasoga [Proverbs of the Basoga]*. Kampala: Milton Obote Foundation.
- Lyavala-Lwanga, E.J.** 1969. *Kiyini Kibi [on language and Lusoga riddles]*. Kampala: Milton Obote Foundation.
- Maho, J.F.** 2009. NUGL Online: The Online version of the New Updated Guthrie List, a Referential Classification of the Bantu Languages. Available online at: <http://goto.glocalnet.net/mahopapers/nuglonline.pdf>.
- Mberamihigo, F.** 2014. *L'expression de la modalité en kirundi : Exploitation d'un corpus électronique*. Unpublished Ph.D. dissertation. Brussels; Ghent: Université libre de Bruxelles; Ghent University.
- Mberamihigo, F., G.-M. de Schryver and K. Bostoen.** 2016. Entre verbe et adverbe : Grammaticalisation et dégrammaticalisation du marqueur épistémique *umeengo/umeenga* en kirundi (bantou, JD62). *Journal of African Languages and Linguistics* 37(2): 247-286.
- Misago, M.-J.** 2018. *Les verbes de mouvement et l'expression du lieu en kirundi (bantou, JD62) : Une étude linguistique basée sur un corpus*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Mwesigwa, R.** s.d. *Mu Bigere Bye. Olugero lw'Abasoga ku kugoberera Yesu*. Jinja: Church of Christ.
- Nabirye, M.** 2008. *Compilation of the Monolingual Lusoga Dictionary*. Unpublished M.A. dissertation. Kampala: Makerere University.
- Nabirye, M.** 2009. *Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwega Olusoga [A Dictionary of Lusoga. For Speakers of Lusoga, and for Those Who Would Like to Learn Lusoga]*. Kampala: Menha Publishers.
- Nabirye, M.** 2016. *A Corpus-based Grammar of Lusoga*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Nabirye, M. and G.-M. de Schryver.** 2011. From Corpus to Dictionary: A Hybrid Prescriptive, Descriptive and Proscriptive Undertaking. *Lexikos* 21: 120-143.
- Nabirye, M., G.-M. de Schryver and J. Verhoeven.** 2016. Illustrations of the IPA: Lusoga (Lutenga). *Journal of the International Phonetic Association* 46(2): 219-228 (+ supplementary audio online).
- NCDC.** 2006. *THEMA News Letter* (Issue 2, December 2006). Kampala: National Curriculum Development Centre.
- Ndimugezi.** 1998–99. *Ndimugezi n'omukobere: The Factfinder* (a short-lived newspaper, with sections in Lusoga). Jinja: Ndimugezi Publications.
- Nshemezimana, E.** 2016. *Morphosyntaxe et structure informationnelle en kirundi : Focus et stratégies de focalisation*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- OmniPage.** 1995–2018. Optical character recognition (OCR) software now available from Nuance Communications. Available online at: <http://www.nuance.com/for-individuals/by-product/omnipage/>.
- Prinsloo, D.J.** 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285-300.
- Prinsloo, D.J. and G.-M. de Schryver.** 2001. Monitoring the Stability of a Growing Organic Corpus, with Special Reference to Sepedi and Xitsonga. *Dictionaries: Journal of The Dictionary Society of North America* 22: 85-129.

- Prinsloo, D.J. and G.-M. de Schryver.** 2005. Managing Eleven Parallel Corpora and the Extraction of Data in All Official South African Languages. Daelemans, W., T. du Plessis, C. Snyman and L. Teck (Eds). 2005. *Multilingualism and Electronic Language Management. Proceedings of the 4th International MIDP Colloquium, 22–23 September 2003, Bloemfontein, South Africa* (Studies in Language Policy in South Africa 4): 100-122. Pretoria: Van Schaik Publishers.
- Rundell, M. and P. Stock.** 1992. The Corpus Revolution 3. A Consideration of the Prospects and Potential of Corpus-and-concordance Lexicography (third article of three). *English Today, The International Review of the English Language* 8(4): 45-51.
- Scott, M.** 1996–2018. WordSmith Tools. Available online at: <http://www.lexically.net/wordsmith/>.
- Sene-Mongaba, B.** 2013. *Le lingála dans l'enseignement des sciences dans les écoles de Kinshasa : Une approche socioterminologique*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Sene-Mongaba, B.** 2015. The Making of Lingala Corpus: An Under-resourced Language and the Internet. *Procedia — Social and Behavioral Sciences* 198: 442-450.
- UBOS.** 2006. *The 2002 Uganda Population and Housing Census, Analytical Report, Population Composition*. Kampala: Uganda Bureau of Statistics.
- UBOS.** 2016. *The National Population and Housing Census 2014 — Main Report*. Kampala: Uganda Bureau of Statistics.
- Wabugoyera, J.B. Kasubi, J.P. Kaluuba, Mukama, M. Maganda and M. Maganda.** 2008. *Ekimuliikirira, August–October 2008*. Jinja: Diocese of Jinja.
- Wambi, M., R. Naigaga and CRC.** 2005. *Idha Tusome [Come and We Read]*. Jinja: Lusoga Language Authority.
- YouTube.** 2005–18. YouTube video-sharing website. Available online at: <https://www.youtube.com>.

Addendum 1: Corpus filename bibliography for the 391 sources in the 1.7m Lusoga corpus

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
1Bakyaba	Ambassador Institute 2012 (Eiterekero: Eriya ku lusozi kalameri) • Biblical documents / Religion • 741 / 380 • W ~ e-Transfer • Internet: Ambassador Institute • 2 pp.
Ababala	Malagala, Stephen 2000s (Ababala Emilimu) • Songs - Traditional / Sensitization • 660 / 434 • O ~ Transcription • Anon.: - • 0:08:08
AbabitAb	Ssajabi, Sophronius 1999 (Ababita Ababiri) • Literature / Fables • 5,063 / 1,825 • W ~ OCR • Jinja: CRC • 38 pp.
Abadhel	Malagala, Stephen 2000s (Abadhelega Emilimu) • Songs - Traditional / Sensitization • 551 / 364 • O ~ Transcription • Anon.: - • 0:06:34
Abalamu	Baisi 2010s (Abalamu Tusaanila Tukole) • Songs - Traditional / Sensitization • 633 / 308 • O ~ Transcription • Anon.: - • 0:08:49
ABamBamu	Mata, Nassani & Isiko 1990s (Bamusabire) • Songs - Traditional / Money • 441 / 153 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:13:00
ABamEita	Mata, Nassani & Isiko 1990s (Eitaka) • Songs - Traditional / Rehabilitation • 252 / 140 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:07:15
ABamKate	Mata, Nassani & Isiko 1990s (Katengeke) • Songs - Traditional / Relationships • 331 / 161 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:10:14
Abantub	Mugwisa 2000s (Abantu Beebisa Bulala) • Songs - Traditional / Rehabilitation • 609 / 359 • O ~ Transcription • Anon.: - • 0:08:57
Abasikaw	Gulere, Cornelius 2011 (Abasikawutu) • Literature / Fables • 1,885 / 851 • W ~ OCR • Internet: Google books • 19 pp.
Abasoga	Salimu 2010 (Abasoga) • Songs - Modern / Sensitization • 793 / 238 • O ~ Transcription • Jinja: CRC • 0:05:51
Abatool	Malagala, Stephen 2000s (Abatoolamu Embuto) • Songs - Traditional / Health • 697 / 391 • O ~ Transcription • Anon.: - • 0:07:46
Abatwes	Malagala, Stephen 2000s (Abatwesimbamu) • Songs - Traditional / Relationships • 377 / 262 • O ~ Transcription • Anon.: - • 0:04:35
Abeelad	Malagala, Stephen 2000s (Abeeladha ku Nsolo) • Songs - Traditional / Rehabilitation • 724 / 375 • O ~ Transcription • Anon.: - • 0:07:11
AEGY1	Various 2010 (Aids Education Group for Youths 1) • Radio talk shows / Health • 3,668 / 1,380 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:45:16
AEGY2	Various 2010 (Aids Education Group for Youths 2) • Radio talk shows / Health • 4,979 / 1,679 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:59:26
AEGY3	Various 2010 (Aids Education Group for Youths 3) • Radio talk shows / Health • 6,126 / 2,152 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:31:15
AEGY4	Various 2010 (Aids Education Group for Youths 4) • Radio talk shows / Health • 1,679 / 801 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:13:52

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
AgakbOmu	Kuunya, Christopher 2012 (Agakuba Omughafu) • Literature - Novels / Life • 47,964 / 12,828 • W ~ OCR • Jinja: Marianum Publishing Company • 310 pp.
Akabend	Mugwisa 2000s (Akabendhe) • Songs - Traditional / Money • 590 / 320 • O ~ Transcription • Anon.: - • 0:09:07
Akalango	Orange 2009 (Akalango ka Orange) • Advertisements / Networking • 347 / 193 • W ~ Translation • Kampala: Orange mobile phone network in Uganda • 1 p.
Akaleed	Malagala, Stephen 2000s (Akaleediyo) • Songs - Traditional / Marriage • 746 / 469 • O ~ Transcription • Anon.: - • 0:07:07
Akaleky	Anon. 2010s (Akalelelo ka Kiyingi) • Songs - Modern / Politics • 501 / 96 • O ~ Transcription • Anon.: - • 0:05:13
akalelel	Anon. 2010s (Leeta Akalelelo) • Songs - Traditional / Money • 677 / 116 • O ~ Transcription • Iganga: - • 0:07:40
AkatAkas	Ssajabi, Sophronius 1999 (Akatabo Akasooka ak'Enfumo edh'Abasoga) • Literature / Fables • 5,831 / 2,107 • W ~ OCR • Jinja: CRC • 38 pp.
akatiko	Geo Bless 2010s (Akatiko) • Songs - Modern / Relationships • 322 / 196 • O ~ Transcription • Iganga: - • 0:03:59
Akatook	Anon. 2010s (Akatooke k'Endala) • Songs - Traditional / Politics • 1,075 / 143 • O ~ Transcription • Anon.: - • 0:08:25
Akeeyo	Bamulumba, Yasiini 2012 (Akeeyo) • Songs - Modern / Rehabilitation • 642 / 317 • O ~ Transcription • internet: Intangible Culture Heritage Conservation Project • 0:08:46
ALwaLwak	Kabugu, Jessica & Kabugu, Milton Peter 1990s (Lwaki Tosulanga y'Ogunhwa) • Songs - Traditional / Rehabilitation • 1,076 / 565 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:11:21
ALwaOmwe	Kabugu, Milton Peter 1990s (Omwenge Taabbu) • Songs - Traditional / Rehabilitation • 717 / 258 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:09:05
ALwaSili	Kabugu, Milton Peter 1990s (Siliimu) • Songs - Traditional / Health • 549 / 309 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:08:21
Amaadhi	Anon. 2010s (Amaadhi) • Songs - Traditional / Relationships • 515 / 274 • O ~ Transcription • Anon.: - • 0:05:14
AmagelM	Gulere, Cornelius 2007 (Amagelo mu Nsiko) • Literature / Fables • 1,185 / 653 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 20 pp.
Amagelom	Gulere, Cornelius 2011 (Amagelo mu Nsiko) • Literature / Fables • 1,084 / 607 • W ~ OCR • Internet: Google books • 18 pp.
AMagEnfu	Kabugu, Milton Peter 1990s (Enfuna y'Esente) • Songs - Traditional / Money • 503 / 265 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:08:26
AMagInha	Kabugu, Milton Peter 1990s (Inhazaala Ghange) • Songs - Traditional / Marriage • 426 / 248 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:10:56
AMagObuf	Kabugu, Milton Peter 1990s (Obufumbo) • Songs - Traditional / Marriage • 1,196 / 551 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:11:46

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
AmagTigm	Kyakulaga, Zion 1999 (Amagezi Tigamalwayo) • Literature / Fables • 4,198 / 1,735 • W ~ OCR • Jinja: CRC • 38 pp.
Amateeka	Justice Law and Order Sector (JLOS) 2008 (Amateeka Agagema ku Kusiba Abantu mu Byalo) • Policy documents - Government / Sensitization • 89 / 61 • W ~ Translation • Kampala: JLOS • 0.5 pp.
Archbis	Orombi, Henry Luke 2009 (Speech of the Archbishop of Uganda during his Visit to the Diocese of Jinja) • Celebrations / Religion • 4,574 / 1,786 • O ~ Transcription • Jinja: Church of Uganda • 1:25:33
Artbase	Artbase 2010s (Artbase Anthem) • Songs - Modern / Inspirational • 322 / 132 • O ~ Transcription • Anon.: - • 0:05:46
Asaanak	Ntende, Monika 2010 (Asaana Kwebaza) • Songs - Gospel / Religion • 404 / 169 • O ~ Transcription • Internet: YouTube • 0:05:47
Ateoba	Kigenyi, Amos 2010 (Ate oba Wankyawa) • Songs - Modern / Relationships • 57 / 29 • O ~ Transcription • Internet: YouTube • 0:04:15
ATirEkya	Mata, Nassani 1990s (Ekyanguza Empale) • Songs - Traditional / Inspirational • 192 / 128 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:57
ATirKawo	Mata, Nassani 1990s (Kawoiwolo) • Songs - Traditional / Relationships • 368 / 222 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:28
ATirMump	Mata, Nassani 1990s (Munpe Omwana) • Songs - Traditional / Marriage • 343 / 180 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:07:08
ATirOmwo	Mata, Nassani 1990s (Omwoyo Fiitina) • Songs - Traditional / Relationships • 171 / 102 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:02:15
ATirSula	Mata, Nassani 1990s (Sulaayi) • Songs - Traditional / Relationships • 436 / 307 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:06:09
ATirTiil	Mata, Nassani 1990s (Tiilime) • Songs - Traditional / Money • 249 / 144 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:04:50
AVATVAT	Kirimungu, Siragi 2000s (VAT) • Songs - Traditional / Sensitization • 553 / 251 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:08:30
AVATVAT2	Kirimungu, Siragi 2000s (VAT Vol. 2) • Songs - Traditional / Sensitization • 278 / 184 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:45
Babalan	Baisi 2010s (Babalanda) • Songs - Traditional / Inspirational • 870 / 296 • O ~ Transcription • Anon.: - • 0:08:39
Bakalakt	Gulere, Cornelius 2006 (Bakalakatana) • Literature / Fables • 3,902 / 1,770 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 32 pp.
Bakulim	Magoola, Racheal 2010 (Bakulimba) • Songs - Modern / Relationships • 173 / 93 • O ~ Transcription • Internet: YouTube • 0:05:00
Bakyali	Musooko 2010s (Bakyali) • Songs - Modern / Relationships • 701 / 396 • O ~ Transcription • Iganga: - • 0:06:43
Balocle	Bujagaali's daughters 2006 (Ennhemba dh'Abalongo) • Interviews / Health • 310 / 185 • O ~ Transcription • Jinja: - • 0:09:20

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Balodis	Bujagaali's client 2006 (Balongo Discussion) • Interviews / Marriage • 1,757 / 719 • O ~ Transcription • Jinja: - • 0:33:24
Bascath	Cultural Research Centre (CRC) 2012 (Basoga Catholics in and Around Kampala) • Biblical documents / Networking • 426 / 203 • W ~ OCR • Nsambya: Diocese of Jinja • 16 pp.
Bbaabba	Kigenyi, Amos 2010 (Bbaabba Toyombesa Maama) • Songs - Modern / Marriage • 311 / 114 • O ~ Transcription • Internet: YouTube • 0:04:28
BBamAkat	Mata, Nassani & Isiko 1990s (Akatooke) • Songs - Traditional / Rehabilitation • 427 / 208 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:12:06
BBamBali	Mata, Nassani & Isiko 1990s (Balizanila) • Songs - Traditional / Relationships • 306 / 141 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:10:15
BBamEndo	Mata, Nassani & Isiko 1990s (Endoola) • Songs - Traditional / Life • 346 / 195 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:08:45
BEL09-Q2	Bujagali Hydropower Project (BHPP) 2009 (Bujagali Project Newsletter Q2 - July, 2009) • Newsletters / Sensitization • 1,645 / 784 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 5 pp.
BEL09-Q3	Bujagali Hydropower Project (BHPP) 2009 (Bujagali Project Newsletter Q3 - 30th September, 2009) • Newsletters / Sensitization • 1,532 / 717 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 4 pp.
BEL09-Q4	Bujagali Hydropower Project (BHPP) 2009 (Bujagali Project Newsletter Q4 - December, 2009) • Newsletters / Sensitization • 1,796 / 804 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 5 pp.
BEL10-Q4	Bujagali Hydropower Project (BHPP) 2010 (Bujagali Project Newsletter Q4 - 31st December, 2010) • Newsletters / Sensitization • 1,991 / 749 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 5 pp.
BEL11-Q1	Bujagali Hydropower Project (BHPP) 2011 (Bujagali Project Newsletter Q1 - 31st March, 2011) • Newsletters / Sensitization • 1,692 / 737 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 5 pp.
BEL11-Q3	Bujagali Hydropower Project (BHPP) 2011 (Bujagali Project Newsletter Q3 - 30th September, 2011) • Newsletters / Sensitization • 1,371 / 585 • W ~ e-Transfer • Internet: Bujagali Energy Ltd • 4 pp.
Betty	Malagala, Stephen 2000s (Betty) • Songs - Traditional / Marriage • 563 / 373 • O ~ Transcription • Anon.: - • 0:05:56
BibChEas	Hughes, Edward 2013 (Embaga ey'Amazuukira Eyasooka) • Biblical documents / Religion • 1,030 / 561 • W ~ e-Transfer • Internet: Bible for Children • 25 pp.
BibChGod	Hughes, Edward 2013 (Katonda nga Bweyatonda Buli Kintu) • Biblical documents / Religion • 921 / 446 • W ~ e-Transfer • Internet: Bible for Children • 26 pp.
BibChHea	Hughes, Edward 2013 (Eigulu, Amaka ga Katonda Agaboneka Obulungi Einho) • Biblical documents / Religion • 928 / 496 • W ~ e-Transfer • Internet: Bible for Children • 22 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
BibChJes	Hughes, Edward 2013 (Okuzaalibwa kwa Yesu) • Biblical documents / Religion • 815 / 449 • W ~ e-Transfer • Internet: Bible for Children • 29 pp.
BibChNoa	Hughes, Edward 2013 (Ebigema ku Noah n'Omwidhuzo ogw'Amaadhi Omukologho) • Biblical documents / Religion • 730 / 413 • W ~ e-Transfer • Internet: Bible for Children • 25 pp.
BibChSad	Hughes, Edward 2013 (Amainhama ag'Okunakughala okw'Omuntu) • Biblical documents / Religion • 744 / 427 • W ~ e-Transfer • Internet: Bible for Children • 25 pp.
Biblest2	Various 2010 (Bible Story 2) • Biblical documents / Religion • 6,344 / 1,917 • O ~ Transcription • Internet: YouTube • 0:44:48
Biblest3	Various 2010 (Bible Story 3) • Biblical documents / Religion • 2,785 / 1,034 • O ~ Transcription • Internet: YouTube • 0:28:09
Birugrd	Various 2011 (Graduation Ceremony in Buwaabe) • Celebrations / Inspirational • 5,858 / 2,114 • O ~ Transcription • Iganga: - • 1:04:38
BLwaBana	Kabugu, Milton Peter 1990s (Banamwandu ni Bamulekwa) • Songs - Traditional / Marriage • 774 / 434 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:09:25
BLwaNgol	Kabugu, Jessica 1990s (Ngoli Namala Naidha Luvanhuma) • Songs - Traditional / Marriage • 854 / 421 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:10:27
BLwaOmun	Kabugu, Jessica 1990s (Omuntu gh'Ensi Muzibu) • Songs - Traditional / Life • 691 / 354 • O ~ Transcription • Jinja: Ali Mukembo and Sons • 0:08:04
BMagEbiz	Kabugu, Milton Peter 1990s (Ebizibu eby'Ensi) • Songs - Traditional / Life • 488 / 259 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:10:09
BMagOmul	Kabugu, Milton Peter 1990s (Omulamu Asaalilwa) • Songs - Traditional / Relationships • 445 / 234 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:09:30
BMagRose	Kabugu, Milton Peter 1990s (Rose Mary) • Songs - Traditional / Relationships • 480 / 250 • O ~ Transcription • Jinja: Sanyu Music Studios • 0:20:13
BTirBand	Mata, Nassani 1990s (Bando Asiliile) • Songs - Traditional / Marriage • 196 / 131 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:02:12
BTirIdha	Mata, Nassani 1990s (Idha Ompelekeleku) • Songs - Traditional / Relationships • 264 / 116 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:04:35
BTirNang	Mata, Nassani 1990s (Nangobi) • Songs - Traditional / Relationships • 448 / 329 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:37
BTirNinz	Mata, Nassani 1990s (Ni Nze Mbeese) • Songs - Traditional / Money • 388 / 236 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:23
BTirObug	Mata, Nassani 1990s (Obugumba) • Songs - Traditional / Marriage • 234 / 124 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:05:15
BTirOmul	Mata, Nassani 1990s (Omukazi Omwenzi) • Songs - Traditional / Money • 251 / 179 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:03:06
BTirWabu	Mata, Nassani 1990s (Wabukala Bando) • Songs - Traditional / Marriage • 244 / 131 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:06:11

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
BucakaML	Various 2013 (Bucaka Mails) • E-mails / Networking • 5,535 / 2,201 • W ~ e-Transfer • Internet: Bucaka • 57 pp.
Bujaga11	Bujagaali & First wife 2006 (Bujagaali Interview 1) • Interviews / Health • 917 / 431 • O ~ Transcription • Jinja: - • 0:32:30
Bujaga12	Bujagaali & First wife 2006 (Bujagaali Interview 2) • Celebrations / Health • 2,280 / 1,056 • O ~ Transcription • Jinja: - • 0:43:40
Bultg10	Various 2010 (Busoga-Bulleting 1) • E-mails / Networking • 4,567 / 2,045 • W ~ e-Transfer • Internet: Yahoo! • 27 pp.
Bultg11.6	Various 2011 (Busoga-Bulleting 2) • E-mails / Networking • 6,842 / 2,716 • W ~ e-Transfer • Internet: Yahoo! • 32 pp.
Bultg11.7	Various 2011 (Busoga-Bulleting 3) • E-mails / Networking • 537 / 261 • W ~ e-Transfer • Internet: Yahoo! • 4 pp.
Bultg12	Various 2012 (Busoga-Bulleting 4) • E-mails / Networking • 3,854 / 1,691 • W ~ e-Transfer • Internet: Yahoo! • 17 pp.
BusogaCh	The Cross-Cultural Foundation of Uganda (CCFU) 2012 (The Uganda Clan Leaders' Charters) • Policy documents - Busoga Kingdom / Sensitization • 1,339 / 787 • W ~ e-Transfer • Kampala: The Cross-Cultural Foundation of Uganda • 86 pp.
Buwaabe	Nabirye, Minah 2009 (Buwaabe Sunday Church Service) • Biblical documents / Religion • 8,788 / 2,936 • O ~ Transcription • Iganga: - • 1:44:26
BuwaabGr	Nabirye, Minah 2010 (Buwaabe Graduation Ceremony) • Celebrations / Inspirational • 16,370 / 4,605 • O ~ Transcription • Iganga: - • 3:51:17
BVATObuf	Kirimungu, Siragi 2000s (Obufumbo Buzibu) • Songs - Traditional / Marriage • 349 / 182 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:14:26
BVATOmwo	Kirimungu, Siragi 2000s (Omwoyo Fiitina) • Songs - Traditional / Relationships • 1,153 / 513 • O ~ Transcription • Jinja: Ali Mukembo Studio • 0:14:05
Bwoteef	Mugwisa 2000s (Bw'oteefaaku) • Songs - Traditional / Sensitization • 428 / 219 • O ~ Transcription • Anon.: - • 0:08:48
Bwozaal	Mugwisa 2000s (Bw'ozaaala n'Abaawo) • Songs - Traditional / Inspirational • 562 / 308 • O ~ Transcription • Anon.: - • 0:09:06
Byaif09	Various 2009 (Busogayaife 1) • E-mails / Networking • 2,020 / 947 • W ~ e-Transfer • Internet: Yahoo! • 21 pp.
Byaif10	Various 2010 (Busogayaife 2) • E-mails / Networking • 29,743 / 9,845 • W ~ e-Transfer • Internet: Yahoo! • 142 pp.
Byaif11.6	Various 2011 (Busogayaife 3) • E-mails / Networking • 24,805 / 8,320 • W ~ e-Transfer • Internet: Yahoo! • 98 pp.
Byaif11.7	Various 2011 (Busogayaife 4) • E-mails / Networking • 5,289 / 2,440 • W ~ e-Transfer • Internet: Yahoo! • 25 pp.
Byaif12	Various 2012 (Busogayaife 5) • E-mails / Networking • 41,983 / 13,218 • W ~ e-Transfer • Internet: Yahoo! • 169 pp.
ByaKfaK1	Gulere, Cornelius 2010 (Bya Kufa Kuleka 5) • Literature / Fables • 16,109 / 5,449 • W ~ OCR • Busembatya: Mpolyabigere RC – RICED Center • 72 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
ChildAct	Uganda Legal Information Institute (ULII) 2012 (Children's Act in Lusoga) • Policy documents - Government / Sensitization • 16,211 / 2,354 • W ~ OCR • Internet: Government of Uganda Parliamentary Act on Human Rights • 68 pp.
Cohen86	Cohen, David 1986 (Towards a Reconstructed Past: Historical texts from Busoga, Uganda) • Academic documents / History • 16,657 / 5,742 • W ~ OCR • London: Oxford University Press • 54 pp.
Communit	Various 2009 (Community Development) • Radio talk shows / Inspirational • 5,164 / 1,955 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:32:17
Diocesan	Cultural Research Centre (CRC) 2003 (Diocesan Family Day) • Biblical documents / Religion • 4,201 / 1,486 • W ~ OCR • Iganga: Diocesan Printery • 48 pp.
Ebibiin	Kirimungu, Siragi 2000s (Ebibiina Biyamba) • Songs - Traditional / Networking • 484 / 208 • O ~ Transcription • Anon.: - • 0:06:48
Ebigkbal	Bujagaali's client 2006 (Ebigema ku Balongo) • Interviews / Marriage • 1,954 / 815 • O ~ Transcription • Jinja: - • 0:14:24
Ebikemo	Soyinka, Wole 2010 (Ebikemo by'Owoluganda Yero) • Literature - Plays / Life • 7,092 / 2,725 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 39 pp.
Ebikete	Gulere, Cornelius 2011 (Ebikete bya Busoga) • Literature / Fables • 2,572 / 1,430 • W ~ OCR • Internet: Google books • 40 pp.
EbikoikE	Cultural Research Centre (CRC) 2002 (Ebikoiko eby'Abasoga) • Literature / Riddles • 6,963 / 2,818 • W ~ OCR • Jinja: CRC • 126 pp.
Ebikoiko	Gulere, Cornelius 2008 (Ebikoiko mu Lusoga) • Literature / Riddles • 7,797 / 2,796 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 35 pp.
Ebikolwa	Wabugoyera; Kasubi, J.B.; Kaluuba, John Patrick; Mukama; Maganda, Matia & Maganda, Matayo 2010 (Ebikolwa bya Sapuli) • Biblical documents / Religion • 17,887 / 6,001 • W ~ OCR • Jinja: Our Lady of Fatima Parish Church • 49 pp.
EbikolwE	Uganda Gazette 2008 (Ebikolwa Eby'ongelwaaku) • Policy documents - Government / Politics • 19,064 / 2,728 • W ~ OCR • Internet: Ministry of Education • 159 pp.
EbikolWK	Gulere, Cornelius 2007 (Ebikolwa bya Wankembo) • Literature / Fables • 1,087 / 588 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 19 pp.
Ebilung	Mugwisa 2000s (Ebilungi Tibikoma) • Songs - Traditional / Sensitization • 535 / 322 • O ~ Transcription • Anon.: - • 0:09:11
EbindKuI	Cultural Research Centre (CRC) 2005 (Ebindi kw'Idembe ery'Obw'omuntu mu Nsi Yoonayoona) • Policy documents - Human rights / Sensitization • 26,520 / 4,891 • W ~ OCR • Kisubi: Marianum Publishing Company • 119 pp.
Ebintub	Malagala, Stephen 2010s (Ebintu Bisingagana) • Songs - Traditional / Life • 815 / 447 • O ~ Transcription • Kamuli: - • 0:09:23
Ebizibu	Mugwisa 2000s (Ebizibu mu Duniya) • Songs - Traditional / Health • 576 / 349 • O ~ Transcription • Anon.: - • 0:09:07
Ebyensi	Crado 2010s (Eby'ensi) • Songs - Modern / Rehabilitation • 474 / 251 • O ~ Transcription • Anon.: - • 0:06:00

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Egyamw	Crado 2010s (Egya Mwete) • Songs - Modern / Health • 314 / 181 • O ~ Transcription • Anon.: - • 0:05:03
Eidemban	Cultural Research Centre (CRC) 2010 (Eidembe ly'Abantu) • Songs - Modern / Sensitization • 3,360 / 1,224 • O ~ Transcription • Jinja: CRC • 0:31:17
Eidembbw	Cultural Research Centre (CRC) 2010 (Eidembe ly'Obw'obuntu) • Songs - Modern / Sensitization • 2,744 / 1,034 • O ~ Transcription • Jinja: CRC • 0:29:04
Eidembun	Anon. 2010 (Eidembe ly'Obuntu) • Songs - Modern / Sensitization • 946 / 240 • O ~ Transcription • Jinja: CRC • 0:07:23
Eifumbi	Malagala, Stephen 2000s (Eifumbilo) • Songs - Traditional / Marriage • 537 / 374 • O ~ Transcription • Anon.: - • 0:06:32
Eighali	Gulere, Cornelius 1998 (Eighali Lirikwisa) • Literature / Fables • 988 / 630 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 7 pp.
Eisomoly	International Centre for Eye Education (ICEE) 2008 (Eisomo ly'Okugezesa Obwangu bw'Enkyukakyuka mu Kubona) • Policy documents - NGOs / Health • 576 / 289 • W ~ Translation • Kampala: International Centre for Eye Education • 0.5 pp.
Eisuubi	Gulere, Cornelius 2013 (Eisuubi Okusaaka Obusomi) • Celebrations / Politics • 955 / 541 • W ~ OCR • Internet: Tarehe sita • 9 pp.
Ekibila	Kabugu, Milton Peter 2010 (Ekibila) • Songs - Modern / Sensitization • 784 / 130 • O ~ Transcription • Jinja: CRC • 0:06:13
Ekidhuub	Gulere, Cornelius 2011 (Ekidhuubo) • Literature / Fables • 1,829 / 1,027 • W ~ OCR • Internet: Google books • 22 pp.
Ekikwek	Malagala, Stephen 2000s (Ekikwekabya) • Songs - Traditional / Marriage • 748 / 470 • O ~ Transcription • Anon.: - • 0:08:36
Ekimlik	Wabugoyera; Kasubi, J.B.; Kaluuba, John Patrick; Mukama; Maganda, Matia & Maganda, Matayo 2008 (Ekimuliikirira, August-October 2008) • Biblical documents / Religion • 18,452 / 4,839 • W ~ OCR • Jinja: Diocese of Jinja • 54 pp.
Ekinait	Malagala, Stephen 2000s (Ekinaita Embwa) • Songs - Traditional / Health • 657 / 378 • O ~ Transcription • Anon.: - • 0:07:52
Ekirangi	Abakulu b'ebika bya Busoga 2009 (Ekirangiriro eri Obusoga n'Ensi Yoonayoona) • Policy documents - Busoga Kingdom / History • 2,710 / 980 • W ~ Translation • Jinja: Katukiro w'olukiiko lw'abakulu b'ebika bya Busoga • 14 pp.
Ekiwandi	International Centre for Eye Education (ICEE) 2010 (Ekiwandiiko Ekilaga Ennamuula y'Amaka) • Policy documents - NGOs / Health • 1,612 / 525 • W ~ Translation • Kampala: International Centre for Eye Education • 5 pp.
EliinaEl	International Centre for Eye Education (ICEE) 2010 (Eliina Elisooka) • Policy documents - NGOs / Health • 195 / 109 • W ~ Translation • Kampala: International Centre for Eye Education • 1 p.
Embeeke	Malagala, Stephen 2000s (Embeekela) • Songs - Traditional / Relationships • 634 / 381 • O ~ Transcription • Anon.: - • 0:07:08

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Empambo	Gulere, Cornelius 2007 (Empambo) • Literature / Fables • 2,335 / 1,391 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 43 pp.
EmpisaB	Gulere, Cornelius 2007 (Empisa n'Obuntubulamu) • Literature / Fables • 1,271 / 776 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 16 pp.
Endagaan	Bible Society Uganda (BSU) 1998 (Endagaano Empyaka) • Biblical documents / Religion • 150,223 / 19,829 • W ~ OCR • Jinja: The Bible Society of Uganda • 518 pp.
EndhesE2	Gulere, Cornelius 2007 (Endheso Ennhimpi) • Literature / Proverbs • 2,659 / 1,877 • W ~ OCR • Busembatya: Lusoga Language Academic Board (LLAB) • 34 pp.
EndhesoD	Lyavala-Lwanga, E.J. 1967 (Endheso dh'Abasoga) • Literature / Proverbs • 16,809 / 7,289 • W ~ OCR • Kampala: Milton Obote Foundation • 97 pp.
Endhesoe	Gulere, Cornelius 2011 (Endheso Ennhimpi) • Literature / Fables • 2,723 / 1,900 • W ~ OCR • Internet: Google books • 40 pp.
Endhesul	Gulere, Cornelius 2012 (Endheso Ensusulemu) • Literature / Proverbs • 2,125 / 1,066 • W ~ OCR • Internet: Anon. • 31 pp.
Endhiiy1	Private Sector Uganda (PSU) 2009 (Endhiya y'Obukodyo bw'Enkulankulana 1) • Policy documents - NGOs / Money • 602 / 299 • W ~ Translation • Kampala: Private Sector Uganda • 5 pp.
Endhiiy2	Private Sector Uganda (PSU) 2009 (Endhiya y'Obukodyo bw'Enkulankulana 2) • Policy documents - NGOs / Money • 602 / 300 • W ~ Translation • Kampala: Private Sector Uganda • 4 pp.
EndyaBul	Gulere, Cornelius 2007 (Endya Bulamu) • Literature / Fables • 1,101 / 644 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 13 pp.
EnfumKay	Kisubi, Alfred James Igaga 2009 (Enfumitiriza Kayingo n'Entegeka luv'Okwaaya y'Oluikiiko lw'Abakulu b'Ebika) • Policy documents - Busoga Kingdom / History • 2,550 / 1,049 • W ~ OCR • Jinja: The Re-Unification of the Clans of Busoga • 14 pp.
Engabo	Gulere, Cornelius 2011 (Engabo ya Busoga) • Literature / Fables • 765 / 431 • W ~ OCR • Internet: Google books • 19 pp.
Engedh	Anon. 2010s (Engeli Dhaimwe) • Songs - Modern / Relationships • 463 / 237 • O ~ Transcription • Anon.: - • 0:05:13
EnhemboM	Cultural Research Centre (CRC) 2008 (Enhembo mu Mikolo Emitukuvu) • Biblical documents / Religion • 3,771 / 1,486 • W ~ OCR • Jinja: CRC • 35 pp.
EnkEkiFn	Cultural Research Centre (CRC) 2000 (Enkabi Ekifiini mu Busoga) • Literature / Proverbs • 7,842 / 2,768 • W ~ OCR • Jinja: CRC • 30 pp.
Ennakun	Mugwisa 2000s (Ennaku Namugalula) • Songs - Traditional / Life • 486 / 300 • O ~ Transcription • Anon.: - • 0:09:51
Ennhemba	Nabirye, Minah 2000 (Ennhemba dh'Olusoga) • Songs - Traditional / Relationships • 653 / 284 • O ~ e-Transfer • Jinja: - • (own writing from memory recollections)
Ennhonh	Mugwisa, Andy Cooke 2010 (Ennhonhi ku Lugyo) • Songs - Traditional / Science • 262 / 161 • O ~ Transcription • Internet: YouTube • 0:04:00

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
EnsambDh	Cultural Research Centre (CRC) 1999 (Ensambo edh'Abasoga) • Literature / Proverbs • 16,121 / 7,810 • W ~ OCR • Jinja: CRC • 90 pp.
Ensiiek	Malagala, Stephen 2000s (Ensi Ekyuse) • Songs - Traditional / Life • 472 / 315 • O ~ Transcription • Anon.: - • 0:06:08
Ensieno	Baisi 2010s (Ensi eno Weetuse) • Songs - Traditional / Rehabilitation • 676 / 347 • O ~ Transcription • Anon.: - • 0:08:55
Ensinzb	Salimu 2010 (Ensi Nzibu) • Songs - Modern / Life • 1,206 / 234 • O ~ Transcription • Jinja: CRC • 0:06:31
Erikwain	Various 2009 (Erikwaine) • Radio talk shows / Politics • 5,885 / 2,077 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:56:23
Esaalmk1	Mbutu, Rose & Musooko, Paulo 2010 (Esaala ey'Amaka 1) • Radio talk shows / Religion • 6,281 / 1,951 • O ~ Transcription • Bugembe: Holy Christ Family Ministry • 0:48:06
Esaalmk2	Mbutu, Rose & Musooko, Paulo 2010 (Esaala ey'Amaka 2) • Radio talk shows / Religion • 6,624 / 2,260 • O ~ Transcription • Bugembe: Holy Christ Family Ministry • 0:54:08
Eyalyaom	Gulere, Cornelius 2007 (Eyalya Omuunhu) • Literature / Fables • 512 / 340 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 10 pp.
Eyeesig	Geo Bless 2010s (Eyeeesigibwa) • Songs - Modern / Relationships • 245 / 133 • O ~ Transcription • Anon.: - • 0:04:16
Ezilamul	Kabugu, Milton Peter 2010 (Ezila Mulungi ku Nsi) • Songs - Modern / Life • 986 / 150 • O ~ Transcription • Jinja: CRC • 0:05:39
Ezirakkw	Gulere, Cornelius 2012 (Ezira Kyetaagisibwa Kwongeraku) • Literature / Fables • 352 / 211 • W ~ OCR • Internet: Tarehe sita • 9 pp.
Facebook	Various 2009 (Posting) • E-mails / Language • 54 / 51 • W ~ e-Transfer • Internet: Facebook • 1 p.
Fiida	Malagala, Fiida 2010s (Fiida) • Songs - Traditional / Marriage • 268 / 190 • O ~ Transcription • Anon.: - • 0:05:02
GulwOLAs	Dhizaala, John Stephen 2011 (Gulaama w'Olulimi Olusoga Asookerwaku) • Academic documents / Language • 11,311 / 3,464 • W ~ OCR • Jinja: CRC • 57 pp.
Gw'olile	Anon. 2010s (Gw'Olilekela Omwana) • Songs - Traditional / Marriage • 244 / 133 • O ~ Transcription • Iganga: - • 0:04:34
Gw'olkba	Kaluuba, John Patrick; Kivuunike, James; Dhizaala, John Stephen & Nabirye, Christine 2010 (Gw'Olekera Abato: Okusoma kuleeta obusobozi) • Literature / Language • 5,766 / 2,170 • W ~ OCR • Jinja: CRC, Marianum Publishing Company and LABE • 55 pp.
HonKiyg	Various 2003 (Hon Kiyingi) • Radio talk shows / Politics • 1,068 / 550 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:11:45
HonKizge	Various 2009 (Hon Kizige) • Radio talk shows / Politics • 7,932 / 2,715 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:09:59

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Idembeb	Kabugu, Milton Peter 2010 (Eidembe ly'Abaana) • Songs - Modern / Sensitization • 599 / 142 • O ~ Transcription • Jinja: CRC • 0:06:06
IdhaTusm	Wambi, M.; Naigaga, R. & CRC 2005 (Idha Tusome) • Academic documents / Language • 3,059 / 1,458 • W ~ OCR • Jinja: Lusoga Language Authority (LULA) • 80 pp.
Immunisa	Various 2009 (Immunization) • Radio talk shows / Health • 6,544 / 2,353 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:07:21
ImweMwOk	Bible Society Uganda (BSU) 1994 (Imwe Mwachhebwa Okumanha) • Biblical documents / Religion • 1,681 / 828 • W ~ OCR • Kampala: The Bible Society of Uganda • 8 pp.
InstallC	Cultural Research Centre (CRC) 2010 (Installation of Rt. Rev. Bishop Charles Martin Wamika as Bishop of Diocese of Jinja) • Biblical documents / Religion • 1,897 / 611 • W ~ OCR • Jinja: Marianum Publishing Company • 48 pp.
IntBilaa	Various 2008 (Introduction Ceremony 1) • Celebrations / Marriage • 11,716 / 3,555 • O ~ Transcription • Iganga: - • 4:35:34
IntHadij	Nabirye, Minah 2008 (Introduction Ceremony 2) • Celebrations / Marriage • 12,098 / 3,558 • O ~ Transcription • Iganga: - • 3:33:41
Isabiry	Malagala, Stephen 2000s (Isabirye ni Bbaabba We) • Songs - Traditional / Marriage • 441 / 301 • O ~ Transcription • Anon.: - • 0:06:15
Isatifik	Foundation for Endangered Languages (FEL) 2012 (Isatifiikeeti Iya Bughanzi) • Celebrations / Inspirational • 41 / 36 • W ~ OCR • Internet: Foundation for Endangered Languages • 1 p.
Isebantu	Kabugu, Milton Peter 2010 (Isebantu) • Songs - Modern / History • 920 / 127 • O ~ Transcription • Jinja: CRC • 0:06:18
Judicial	Various 2010 (Judicial Service Commission) • Radio talk shows / Sensitization • 7,721 / 2,461 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:00:41
Kabili	Malagala, Stephen 2000s (Kabili Ndeese) • Songs - Traditional / Sensitization • 320 / 228 • O ~ Transcription • Anon.: - • 0:04:03
Kabindi1	Kabindi, Erukaana 2008 (Kabindi Interview 1) • Interviews / Health • 4,313 / 1,504 • O ~ Transcription • Bugiri: - • 0:40:05
Kabindi2	Kabindi, Erukaana 2009 (Kabindi Interview 2) • Interviews / Health • 2,549 / 1,026 • O ~ Transcription • Bugiri: - • 0:44:49
KadokInt	Kadooko, John 2012 (Interview on the Kyabazingaship and the Status of Lusoga) • Interviews / History • 9,414 / 2,999 • O ~ Transcription • Jinja: - • 1:35:07
Kaibutag	Kai Butagaya women 2010 (Katonda n'Agaba) • Songs - Traditional / Gratitude • 366 / 110 • O ~ Transcription • Internet: YouTube • 0:04:28
Kaleebi	Kaleebi, George 2010 (Kaleebi) • E-mails / Networking • 1,437 / 809 • W ~ e-Transfer • Local: Self • 7 pp.
Kalikub	Anon. 2010s (Kali Kubayiiga) • Songs - Modern / Health • 510 / 111 • O ~ Transcription • Anon.: - • 0:06:00

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
KamuliTC	Various 2009 (Kamuli Town Council) • Radio talk shows / Sensitization • 5,369 / 1,774 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:55:05
KasreeOl	Various 2009 (Posting) • E-mails / Language • 71 / 63 • W ~ e-Transfer • Internet: Facebook • 1 p.
Katondk	Baisi 2010s (Katonda ky'Akuwa Osiima) • Songs - Traditional / Gratitude • 539 / 255 • O ~ Transcription • Anon.: - • 0:08:25
Kawoiwol	Gulere, Cornelius 2007 (Kawoiwolo Ayenda Kubayiza) • Literature / Fables • 1,123 / 670 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 18 pp.
KayondOm	Gulere, Cornelius 2011 (Kayondo Omuyondho) • Literature - Booklets / Sensitization • 709 / 315 • W ~ e-Transfer • Busembatya: Mpolyabigere RC – RICED Center • 11 pp.
Kibbaaly	Various 2009 (Kibbaalya) • Radio talk shows / Politics • 9,707 / 3,059 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:21:06
Kibumb	Anon. 2010 (Kibumba) • Songs - Traditional / Religion • 675 / 142 • O ~ Transcription • Internet: YouTube • 0:08:30
Kilikum	Anon. 2010s (Kili ku Mwino) • Songs - Traditional / Life • 306 / 141 • O ~ Transcription • Iganga: - • 0:05:12
Kintu	Cultural Research Centre (CRC) 1998 (Kintu) • Literature - Plays / History • 3,785 / 1,503 • W ~ OCR • Jinja: CRC • 34 pp.
Kiriggwa	Kiriggwajjo 2007 (Akalango ka Kiriggwajjo) • Advertisements / Networking • 87 / 61 • W ~ Translation • Kampala: - • 0.5 pp.
Kisaati	Various 2010 (Kisaati Kawooya Mugainho) • Radio talk shows / Politics • 5,155 / 2,041 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:44:45
KisambIA	Kisambira, Amurafeeri 2004 (Kisambira Amurafeeli) • Academic documents / Language • 219 / 161 • W ~ Retyping of hand-written document • Kampala: - • 2 pp.
Kisiki	Gulere, Cornelius 2006 (Kisiki) • Literature / Fables • 954 / 554 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 7 pp.
KiyinKbi	Lyavala-Lwanga, E.J. 1969 (Kiyini Kibi) • Literature / Language • 19,256 / 7,737 • W ~ Retyping of image • Kampala: Milton Obote Foundation • 123 pp.
KKhst	Kirunda, Kivejinja 2012 (Interview on the History of Busoga and Lusoga) • Interviews / History • 8,296 / 2,659 • O ~ Transcription • Kampala: - • 1:25:07
KKspb	Kirunda, Kivejinja 2012 (Interview on the Sapoba Legacy) • Interviews / History • 5,801 / 2,239 • O ~ Transcription • Kampala: - • 0:56:10
Kodh'eyo	Various 1997–1998 (1997) (Kodh'eyo: Busoga etebenkere) • Journalism / Networking • 185,843 / 45,945 • W ~ OCR & Retyping • Kampala: Kodh'eyo Publications • 341 pp.
Kolatug	Anon. 2010s (Kola Tugyeyo) • Songs - Modern / Relationships • 158 / 65 • O ~ Transcription • Anon.: - • 0:04:15
KufaLeka	Gulere, Cornelius 2006 (Kufa, Leka Kweghaana) • Literature / Fables • 1,091 / 653 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 11 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Kuwagi1	Kigenyi, Amos 2000s (Kuwagila Kaguta) • Songs - Modern / Politics • 161 / 100 • O ~ Transcription • Anon.: - • 0:04:50
Kyandib	Salimu 2000s (Kyandibaile Kilungi) • Songs - Traditional / Life • 683 / 360 • O ~ Transcription • Anon.: - • 0:12:35
Ky'oyend	Anon. 2010 (Ky'oyenda) • Songs - Modern / Inspirational • 357 / 173 • O ~ Transcription • Internet: YouTube • 0:03:31
LandPoli	The Uganda National Land Policy 2010 (Ekighandiiko Ekigema ku Itaka) • Policy documents - NGOs / Sensitization • 7,602 / 2,680 • W ~ e-Transfer • Kampala: The Uganda Land Alliance, supported by Concern Worldwide • 40 pp.
Lexiko09	Nabirye, Minah 2009 (Eiwanika ly'Olusoga Elyasookela Ilala) • Academic documents / Language • 323 / 187 • W ~ Own writing • Stellenbosch: WAT • 0.5 pp.
Lexiko10	Nabirye, Minah 2010 (Eiwanika ly'Olusoga Lizuuseeku Omukozesa Ataali Mutuubilile) • Academic documents / Language • 332 / 203 • W ~ Own writing • Stellenbosch: WAT • 0.5 pp.
Lexiko11	Nabirye, Minah 2011 (Okulondoola Engeli Eitu ly'Olusoga bwe Linaatuusibwa mu Iwanika: Omutindo ogulaga olulimi bwe luli, bwe luteekwa okuba oba bwe lube lutwalibwe) • Academic documents / Language • 386 / 223 • W ~ Own writing • Stellenbosch: WAT • 0.5 pp.
Lexiko13	Nabirye, Minah 2013 (Okuta Eiwanika ly'Olusoga mu Mbeela y'Omutegekozawiso Ogosomwa ku Kompyuta: Ebizibu n'ebiluubililwa) • Academic documents / Language • 174 / 121 • W ~ Own writing • Stellenbosch: WAT • 0.5 pp.
Lukabyo	Various 2012 (Lukabyo Eulogy) • Biblical documents / Religion • 358 / 218 • W ~ OCR • Ibulanku: - • 12 pp.
LusgOls2	Gulere, Cornelius 2012 (Olusoga Olusookelwaaku (Level 2)) • Academic documents / Language • 4,570 / 2,022 • W ~ OCR • Busembatya: Lusoga Language Academic Board (LLAB) • 37 pp.
LusHymns	Cultural Research Centre (CRC) 2012 (Lusoga Hymns) • Biblical documents / Religion • 477 / 299 • W ~ OCR • Jinja: Diocese of Jinja • 4 pp.
LusLdPr	Bible Society Uganda (BSU) 2012 (The Lord's Prayer in Lusoga) • Biblical documents / Religion • 114 / 91 • W ~ OCR • Internet: Bible Society Uganda • 1 p.
Lusmades	Kagoya, Michelle Johnson 2011 (Lusoga Made Simple) • Academic documents / Language • 2,181 / 800 • W ~ OCR • Jinja: CRC • 134 pp.
LusMath1	Gulere, Cornelius 2006 (Lusoga Mathematics Primer 1) • Academic documents / Science • 5,081 / 259 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 105 pp.
Lusterm1	Jore, Nathan D. 2011 (Obutonde Okutuuka ku Kuva Ekiketozo 1) • Biblical documents / Religion • 7,046 / 2,258 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 84 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Lusterm2	Jore, Nathan D. 2011 (Ensi Ensubize Okutuuka ku Banabbi Ekiketezo 2) • Biblical documents / Religion • 9,937 / 3,094 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 84 pp.
Lusterm3	Jore, Nathan D. 2011 (Obulamu bwa Yesu Kiketezo 3) • Biblical documents / Religion • 6,032 / 2,328 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 75 pp.
Lusterm4	Jore, Nathan D. 2011 (Amagezi Ekiketezo 4) • Biblical documents / Religion • 4,702 / 1,913 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 69 pp.
Lusterm5	Jore, Nathan D. 2011 (Enono Ekiketezo 5) • Biblical documents / Religion • 6,481 / 2,356 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 89 pp.
Lusterm6	Jore, Nathan D. 2011 (Obuwereza Ekiketezo 6) • Biblical documents / Religion • 6,599 / 2,675 • W ~ e-Transfer • Plymouth, MN: Ambassador Institute • 96 pp.
Luthour	Lutheran Church Ministry 2010 (Lutheran Hour) • Radio talk shows / Religion • 7,604 / 2,441 • O ~ Transcription • Jinja: NBS FM • 0:59:14
LwakAbTb	Cultural Research Centre (CRC) 2000 (Lwaki Abakazi Tibabeeda Mulambo) • Literature / Fables • 8,086 / 2,615 • W ~ OCR • Jinja: CRC • 77 pp.
M&G-Mkwa	Various 2009 (Introduction Ceremony 3) • Celebrations / Marriage • 11,194 / 3,520 • O ~ Transcription • Kampala: - • 2:00:00
MAAppdx3	Nabirye, Minah 2008 (Test 1A Questionnaire) • Academic documents / Language • 657 / 343 • W ~ Own writing • Kampala: Makerere Institute of Languages • 4 pp.
MAAppdx6	Nabirye, Minah 2008 (Okugezesa Eiwaniika ly'Olusoga) • Academic documents / Language • 696 / 363 • W ~ Own writing • Kampala: Makerere Institute of Languages • 4 pp.
MagezinK	Gulere, Cornelius 2007 (Magezi ni Kasilu) • Literature / Fables • 2,094 / 907 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 17 pp.
Mail13-5	Various 2013 (Lusoga Mails 1) • E-mails / Networking • 13,538 / 5,367 • W ~ e-Transfer • Internet: Yahoo! • 69 pp.
Mail13-6	Various 2013 (Lusoga Mails 2) • E-mails / Networking • 9,446 / 4,079 • W ~ e-Transfer • Internet: Yahoo! • 31 pp.
MaisoTig	Gulere, Cornelius 2006 (Maiso Tigalya Guba Mwoyo) • Literature / Fables • 1,377 / 855 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 10 pp.
MarikoA	Bible Society Uganda (BSU) 1996 (Mariko. Amawulire Amalungi mu Lusoga) • Biblical documents / Religion • 12,416 / 3,743 • W ~ OCR • Kampala: The Bible Society of Uganda • 54 pp.
Mazima	Various 2010 (Mazima) • Radio talk shows / Politics • 7,423 / 2,375 • O ~ Transcription • Jinja: NBS FM • 1:11:16
MenhaW1	Nabirye, Minah 2008 (Website Information) • Academic documents / Language • 1,674 / 782 • W ~ e-Transfer • Internet: Menha Publishers • 2 pp.
MenhaW2	Nabirye, Minah 2010 (Engeli Kampuni bwe Yatandiika) • Academic documents / Language • 1,140 / 575 • W ~ e-Transfer • Internet: Menha Publishers • 3 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Missa1	Cultural Research Centre (CRC) 2012 (Missa mu Lusoga Ebiseera eby'Omwaka n'Enaku edh'Abatuukirivuu) • Biblical documents / Religion • 31,122 / 4,019 • W ~ OCR • Jinja: Diocese of Jinja • 195 pp.
Missa2	Cultural Research Centre (CRC) 2012 (Missa mu Lusoga Ebiseera eby'Amatuuka n'Amazaalibwa) • Biblical documents / Religion • 12,344 / 2,249 • W ~ OCR • Jinja: Diocese of Jinja • 65 pp.
Missa3	Cultural Research Centre (CRC) 2012 (Missa mu Lusoga Ensengeka y'Emikolo gya Wiiki Entukuvu) • Biblical documents / Religion • 7,165 / 1,887 • W ~ OCR • Jinja: Diocese of Jinja • 38 pp.
Missa4	Cultural Research Centre (CRC) 2012 (Missa mu Lusoga Ebiseera eby'Amazuukira) • Biblical documents / Religion • 10,593 / 1,489 • W ~ OCR • Jinja: Diocese of Jinja • 73 pp.
Missa5	Cultural Research Centre (CRC) 2012 (Missa mu Lusoga Ebiseera eby'Ekisiibo) • Biblical documents / Religion • 7,979 / 1,639 • W ~ OCR • Jinja: Diocese of Jinja • 53 pp.
MpeeBulm	Ministry of Health 2010 (Mpeereza ya Bulamu) • Policy documents - Government / Health • 9,773 / 2,305 • W ~ OCR • Internet: Ministry of Health • 29 pp.
Mpuuta	Malagala, Stephen 2010s (Mpuuta na Mwogo) • Songs - Traditional / Marriage • 951 / 497 • O ~ Transcription • Kamuli: - • 0:09:23
MuBigrBy	Mwesigwa, Roy 2000 (Mu Bigere Bye. Olugero lw'Abasoga ku kugoberera Yesu) • Biblical documents / Religion • 25,870 / 5,519 • W ~ OCR • Jinja: Church of Christ • 111 pp.
Mukamug	Mukaabya, Willy 2010 (Muka Mugandawo Twala Butwale) • Songs - Modern / Marriage • 521 / 252 • O ~ Transcription • Internet: YouTube • 0:07:37
Mukazimk	Anon. 2010s (Omukazi Muka Beene) • Songs - Modern / Rehabilitation • 365 / 169 • O ~ Transcription • Iganga: - • 0:04:22
Muko	Mukaabya, Willy 2010 (Muko) • Songs - Modern / Marriage • 586 / 350 • O ~ Transcription • Internet: YouTube • 0:08:23
Mulinaa	Mugwisa 2000s (Mulinaanwa) • Songs - Traditional / Networking • 476 / 179 • O ~ Transcription • Anon.: - • 0:09:17
Musoke	Various 2010 (Eby'omu Ndhu) • Radio talk shows / Marriage • 7,407 / 2,514 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:04:40
MutMalym	Bible Society Uganda (BSU) 2012 (Mutendwa Malyaamu) • Biblical documents / Religion • 55 / 47 • W ~ OCR • Internet: Bible Society Uganda • 1 p.
Muwuliil	Kigenyi, Amos 2010 (Muwuliile Bulungi) • Songs - Modern / Gratitude • 297 / 79 • O ~ Transcription • Internet: YouTube • 0:04:45
Mwebale	Mugwisa 2000s (Mwebale Ssaba) • Songs - Traditional / Gratitude • 560 / 341 • O ~ Transcription • Anon.: - • 0:07:43
Mwewewo	Geo Bless 2010s (Mwewewo) • Songs - Modern / Marriage • 288 / 202 • O ~ Transcription • Anon.: - • 0:04:39

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
MwidhTgT	Kasozi, John 2000 (Mwidhe Tugye Tusenge) • Literature / Religion • 1,568 / 604 • W ~ Retyping of image • Jinja: Diocese of Jinja • 12 pp.
MwidTufm	Ssajabi, Sophronius 1999 (Mwidhe Tufume) • Literature / Fables • 5,867 / 2,087 • W ~ OCR • Jinja: CRC • 62 pp.
Mwino	Various 2010 (Mwino Akuwa y'Owa) • Radio talk shows / Politics • 9,211 / 2,666 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:19:37
Mwinoak	Anon. 2000s (Mwino Akuwa y'Owa) • Songs - Traditional / Networking • 151 / 76 • O ~ Transcription • Anon.: - • 0:08:20
NAAD113	Various 2009 (NAADS 1) • Radio talk shows / Science • 5,797 / 2,092 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:51:05
NAAD1130	Various 2009 (NAADS 2) • Radio talk shows / Science • 6,357 / 2,087 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:55:56
NabM10	Nabirye, Minah 2010 (Nabirye 1) • E-mails / Networking • 2,530 / 1,172 • W ~ e-Transfer • Local: Self • 8 pp.
NabM11.6	Nabirye, Minah 2011 (Nabirye 2) • E-mails / Language • 840 / 533 • W ~ e-Transfer • Local: Self • 3 pp.
NabM11.7	Nabirye, Minah 2011 (Nabirye 3) • E-mails / Language • 472 / 287 • W ~ e-Transfer • Local: Self • 2 pp.
Nakoowa	Crado 2010s (Nakoowa) • Songs - Modern / Marriage • 242 / 184 • O ~ Transcription • Anon.: - • 0:04:06
Nantaga	Mugwisa 2000s (Nantagalagilwa) • Songs - Traditional / Marriage • 434 / 223 • O ~ Transcription • Anon.: - • 0:07:48
Nantameg	Gulere, Cornelius 2007 (Nantamegwa) • Literature - Plays / Life • 9,702 / 3,522 • W ~ OCR • Busembatya: Lusoga Language Academic Board (LLAB) • 48 pp.
NantamuD	Nantamu, Dyogo Peter 2011 (Factors Associated With Male Involvement in Maternal Health Care Services in Jinja District, Uganda) • Academic documents / Health • 1,328 / 634 • W ~ e-Transfer • Kampala: Makerere University School of Public Health • 108 pp.
Ndimbonk	Gulere, Cornelius 2006 (Ndimubonakuuli) • Literature / Fables • 1,192 / 714 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 11 pp.
Ndimugez	Various 1998–1999 (1998) (Ndimugezi n'Omukobere: The factfinder) • Journalism / Networking • 15,821 / 7,159 • W ~ OCR & Retyping • Jinja: Ndimugezi Publications • 42 pp.
Ndinimuk	Gulere, Cornelius 2011 (Ndi ni Mukazi Wange) • Literature / Riddles • 934 / 509 • W ~ OCR • Internet: Google books • 12 pp.
Ngulina	Gulere, Cornelius 2011 (Lusoga: Nguli namanha) • Literature / Fables • 1,445 / 531 • W ~ OCR • Internet: Google books • 21 pp.
Nibwonv	Kigenyi, Amos 2010 (Ni bw'Onvuma Agaiso) • Songs - Modern / Relationships • 393 / 107 • O ~ Transcription • Internet: YouTube • 0:04:01
Nkontam	Gulere, Cornelius 2007 (Nkontamuti) • Literature / Fables • 1,082 / 585 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 20 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Nsangail	Magoola, Racheal 2010s (Nsangaile) • Songs - Traditional / History • 311 / 163 • O ~ Transcription • Iganga: - • 0:06:09
Nsobola	Gulere, Cornelius 2011 (Nsobola Nsobola) • Literature / Fables • 979 / 596 • W ~ OCR • Internet: Google books • 16 pp.
Obbangai	Afrigo Band 2010 (Obbangaina) • Songs - Modern / Relationships • 232 / 109 • O ~ Transcription • Internet: YouTube • 0:04:07
Obufumbo	Gulere, Cornelius 2012 (Obufumbo) • Literature - Booklets / Sensitization • 1,306 / 669 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 8 pp.
Obughan	Baisi 2010s (Obughangwa Bwaife) • Songs - Traditional / Sensitization • 226 / 148 • O ~ Transcription • Anon.: - • 0:06:42
Obugumb	Anon. 2000s (Obugumba Buluma) • Songs - Traditional / Marriage • 297 / 154 • O ~ Transcription • Anon.: - • 0:07:43
Obukyay	Geo Bless 2010s (Obukyayi) • Songs - Modern / Relationships • 340 / 185 • O ~ Transcription • Anon.: - • 0:04:37
Obululu	Malagala, Stephen 2000s (Obululu) • Songs - Traditional / Politics • 678 / 426 • O ~ Transcription • Anon.: - • 0:06:51
Obwende	Malagala, Fiida 2010s (Obwende Mpisa) • Songs - Traditional / Marriage • 327 / 249 • O ~ Transcription • Anon.: - • 0:05:51
Ogunguma	Babirye, Judith 2010 (Ogungumale Kibbumba) • Songs - Gospel / Religion • 357 / 77 • O ~ Transcription • Internet: YouTube • 0:03:16
Ogusolo	Gulere, Cornelius 2011 (Ogusolo n'Ekikaadho) • Literature / Fables • 1,030 / 594 • W ~ OCR • Internet: Google books • 15 pp.
Okozeewo	Crado 2010s (Okozeewo Ki) • Songs - Modern / Gratitude • 536 / 306 • O ~ Transcription • Anon.: - • 0:06:08
Okukonk	Mugwisa 2000s (Okukonkona Embaile) • Songs - Traditional / Inspirational • 464 / 280 • O ~ Transcription • Anon.: - • 0:08:59
Okukyala	Various 2011 (Okukyala kw'Abasiki e Buwaabe) • Celebrations / Politics • 17,052 / 4,973 • O ~ Transcription • Iganga: - • 3:28:26
OkusanT1	Gulere, Cornelius 2007 (Okusanhusa Tikwesanusu 1) • Literature / Language • 614 / 357 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 15 pp.
OkusanT2	Gulere, Cornelius 2007 (Okusanhusa Tikwesanusu 2) • Literature / Language • 219 / 144 • W ~ OCR • Internet: Google books • 15 pp.
OkusBkUg	Mbowa, Rose 2013 (Okusaaka kwa Bakazi ba Uganda) • Policy documents - Human rights / Sensitization • 605 / 363 • W ~ OCR • Internet: Human Rights Advocacy • 5 pp.
Okuwasa	Geo Bless 2010s (Okuwasa) • Songs - Modern / Money • 408 / 276 • O ~ Transcription • Anon.: - • 0:04:53
OkwEniBz	Uganda Human Rights Commission (UHRC) 2008 (Okwanganga Ennimi dh'Obuzaale) • Policy documents - Endangered languages / Language • 1,857 / 801 • W ~ OCR • Internet: Anon. • 11 pp.

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Olimumd	Salimu 2010 (Oli mu Ndoolo) • Songs - Modern / Sensitization • 934 / 188 • O ~ Transcription • Jinja: CRC • 0:06:08
Olugelsk	Various 2010 (Bible Story 1) • Biblical documents / Religion • 691 / 358 • O ~ Transcription • Internet: YouTube • 0:08:48
Olukiiko	Wangoola, Paulo 2009 (Olukiiko Oluluubililia Okugaita Eitwale Iya Busoga) • Policy documents - Busoga Kingdom / History • 1,710 / 764 • W ~ Translation • Kampala: Task Force on the Principled Cultural Unity - Busoga Kingdom • 6 pp.
Olumbe	Malagala, Stephen 2000s (Olumbe) • Songs - Traditional / Health • 467 / 311 • O ~ Transcription • Anon.: - • 0:06:01
Olumbel	Anon. 2000s (Olumbe Lulaile) • Songs - Traditional / Health • 507 / 215 • O ~ Transcription • Anon.: - • 0:08:50
Olumbes	Anon. 2000s (Olumbe Siliimu) • Songs - Traditional / Health • 423 / 242 • O ~ Transcription • Anon.: - • 0:09:34
Olusoga1	Gulere, Cornelius 2007 (Olusoga Olusookelwaku (Level 1)) • Literature / Language • 2,800 / 1,380 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 33 pp.
Olusoga3	Gulere, Cornelius 2007 (Olusoga Olusookelwaku (Level 3)) • Literature / Language • 3,015 / 1,418 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 36 pp.
Omudaal	Salimu 2000s (Omudaala) • Songs - Traditional / Life • 702 / 358 • O ~ Transcription • Anon.: - • 0:11:37
Omugole	Bukenya, Austin 2007 (Omugole) • Literature - Plays / Marriage • 11,898 / 3,753 • W ~ OCR • Busembatya: Lusoga Language Academic Board (LLAB) • 45 pp.
Omukazi	Gulere, Cornelius 2010 (Omukazi) • Literature / Fables • 227 / 154 • W ~ OCR • Internet: Anon. • 1 p.
Omukonk	Geo Bless 2010s (Omukonkoonhia) • Songs - Modern / Relationships • 272 / 162 • O ~ Transcription • Anon.: - • 0:04:38
Omulam	Mugwisa 2000s (Omulam Tiyeesigika) • Songs - Traditional / Relationships • 425 / 254 • O ~ Transcription • Anon.: - • 0:07:58
Omulilo	Gulere, Cornelius 2007 (Omulilo) • Literature / Fables • 629 / 398 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 20 pp.
Omulyam	Kirimungu, Siragi 2000s (Omulya Mmele) • Songs - Traditional / Life • 493 / 284 • O ~ Transcription • Anon.: - • 0:07:59
Omumbeed	Gulere, Cornelius 2007 (Omumbeedha Omutuufu) • Literature / Fables • 728 / 458 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 15 pp.
Omusaad	Anon. 2000s (Omusaadha Awalamula Egaali) • Songs - Traditional / Money • 135 / 55 • O ~ Transcription • Anon.: - • 0:06:47
OmusAkb	Luboga, Sam 2012 (Omusoga Akoba) • Literature / Fables • 1,839 / 1,216 • W ~ OCR • Busembatya: Lusoga Language Academic Board (LLAB) • 15 pp.
Omutmuz	Salimu 2010 (Omuntu Muzibu) • Songs - Modern / Life • 683 / 264 • O ~ Transcription • Jinja: CRC • 0:05:07
OmuvangL	Kirimungu, Siragi 2000s (Omuvangano) • Songs - Traditional / Rehabilitation • 353 / 175 • O ~ Transcription • Anon.: - • 0:07:01

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
OmuvngMB	Cultural Research Centre (CRC) 1999 (Omuvangano mu Busoga) • Biblical documents / Religion • 6,927 / 2,304 • W ~ OCR • Jinja: CRC • 28 pp.
Omuzail	Mugwisa 2000s (Omuzaila Muwe Ekitibwa) • Songs - Traditional / Marriage • 344 / 168 • O ~ Transcription • Anon.: - • 0:07:28
OmwanaK	Gulere, Cornelius 2007 (Omwana Kwania) • Literature / Fables • 253 / 190 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 26 pp.
Omwenge	Salimu 2000s (Omwenge Seneta) • Songs - Traditional / Rehabilitation • 531 / 299 • O ~ Transcription • Anon.: - • 0:11:20
Omwenska	Kabugu, Milton Peter 2010 (Omwenkanonkano) • Songs - Modern / Sensitization • 884 / 167 • O ~ Transcription • Jinja: CRC • 0:07:14
OrderoM1	Cultural Research Centre (CRC) 2012 (Order of Mass 1) • Biblical documents / Religion • 1,846 / 903 • W ~ OCR • Namwendwa: Diocese of Jinja • 15 pp.
OrderoM2	Cultural Research Centre (CRC) 2012 (Order of Mass 2) • Biblical documents / Religion • 872 / 434 • W ~ OCR • Jinja: Diocese of Jinja • 16 pp.
Otabona	Mata, Nassani 2010 (Otabona Bukaile na Nvu) • Songs - Traditional / Inspirational • 574 / 179 • O ~ Transcription • Internet: YouTube • 0:07:08
Otawuli	Kirimungu, Siragi 2000s (Otawulila Boogezi) • Songs - Traditional / Life • 813 / 409 • O ~ Transcription • Anon.: - • 0:12:42
Otelaok	Gulere, Cornelius 2011 (Otela Okwila) • Literature / Fables • 315 / 229 • W ~ OCR • Internet: Google books • 14 pp.
Owayang	Baisi 2010s (Owayanga) • Songs - Traditional / Health • 358 / 219 • O ~ Transcription • Anon.: - • 0:08:14
PEAP	Ministry of Finance, Planning and Economic Development 2005 (Poverty Eradication Action Plan (PEAP)) • Policy documents - Government / Sensitization • 2,476 / 1,022 • W ~ Translation • Kampala: Makerere Institute of Languages • 12 pp.
Petsg091	Various 2009 (Lusoga Songs Performed in Twin Ceremonies 1) • Songs - Traditional / Marriage • 3,214 / 864 • W ~ Translation • Jinja: - • 13 pp.
Petsg092	Various 2010 (Lusoga Songs Performed in Twin Ceremonies 2) • Songs - Traditional / Marriage • 1,201 / 444 • W ~ e-Transfer • Jinja: - • 14 pp.
PFExtaud	Nabirye, Minah 2012 (Phonetics Fieldwork - Extra Audio files) • Interviews / Language • 52,098 / 10,051 • O ~ Transcription • Busoga: - • 8:55:38
PIbaale1	Pastor Ibaale 2010 (Pastor Ibaale 1) • Radio talk shows / Religion • 6,352 / 1,935 • O ~ Transcription • Jinja: NBS FM • 0:59:24
PIbaale2	Pastor Ibaale 2010 (Pastor Ibaale 2) • Radio talk shows / Religion • 5,728 / 1,955 • O ~ Transcription • Jinja: NBS FM • 1:01:23
PIbaale3	Pastor Ibaale 2010 (Pastor Ibaale 3) • Radio talk shows / Religion • 5,676 / 1,941 • O ~ Transcription • Jinja: NBS FM • 0:59:42
PIbaale4	Pastor Ibaale 2010 (Pastor Ibaale 4) • Radio talk shows / Religion • 6,472 / 2,051 • O ~ Transcription • Jinja: NBS FM • 1:00:29

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
Pililya	Geo Bless 2010s (Pililya) • Songs - Modern / Inspirational • 314 / 183 • O ~ Transcription • Anon.: - • 0:05:13
P1101021	Various 2010 (Plan Water and Sanitation 1) • Radio talk shows / Health • 9,534 / 2,843 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:29:37
P1101215	Various 2010 (Plan Water and Sanitation 2) • Radio talk shows / Health • 5,500 / 1,961 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:49:26
P1101216	Various 2010 (Plan Water and Sanitation 3) • Radio talk shows / Health • 5,900 / 1,990 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:50:48
Priest01	Cultural Research Centre (CRC) 1998 (Priestly Ordination of Rev. Richard Kayaga Gonza, Rev. Silvester Makwali) • Biblical documents / Religion • 517 / 252 • W ~ OCR • Bugembe: Diocese of Jinja • 26 pp.
Priest02	Cultural Research Centre (CRC) 2003 (The Priestly Ordination for Rev. Deacon Mbaziira Henry Jude, Rev. Deacon Musana Paul) • Biblical documents / Religion • 2,071 / 794 • W ~ OCR • Bugembe: Diocesan Printery • 34 pp.
Priest03	Cultural Research Centre (CRC) 2003 (Priestly Ordination of Rev. Serapio Kasuura Wamara Araali) • Biblical documents / Religion • 1,343 / 640 • W ~ OCR • Bugembe: Diocese of Jinja • 31 pp.
Priest04	Cultural Research Centre (CRC) 2005 (Priestly Ordination of Deacon Mwangi Simon Gitua, Deacon Mugabe Paschal Atwooki, Deacon Jenga Fred) • Biblical documents / Religion • 1,446 / 575 • W ~ OCR • Jinja: Little Sisters of St. Francis • 20 pp.
Publiche	School of Public Health (SPH) 2009 (Baseline Survey on Institutional Deliveries 1) • Academic documents / Health • 2,870 / 985 • W ~ Translation • Kampala: Makerere University School of Public Health • 28 pp.
Queenw	Crado 2010s (Queen Wange) • Songs - Modern / Relationships • 210 / 101 • O ~ Transcription • Anon.: - • 0:04:26
SafedelQ	School of Public Health (SPH) 2010 (Baseline Survey on Institutional Deliveries 2) • Academic documents / Health • 3,339 / 1,182 • W ~ Translation • Kampala: Makerere University School of Public Health • 28 pp.
Safedelv	Various 2010 (Safe Deliveries) • Radio talk shows / Health • 8,453 / 2,661 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:11:05
Sente	Salimu 2000s (Sente) • Songs - Traditional / Money • 868 / 448 • O ~ Transcription • Anon.: - • 0:13:32
Sentene	Kirimungu, Siragi 2000s (Sente n'Ekola) • Songs - Traditional / Money • 755 / 394 • O ~ Transcription • Anon.: - • 0:08:33
Soyabean	Anon. 2010 (Soya Bean in Lusoga) • Literature - Booklets / Science • 2,287 / 968 • W ~ Retyping of image • Internet: - • 32 pp.
StarEC1	Various 2010 (Star EC 1) • Radio talk shows / Health • 6,997 / 2,224 • O ~ Transcription • Jinja: NBS FM • 0:52:46
StarEC2	Various 2010 (Star EC 2) • Radio talk shows / Health • 7,364 / 2,208 • O ~ Transcription • Jinja: NBS FM • 0:58:01

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
StarEC3	Various 2010 (Star EC 3) • Radio talk shows / Health • 6,603 / 2,296 • O ~ Transcription • Jinja: NBS FM • 0:53:05
StarEC4	Various 2010 (Star EC 4) • Radio talk shows / Health • 6,430 / 2,312 • O ~ Transcription • Jinja: NBS FM • 0:58:35
StarEC5	Various 2010 (Star EC 5) • Radio talk shows / Health • 7,149 / 2,208 • O ~ Transcription • Jinja: NBS FM • 0:58:28
StarEC6	Various 2010 (Star EC 6) • Radio talk shows / Health • 6,939 / 2,117 • O ~ Transcription • Jinja: NBS FM • 1:00:36
StarEC7	Various 2011 (Star EC 7) • Radio talk shows / Health • 7,103 / 2,147 • O ~ Transcription • Jinja: NBS FM • 0:58:09
Strides1	Various 2010 (Strides 1) • Radio talk shows / Sensitization • 6,859 / 2,257 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:59:31
Strides2	Various 2010 (Strides 2) • Radio talk shows / Health • 7,285 / 2,276 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 1:04:51
Sunpanel	Various 2011 (Sunday Panel) • Radio talk shows / Politics • 11,466 / 3,460 • O ~ Transcription • Jinja: NBS FM • 1:28:50
ThanksGv	Cultural Research Centre (CRC) 2012 (Thanksgiving) • Biblical documents / Religion • 659 / 353 • W ~ OCR • Jinja: Diocese of Jinja • 32 pp.
Tubeepn	Geo Bless 2010s (Tubeepene) • Songs - Modern / Politics • 452 / 265 • O ~ Transcription • Anon.: - • 0:04:40
Tuboin1	Salimu 2000s (Tuboineboine 1) • Songs - Traditional / Life • 564 / 343 • O ~ Transcription • Anon.: - • 0:09:55
Tuboin2	Salimu 2000s (Tuboineboine 2) • Songs - Traditional / Life • 745 / 414 • O ~ Transcription • Anon.: - • 0:11:09
Tusanga	Anon. 2000s (Tusangaile) • Songs - Traditional / Relationships • 730 / 301 • O ~ Transcription • Anon.: - • 0:10:42
Twaghaya	Baisi 2010s (Twaghanga) • Songs - Traditional / Rehabilitation • 554 / 293 • O ~ Transcription • Anon.: - • 0:10:09
Twebaze	Salimu 2000s (Twebaze Katonda) • Songs - Traditional / History • 301 / 171 • O ~ Transcription • Anon.: - • 0:08:20
Twekubi	Geo Bless 2010s (Twekubile Dance) • Songs - Modern / Inspirational • 355 / 179 • O ~ Transcription • Anon.: - • 0:03:59
TwireKBU	Ssajabi, Sophronius 1999 (Twire ku Butaka) • Literature / Fables • 5,108 / 1,838 • W ~ OCR • Jinja: CRC • 58 pp.
TwoLusFb	Gumbo, E. & Kafuho, E. 1946 (Two Lusoga Fables) • Literature / Fables • 1,325 / 791 • W ~ Retyping of image • Kampala: The Uganda Journal • 16+24 pp.
Vooto	Anon. 2010s (Vooto) • Songs - Modern / Relationships • 338 / 70 • O ~ Transcription • Iganga: - • 0:05:06
Waalink	Magoola, Racheal 2010 (Waalinkobye) • Songs - Modern / Relationships • 289 / 129 • O ~ Transcription • Internet: YouTube • 0:04:58

Filename	Author or Performer Year or Period (Title) • Genre / Topic • Tokens / Types • W(ritten) or O(ral) ~ Source • Place: Publisher • Pages or Length of recording
WalgunD	Gulere, Cornelius 2007 (Walugundhu) • Literature / Fables • 380 / 280 • W ~ OCR • Internet: Mpolyabigere RC – RICED Center • 9 pp.
WangoInt	Wangoola, Paulo 2012 (Interview on the Evolution of the Dialects of Busoga) • Interviews / History • 7,284 / 2,312 • O ~ Transcription • Kampala: - • 1:18:05
Wankoko	Gulere, Cornelius 2007 (Wankoko ni Wamusota) • Literature / Fables • 554 / 288 • W ~ e-Transfer • Cape Town: CASAS (pre-publication) • 15 pp.
Water1	Various 2009 (District Water 1) • Radio talk shows / Health • 5,077 / 1,743 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:44:49
Water2	Various 2009 (District Water 2) • Radio talk shows / Health • 5,138 / 1,705 • O ~ Transcription • Kamuli: Kamuli Broadcasting Service • 0:45: 31
Weebale	Kiirya, Maurice 2010s (Weebale Okundoga) • Songs - Modern / Relationships • 274 / 131 • O ~ Transcription • Iganga: - • 0:04:18
WorCulLe	Federation of Female Lawyers (FIDA) and Plan Uganda 2010 (Enkolagana y'Okutumbula Eidembe) • Policy documents - NGOs / Marriage • 4,936 / 1,562 • W ~ Translation • Kampala: FIDA and PLAN Uganda • 16 pp.
WSGextr	Nabirye, Minah 2009 (Ebikookelo mu Eiwaniika ly'Olusoga) • Academic documents / Language • 12,665 / 4,039 • W ~ e-Transfer • Kampala: Menha Publishers • 79 pp.

Corpus-driven Bantu Lexicography Part 2: Lemmatisation and Rulers for Lusoga

Gilles-Maurice de Schryver, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of African Languages, University of Pretoria, Pretoria, South Africa (gillesmaurice.deschryver@UGent.be)*

and

Minah Nabirye, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of Teacher Education and Development Studies, Kyambogo University, Kampala, Uganda (minah.nabirye@UGent.be)*

Abstract: This article is the second in a trilogy that deals with corpus-driven Bantu lexicography, which is illustrated for Lusoga. The focus here is on the macrostructure and in particular on the building of a lemmatised frequency list directly within a dictionary-writing system. The programming code for the parts of the lemmatisation that may be automated is included as addenda. A second focus is on the embedded part-of-speech and alphabetical rulers, for which it is shown how these may be used to plan the actual compilation of the dictionary entries.

Keywords: BANTU, LUSOGA, CORPUS LEXICOGRAPHY, LEMMATISATION, LEMMATISED FREQUENCY LIST, PART-OF-SPEECH RULER, ALPHABETICAL RULER, MULTIDIMENSIONAL LEXICOGRAPHIC RULER, DICTIONARY PLANNING, DICTIONARY-WRITING SYSTEM, TLEX, TSHWANELEX

Obufunze: Omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu. Ekitundu 2: Okugelaagelania eigambowaziso n'enta dha namugelo waalyo mu walifu w'Olusoga. Olupapula luno n'olwo'kubili mu nteeko y'okulaga omusomo gw'omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu ogulaga omulimu ogw'akolebwa ku Lusoga. Mu lupapula luno eisila liteebwa ku muteeko gw'omutindiigo okusingila ilala ku kuzimba olukalala lwa namungi w'ebigambowazo mu muteeko ogukozesebwa okuwandiika amawanika. Namugelo w'okutegekuza ebitundu by'okugambowaza ebisobola okuba mu mbeela ya kaneetindiigo bilagibwa mu kikugilo. Eisila ely'okubili lili ku mbu dh'ebigambo edh'ennimbyo n'engeli ye dhilagibwa mu nsengeka ya walifu ng'olupapula luno kwe lusenziila okuwa endowooza ekoba nti ebintu bino ebibili bisobola okukozesebwa okutaawo omusingi gw'okwingiza ebigambo mu iwanika.

Ebigambo ebikulu: BANTU, LUSOGA, EITU LY'ANAMAWAIKA, OKUGAMBOWAZA, OLUKALALA LWA NAMUNGI W'EBIGAMBOWAZO, ENNEYOLEKA Y'EMBU, ENNEYOLEKA YA WALIFU, Omutengo gw'ENNEYOLEKA YA NAMAWANIKA, ENTEGEKA Y'EIWANIKA, ENGELI EDHIKOZESEBWA OKUWANDIIKA AMAWANIKA, TLEX, TSHWANELEX

1. Goal of the present study

This article is concerned with the use of corpora to successfully kickstart Bantu-language dictionary projects. Considering the traditional lexicographic distinction between the macrostructural and the microstructural level, this therefore means that the present study will focus on the design of the macrostructure of a Bantu-language dictionary, for which Lusoga will serve as an example. The major reference for any corpus-based macrostructural issues in Bantu lexicography is de Schryver and Prinsloo (2000). A year later, de Schryver and Prinsloo (2001) looked at the difference between intuition-based and corpus-based designs of various lemma-sign lists, as found in and for Northern Sotho dictionaries. While a single study on how to draw up a dictionary's macrostructure may suffice for a disjunctively-written Bantu language like Northern Sotho, much more guidance is certainly needed for the conjunctively-written ones.¹ To date, there seems to be just one such published study, for Southern Ndebele (de Schryver 2003). In our case study for Lusoga below, which is based on Nabirye (2016), we will further develop the proposals from the 2003 study, and will in effect offer a hands-on method which may be performed directly within a dictionary-writing system. The programming code needed for the actual lumping of all the members of each single lemma, as well as for the summations of the underlying corpus frequencies, and the calculation of the frequency bands, will be presented as addenda.

As a supplementary objective, we will want to uncover the relationships between lemmatised frequency lists of conjunctive Bantu languages, and their unlemmatised counterparts. While lemmatised and unlemmatised frequency lists may be near-identical for a disjunctive Bantu language like Northern Sotho (Prinsloo and de Schryver 2007), this is certainly not the case for a conjunctive one like Lusoga. This part of the study will inevitably also require a consideration of two types of rulers: 'part-of-speech rulers' and 'alphabetical rulers' (aka 'multidimensional lexicographic rulers') (de Schryver 2013). In order to put our results in perspective, comparisons will furthermore be made with comparable data freshly drawn from the *Oxford Bilingual School Dictionary: Zulu and English* (de Schryver 2010a).

2. Automated vs. manual, and semi-manual lemmatisation

How does one begin analysing a corpus with the aim of compiling a dictionary of the language covered by that corpus? Modern dictionary-makers will want to start from a lemmatised frequency list derived from that corpus, with which they can set out to build the macrostructure of their dictionaries. A good entry point for the concept of lemmatisation in the field of computational and corpus linguistics remains Kilgarriiff's:

By 'lemmatised', we mean two things. First, for verbal *aim*, the count will consider all instances of *aim*, *aims*, *aiming*, *aimed*; and second, it will exclude all non-

verbal instances, so nominal *aim* and *aims* will not be counted. The count will be of verbal instances only of any of the four forms.
(Kilgarriff 1997: 139)

In other words, the idea is to take a list of orthographic words, each with their type frequency as counted in a corpus, and to turn that list into its lemmatised counterpart, now with summed frequencies and a part of speech for each lemma. The result is a so-called 'lemmatised frequency list'.

While automatic lemmatisers capable of processing raw corpus data may be available for several of the world's major languages, no such software has of course been written for Lusoga. Actually, for the Bantu languages as a whole, only Swahili has been provided with working tools for this task, by Hurskainen (1992, 2016) who uses a rule-driven approach, and by the AfLaT team (De Pauw et al. 2006) who use a data-driven approach. The AfLaT team also developed small data-driven part-of-speech taggers for Northern Sotho, Zulu and Cilubà (De Pauw et al. 2012), while a team at the University of South Africa (UNISA) built broad-coverage finite-state morphological analysers for Xhosa, Swati and Southern Ndebele (Bosch et al. 2008) by adapting an existing prototype morphological analyser for Zulu (Bosch and Pretorius 2003, 2004).

In his MA, de Schryver (1999: 118-129) proposed a low-key, fully manual approach to the lemmatisation task of a Bantu language, which he successfully applied to Cilubà for the compilation of a set of bilingual Cilubà-Dutch dictionaries (de Schryver and Kabuta 1997, 1998). His basic assumption was that there is no need to lemmatise an entire corpus, as only the frequent orthographic word forms are needed as lemma signs in a general-language dictionary. Taking into account the Zipfian distribution of corpus frequencies (Zipf 1935, Kilgarriff 1997: 136-137), it is indeed clear that the lemmatised forms of low-frequency orthographic words and hapaxes hardly make a dent in what is frequent. De Schryver explained his approach as follows, after having used WordSmith Tools (Scott 1996–2018) to calculate the frequency of all the orthographic words in a 300 000-word corpus of Cilubà:

[...] we simply went through the first 1,000 items of the [WordSmith Tools output, ranked in descending frequency order] and lemmatised 'by hand.' For nouns this meant that, when we encountered a singular form, we added the frequency of the plural form (or vice versa), where relevant. For verbs this meant that we kept track of those verbs we had already encountered and added the frequency of every single 'conjugated form' we encountered subsequently. Also, for very frequent verbs we brought together the frequencies of the entire paradigm. In addition to this 'true lemmatisation' we joined divergent orthographies — and this for all possible parts of speech.
(de Schryver 1999: 125)

To move from a lemmatised frequency list to the actual macrostructure, de Schryver (1999: 127-128) further stipulated that candidate lemma signs should occur 'in a sufficient variety of sources' (Sinclair 1995: ix), or as put by Knowles:

[...] a word must occur evenly in a large number of the stratified sub-samples rather than excessively often in a small number of them, given that these two very different cases could show identical 'total-corpus' frequencies. (Knowles 1983: 188)

Finally, and in imitation of Kilgarriff (1997), de Schryver (1999: 150-152) also marked the frequent lemma signs in his dictionary, using three frequency bands which had been directly derived from the top ranks as seen in his lemmatised frequency list.

In de Schryver (2003) a suggestion was made to enlist the power of spreadsheet software for the same task, where it was illustrated for Southern Ndebele. In the latter article, a four-step methodology was introduced to go from a raw corpus (i.e., a corpus without any linguistic annotations) to a lemmatised frequency list (i.e., the list of candidate dictionary citation forms together with summed frequencies, ordered from most to lesser frequent). The steps themselves have been summarised as follows:

In Step 1 top-frequency words are extracted from a corpus of running text. This step can be performed with versatile corpus query software such as WordSmith Tools. In Step 2 the dictionary-citation forms are isolated from each of the top-frequency items; in Step 3 the dictionary-citation forms that are equal as well as their corresponding frequencies are brought together; and in Step 4 frequency bands are added to the lemma-sign list. Steps 2 to 4 can easily be performed with spreadsheet software such as Microsoft Excel. (de Schryver 2003: 22-23)

Observe that in this four-step methodology, parts of speech were not taken into account, as they should have been. This 'error'² has been corrected in the method to be explained now.

Over the subsequent years, the use of spreadsheet software morphed into using the dictionary application TshwaneLex (TLex) (Joffe and de Schryver 2002–18) to undertake Steps 2 to 4. When using TLex to lemmatise corpus data, orthographic words together with their frequencies and their spread across the corpus texts constitute the input, while the output consists of the lemma signs, with frequencies, parts of speech, ranks and frequency bands, and, optionally, main meanings. In effect, the Bantu to English sides of the school dictionaries for Northern Sotho, Zulu and Xhosa published by Oxford University Press Southern Africa (OUPSA) (de Schryver 2007, 2010a, de Schryver and Reynolds 2014) have all used TLex to draw up the macrostructure along these lines.³

Even though an in-depth analysis was undertaken of the compilation of the OUPSA Zulu school dictionary, the creation of its macrostructure was not discussed as part of that analysis: 'Detailing how the Zulu lemma list was created would need at least one other paper-length treatment' (de Schryver 2010b: 166). By explaining how Steps 2 to 4 may be performed within TLex in the present article (as will be done in §3 below), we will (finally) have begun dealing with this issue in the scientific literature of our discipline.

3. From corpus to lemmatised frequency list

As was seen in Part 1 of the present series of three articles, a Lusoga corpus of 1.7 million words (tokens) contains approximately 200 000 orthographically different words (types), and it is the latter that need to be lemmatised. Two hundred thousand words are still too many to look at manually, so, as a proxy, the idea is again to work with the top-frequent orthographic words only, and thus also to lemmatise only that top section. In practical terms one chooses a cut-off frequency, and focuses on all the types with a frequency at and above that threshold. We decided to work through about 10 000 types, which corresponded to a cut-off frequency of 12 in the 1.7m Lusoga corpus.

By lemmatising the top 10 000 orthographic words in a Lusoga corpus, all the common 'words' of the language will be known: each will have been given a part-of-speech tag, as well as a relative frequency (and in the approach that will be suggested, also a brief meaning). The term *word* was placed between quotes, as we are referring here to the component known to computational linguists as the *lemma*, to dictionary-makers as the *dictionary citation form*, to metalexigraphers as the *lemma sign*, and to Bantuists most likely as the *stem*.

The full 1.7m Lusoga corpus was loaded into WordSmith Tools, and with its *WordList* tool a wordlist of all the orthographic words in the corpus, together with their respective frequencies and the number of files each orthographic word occurs in, was generated. This information was imported into TLex, using its *Import* function. The approach from then onwards was to go down the frequency list in TLex, down to frequency 12, and to add for each orthographic word the following: the lemmatised form, the part of speech, and a brief meaning — all in dedicated slots in the dictionary-writing system. Differences in orthography were taken care of on the fly, as a uniform spelling was pursued in the slot for the lemma. See Figure 1 for a screenshot of the first step: the orthographic form from the corpus is in dark blue at the beginning of each entry; the lemmatised form follows in black and between square brackets; the part of speech is in pink and italics; the brief meaning(s) of the lemma is/are in green; the frequency of the orthographic form is in red and italics preceded by 'freq.'; the rank is in light blue and preceded by 'rank'; and the number of files in which the orthographic form was found is in black preceded by a hashtag and the word 'texts'.

As we proceeded down the frequency list,⁴ the *fanouts* tool of TLex enabled us to preview those unlemmatised forms that would eventually be brought together under a single lemma. In the DTD (i.e., Document Type Definition (Joffe and de Schryver 2005)) one may actually choose which field to use for that, typically the field for the TEs (i.e., the translation equivalents), but at times using the lemma field for fanouts is also handy. The latter is done in Figure 2. Regardless of which one is used for fanouts, during actual lemmatisation the software will need to take the lemma *in combination with* the part of speech into account.

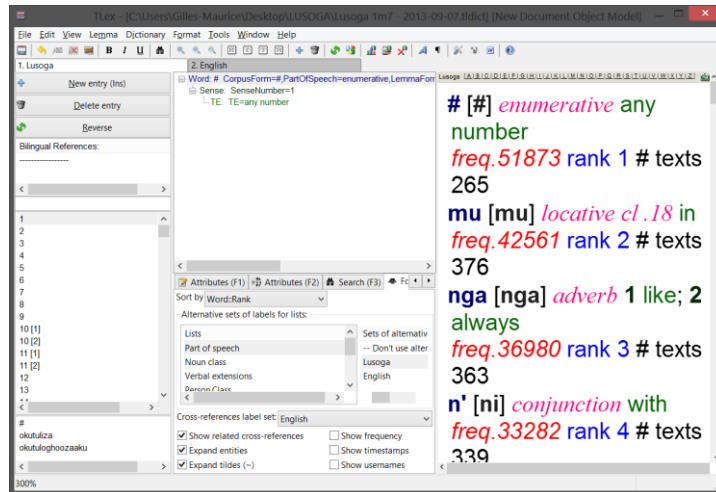


Figure 1: Lemmatising the 1.7m Lusoga corpus in TLex: going down the unlemmatised frequency list

In Figure 2 we went back to the infinitive form for the verb 'to come'. All other entries where we added **-idha** as a lemma are automatically brought together by the fanouts tool. They are all verbs, and they will indeed all be merged into a single **-idha**, and their respective frequencies will all be summed.

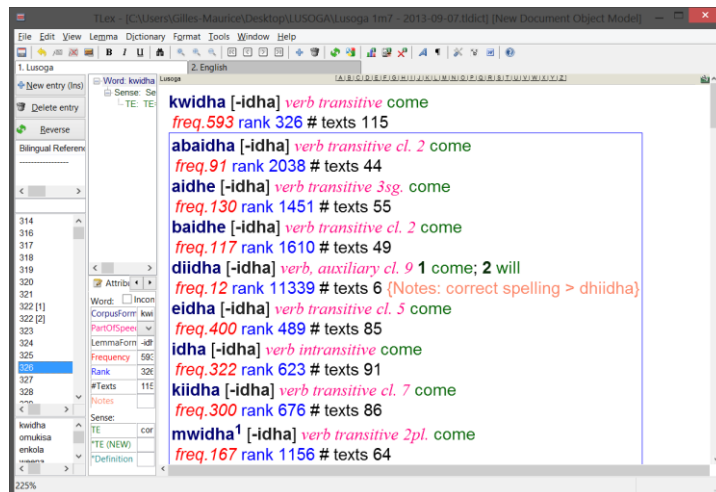


Figure 2: Lemmatising the 1.7m Lusoga corpus in TLex: the fanouts tool brings all the entries with the same lemma together

Contrast this with the material seen in Figure 3, where the orthographic forms with **-kazi** as the lemma are brought together. Given that there are both nominal and adjectival forms, these two word classes will need to be kept separate from one another when the material is eventually merged.

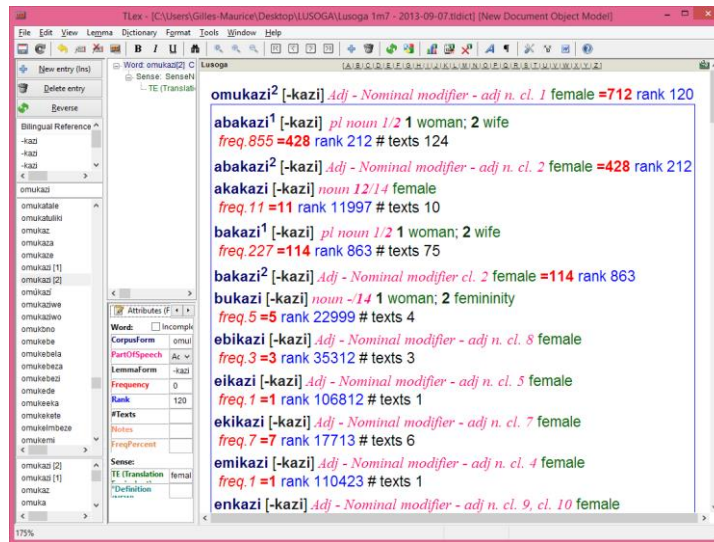


Figure 3: Lemmatising the 1.7m Lusoga corpus in TLex: the combination 'lemma & part of speech' will eventually be used to bring related forms together

Figure 2 illustrates that notes could additionally be attached to any entry; seen in orange and between curly brackets. Figure 3 illustrates another aspect, namely that for closed-class sets such as pronouns and adjectives, all the forms were considered in which the respective stems occurred in the 1.7m Lusoga corpus, and not only those with a frequency of at least 12. This could simply be achieved by doing field-specific searches across the entire TLex database, given that the full wordlist had been imported. This change in approach meant that the frequencies of the resulting lemma signs of these closed-class items were slightly raised. This was a trade-off, but with the advantage that the full picture became available for each of these closed-class items.⁵

Implicit in Figure 3, given the raised homonym numbers, is the fact that many entries had to be split up in two or more parts, typically because they could be assigned to different parts of speech, and/or because they had unrelated translation equivalents. Such entries were duplicated, and their frequencies were redistributed based on a quick and rough corpus sample.⁶ In Figure 3, **omukazi¹** (not shown) is the noun 'woman; wife'.

This lemmatisation phase took us about one month. A total of 10 318 items were eventually tagged,⁷ which corresponds to just over 5% of the types in the

1.7m Lusoga corpus, but it also corresponds to well over 80% of the tokens. Eighty percent of the word forms in the 1.7m Lusoga corpus were accordingly seen by only looking at 5% of it.

Three Lua scripts were then written which run in TLex to actually perform the lemmatisation: (i) to bring the 'lemma – part-of-speech' pairs together, see Addendum 1; (ii) to sum the frequencies of all the members of each of these pairs and to calculate the new ranks, see Addendum 2; and (iii) to use the latter ranks to group the lemma signs into frequency bands, see Addendum 3. A random section of the outcome, ranks 500 to 510, is summarised in Table 1.

Table 1: Lemmatised frequency list for Lusoga, ranks 500-510, derived from the top 10 000 types in the 1.7m Lusoga corpus

Lemma	Part of speech	Meaning	Freq.	Rank	Freq. band
-lim-	<i>verb</i>	dig; farm	296	500	①
-goloza	<i>noun 5/6</i>	county	295	501	②
-ikiliza	<i>noun 1/2</i>	believer; saint	295	502	②
nkani	<i>connective</i>	at least	295	503	②
ee	<i>ideophone</i>	wonder	293	504	②
-lundi	<i>pl noun 3/4</i>	instances	293	505	②
-idhukil-	<i>verb</i>	remember; recall	292	506	②
-taama	<i>noun 9/10</i>	sheep	291	507	②
-teekw-	<i>verb, modal</i>	must	290	508	②
nguli	<i>connective</i>	if	288	509	②
-wanika	<i>noun 5/6</i>	treasury; mortuary; dictionary	286	510	②

Regarding these three Lua scripts, it is important to point out that they may be re-run at any time, with changing data, even (also!) during actual dictionary compilation, down to the very last day of preparing an actual dictionary. Specifically with regard to the third Lua script, the one which adds the frequency bands, it is moreover trivial to change the values, which are set here to mark the top 500 lemma signs with ①, the next 500 with ②, the third 500 with ③, and no symbol for the remainder.

Table 1, which summarises data (al)ready in TLex, can also be seen as the start-pack of a (bilingual) Lusoga dictionary. This, of course, is no coincidence.

To develop the potential of this material further, the next two sections (§4 and §5) are structured in the same way, based on the fact that the lemmatised frequency list that was built directly with and into TLex embeds both part-of-speech data as well as alphabetical information: first, a type of ruler is introduced theoretically; then, a practical one is built for Lusoga; followed by a comparison with an equivalent Zulu ruler; ending with the use of such a ruler in the planning of the actual compilation of a future (bilingual) Lusoga dictionary.

4. From lemmatised frequency list to part-of-speech distributions

4.1 Part-of-speech rulers

As shown by de Schryver (2013), the relative size of each word class does not constitute a fixed percentage across corpora of the same language. Intuitively, it is clear that a large general-language corpus will proportionally contain more nouns and verbs than a smaller one (Hanks 2001). The trend, it turns out, is asymptotic, and from a few thousand items onwards one gets a good idea of the *direction* of the distribution of the various word classes. This may be illustrated with data taken from the unlemmatised version of the 100m *British National Corpus* (BNC 1994–2018), as shown in Figure 4.

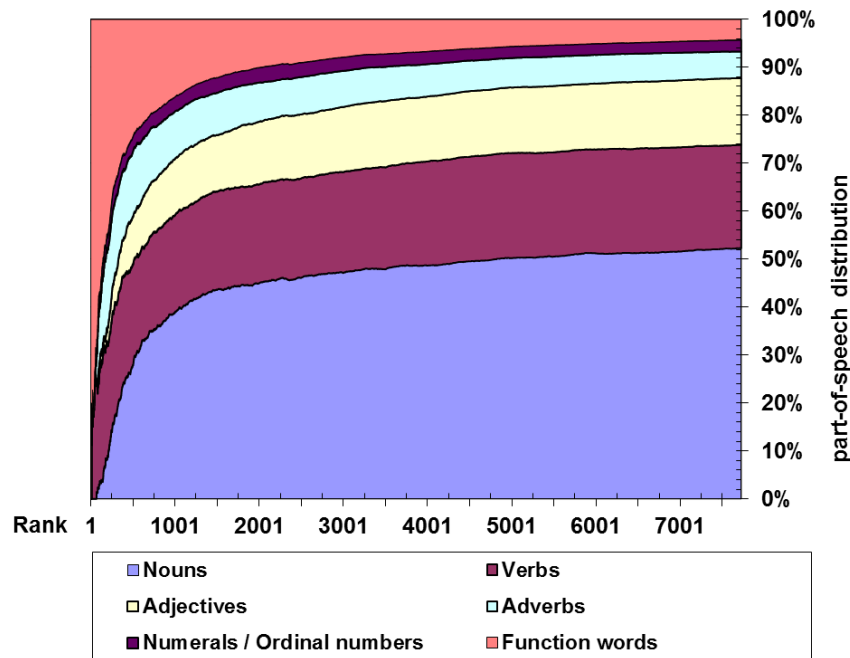


Figure 4: Part-of-speech distribution of the top 7 000+ types in the unlemmatised 100m *British National Corpus* [taken from de Schryver (2013: 1387)]

With regard to the data in Figure 4, de Schryver argues:

One may clearly deduce from this graph that function words and verbs dominate the top-frequent ranks in an English corpus. The percentage of nouns grows steadily as one goes down the frequency list. At the 1,000+ mark the overall percentage of nouns already stands at 40 %, that of the verbs at 20 %, while the

function words shrank to 16 % of the total (whereas these still represented roughly two thirds at the 100 mark). [...] The allocation to the nouns at the 7,000+ mark [...] stands at 52 %, that to the verbs grew to 22 %, while the function words shrank to a mere 4% of the total. These graphs can be extended down to any rank, while the same type of calculations can of course also be performed on lemmatized frequency lists, with similar results. (de Schryver 2013: 1386-1388)

What is important to remember from this is that there are as many part-of-speech rulers as there are numbers of lemma signs in a dictionary; each dictionary has a different distribution. Indeed, looking up from any rank in a graph like Figure 4, one obtains a different part-of-speech ruler.

4.2 Towards a part-of-speech ruler for Lusoga

The distribution of the main parts of speech in the lemmatised frequency list derived from the top section of the 1.7m Lusoga corpus is shown in Table 2 and Figure 5.

Table 2: Statistics for the distribution of the parts of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

Rank	Part of speech	Lemmatised	% = POS-ruler
1	noun	2 440	57.41%
2	verb	1 113	26.19%
3	pronoun	156	3.67%
4	quantifier	143	3.36%
5	adjective	117	2.75%
6	locative	75	1.76%
7	connective	68	1.60%
8	interjection	54	1.27%
9	ideophone	49	1.15%
10	adverb	35	0.82%
SUM		4 250	100.00%

As can be seen, the main part of speech of Lusoga is the noun, which accounts for 57% of all the lemma signs. The second most frequent part of speech is the verb, covering 26%. Nouns and verbs make up a staggering 83% of all the lemma signs in Lusoga. The third most frequent group are the various pronouns (4% of the total), followed by the quantifiers (3%), adjectives (3%) and locatives (2%). The remaining 5% is made up of connectives (2%), interjections (1%), ideophones (1%) and adverbs (1%). A comparison with the values seen in Figure 4 is tempting, but faces at least two problems.

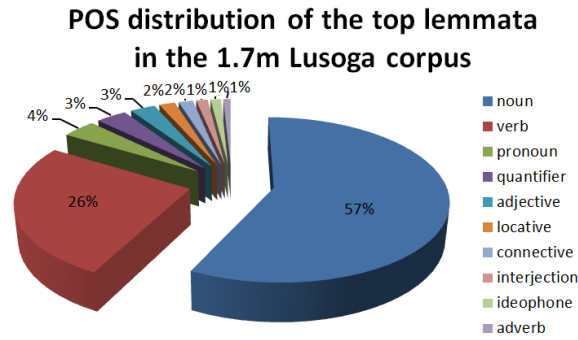


Figure 5: Pie chart showing the distribution of the parts of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

The first challenge is that the distributions across languages that belong to two very different language families are being compared. Even so, at the right-hand side of the graph seen in Figure 4, nouns and verbs already make up 74% of the total in English. The second challenge is that an unlemmatised distribution is compared to a lemmatised one. Indeed, as may be seen from Table 3, the original unlemmatised top-frequent 10 318 orthographic word forms (which includes some lower-frequent word forms from the closed-class parts of speech), as taken from the 1.7m Lusoga corpus, yielded a lemmatised frequency list of just 4 250 items.

Table 3: Statistics for the distribution of the parts of speech in the unlemmatised vs. lemmatised frequency lists derived from the top 10 000 types in the 1.7m Lusoga corpus

Part of speech	Unlemmatised	%	Lemmatised	%
verb	4 444	43.07%	1 113	26.19%
noun	3 622	35.10%	2 440	57.41%
adjective	1 105	10.71%	117	2.75%
pronoun	460	4.46%	156	3.67%
quantifier	231	2.24%	143	3.36%
locative	187	1.81%	75	1.76%
adverb	98	0.95%	35	0.82%
connective	68	0.66%	68	1.60%
interjection	54	0.52%	54	1.27%
ideophone	49	0.47%	49	1.15%
SUM	10 318	100.00%	4 250	100.00%

Expressed as a percentage of the total, three categories especially change their allocation drastically after lemmatisation. While verbs make up 43% of all the

top orthographic types in this Lusoga corpus, they only make up 26% after lemmatisation. Nouns do the reverse: they make up 35% of all the top orthographic types, but reach a massive 57% after lemmatisation. Adjectives go from nearly 11% down to about 3%. Unlemmatised and lemmatised part-of-speech distributions are thus different, as shown graphically in Figures 6 vs. 7.⁸

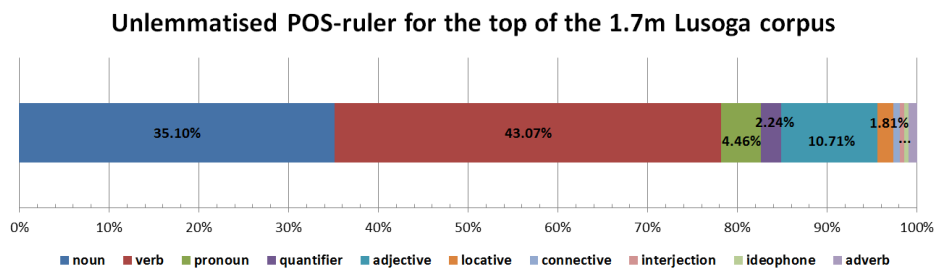


Figure 6: Part-of-speech ruler for the unlemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

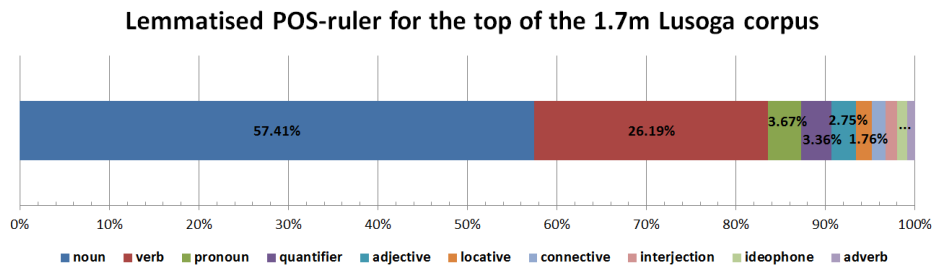


Figure 7: Part-of-speech ruler for the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

4.3 Contrasting part-of-speech rulers for Lusoga and Zulu

In order to judge whether the data seen in Table 2 and Figure 5 is plausible, it is instructive to compare the part-of-speech distribution for the Lusoga lemma signs with that for Zulu, as described in the corpus-based Zulu mini-grammar included in the *Oxford Bilingual School Dictionary: Zulu and English* (de Schryver 2010a: S13-S26) and summarised in Figure 8. On the Zulu to English side, this dictionary contains about 5 000 lemma signs (which were derived from the top section of a 7.5m general + 1m textbook Zulu corpus). This order of magnitude allows for comparisons with the 4 250 lemmatised forms which were obtained for Lusoga. While there are differences in the lemmatisation approach between the two languages, and even differences in categorising and naming the word classes, the overall picture seen for Zulu *may* be compared with that for

Lusoga. At that point one realises that the two distributions are indeed rather similar, especially as regards nouns, with an allocation of 57% in Lusoga vs. 58% in Zulu. However, one does notice that there seems to be an exceptionally high number of verbs in Lusoga (26%) as compared to verbs in Zulu (16%).

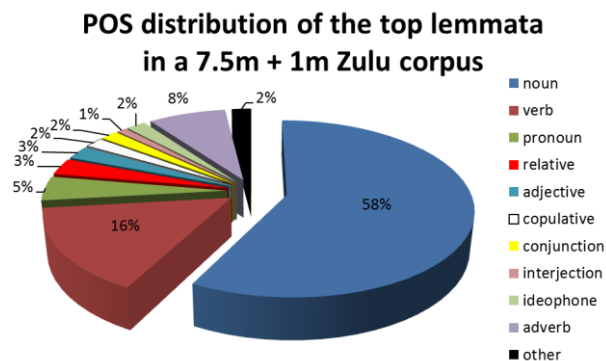


Figure 8: Part-of-speech distribution of the lemma signs in a corpus-based Zulu dictionary derived from the top types in a 7.5m general + 1m textbook Zulu corpus [adapted from de Schryver (2010a: S15)]

In these distributions, there are about ten main parts of speech ('main', as there are a number of sub-types as well) for both Lusoga and Zulu, but this could have been very different. The monolingual Zulu dictionary completed by the Zulu National Lexicography Unit (Mbatha 2006), for instance, uses just *four* parts of speech, following notions expounded in the PhD of Nkabinde (1975). Given the OUPSA Zulu school dictionary was meant to be as user-friendly as possible, such a drastic reduction of word classes was not entertained. The same holds for our decision regarding the word classes in Lusoga.

4.4 Using a part-of-speech ruler for Lusoga in dictionary planning

Using actual counts, Figures 6 and 7 can also be depicted as Figures 9 and 10 respectively. Of the two part-of-speech rulers, the lemmatised one is the most useful to support dictionary-making, hence Figure 10. The choice to lemmatise the top 10 000 orthographic words from the 1.7m Lusoga corpus was made in an attempt to arrive at a list of between 4 000 and 5 000 candidate lemma signs; we arrived at 4 250. If conceived in the way the OUPSA bilingual school dictionaries were conceived, then room must also be left for the inclusion of specialised vocabulary in the macrostructure, which is to be extracted from a separate, purpose-built specialised corpus. For Zulu, see de Schryver (2010b: 169), a concept based on the earlier de Schryver and Prinsloo (2003), where it was exemplified for Afrikaans. Basically, the Lusoga part-of-speech ruler seen in Figure 10 tells us that for a Lusoga dictionary of about 5 000 lemma signs, there

should/will be 2 440 nouns, 1 113 verbs, etc. down to 49 ideophones and 35 adverbs taken from the general language.

POS counts in the unlemmatised top of the 1.7m Lusoga corpus

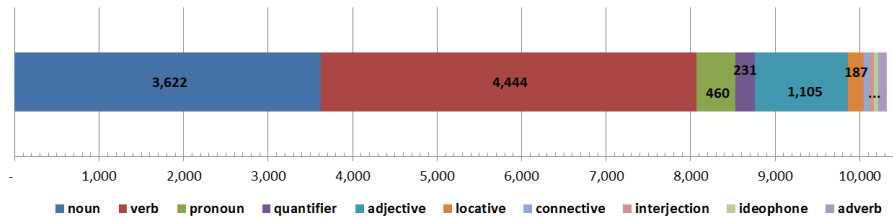


Figure 9: Counts per part of speech in the unlemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

POS counts in the lemmatised top of the 1.7m Lusoga corpus

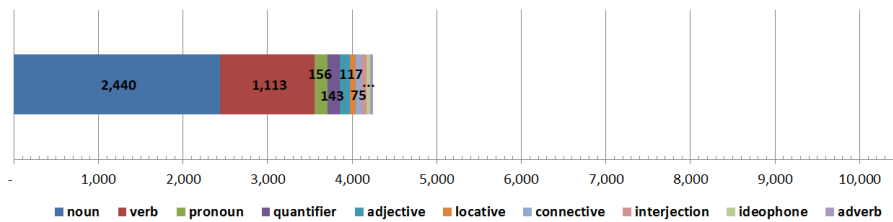


Figure 10: Counts per part of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

Knowing the (approximate) size of each word class in advance truly helps planning the actual dictionary work: equivalent and comparable chunks of the data may for instance be distributed to different team members, time extrapolations for the total work involved may be based on samples that were compiled for the different word classes, and dictionary-making itself may be organised and proceed 'by word class'. The latter has turned out to be an extremely important concept in Bantu lexicography, and may be spotted in the literature from article titles that refer to 'the lemmatisation of'-formula (de Schryver et al. 2004: 37). Taking Zulu as an example, the lemmatisation of nouns (Mpungose 1998, Prinsloo 2011), verbs (Prinsloo 2011), adjectives (de Schryver 2008b), pronouns (de Schryver 2008a, de Schryver and Wilkes 2008) and ideophones (de Schryver 2009), have all received attention in dedicated lexicographic studies, as have the treatment of terminological (Khumalo 2015) and cultural (Prinsloo and Bosch 2012) vocabulary.

Many problems in Bantu lexicography are part-of-speech dependent and need unique solutions that are different from one part of speech to the next.

Working through batches of a single word class during actual dictionary compilation therefore has ample advantages. In a dictionary-writing system like TLex, this is moreover fully supported: the part-of-speech tags that have been attached to the candidate lemma signs following lemmatisation (cf. §3) may first be used to isolate each word class as a group using the *Filter* tool, and that subset of the data may then be combined with any other filter parameters to allow for focused dictionary compilation.

5. From lemmatised frequency list to alphabetical distributions

5.1 Alphabetical rulers (aka 'multidimensional lexicographic rulers')

Some printed dictionaries have a thumb index per alphabetical category, either physically cut out in the pages or painted directly on the surface of the fore-edge, showing the progression of the different alphabetical categories, often in ladderised form. An alphabetical ruler is exactly that: an instrument which represents the relative allocation to each stretch of the alphabet. As a metalexicographical concept, such rulers were first introduced for Afrikaans (Prinsloo and de Schryver 2002a, 2003, de Schryver 2005, Prinsloo 2010, Taljard et al. 2017) and subsequently designed for all other official South African languages (de Schryver 2003, Prinsloo 2004, Prinsloo and de Schryver 2005, 2007).⁹ Such rulers may be built from dictionary data, corpus data, or both. They may also be built to reflect the general language, or else a specific specialised domain of the language. In contrast to a part-of-speech ruler, an alphabetical ruler does not vary with corpus or dictionary sizes. The series of percentages per alphabetical stretch, for instance per alphabetical category, is very stable indeed, and the only difference one observes is between its lemmatised and unlemmatised versions.

Initially a 'measurement instrument', it quickly became clear that a ruler of this sort is also an 'evaluation instrument', as well as a 'prediction instrument', and ultimately even a 'management instrument' (de Schryver 2013). Given the many ways in which it can be used, such rulers have also been termed 'multidimensional lexicographic rulers'. Of the various uses, the one that interests us in the present contribution is as a prediction instrument, more specifically with the aim of predicting features of the compilation of a new Lusoga dictionary.

5.2 Towards an alphabetical ruler for Lusoga

From all the types in the full 1.7m Lusoga corpus as well as the unlemmatised and lemmatised frequency lists derived from the top 10 000 types (cf. §3), one can straightforwardly derive the data presented in Table 4. The three series of percentages represent general-language alphabetical rulers, and this in two unlemmatised environments and one lemmatised environment respectively.

Comparing the three distributions with one another, it is clear that there is a good correlation between the two unlemmatised ones, but no correlation between either of the unlemmatised distributions and the lemmatised one.¹⁰

Table 4: Statistics for the distribution of the alphabetical categories in the 1.7m Lusoga corpus as well as the unlemmatised and lemmatised frequency lists derived from the top 10 000 types

Section	Unlemmatised		Unlemmatised		Lemmatised	
	all corpus types	%	top corpus types	%	lemma signs from top	% = ABC-ruler
A	20 569	10.55%	1 152	11.16%	147	3.46%
B	25 030	12.83%	1 265	12.26%	368	8.66%
C	1 150	0.59%	5	0.05%	5	0.12%
D	3 089	1.58%	106	1.03%	83	1.95%
E	19 569	10.03%	1 354	13.12%	233	5.48%
F	643	0.33%	18	0.17%	78	1.84%
G	6 699	3.43%	260	2.52%	297	6.99%
H	830	0.43%	28	0.27%	24	0.56%
I	1 959	1.00%	187	1.81%	198	4.66%
J	309	0.16%	6	0.06%	5	0.12%
K	20 110	10.31%	1 116	10.82%	529	12.45%
L	4 462	2.29%	267	2.59%	338	7.95%
M	13 373	6.86%	933	9.04%	257	6.05%
N	14 425	7.40%	664	6.44%	277	6.52%
O	27 210	13.95%	1 720	16.67%	82	1.93%
P	1 126	0.58%	39	0.38%	84	1.98%
Q	36	0.02%	0	0.00%	0	0.00%
R	756	0.39%	3	0.03%	3	0.07%
S	2 032	1.04%	86	0.83%	374	8.80%
T	16 685	8.56%	453	4.39%	298	7.01%
U	415	0.21%	13	0.13%	14	0.33%
V	306	0.16%	10	0.10%	55	1.29%
W	4 028	2.07%	211	2.04%	202	4.75%
X	16	0.01%	0	0.00%	0	0.00%
Y	9 978	5.12%	411	3.98%	200	4.71%
Z	227	0.12%	11	0.11%	99	2.33%
SUM	195 032	100.00%	10 318	100.00%	4 250	100.00%

The only alphabetical ruler that is relevant to lexicographic work for a Bantu language is obviously the lemmatised one, except, perhaps, for those rare cases where full orthographic words are presented as lemma signs, including for all the verbs, as has been done for an experimental online Swahili dictionary

(Hillewaert and de Schryver 2004). Therefore, 'the' alphabetical ruler for Lusoga is as shown in Figure 11.¹¹

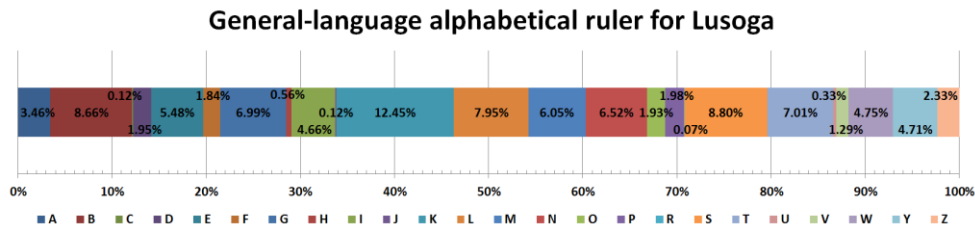


Figure 11: General-language alphabetical ruler based on the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

5.3 Contrasting alphabetical rulers for Lusoga and Zulu

The alphabetical ruler for Lusoga may be compared to the alphabetical ruler for Zulu that was used for the OUPSA Zulu school dictionary (de Schryver 2010a), shown in Figure 12.

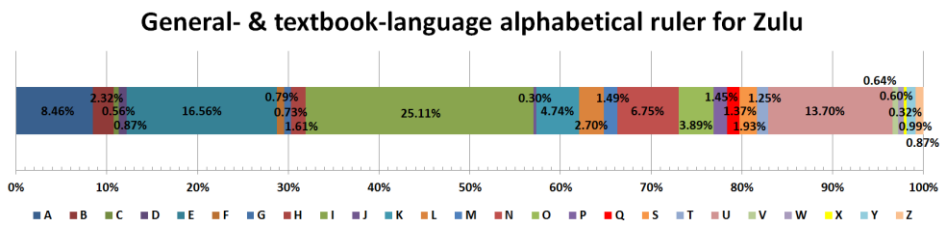


Figure 12: Alphabetical distribution of the lemma signs in a corpus-based Zulu dictionary derived from the top types in a 7.5m general corpus + 1m textbook Zulu corpus

As one may see, the two alphabetical rulers look very different indeed. This is because a decision was made in the Zulu dictionary to present full words for all parts of speech except verbs, on that account breaking with the stem tradition for this language. As a result of Zulu's pre-prefixes especially at nouns, the alphabetical categories A, I and U are massive, as is the alphabetical category E which contains the many locativised nouns for which the 'e-/o-...-ini locativisation strategy' was used (de Schryver and Gauton 2002).

Atypical alphabetical distributions such as the one seen in Figure 12 should remind every prospective compiler of a Bantu-language dictionary that careful thought should be put into who the envisaged target user group is. Reasoning back from the target user group, this then leads to a decision on pres-

entation. Given that the Zulu dictionary was meant for school-going pupils, the goal was to present the material in as user-friendly a manner as possible, hence the decision to present words rather than stems for most parts of speech. Reasoning further back, from presentation to the actual lemmatisation required to achieve that presentation, one realises that there is always a direct link between target user group and lemmatisation approach, and vice versa. Relating this to the candidate Lusoga lemma-sign list means that the target user group envisaged is one that will be able to handle the lookup of word stems.

5.4 Using an alphabetical ruler for Lusoga in dictionary planning

Although the backbone of an alphabetical ruler is merely a single list of percentages totalling one hundred, it is a powerful instrument. From §5.2 it follows that the distribution of the number of (general-language) lemma signs per alphabetical category in Lusoga is not only according to the alphabetical ruler, but even the exact counts for each category are a given, and may be depicted as shown in Figure 13.

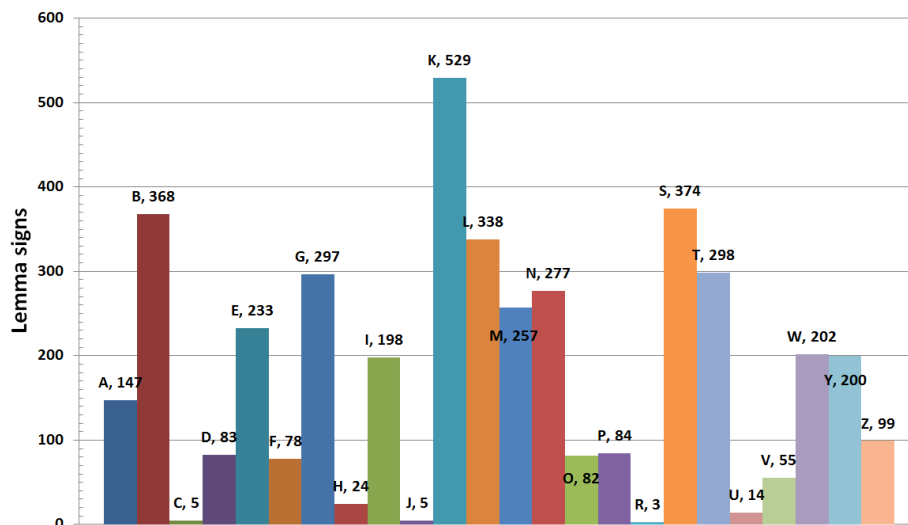


Figure 13: Distribution of the (general-language) lemma signs per alphabetical category in a planned Lusoga dictionary (sum: 4 250 lemma signs)

What is more, the actual lemma signs themselves are waiting in TLex, together with a brief preliminary meaning for each.

The alphabetical ruler may also be used to do some advance planning as far as dictionary size is concerned. Suppose a dictionary publisher envisages a central text for one side of the dictionary of 350 pages, then this ruler may

straightforwardly be used to predict the page allocation to each alphabetical category, as shown in Figure 14. Evidently, the presentation shown in Figure 14 is none other than the alphabetical ruler itself, hence Figure 11, now with a different *x*-axis.

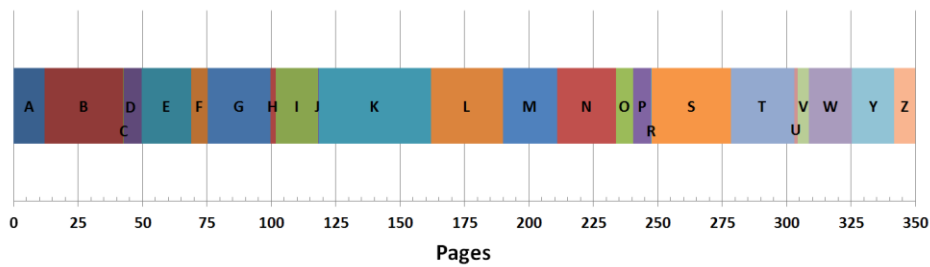


Figure 14: Distribution of the number of pages per alphabetical category in a planned Lusoga dictionary (aim: 350 pages for one side)

As a last example of the use of an alphabetical ruler as a prediction instrument, suppose the dictionary team wishes to work 'through the alphabet' (rather than, say, by word class), and that two years are available for the compilation of the central text, then Figure 15 predicts in which week which alphabetical category should be reached.

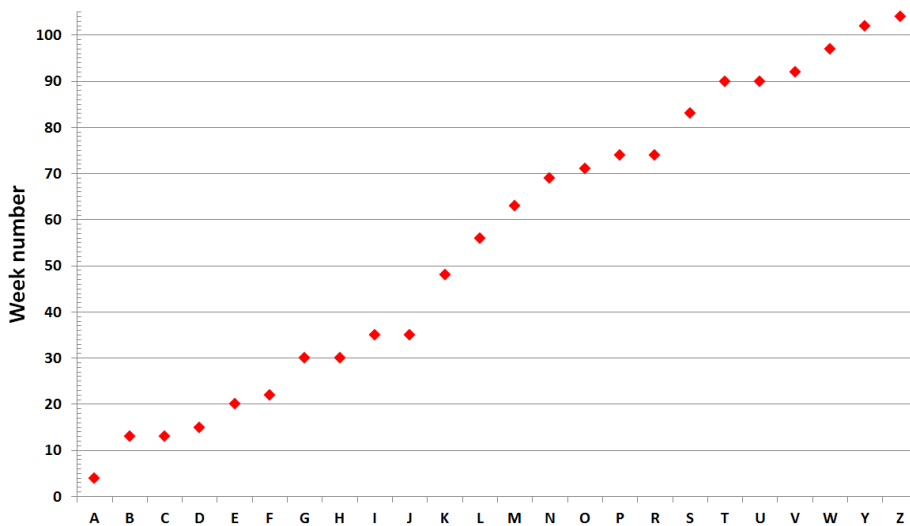


Figure 15: Projected progress through the alphabet for a planned Lusoga dictionary (aim: 2 years, or 104 weeks)

The underlying data for Figures 13 to 15 is shown in Table 5, but it should be clear that the alphabetical ruler may be used in any other creative way; for some of these, see the references in §5.1.

Table 5: Multidimensional predictions on lemma-sign, page and time levels for a planned Lusoga dictionary, using an alphabetical ruler for Lusoga

Section	ABC-ruler	Lemma signs	Pages	Reached in week	Days	...
A	3.46%	147	12.1	4	18.0	
B	8.66%	368	30.3	13	45.1	
C	0.12%	5	0.4	13	0.6	
D	1.95%	83	6.8	15	10.2	
E	5.48%	233	19.2	20	28.6	
F	1.84%	78	6.4	22	9.6	
G	6.99%	297	24.5	30	36.4	
H	0.56%	24	2.0	30	2.9	
I	4.66%	198	16.3	35	24.3	
J	0.12%	5	0.4	35	0.6	
K	12.45%	529	43.6	48	64.8	
L	7.95%	338	27.8	56	41.4	
M	6.05%	257	21.2	63	31.5	
N	6.52%	277	22.8	69	34.0	
O	1.93%	82	6.8	71	10.1	
P	1.98%	84	6.9	74	10.3	
R	0.07%	3	0.2	74	0.4	
S	8.80%	374	30.8	83	45.8	
T	7.01%	298	24.5	90	36.5	
U	0.33%	14	1.2	90	1.7	
V	1.29%	55	4.5	92	6.7	
W	4.75%	202	16.6	97	24.8	
Y	4.71%	200	16.5	102	24.5	
Z	2.33%	99	8.2	104	12.1	
SUM	100.00%	4 250	350		521	

6. Discussion

In this article we have illustrated how a lemmatised frequency list may be built directly within a dictionary-writing system like TLex, using as input plain orthographic words with occurrence frequencies as generated by corpus-query software like WordSmith Tools. These specific software programs are not crucial to the procedure, but they have been employed a number of times now and have proven their worth. Comparable programs will also do; what is important

to remember from the text is the necessary steps. The procedure is a mostly manual process, which needs to take the future target user group into account, and a process whereby all details are logged so that instant use may be made of two types of rulers: a part-of-speech ruler and an alphabetical ruler. A Lusoga corpus that was presented in the first of our three linked articles was processed to demonstrate the actual workings, and comparisons were also made with a completed Zulu dictionary project.

Honesty compels us to admit that the procedure described is the 'ideal' one, however. In actual practice, given that corpus data had to be analysed before it could be *explained* — and that the part-of-speech tagging and lemmatisation were merely the first steps of the analysis — even a seemingly basic task such as pinpointing the part(s) of speech of an orthographic word form was not that trivial. To start any analysis one needs a way to create order first, by grouping related material. But from the moment one starts to group material, one has already made a decision on how to analyse that material, as part-of-speech assignment is dependent on the framework or theory of the analysis. Conversely, without any advance decisions, one cannot begin to group and so can never get to any analysis. This chicken-and-egg conundrum was partly solved by falling back on received knowledge regarding the Bantu languages, as for instance summarised in handbooks such as that of Nurse and Philippson (2003) or the earlier ones of Guthrie (1948, 1953), Doke (1954) and Bryan (1959). Furthermore, as the analysis of the corpus material proceeded, we *did* go back to material that had already been completed in the TLex file, retagged some of the material, and reran the Lua scripts in order to generate an 'update' of the lemmatised frequency list.

Reformulated, even the mere act of labelling certain word forms as demonstratives or possessives, and considering these under the wider umbrella of pronouns, already crosses the line from analysis to explanation. That said, despite the received knowledge, we have tried to stick as much as possible to what we could observe in the corpus data, by also looking at the wider context and thus by avoiding limiting our look at words in isolation. With this we are now ready for the next step, the actual explanation of the material.

Acknowledgements

The research for this article was funded by the Special Research Fund of Ghent University. Thanks are due to the two anonymous referees.

Endnotes

1. For more on the difference between conjunctive and disjunctive writing systems in Bantu, see Prinsloo and de Schryver (2002b).

2. Whether or not this is an error actually depends on the lemmatisation strategy chosen. In Nguni lexicography, there is a 'stem tradition' (Ziervogel 1965, Van Wyk 1995), so if one also presents both nouns and verbs under the same stems (where relevant), then one could indeed lump their frequencies as well. Conversely, there is an argument to be made to keep the frequencies of different parts of speech separate, thereby leaving some presentation options open until actual dictionary compilation. In this regard, Prinsloo (1991), in the very-first exploratory study of the use of frequency counts for Bantu-language dictionary-making, did point out: 'It is very important to note that the interpretation of the output of a word frequency study is closely related to the lexicographical approach and the editorial policy from which the lexicographer embarked' (Prinsloo 1991: 59). The section from which this sentence is taken, 'Frequency studies in perspective' (Prinsloo 1991: 59-60), actually deals with lemmatisation options/decisions, even though Prinsloo does not use the term nor concept of lemmatisation.
3. Incidentally, the grammars included as middle matter in these dictionaries are furthermore the first corpus-based mini-grammars for any Bantu language, as described in de Schryver and Taljard (2007) for Northern Sotho, and de Schryver (2010b) for Zulu.
4. This is shown quite literally in Figure 1, where the data is sorted on the field 'Rank', so one truly moves from most frequent to least frequent. Another option is to use filters to extract the top-frequent section from the database, to work on in alphabetical order (or in any other, even random, order).
5. For more on the advantages, see for instance de Schryver et al. (2004), de Schryver (2008a, 2008b), de Schryver and Wilkes (2008) and de Schryver (2009).
6. When quick-and-rough frequencies were not provided, a Lua script (cf. further) would take care of this aspect, by automatically distributing the frequencies equally as a first approach (subject to correction later).
7. Junk was not tagged but deleted. Material with a poor spread across the sources was flagged as such, indicating that it may require a label.
8. The Pearson product moment correlation coefficient r between the unlemmatised and lemmatised part-of-speech distributions is 0.85.
9. The concept of an alphabetical ruler may be traced back to the 'block system of distribution of dictionary entries by initial letters' prepared for English by Edward L. Thorndike during the 1950s (Landau 2001: 360-362). Thorndike divided the alphabet into 105 blocks: 6 for A (A1: a-adk, A2: adl-alh, A3: ali-angk, ...), ... 1 for J (J50: j-jz), ... 3 for W (... , W104: wit-wz) and 1 for XYZ (XYZ105: x-zz). With approximately the same weight assigned to each of those blocks, this series supposedly reflects the 'distribution of lexical units throughout the alphabet'. See also Jackson (2002: 163-164), Moon (2004: 649-650) and Svensén (2009: 406).
10. The Pearson product moment correlation coefficient r between the two unlemmatised alphabetical distributions is an excellent 0.97; while it is just 0.56 between the full unlemmatised distribution and the lemmatised distribution, and 0.49 between the top unlemmatised distribution and the lemmatised distribution.
11. Observe that the letters c, j, q, r and x are not native to Lusoga, but may appear in borrowed abbreviations, place names and surnames, and the like.

References

- BNC. 1994–2018. British National Corpus. Available online at: <http://www.natcorp.ox.ac.uk/>.
- Bosch, S.E. and L. Pretorius. 2003. Building a Computational Morphological Analyser/Generator for Zulu Using the Xerox Finite-State Tools. *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics*: 27-34. Budapest: ACL.
- Bosch, S.E. and L. Pretorius. 2004. Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development. *Proceedings of the 4th International Language Resources and Evaluation Conference*: 1251-1254. Lisbon: ARTIPOL.
- Bosch, S.E., L. Pretorius and A. Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies* 17(2): 66-88.
- Bryan, M.A. 1959. *The Bantu Languages of Africa* (Handbook of African Languages 4). London: Oxford University Press (for the International African Institute).
- De Pauw, G., G.-M. de Schryver and J. van de Loo. 2012. Resource-light Bantu Part-of-speech Tagging. De Pauw, G., G.-M. de Schryver, M.L. Forcada, K. Sarasola, F.M. Tyers and P.W. Wagacha (Eds). 2012. *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL 8 — AfLaT 2012)*: 85-92. Istanbul: European Language Resources Association.
- De Pauw, G., G.-M. de Schryver and P.W. Wagacha. 2006. Data-driven Part-of-Speech Tagging of Kiswahili. Sojka, P., I. Kopeček and K. Pala (Eds). 2006. *Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic, September 11–15, 2006, Proceedings* (Lecture Notes in Artificial Intelligence (LNAI), subseries of Lecture Notes in Computer Science (LNCS) 4188): 197-204. Berlin: Springer-Verlag.
- De Pauw, G., G.-M. de Schryver and P.W. Wagacha. 2006–18. AfLaT — African Language Technology. Available online at: <http://aflat.org/>.
- de Schryver, G.-M. 1999. *Bantu Lexicography and the Concept of Simultaneous Feedback, Some Preliminary Observations on the Introduction of a New Methodology for the Compilation of Dictionaries with Special Reference to a Bilingual Learner's Dictionary Cilubà-Dutch*. Unpublished M.A. dissertation. Ghent: Ghent University.
- de Schryver, G.-M. 2003. Drawing up the Macrostructure of a Nguni Dictionary, with Special Reference to isiNdebele. *South African Journal of African Languages* 23(1): 11-25.
- de Schryver, G.-M. 2005. Concurrent Over- and Under-treatment in Dictionaries — The *Woordeboek van die Afrikaanse Taal* as a Case in Point. *International Journal of Lexicography* 18(1): 47-75.
- de Schryver, G.-M. 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford*. Cape Town: Oxford University Press Southern Africa.
- de Schryver, G.-M. 2008a. The Lexicographic Treatment of Quantitative Pronouns in Zulu. *Lexikos* 18: 92-105.
- de Schryver, G.-M. 2008b. A New Way to Lemmatize Adjectives in a User-friendly Zulu–English Dictionary. *Lexikos* 18: 63-91.
- de Schryver, G.-M. 2009. The Lexicographic Treatment of Ideophones in Zulu. *Lexikos* 19: 34-54.
- de Schryver, G.-M. 2010a. *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford*. Cape Town: Oxford University Press Southern Africa.

- de Schryver, G.-M.** 2010b. Revolutionizing Bantu Lexicography — A Zulu Case Study. *Lexikos* 20: 161-201.
- de Schryver, G.-M.** 2013. Tools to Support the Design of a Macrostructure. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (Handbooks of Linguistics and Communication Science, HSK 5.4): 1384-1395. Berlin: Walter de Gruyter.
- de Schryver, G.-M. and R. Gauton.** 2002. The Zulu Locative Prefix ku- Revisited: A Corpus-based Approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201-220.
- de Schryver, G.-M. and N.S. Kabuta.** 1997. *Lexicon Cilubà–Nederlands, Een circa 2500-lemma's-tellend strikt alfabetisch geordend vertalend aanleerderslexicon met decodeer-functie ten behoeve van studenten Afrikaanse Talen & Culturen aan de Universiteit Gent* (Recall Linguistics Series 1). Ghent: Recall.
- de Schryver, G.-M. and N.S. Kabuta.** 1998. *Beknopt woordenboek Cilubà–Nederlands & Kalombodi-mfündilu kàà Cilubà (Spellingsgids Cilubà), Een op gebruiks-frequentie gebaseerd vertalend aanleerderslexicon met decodeerfunctie bestaande uit circa 3.000 strikt alfabetisch geordende lemma's & Mfündilu wa myakù idi itàmba kumwènèka (De orthografie van de meest gangbare woorden)* (Recall Linguistics Series 12). Ghent: Recall.
- de Schryver, G.-M. and D.J. Prinsloo.** 2000. 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.** 2001. Corpus-based Activities versus Intuition-based Compilations by Lexicographers, the Sepedi Lemma-sign List as a Case in Point. *Nordic Journal of African Studies* 10(3): 374-398.
- de Schryver, G.-M. and D.J. Prinsloo.** 2003. Compiling a Lemma-sign List for a Specific Target User Group: The Junior Dictionary as a Case in Point. *Dictionaries: Journal of The Dictionary Society of North America* 24: 28-58.
- de Schryver, G.-M. and M. Reynolds.** 2014. *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama sesikolo*. Cape Town: Oxford University Press Southern Africa.
- de Schryver, G.-M. and E. Taljard.** 2007. Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho. *Lexikos* 17: 37-55.
- de Schryver, G.-M., E. Taljard, M.P. Mogodi and S. Maepa.** 2004. The Lexicographic Treatment of the Demonstrative Copulative in Sesotho sa Leboa — An Exercise in Multiple Cross-referencing. *Lexikos* 14: 35-66.
- de Schryver, G.-M. and A. Wilkes.** 2008. User-friendly Dictionaries for Zulu: An Exercise in Complexicography. Bernal, E. and J. DeCesaris (Eds). 2008. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)* (Sèrie Activitats 20): 827-836. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Doke, C.M.** 1954. *The Southern Bantu Languages* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).
- Guthrie, M.** 1948. *The Classification of the Bantu Languages* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).

- Guthrie, M.** 1953. *The Bantu Languages of Western Equatorial Africa* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).
- Hanks, P.** 2001. The Probable and the Possible: Lexicography in the Age of the Internet. Lee, S. (Ed.). 2001. *ASIALEX 2001 Proceedings, Asian Bilingualism and the Dictionary*: 1-15. Seoul: Center for Linguistic Informatics Development, Yonsei University.
- Hillewaert, S. and G.-M. de Schryver.** 2004. Online Kiswahili (Swahili) — English Dictionary. Available online at: <http://africanlanguages.com/swahili/>.
- Hurskainen, A.** 1992. A Two-level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.
- Hurskainen, A.** 2016. Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version. Available online at: <http://urn.fi/urn:nbn:fi:lb-2016011301>.
- Jackson, H.** 2002. *Lexicography: An Introduction*. London: Taylor & Francis Routledge.
- Joffe, D. and G.-M. de Schryver.** 2002–18. TLex Suite — Dictionary Compilation Software. Available online at: <http://tshwanedje.com/tshwanelex/>.
- Joffe, D. and G.-M. de Schryver.** 2005. Representing and Describing Words Flexibly with the Dictionary Application TshwaneLex. Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan and Y.Y. Tan (Eds). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1–3 June 2005, M Hotel, Singapore*: 108-114. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.
- Khumalo, L.** 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos* 25: 495-506.
- Kilgarriff, A.** 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.
- Knowles, F.** 1983. Towards the Machine Dictionary, 'Mechanical' Dictionaries. Hartmann, R.R.K. (Ed.). 1983. *Lexicography: Principles and Practice*: 181-193. London: Academic Press.
- Landau, S.I.** 2001. *Dictionaries: The Art and Craft of Lexicography (2nd edition)*. Cambridge: Cambridge University Press.
- Mbatha, M.O.** 2006. *Isichazamazwi sesiZulu*. Pietermaritzburg: New Dawn Publishers.
- Moon, R.** 2004. Cawdrey's A Table Alphabeticall: A Quantitative Approach. Williams, G. and S. Vessier (Eds). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 639-650. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.
- Mpungose, M.H.** 1998. Analysis of the Word-initial Segment with Reference to Lemmatising Zulu Nasal Nouns. *Lexikos* 8: 65-87.
- Nabirye, M.** 2016. *A Corpus-based Grammar of Lusoga*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Nkabinde, A.C.** 1975. *A Revision of the Word Categories in Zulu*. Unpublished PhD dissertation. Pretoria: UNISA.
- Nurse, D. and G. Philippson (Eds).** 2003. *The Bantu Languages* (Language Family Series 4). London: Routledge.
- 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.** 2004. Revising Matumo's *Setswana–English–Setswana Dictionary*. *Lexikos* 14: 158-172.

- Prinsloo, D.J.** 2010. Die verifiëring, verfyning en toepassing van leksikografiese liniale vir Afrikaans. *Lexikos* 20: 390-409.
- Prinsloo, D.J.** 2011. A Critical Analysis of the Lemmatisation of Nouns and Verbs in isiZulu. *Lexikos* 21: 169-193.
- Prinsloo, D.J. and S.E. Bosch.** 2012. Kinship Terminology in English–Zulu / Northern Sotho Dictionaries — A Challenge for the Bantu Lexicographer. Fjeld, R.V. and J.M. 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.** 2002a. 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.** 2002b. Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. *Nordic Journal of African Studies* 11(2): 249-265.
- Prinsloo, D.J. and G.-M. de Schryver.** 2003. Effektiewe vordering met die *Woordeboek van die Afrikaanse Taal* soos gemeet in terme van 'n multidimensionele Liniaal. Botha, W. (Ed.). 2003. *'n Man wat beur: Huldigingsbundel vir Dirk van Schalkwyk*: 106-126. Stellenbosch: Bureau of the WAT.
- Prinsloo, D.J. and G.-M. de Schryver.** 2005. Managing Eleven Parallel Corpora and the Extraction of Data in All Official South African Languages. Daelemans, W., T. du Plessis, C. Snyman and L. Teck (Eds). 2005. *Multilingualism and Electronic Language Management. Proceedings of the 4th International MIDP Colloquium, 22–23 September 2003, Bloemfontein, South Africa* (Studies in Language Policy in South Africa 4): 100-122. Pretoria: Van Schaik Publishers.
- Prinsloo, D.J. and G.-M. de Schryver.** 2007. Crafting a Multidimensional Ruler for the Compilation of Sesotho sa Leboa Dictionaries. Mojalefa, M.J. (Ed.). 2007. *Rabadia Ratšhatšha: Studies in African Language Literature, Linguistics, Translation and Lexicography*: 177-201. Stellenbosch: SUN PReSS.
- Scott, M.** 1996–2018. WordSmith Tools. Available online at: <http://www.lexically.net/wordsmith/>.
- Sinclair, J.M.** 1995. *Collins Cobuild English Dictionary*. London: HarperCollins.
- Svensén, B.** 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- Taljar, E., D.J. Prinsloo and N. Bosman.** 2017. Honderd jaar *Afrikaanse Woordelys en Spelreëls* — 'n oorsig en waardering. Deel 2: Die gebruiker in fokus. *Tydskrif vir Geesteswetenskappe* 57(2.1): 302-322.
- Van Wyk, E.B.** 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5: 82-96.
- Ziervogel, D.** 1965. Die probleme van leksikografie in die Suid-Afrikaanse Bantoetale. *Taalfasette* 1(1): 45-53.
- Zipf, G.K.** 1935. *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin Co.

Addendum 1: Lua script: GenerateSecondSide.lua

```
-- 2013-09 Lusoga collapse LemmaForm into second 'side' of dictionary database
-- and add up frequencies
-- David Joffe

-- 'CONSTANTS' (script configuration - if e.g. attribute names change, update
-- script here)
CFG =
{
  ATTR_POS      = "PartOfSpeech",
  -- Part of speech
  ATTR_LEMMAFORM = "LemmaForm",
  ATTR_FREQ     = "CalculatedFrequency",
  -- Recalculated frequency attribute (that incorporates homonym percentage).
  -- For reading the value from the corpus list.
  ATTR_FREQUENCY = "Frequency",
  -- Actual "Frequency" attribute (not re-calculated one that incorporates
  -- percentage). For setting frequency on created entries.
  ATTR_INCOMPLETE = "Incomplete",
  -- Section 0-based index with source list (i.e. corpus forms)
  SECTION_SRC    = 0,
  -- Section 0-based index for creating collapsed forms (e.g. "-ba")
  SECTION_DEST   = 1
}

local DOC=tApp():GetCurrentDoc();
if DOC==nil then return "";end

-- STATS
local nNumForms=0;
local nNumCreated=0;
local nNumExistingModified=0;
local SECTION=DOC:GetDictionary():GetLanguage( CFG.SECTION_SRC );
local SECTIONDEST=DOC:GetDictionary():GetLanguage( CFG.SECTION_DEST );
local i;
local data={}
for i=0,SECTION:GetNumEntries()-1,1 do
  local ENTRY=SECTION:GetEntry(i);
  local bDoEntry=false;
  local incomplete= tQuery(ENTRY,"/@"..CFG.ATTR_INCOMPLETE);
  if (incomplete=="") or (incomplete=="0") then
    bDoEntry = true;
  end
end
```

```
if (bDoEntry) then
  local pos      = tQuery(ENTRY,"/@"..CFG.ATTR_POS);
  local lemmaform = tQuery(ENTRY,"/@"..CFG.ATTR_LEMMAFORM);
  local freqs    = tQuery(ENTRY,"/@"..CFG.ATTR_FREQ);
  -- Can return nil on empty string, so check for nil next and set
  -- to 0 in that case
  local freq = tonumber(freqs);
  if (freq==nil) then
    freq= 0;
  end

  tLuaLog("FORM:"..lemmaform
    .. "(" .. ENTRY:GetLemmaSign()..") pos="..pos.." freq="..freq)

  -- Make a unique string that is the combination of LemmaForm and the
  -- partofspeech (e.g. "-ba_$$$_noun") .. this separator string must just
  -- be some string that doesn't occur in the actual data ever, but other
  -- than that it's arbitrary
  if ( data[ lemmaform .. "_$$$_" .. pos ] == nil ) then
    --data[ lemmaform .. "_$$$_" .. pos ] =
    --{ lemmaform, pos, tonumber(freq) };--ADD NEW
    data[ lemmaform .. "_$$$_" .. pos ] = { }
    data[ lemmaform .. "_$$$_" .. pos ][1] = lemmaform;
    data[ lemmaform .. "_$$$_" .. pos ][2] = pos;
    data[ lemmaform .. "_$$$_" .. pos ][3] = freq;
  else
    -- ADD UP FREQUENCIES (TO EXISTING) (note Lua arrays = 1-based index)
    data[ lemmaform .. "_$$$_" .. pos ][3] =
      data[ lemmaform .. "_$$$_" .. pos ][3] + tonumber(freq);
  end
  nNumForms = nNumForms + 1;
end--bDoEntry
end

for key,value in pairs(data) do
  local lemmaform = value[1];
  local pos = value[2];
  local freq = value[3];
  tLuaLog("FINAL:"..lemmaform.." "..pos.." "..freq)

  -- See if there is an existing entry of this form and part of speech
  local ENTRY = nil;
  local CURRENT = SECTIONDEST:FindEntries( lemmaform );
  for i=0,CURRENT:size()-1,1 do
```

```
    if (tQuery(CURRENT[i],"/@"..CFG.ATTR_POS) == pos) then
        ENTRY = CURRENT[i];
    end
end

-- If no existing entry, create a new one
if ENTRY==nil then
    local NODE = DOC:AllocateElementByID(NODE_ENTRY,true);-- Alloc new entry
    ENTRY = tolua.cast(NODE, "tcEntry");
    ENTRY:SetLemmaSign( lemmaform );

    SECTIONDEST:InsertEntry(ENTRY);

    nNumCreated = nNumCreated+1;
else
    nNumExistingModified = nNumExistingModified + 1;
end

-- Set frequency, POS etc.
local ATTR_FREQ=ENTRY:GetElement():FindAttributeByName(CFG.ATTR_FREQUENCY);
if (ATTR_FREQ~=nil) then
    ENTRY:SetAttributeI( ATTR_FREQ, freq );
end

local ATTR_POS = ENTRY:GetElement():FindAttributeByName(CFG.ATTR_POS);
if (ATTR_POS~=nil) then
    ENTRY:SetAttributeDisplayByString( ATTR_POS, pos, false,
        "___prevent_unintentional_list_string_splitting___" );
end
end
data=nil;
Evt_LemmasInserted:Trigger(nil, SECTIONDEST);--Update UI etc.
DOC:SetDirty();

local sRetMessage =
    "FORMS: ".. nNumForms ..
    " CREATED: " .. nNumCreated..
    " EXISTING_UPDATED: " .. nNumExistingModified
    ;
return sRetMessage;
```

Addendum 2: Lua script: AssignRankBasedOnSortBy.lua

```
-- 2013-10 Assign numerical 'rank' based on sort order
-- (sort order defined by e.g. FIRST selecting F4 SortBy
-- 'Word::Frequency' in second section just before running this script)
-- David Joffe

-- 'CONSTANTS' (script configuration - if e.g. attribute names change,
-- update script here)
CFG =
{
  ATTR_RANK = "Rank", -- Rank
  SECTION = 1 -- Section 0-based index for generating ranks
}

local DOC=tApp():GetCurrentDoc();
if DOC==nil then return "No document"; end

-- STATS
local SECTION=DOC:GetDictionary():GetLanguage( CFG.SECTION );
if (SECTION==nil) then return "Invalid section index"; end

local SECTWND = tFrameWindow():GetLanguageWindow(SECTION);
if (SECTWND==nil) then
  return "No section window for section (try go out of expanded view mode)";
end

-- By default F4 SortBy puts highest frequency at bottom, so if so, invert rank
-- values set as we loop across entries (e.g. rank '1' would be the bottom entry
-- in the list if this is set to true)
local bInvertOrdering=true;

-- Iterate through (NB) the SECTION WINDOW entry list - so e.g. SortBy may be in
-- effect
local i;
local Attr=nil;
for i=0,SECTWND:GetNumLemmaListEntries()-1,1 do
  local ENTRY=SECTWND:GetLemmaListEntry(i);

  if Attr==nil then
    Attr = ENTRY:GetElement():FindAttributeByName( CFG.ATTR_RANK );
    if Attr==nil then return "Rank attribute not found"; end
  end
end
```



```
if bInvertOrdering then
  ENTRY:SetAttributeDisplayByString( Attr,
    SECTWND:GetNumLemmaListEntries() - i, false,
    "___prevent_unintentional_list_string_splitting___" );
else
  ENTRY:SetAttributeDisplayByString( Attr,
    i+1, false,
    "___prevent_unintentional_list_string_splitting___" );
end
end

-- Update user interface etc.
Evt_LemmasInserted:Trigger(nil, SECTION);
DOC:SetDirty();

return "";
```

Addendum 3: Lua script: AssignFrequencyBandBasedOnRank.lua

```
--Frequency bands
local RANKS={
  {500,"1"},
  {1000,"2"},
  {1500,"3"},
  {5000,""},
  {999999999999999999,"LOW"}
};

--check that there is a global document object
--(i.e. make sure we actually have a dictionary open)
if g_pDoc == nil then
  --No document is loaded - exit
  return "Failed - No document loaded";
end

local LEMARRAY = {};
local LEMMAELEM = g_pDoc:GetDictionary():GetDTD():FindElementByName("Word");
local LEMMAFREQATTR = LEMMAELEM:FindAttributeByName("Frequency");
local LEMMAFREQBANDATTR = LEMMAELEM:FindAttributeByName("FrequencyBand");
local LEMMARANKATTR = LEMMAELEM:FindAttributeByName("Rank");

-- Change this 0 to 1 if doing a bilingual 2nd section (right half):
local LANG = g_pDoc:GetDictionary():GetLanguage(0);
local COUNT = 0;
for i=0,LANG:GetNumChildren()-1,1 do
  local LEMMA = LANG:GetChild(i);
  local FREQ = LEMMA:GetAttributeIIntValue(LEMMAFREQATTR);
  table.insert(LEMARRAY,{freq = FREQ,lem = LEMMA});
end

table.sort(LEMARRAY,
  function (a, b)
    return a["freq"] > b["freq"]
  end)

for j,k in pairs(LEMARRAY) do
  for v,d in pairs(RANKS) do
    if COUNT < d[1] then
      --First have to clear the original selection (in case we want an
      --empty value for any of the ranges) - SetAttribute*() does not
      --set the list value at all if passed a null string
```

```
k["lem"]:SetAttributeListID(LEMMAFREQBANDATTR,0);
k["lem"]:SetAttributeDisplayByString(LEMMAFREQBANDATTR,d[2]);
--k["lem"]:SetAttributeDisplayByString(LEMMARANKATTR,COUNT+1);
break;
end
end
COUNT = COUNT + 1;
end

g_pDoc:SetDirty();

--script terminated without error
return "done";
```

Corpus-driven Bantu Lexicography Part 3: Mapping Meaning onto Use in Lusoga

Gilles-Maurice de Schryver, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of African Languages, University of Pretoria, Pretoria, South Africa (gillesmaurice.deschryver@UGent.be)*

and

Minah Nabirye, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of Teacher Education and Development Studies, Kyambogo University, Kampala, Uganda (minah.nabirye@UGent.be)*

Abstract: This article is the third instalment in a trilogy of studies that deal with corpus-driven Bantu lexicography as applied to Lusoga. Having dealt with corpus-building in Part 1, and macro-structural aspects in Part 2, we now focus on the microstructure of a dictionary and in particular on the concept of Mapping Meaning onto Use. The starting point is Patrick Hanks's book chapter by the same title, which we transpose to a study of the high-frequent motion verb *-v-* in Lusoga. Our detailed analysis is as much practical as it is methodological.

Keywords: BANTU, LUSOGA, CORPUS LEXICOGRAPHY, DISTRIBUTIONAL CORPUS ANALYSIS, MAPPING MEANING ONTO USE, MEANING POTENTIALS, MOTION VERBS

Obufunze: Omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu. Ekitundu 3: Okukwanaganika amakulu n'enkozesa mu Lusoga.

Olupapula luno n'olwokusatu mu nteeko y'okulaga omusomo gw'omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu ogulaga omulimu ogw'akolebwa ku Lusoga. Oluvainhuma lw'okwandhula engeli eitu ly'Olusoga mu Kitundu 1 n'omuteeko gw'omutindiigo ogusinziilwaku okuzimba olukala lwa namungi w'ebigambowazo mu Kitundu 2, buti eisila liize ku kulaga ngeli amakulu g'ebigambo ye gakwanaganizibwa n'enkozesa. Omusingi gw'eisomo elilagibwa mu kitundu kino gw'atebwawo Patrick Hanks. Ensonga enkulu dhe yataaku eisila dhilondoolebwa okusinziila ku kigelo kya namungi w'ennhingizo entabaazi (*o*)ku.v.a. Olupapula luno lugelaagelania engeli ennhingizo eno bwe yaingizibwa mu *Eitwanika ly'Olusoga* elitaasinziililwa ku itu lya bigambo n'engeli gye yandibaile esengekebwa singa eitu n'ebigelo by'emiwendo egilagibwamu byali bikozezebwa. Eby'asoboka n'ebitaasoboka bigelaagelanzibwa n'ebigendelelwa by'omusingi gw'eisomo ly'eitu lya namawanika mu mpandiika y'amawanika.

Ebigambo ebikulu: BANTU, LUSOGA, EITU LYA NAMAWAIKA, ENNEKEENEENIA Y'EBIGELO BY'EMIWENDO EBILAGIBWA MU ITU LY'OLULIMI, OKUKWANAGANIKAMA-AMAKULU N'ENKOZESA, AMAKULU AGASOBOKA, KINANTABILA OMUTABAAZI

1. Goal of the present study

In this article we wish to investigate how meaning potentials may be drawn from usages as found in a Bantu-language corpus, through an approach known as 'mapping meaning onto use' (Hanks 2002), as applied in the ongoing compilation of a new Lusoga dictionary. With this topic we are squarely dealing with a dictionary's microstructure, although the method may of course be used (and *is* used) in the field of Bantu corpus linguistics more generally, as may be seen from the recent PhDs of Nabirye (2016) for Lusoga, Kawalya (2017) for Luganda, and Mberamihigo (2014), Nshemezimana (2016) and Misago (2018) for Kirundi.

The major reference for any corpus-based microstructural issues in Bantu lexicography is de Schryver and Prinsloo (2000). In the academic literature, the attention paid to the microstructural level is far more extensive than that paid to the macrostructural level, even in articles that aim to give a perspective on both (Prinsloo and de Schryver 2001, de Schryver 2008) or in articles that take the 'lemmatisation of ...'-formula as a point of departure (de Schryver et al. 2004: 37), which is at heart macrostructural in nature but typically develops into a discussion of microstructural aspects. This may briefly be illustrated with dictionary research undertaken for Northern Sotho.

The 'lemmatisation of ...'-formula may be found in the numerous corpus-based lexicographic studies for the various word classes and other word sets of Northern Sotho, including: reflexives (Prinsloo 1992), verbs (Prinsloo 1994, Prinsloo and Gouws 1996, de Schryver and Prinsloo 2001), adjectives (Gouws and Prinsloo 1997), nouns (Prinsloo and de Schryver 1999, Bosch and Prinsloo 2002), days (de Schryver and Lepota 2001), loan words (Nong et al. 2002), copulatives (Prinsloo 2002), terms (Prinsloo and de Schryver 2002, Taljard and de Schryver 2002), adverbs (Prinsloo 2003), demonstrative copulatives (de Schryver et al. 2004), concords and pronouns (Prinsloo and Gouws 2006), and kinship terms (Prinsloo 2012, Prinsloo and Bosch 2012, Prinsloo 2014b). The opposite also occurs, namely when a primarily microstructural aspect impacts the macrostructure, again with examples for Northern Sotho: left-expanded microstructures (Gouws and Prinsloo 2005), reversibility (de Schryver 2006), communicative equivalence (Prinsloo 2006), and paradigms (Prinsloo 2014a). It has furthermore been noted that the distinction between the macrostructural and microstructural levels tends to disappear in a digital dictionary environment, as has also been illustrated abundantly for Northern Sotho (Prinsloo 2005, Prinsloo et al. 2012, Prinsloo et al. 2014, Prinsloo et al. 2017). Lastly, dictionary reviews, of for instance the corpus-based *Oxford Bilingual School Dictionary: Northern Sotho and English* (de Schryver 2007), likewise tend to focus on microstructural aspects (Prinsloo 2009, Chabata and Nkomo 2010, Faaß 2010, Klein 2010a, b, Madiba and Nkomo 2010, Kosch 2013).

While the use of a corpus to create the microstructure of a Bantu-language dictionary is thus arguably not a novel undertaking in the field, we do add to

the existing studies: (i) a theoretical framework for the current practice,¹ and (ii) a detailed analysis of how one actually goes from concordance lines to dictionary lines. In the process we will also explore two further issues, namely: (i) the differences between the use of a corpus and a manual effort, and (ii) the potential enhancement of illustrative material through the exploitation of corpus metadata.

2. On methods and theoretical models

2.1 Corpus linguistics

The description of any language — whether in dictionaries, grammars or other reference works — should be based on real usage of that language. While one could claim that this ought to be the obvious approach, even a cursory look at much of the output by linguists shows otherwise. As adherents of the work of Patrick Hanks, we find the following quote most appropriate:

[...] the literature of twentieth-century linguistics is strewn with examples of self-fulfilling theoretical prophecies, in which bizarre examples are first invented, then judged to be acceptable (according to the researcher's intuitions), and then presented as evidence for conclusions about some aspect of the nature of language or linguistic rules. (Hanks 2013: 307)

In order to be able to describe 'real' language,² large quantities of actual occurrences of that language are first collected, and then brought together in what is known as 'an electronic corpus'. Dedicated corpus-query software, such as WordSmith Tools (Scott 1996–2018), is used to search and help quantify the hard evidence found in a corpus. At that point, and only at that point, does the researcher *explain* that evidence:

There is a huge difference between consulting one's intuitions to *explain* data and consulting one's intuitions to *invent* data. Every scientist engages in introspection to explain data. No reputable scientist (outside linguistics) invents data in order to explain it. It used to be thought that linguistics is special — that an exception could be made in the case of linguistics — but comparing the examples invented by linguists with the actual usage found in corpora shows that this is not justifiable. (Hanks 2013: 20)

To an increasing number of researchers in the language sciences the power of natural language data is compelling indeed, and for major languages this has given rise to the vibrant field of corpus linguistics, for which Sinclair (1966) may be considered the pioneering study.³ Now half a century on, the field of corpus linguistics is booming; the *International Journal of Corpus Linguistics*, for instance, celebrated its 20th anniversary in 2015.

Crucial for corpus linguistics is to have access to a fair amount of textual data — at least a million running words, although for major languages corpora

of several billion words are not uncommon (Kilgarriff 2003–18). For languages of limited diffusion — be those minor, minority, endangered or simply neglected languages — the lack of sufficient textual data is typically the bottleneck. Billion-word corpora are obtained by crawling the web (de Schryver 2002), a type of corpus-building effort for which most aspects are automated. Transcribing naturally-occurring speech, the default for documentary linguists, is known to be both time-consuming and costly. However, for more and more formerly under-resourced languages, written material is becoming available online (Scannell 2003–18), and for those languages the prospect of applying techniques from the field of corpus linguistics comes into view.

2.2 Bantu corpus linguistics (BCL)

The prospect of applying techniques from the field of corpus linguistics has now become a reality for a good number of Bantu languages. For Lusoga in particular, corpus-building efforts have been described in Part 1 of the present series of three articles. There it was shown that, in addition to an oral component of over half a million words in the 1.7m Lusoga corpus, about a quarter of a million words were found on the Internet, the rest of the corpus being mainly the result of the digitalisation of printed materials.

The field of Bantu corpus linguistics is about two decades old, and is reckoned to have begun with de Schryver's (1999) corpus take on the phonetics of Cilubà. Subsequently, and together with colleagues from South Africa, de Schryver effectively established BCL as a feasible research methodology. While de Schryver was at the University of Pretoria, corpus-based linguistics was undertaken for Zulu (de Schryver and Gauton 2002, Gauton et al. 2004) and for Northern Sotho (Taljard and de Schryver 2002, de Schryver and Taljard 2006). Related work was also done at the universities of Helsinki and Dar es Salaam on Swahili (Sewangi 2000, 2001, Toscano and Sewangi 2005). This early work tended to be corpus-based (i.e. studies for which a corpus is used as one source of evidence in addition to others), in contrast to more recent studies which tend to be corpus-driven (i.e. studies in which a corpus itself is considered to be the sole source of hypotheses about language) — a distinction we owe to Tognini-Bonelli (2001).

The team at the University of Pretoria has since furthered the field of BCL, as may be seen in studies on Northern Sotho (Taljard 2006, de Schryver and Taljard 2007, Taljard 2012, Taljard and de Schryver 2016). Meanwhile at BantUGent (i.e., the UGent Centre for Bantu Studies), an increasing number of research articles includes aspects of BCL, as seen in studies on Lusoga (de Schryver and Nabirye 2010, Nabirye and de Schryver 2011, Nabirye 2016), on Cilubà (De Kind and Bostoen 2012, Dom et al. 2015), on Kirundi (Bostoen et al. 2012, Mberamihigo 2014, Lafkioui et al. 2016, Mberamihigo et al. 2016, Nshemezimana 2016, Nshemezimana and Bostoen 2016, Devos et al. 2017, Misago 2018), on Swahili (Devos and de Schryver 2013, 2016), on Kikongo (De Kind et al. 2013, Bostoen

and de Schryver 2015, De Kind et al. 2015), and on Luganda (Kawalya et al. 2014, Kawalya 2017, Kawalya et al. 2018). Not all of these studies are truly corpus-based, let alone corpus-driven, as some of them are closer to being 'corpus-illustrated' (Tummers et al. 2005) or even tend to use their corpora as fish ponds:

Some famous and influential linguists have simply denied the relevance of corpus evidence to linguistic theory. Others have in recent years treated corpora as 'fish ponds' in which to angle for fish that will fit independently conceived hypotheses and theories. Fish that don't fit the theory are thrown back into the pond. [Note: I owe this metaphor to John Sinclair, in conversation some years ago.] (Hanks 2013: 7, 431)

On the relationship between corpus-driven and fish-pond linguistics, Hanks furthermore points out:

Corpus-driven research [...] attempts to approach corpus evidence with an open mind and to formulate hypotheses and indeed, if necessary, a whole theoretical position on the basis of the evidence found. If work is merely 'corpus-based', [Tognini-Bonelli] argues, it risks missing important insights. A truly empirical linguist (or lexicographer) is 'driven' by the data in the corpus. [... The fish pond] analogy is no doubt unfair, for even Tognini-Bonelli, Sinclair, Stubbs, Hanks, and other empirical linguists cannot avoid making some theoretical assumptions as a starting point and using examples selectively, not merely randomly. However, a corpus-driven linguist holds her or his theoretical assumptions lightly and is ready to reconsider them in the light of accumulated evidence. (Hanks 2012: 417)

Therefore, whenever possible, any future studies for Bantu languages should aim to be *driven* by corpus data. This, too, is valid for the field of lexicography, in our case for the compilation of Lusoga dictionaries.

2.3 Distributional corpus analysis (DCA)

For each aspect for which a corpus is used, a corpus analyst first takes stock of the evidence through an approach that has been termed 'distributional corpus analysis'. Geeraerts (2009: 422-423) proposes to view DCA of the Sinclair-type as a neostructuralist approach to lexical semantics, with, as its main characteristic, the 'radical usage-based rather than system-based approach: it considers the analysis of actual linguistic behaviour to be the ultimate methodological foundation of linguistics' (Geeraerts 2010: 168). Hanks, however, takes issue with Geeraerts's view of DCA as primarily a method, not a model, and comments:

This is odd, because examination of the work of corpus analysts such as Sinclair, Hoey, Wray, Stubbs, Moon, Partington, Semino, McEnery, Hanks, and others would show that corpus analysis lends support to a model of linguistic behav-

our founded on prototypical usage — and Geeraerts himself is a proponent of the theory of conceptual prototypes. (Hanks 2015: 102-103)

Entering the fray on whether or not corpus linguistics is more than a methodology goes beyond the scope of the present study. It is certain, however, that in the field of Bantu lexicography, we do use DCA as a method to arrive at various distributions (of homonyms, of meaning potentials, etc.). We nonetheless also like to believe that corpus linguistics is a/our theoretical model.

2.4 Mapping meaning onto use

The various lexicographic uses of a corpus on the macrostructural level have been described, and were illustrated for Lusoga, in Part 2 of the present series of three articles. When querying a corpus in order to compile a dictionary's microstructure, there are at least five uses of that corpus: (i) to map meaning potentials, (ii) to verify and support mother-tongue intuitions, (iii) to study various distributions, (iv) as a source of examples, and (v) to provide overall counts. Working briefly through this list, from last to first, and with a focus on our Lusoga case study, we can note the following. As far as corpus counts are concerned, these are a natural by-product of the steps described in Part 2. There, it was shown that the output of the lemmatisation effort consists of 'skeleton dictionary articles', each with a lemma, part of speech, frequency, rank, frequency band and (optionally) a short meaning. The relative frequency of each candidate lemma sign is, in other words, known at the start of the compilation of each dictionary article.

Each meaning potential that will eventually be singled out is ideally also illustrated with one or more of the corpus lines that were studied to arrive at that meaning. It is a good idea to include information on the source (cf. the *Filename* in Part 1) in one way or another, with the aim to either show it overtly in 'the' or in 'one of several' final lexicographic products, or to only keep it on file for the dictionary-makers while hiding it from the target users, so that the evidence may always be traced.

As one works through the corpus lines, one is bound to begin sorting and grading the evidence, whereby one automatically ends up drawing up distributions, which may again either be used implicitly or explicitly in the actual dictionary/-ies.

Regarding intuition, it has already been pointed out that the corpus analyst needs her or his own intuition to explain data, but in order to wade through the mass of data beyond the word level, intuition is also an excellent trait to start *exploring* the corpus with. It is good to make ample use of it, but subsequently one should always stick to the principles of corpus-driven analysis in explaining the evidence. What exists is mentioned, what doesn't appear in the corpus (when expected on intuition) may or may not be pointed out. Of

course the latter does not mean that something definitely cannot occur and/or would be ungrammatical, as 'no amount of corpus evidence will provide negative evidence — evidence for what *cannot* occur' (Hanks 2013: 415). This is not a problem, as 'being able to make predictions about probable usage is much more useful than speculating about the boundaries of possibility' (Hanks 2013: 415).⁴

As regards the meaning, it may come as a surprise to non-lexicographers but it is well-known to lexicographers: no single mother-tongue speaker knows 'all the words' of her or his language (a feature lexicographers make you believe they possess; after all, aren't they supposed to say something about every word of a language?). As a matter of fact, corpus data continuously challenges what one assumes one knows about words and their meanings. Meanings, in short, can only sensibly be derived from their uses as seen in a corpus, through a principle known as Mapping Meaning onto Use (Hanks 2002), which uses the technique of Corpus Pattern Analysis (Hanks 2004), itself based on the Theory of Norms and Exploitations (Hanks 2013). Reference is made to these seminal works for the full theoretical framework. The problem has been stated by Hanks as follows:

Existing dictionaries may be guilty of sins of omission (e.g. in accounting for pragmatics and function words), but they are equally guilty of sins of commission. They can make things seem even more complicated than they really are. In part, this is because the structure of a traditional dictionary entry is dictated by meanings not by use. Word meaning (if such a thing exists at all) is extremely vague and unstable. A word can have about as many senses as a lexicographer cares to perceive. (Hanks 2002: 159)

To which Hanks proposes the following solution:

[...] the lexicographer must first group the corpus evidence for each word according to the contexts in which it occurs, and then decide to what extent it is possible to group different contexts together (on the grounds that they express what is essentially the same meaning), and to what extent it is necessary to make distinctions. ¶ With the advent of large corpora, it is possible to be much more precise about the typical contexts in which a word is used, and to associate different meanings with different contexts. The crucial point here is to choose, as an organizing principle for the dictionary entry, *context* (which is objectively observable and measurable) rather than *meaning* (which is opaque and depends on the perceptions of the definer). Lexicographers should think first in terms of syntax and context (or, more strictly, syntagmatics), rather than directly in terms of semantics. They can thus approach meaning indirectly, through syntagmatic analysis, according to a motivated grouping of the evidence. (Hanks 2002: 159-160)

In short, then, and with reference to our new dictionary project for Lusoga, in addition to the brief meanings as may already be logged following lemmatisation in the dictionary writing system (i.e., the TLex file (Joffe and de

Schryver 2002–18)), the main use of a corpus on the microstructural level is to say more about word meanings in context.

3. A case study for Lusoga

3.1 Choosing the Lusoga case study

We now wish to illustrate the mapping of meaning onto use for Lusoga lexicography. Compared to working *on* English and writing about the process *in* English, which is already quite hard enough, we have the additional problem that we need to translate everything *out of* Lusoga and *into* English for the reader to be able to follow. Hanks's (2002) article on the topic, which also bears the title 'Mapping Meaning onto Use', has been summarised as follows:

Hanks presents his own corpus analyses of *lean* and *tank* for lexicographical purposes. Rare are such detailed accounts in which the reader is led by the hand and allowed to see how the master cuts his way through the corpus vines. The latter, including their analyses, are displayed in full as addenda, hereby allowing the reader to appreciate the hesitations — about which Hanks is quite open — even more. Once the path has been cut, once Hanks unspun the hanks, the reader is offered the view that syntagmatics in tandem with 'perceived meaning' ought to be the organising principle of dictionary entries for verbs and adjectives. The organisation for nouns is similar, but slightly more complicated. (de Schryver 2005: 423)

In other words, just two words are used to illustrate the process, one verb (*lean*) and one noun (*tank*). For reasons of space, and given that we also need to translate our material, we will limit our current analysis for Lusoga to just one verb. For an idea of the issues involved in undertaking a study of the Lusoga noun using a corpus, see de Schryver and Nabirye (2010), which contains a section on the semantic import of the noun in Lusoga.

The Lusoga verb chosen for the present case study is the motion verb *-v-*. The root of this verb consists of just one letter, the letter 'v', which immediately indicates the additional difficulty of merely *finding* this verb in a raw corpus, thus one without any morphological analysis, which the 1.7m Lusoga corpus was before lemmatisation. We, however, took up the challenge.

3.2 The verb *-v-* in the monolingual Lusoga dictionary

To begin the discussion in a practical way, we will be employing a shortcut, by translating the relevant information gleaned from the *Eiwanika ly'Olusoga* (Nabirye 2009b), which is a monolingual dictionary of Lusoga, compiled *without* access to a corpus. This dictionary has also been digitised (Nabirye and de Schryver 2013), and is available on disc as well as freely online from <http://menhapublishers.com/dictionary/>. In that dictionary, the verb *-v-* is to

be found on page 379, as two homonymous forms, and as two lemma signs with the locative enclitics *-ku* and *-mu* respectively. This page is shown in Addendum 1, while the slightly edited and reformatted online data is shown in Table 1, on the left.

Table 1: The dictionary articles for the verbs *-v⁻¹*, *-v⁻²*, *-vaaku* and *-vaamu* in the *Eiwanika* *ly'Olusoga* (Nabirye 2009b), together with translations

e-Eiwanika	Translation
<p>(o)ku.v.a¹ [(o)kúvǎ] <i>kt.[L]</i> [-viile] [nviile] bl: [Lg: okuva]</p> <p>1. Okusimbuka mu kifo ekilala waayolekela ekindi. gez: <i>Nva Mayuge.</i></p> <p>2. Okusibuka. gez: <i>Nva Iganga.</i></p> <p>3. Okulekelela ekintu ky'obaile okola. gez: <i>Ebyo nabiviileku.</i></p> <p>4. Okuseguka mu ngila oba mu kifo. gez: <i>Leka nkuviile ofune eidembe.</i></p> <p>ssk:</p> <ul style="list-style-type: none"> • Okuva ku luguudo: <i>Okwonooneka / Okuva ku mulembe</i> <p>(g)gl:</p> <ul style="list-style-type: none"> ◆ Awava akwita n'awava akukobela ◆ Awava ennume waila nnume ◆ Awava mwino tiwaila mwino: Awava eliino waila ilibu ◆ Awava mwino tiwaila mwino: Awava eliiso waila itulu ◆ Akaviile mu igi tikatya ikoli ◆ Edhiva okulala n'embilo ◆ Empambo eva ku kiwalo ◆ Ennhonhi eva ewala temala mutonto ◆ Ensanafu eva ku mugendelo telwa kufuuka kabasa ◆ Atava ku mulungi afa t'awoza ◆ Ka nduviile ku ntobo oti n'omuyala atuuse we bafumba ◆ Olusubi olulala we luva ku ndhu tatoonha ◆ Omukazi omulungi nnilimo ya ngila buli avaayo agyegwaniza 	<p>[lemma sign, part of speech, morphological information, cognate in Luganda]</p> <p>1. to depart. e.g. <i>I come from Mayuge.</i></p> <p>2. to come from. e.g. <i>I hail from Iganga.</i></p> <p>3. to abandon. e.g. <i>I gave up on those things.</i></p> <p>4. to make way. e.g. <i>Let me pave the way for you so that you get peace.</i></p> <p>[combination(s) with the lemma]</p> <ul style="list-style-type: none"> • to go off the road: <i>to be completely damaged and unusable / to be out of fashion</i> <p>[proverbs]</p> <ul style="list-style-type: none"> ◆ The person who warns you comes from the same place as the person who will kill you ◆ Where a male leaves another male will take over that place ◆ The gap that your friend leaves is not filled by another friend: A gap takes the place of a tooth that has left ◆ The gap that your friend leaves is not filled by another friend: Blindness takes the place of the eye that has left ◆ The one that has just come from an egg does not fear an eagle ◆ The steps you take one after the other develop into running ◆ A wise lesson is learned from the cradle ◆ The bird that comes from far away does not finish up the edible fruit ◆ The safari ant that leaves the trail does not take long to turn into a traitor ◆ The one who does not let a beautiful one alone dies while still giving explanations ◆ Let me start from the very beginning like the hungry person who has arrived at the place where food is being cooked ◆ When one blade of grass falls off the house the house does not leak

<ul style="list-style-type: none"> ◆ Omukwano guva mu ngabo ◆ Omukwano guva mu ngila gwatuuka eka ◆ Omusaadha kikele kiva kyonka mu bwina ◆ Va we ndi takulwania ◆ Va ku ntebe ya lata awulilila ku ise ◆ W'ova tosoile w'otela okwila <p>bbgz: <i>Okuviila, Okuviisa.</i></p>	<ul style="list-style-type: none"> ◆ A beautiful woman is a garden along the road: whoever comes by wants it for himself ◆ Friendship comes from the shield (sharing) ◆ Friendship comes from the road and it is brought home ◆ A man is a frog, which comes out of the hole by itself ◆ (The one who says that) 'Go away from where I am' should not make you fight ◆ (The one who says that) 'Go away from father's chair' has heard it from his father ◆ The place that you leave without picking a quarrel is the one where you always return [lemma plus verbal extensions] <p><i>okuviila</i> [+ APPL ext.], <i>okuviisa</i> [+ CAUS ext.]</p>
<p>(o)ku.v.a² [(o)kúv^á] <i>kt.[L]</i> [-viile] [nviile] bl: [Lg: okuva] Okutandiikila mu kifo ekilala okutuuka ku kifo ekindi. gez: <i>Ennhandha ya Nalubaale eva Idhindha.</i></p>	<p>[lemma sign, part of speech, morphological information, cognate in Luganda] to start at a given point and move in the direction of another. e.g. <i>Lake Victoria starts in Jinja.</i></p>
<p>(o)ku.v.a.a.ku [(o)kúv^ááku] <i>kt.[T]</i> [-viileku] [nviileku] bl: [Lg: okuvaako]</p> <ol style="list-style-type: none"> 1. Okuzima. gez: <i>Amasaanhalaze gaviileku.</i> 2. Okuleka. gez: <i>Omwenge nguviileku. / Oyo namuvaaku naafuna owundi.</i> <p>bbgz: <i>Okuviilwaku, Okuviisaaku.</i></p>	<p>[lemma sign, part of speech, morphological information, cognate in Luganda]</p> <ol style="list-style-type: none"> 1. to go/turn off. e.g. <i>The electricity has gone off.</i> 2. to let alone/put aside. e.g. <i>I have put alcohol drinking aside. / I left that person alone and got another.</i> <p>[lemma plus verbal extensions] <i>okuviilwaku</i> [+ APPL + PASS ext.], <i>okuviisaaku</i> [+ CAUS ext.]</p>
<p>(o)ku.v.a.a.mu [(o)kúv^áámu] <i>kt.[T]</i> [-viilemu] [nviilemu] bl: [Lg: okuvaamu]</p> <ol style="list-style-type: none"> 1. Obutatuukiliza kye wasuubiza omuntu. gez: <i>Tubaile tusuubiila nti agya kutuyamba aye atuviilemu.</i> 2. Okulyamu olukwe. gez: <i>Gwetwateesa naye mwene neeyatuvaamu.</i> 3. Okuwa oba okumaliliza. gez: <i>Bw'obifumba bivaamu bulungi.</i> 	<p>[lemma sign, part of speech, morphological information, cognate in Luganda]</p> <ol style="list-style-type: none"> 1. to not fulfil what is expected of you. e.g. <i>The one who promised to help us has failed us.</i> 2. to betray. e.g. <i>The actual person we planned with is the one who betrayed us.</i> 3. to turn out well. e.g. <i>When you cook them they come out very well.</i> 4. to make a loss. e.g. <i>I have come out with nothing.</i> 5. to not be properly fixed. e.g. <i>These shoes do not fit.</i>

<p>4. Okubula ky'ofuna oba ky'ogobolola mu kintu. gez: <i>Nze nviiliilemu awo.</i></p> <p>5. Okusagala. gez: <i>Eno engaito evaamu.</i></p> <p>bbgz: <i>Okuviilamu, Okuviisamu.</i></p>	<p>[lemma plus verbal extensions] <i>okuviilamu</i> [+ APPL ext.], <i>okuviisamu</i> [+ CAUS ext.]</p>
---	--

Intuition combined with the fieldwork that led to the dictionary data seen in Table 1 clearly indicate that the verb(s) *-v-*, without and with locative enclitics, is/are indeed quite polysemous.

3.3 The verb *-v-* in the Lusoga lemmatised frequency list

From the 1.7m Lusoga corpus (cf. Part 1), a lemmatised frequency list was created (cf. Part 2). Perusing it, we notice that the data for the verbal lemma *-v-* was not split into two. Deciding whether or not to create two homonyms for *-v-* was not feasible during lemmatisation, where the focus was literally on lemmatisation and part-of-speech assignment, not on any detailed studies of usage leading to meaning. When it comes to the verbal forms with locative enclitics, however, we find not just *-vaaku* (with an enclitic from cl. 17) and *-vaamu* (cl. 18) in the lemmatised frequency list, but also *-vaawo* (cl. 16) and *-vaayo* (cl. 23). From a frequency point of view, then, one can say that the latter two locativised verbs were 'overlooked' during the manual (i.e., non-corpus) effort to compile the monolingual Lusoga dictionary. Also overlooked in the *Eiwanika ly'Olusoga* is the deverbative noun *-vo* in cl. 14, which does have a respectable frequency in the lemmatised frequency list. These six lemmas are listed in Table 2, together with their lemma frequencies, lemma ranks, lemma frequency bands, as well as number of formatives.

Table 2: The lemmas *-v-*, *-vaawo*, *-vaaku*, *-vaamu*, *-vaayo* and cl. 14 *-vo* in the lemmatised frequency list derived from the 1.7m Lusoga corpus

Lemma	Part of speech	Freq.	Rank	Freq. band	# formatives
<i>-v-</i>	<i>verb</i>	6 611	21	①	67
<i>-vaawo</i>	<i>locativised verb (cl. 16)</i>	15	3 679	-	1
<i>-vaaku</i>	<i>locativised verb (cl. 17)</i>	14	3 852	-	1
<i>-vaamu</i>	<i>locativised verb (cl. 18)</i>	242	571	②	8
<i>-vaayo</i>	<i>locativised verb (cl. 23)</i>	281	518	②	23
<i>-vo</i>	<i>deverbative noun, in cl. 14</i>	40	2 096	-	2

The formative (or underlying) data that led to the six lemmas listed in Table 2 is presented in Addendum 2. For the verb *-v-*, for instance, 67 types were

frequent enough — meaning that their frequency was at least 12 in the 1.7m Lusoga corpus (cf. Part 2, §3) — and the frequencies of these 67 all contribute to the total frequency of the lemma *-v-*, being 6 611, which turns out to be one of the most frequent lemmas in the language, with rank 21. From Table 2 one may further conclude that given that *-vaaku* was entered in the *Eiwanika ly'Olusoga*, *-vaawo* with a similar frequency and cl. 14 *-vo* should indeed have been entered as well, and especially the top-frequent *-vaayo*, the 518th-most-frequent lemma overall in Lusoga.⁵

3.4 The verb *-v-* in the 1.7m Lusoga corpus

3.4.1 Mapping steps and sampling procedure

We are now in a position to study the Lusoga corpus evidence for *-v-*. The steps of the procedure to map meaning onto use have been enumerated as follows by Hanks, with reference to his case study of English *lean*:

Working with a 500-line sample, we sort all the occurrences into different categories, first on broad syntactic grounds (separating adjectives from the verbs), then into more delicate semantic and syntactic frames (e.g. separating 'lean meat' from 'lean businesses') and finally making more subtle distinctions on semantic grounds (e.g. separating different meanings of 'lean on someone', according to the perceived purpose of the person doing the leaning, i.e. reliance or choice). [...] It should be emphasized that the level of detail used in categorization of corpus lines is a matter of choice and judgement: even more delicate subcategorization is possible, or different patterns may be lumped together in a single category. (Hanks 2002: 165-166, our underlining)

Without any further information, sampling the raw Lusoga corpus in search of *-v-* is obviously hard. However, once one realises that one has the underlying forms which led to each lemma at hand, the process is actually perfectly doable. According to the data presented in Addendum 2, the most frequent formatives for the lemma *-v-* are *okuva* (freq. 2 668), *ava* (freq. 389), *ova* (freq. 325), *kuva* (freq. 267), *yava* (freq. 188), *nva* (freq. 162), etc. In other words, one may simply instruct WordSmith Tools to search for any or all of such frequent types at the same time (by simply placing slashes between the various forms), with or without a randomiser (for instance, to limit the output to a sample of 100 lines), to then study the concordance lines. As an alternative, adding a verbal extension, such as an applicative, or the perfect, and searching for *-viil-* rather, is also an option.

3.4.2 The verbs *-v-1*, *-v-2*, the connective *kye-SM-va*, and the adverb *kuva*

After a careful study of several hundreds of concordance lines for *-v-*, we concluded that the various uses are indeed best presented in two separate,

homonymous, dictionary entries. Given that we are describing the evidence in English, there may be a tendency to let the English categories influence the Lusoga evidence. We have avoided that, just as it is good practice in bilingual lexicography not to allow the target language to 'pull' or 'distort' the source language analysis (Atkins 1996: 8).

The various verbal uses as seen in the corpus lead to the meaning potentials listed below, ordered from more to lesser frequent, and grouped around usages that have to do with movement, vs. usages that have to do with projection and direction. Adding an addendum with the many concordance lines will not be beneficial to the reader; instead, we add a glossed example for each use. (For the abbreviations in the glosses, see the explanations at the end.)

*okuva*¹ [move senses]

1. to leave, to depart, to go away
2. to hail (from)
3. to abandon
4. to make way, to move away
5. to result, to come out
6. to spend (time)

1. to leave, to depart, to go away

Kuba ye bwe yava e Makeerere nga amaze okufuna diguli, yafulumamu bufulume yaagya ku mawanga.

kuba	ye	bwe	a-a-v-a	e	makeerere	nga
because	him	when	SM₁-PST-leave-FV	LOC ₂₃	9.Makeerere	CON
a-mal-ile		oku-fun-a	diguli		a-a-fulum-a-mu	
SM ₁ -finish-PERF		15-get-FV	9.degree		SM ₁ -PST-exit-FV-ENCL ₁₈	
bu-fulum-e	a-a-gi-a		ku		ma-wanga	
14-exit-DEV	SM ₁ -PROG-go-FV		LOC ₁₇		6-tribe	

'Because for him when he left Makerere after getting his degree, he just left and went abroad.'

[File ID: PFExtaud | O • Interviews • Language • 2012]

2. to hail (from)

Oviliile wa mu bufunze?

o-v-il-ile	wa	mu	bu-funz-e
SM_{2SG}-hail-APPL-PERF	INTER	LOC ₁₈	14-brief-DEV

'Where did you hail from, in brief?'

[File ID: Ebintub | O • Songs - Traditional • Life • 2010s]

3. to abandon

Omuntu bwe yeetukuza n'ava mu bibi byonabyona, afuuka ekibya ekikozesebwa emilimo egya ghaigulu.

o-mu-ntu bwe a-e-etukul-a ni **a-v-a** mu
AUG-1-person if SM₁-RFL-clean-FV CON **SM₁-abandon-FV** LOC₁₈
bi-bi bi-ona-bi-ona a-fuuk-a e-ki-bya
AP₈-bad PP₈-INC-PP₈-INC SM₁-transform-FV AUG-7-plate
eki-koz-is-ibw-a e-mi-lim-o e-gi-a ghaigulu
SREL₇-use-CAUS-PASS-FV AUG-4-work-DEV AUG-CP₄-of above
'If a person becomes holy and **s/he abandons** all forms of sinful states, s/he becomes a vessel that can be used to do jobs of a high rank.'⁶
[File ID: Endagaan | W • Biblical documents • Religion • 1998]

4. to make way, to move away

Nva ni mu maiso tutasambaganilagho [...]
n-v-a ni mu ma-iso
SM_{1SG}-move_away-FV even LOC₁₈ 6-eye
tu-ta-samb-agan-il-a-gho
SM_{1PL}-NEG_B-kick-REC-APPL-FV-ENCL₁₆
'**Move away** even from my presence; let us not kick each other from here [...]'
[File ID: AkatAkas | W • Literature • Fables • 1999]

5. to result, to come out

Ebiviile mu kubuuzibwa [...]
ebi-v-ile mu ku-buuz-ibw-a
SREL₈-result-PERF LOC₁₈ 15-question-PASS-FV
'**What came out** of the examination [...]'
[File ID: Missa4 | W • Biblical documents • Religion • 2012]

6. to spend (time)

Aye nga Abadiope **baviile** ekiseela nga nga beesabila ela nga bamba nti oba twena
tuliba awo twafuna ku masaanhalaze.
aye_nga a-ba-diope **ba-v-ile** e-ki-seela nga nga
but AUG-2-diope **SM₂-spend-PERF** AUG-7-period ADV ADV
ba-e-sab-il-a ela nga ba-emb-a nti oba
SM₂-RFL-request-APPL-FV CON ADV SM₂-sing-FV that MOD
tu-ena tu-li-b-a a-wa-o tu-a-fun-a
SM_{1PL}-INC SM_{1PL}-FUT₂-be-FV AUG-PP₁₆-DEM_B SM_{1PL}-PROG-get-FV
ku ma-saanhalaze
LOC₁₇ 6-electricity
'But the Badiope **have spent** a long time nagging so that we might also be there (one day) and we (finally) get a bit of electricity.'
[File ID: Mwino | O • Radio talk shows • Politics • 2010]

*okuva*² [projection and direction senses]

1. to start (and continue onwards)
2. to be the source (of), to emanate (from)

1. to start (and continue onwards)

Ate **kiiviila** ilala ku Bbaibbuli wano Yesu bwe yagyanga nga alonda abayigilizwa yaabaagananga na ki? Ni profession dhaibwe.

ate **ki-v-il-a** ilala ku bbaibbuli wa-no yesu
 CON **SM7-start-APPL-FV** INTENS LOC₁₇ 9.Bible PP₁₆-DEMA 1.Jesus
 bwe a-a-gi-ang-a nga a-lond-a a-ba-yigilizwa
 ADV SM₁-PST-go-HAB-FV ADV SM₁-pick-FV AUG-2-disciple
 a-a-ba-agan-ang-a na ki ni profession
 SM₁-PST-OM₂-find-HAB-FV CON INTER CON 10.profession
 dhi-a-ibwe
 CP₁₀-of-POSS_{2PL}

'And **it** really **starts from** the Bible here when Jesus used to go and pick disciples, he usually found they were with what? With their own professions.'
 [File ID: Luthour | O • Radio talk shows • Religion • 2010]

2. to be the source (of), to emanate (from)

NBS ni radio ekutuusaaku amawulile agaba gaakagwawo ate **okuviila** ilala ku bantu abatuufu beene.

NBS ni radio e-ku-tuus-a-ku a-ma-wulile
 9.NBS COP 9.radio SREL₉-OM_{2SG}-bring-FV-ENCL₁₇ AUG-6-news
 aga-b-a ga-aka-gw-a-wo ate **oku-v-il-a**
 SREL₆-be-FV SM₆-APERF-fall-FV-ENCL₁₆ CON **15-emanate-APPL-FV**
 ilala ku ba-ntu aba-tuuf-u ba-ene
 INTENS LOC₁₇ 2-person SREL₂-right-DEV PP₂-RFL

'NBS is the radio that brings you fresh news and on top of that **emanating from** the real right people.'
 [File ID: Mazima | O • Radio talk shows • Politics • 2010]

Combinations

Three combinations appear frequently in the concordance lines, the first derived from *-v*⁻¹, sense 1.

-v- + LOC + maiso = to die

Nga kitalo muna omukaile **okutuwa ku maiso!**

nga kitalo mu-na o-mu-kaile **oku-tu-v-a** ku ma-iso
 as sad 1-AFP AUG-1-old_person **15-OM_{1PL}-leave-FV** LOC₁₇ **6-eye**

'As it is sad my counterpart for the old person **to die!**'
 [File ID: Byaif12 | W • E-mails • Networking • 2012]

The next two frequent combinations are derived from *-v-2*, sense 1, and have to do with measuring, either space or time.

***-v- + -tuuk-* = from ... up to** [measuring space]

[...] *n'abali kwonoona kuva itale ghano ghati okutuukila ghano.*

ni aba-li ku-yoonon-a ku-v-a itale gha-no
COP SREL₂-be 15-spoil-FV 15-start-FV above PP₁₆-DEM_A
ghati oku-tuuk-il-a gha-no
here 15-reach-APPL-FV PP₁₆-DEM_A

'[...] they are the ones who are spoiling from above here like this up to here.'
[File ID: Okukyal | O • Celebrations • Politics • 2011]

***-v- + paka* = from ... up to, from ... until** [measuring time]

Nkola kuva saawa ina paka musanvu ogw'obwile.

n-kol-a ku-v-a saawa ina paka musanvu o-gu-a
SM_{1SG}-work-FV 15-start-FV 9.time ten up_to 3.one AUG-CP₃-of
obu-ile
14-night

'I work from 10:00 until 1:00 o'clock in the night.'

[File ID: BuwaabGr | O • Celebrations • Inspirational • 2010]

Other word classes

Addendum 2 indicates that, among the formatives of the verb *-v-*, one also finds the forms *kyava*, *kyebaava*, *kyenva*, *kyetuva* and *kyeyava*. These words actually belong to a different word class, as these are connectives which are built according to a fixed formula, combining the object relative of class 7, followed by a subject marker, and then *-v-1*, sense 5.

***kye-SM-va* (connective) = that is why**

Buti kyenva tyayenze kufuna batoototo kuba boona baidha kuba bakyotala [...]

buti kye-n-v-a ti-a-yend-ile ku-fun-a
now OREL₇-SM_{1SG}-result-FV NEG_A-PST-want-PERF 15-get-FV
ba-toototo kuba ba-ona ba-idh-a ku-b-a ba-kyotala
AP₂-young because PP₂-INC SM₂-come-FV 15-be-FV 2-half-caste

'Now that is (the reason) why I did not want to get young ones because even they will be half-castes [...]

[File ID: PFExtaud | O • Interviews • Language • 2012]

From *-v-2*, sense 1, the adverb *kuva* is derived.

***kuva* (adverb) = since**

Nnhweileku aye kuva nkyo nkaali kulyaku.

n-nhw-ile-ku aye **ku-v-a** n-kyo n-kaali
SM_{1SG}-drink-PERF-ENCL₁₇ but **15-start-FV** 9-morning SM_{1SG}-not
ku-li-a-ku
15-eat-FV-ENCL₁₇
'I have drunk a bit but **since** morning I have not eaten at all.'
[File ID: AgakbOmu | W • Literature - Novels • Life • 2012]

3.4.3 The locativised verb *-vaawo*

When the class 16 locative enclitic *-wo* is suffixed to the base verb *-v*⁻¹, a new use that was not seen for the base verb is found (1. below), together with the main use as also seen for the base verb (2. below).⁷

okuvaawo < *okuva*¹

1. to stop existing, to die (out)
2. to leave, to depart, to go away

1. to stop existing, to die (out)

*Oyenda toyenda eliyo ebiidha **okuvaawo**.*
o-yend-a ti-o-yend-a e-li-yo ebi-idh-a
SM_{2SG}-want-FV NEG_A-SM_{2SG}-want-FV SM₂₃-be-ENCL₂₃ SREL₈-come-FV
oku-v-a-wo
15-die_out-FV-ENCL₁₆
(Whether) you want or do not want, there are things which will **die out**.
[File ID: Musoke | O • Radio talk shows • Marriage • 2010]

2. to leave, to depart, to go away

*Bakaile baife abalungi **mutavaawo** tuli kwila.*
ba-kaile ba-a-ife aba-lungi **mu-ta-v-a-wo**
2-parent CP₂-of-PERS_{1PL} SREL₂-good **SM_{2PL}-NEG_B-go_away-FV-ENCL₁₆**
tu-li ku-il-a
SM_{1PL}-be 15-come_back-FV
'Our good elders, **do not go away**, we are coming back.'
[File ID: Luthour | O • Radio talk shows • Religion • 2010]

3.4.4 The locativised verb *-vaaku*

When the class 17 locative enclitic *-ku* is suffixed to the base verb *-v*⁻¹, numerous new uses that were not seen for the base verb are found (all but one below), together with one main use as also seen for the base verb (2. below).

*okuvaaku < okuva*¹

1. to go off, to turn off
2. to abandon
3. to trigger, to cause
4. to let aside, to give up
5. to lose
6. to stop
7. to not disturb, to leave alone
8. to finish
9. to come a (little) bit

1. to go off, to turn off

Eeh! Amasaanhalaze gaviileku [...]

eeh a-ma-saanhalaze ga-v-ile-ku
INTERJ AUG-6-electricity SM₆-go_off-PERF-ENCL₁₇

'Eeh! The electricity has gone off [...]

[File ID: PFExtaud | O • Interviews • Language • 2012]

2. to abandon

Oba ti na kindi okutuviilaku ilala [...]

oba ti-na-ki-ndi oku-tu-v-il-a-ku
or NEG_A-MODF-PP₇-EXC 15-OM_{1PL}-abandon-APPL-FV-ENCL₁₇

ilala

INTENS

'Or perhaps even to abandon us completely [...]

[File ID: StarEC3 | O • Radio talk shows • Health • 2010]

3. to trigger, to cause

Ekilivoilaku baana okutuluguunhizibwa nnankani, buvunaanhizibwa bwaiife.

eki-li-Ø-v-il-a-ku ba-ana

SREL₇-PROG-15-cause-APPL-FV-ENCL₁₇ 2-child

oku-tuluguunh-is-ibw-a nnankani bu-vunaanhizibwa bu-a-ife

15-abuse-CAUS-PASS-FV UNS 14-responsibility CP₁₄-of-PRON_{1PL}

'What is causing the children to be abused, (is) something (that is) our responsibility.'

[File ID: P1101216 | O • Radio talk shows • Health • 2010]

4. to let aside, to give up

Mwana wange, kulwaki tobivaaku?

mu-ana a-a-nge ku-lu-a-ki

1-child CP₁-of-POSS_{1SG} PP₁₇-CP₁₁-of-INTER

ti-o-bi-v-a-ku

NEG_A-SM_{2SG}-OM₈-give_up-FV-ENCL₁₇

'My child why **don't you give them up?**'

[File ID: AbabitAb | W • Literature • Fables • 1999]

5. to lose

*Tusaasilaku abantu **abaaviilwaku** abantu baibwe modulo oti n'eyo.*

tu-saasil-a-ku a-ba-ntu

SM_{1PL}-sympathise-FV-ENCL₁₇ AUG-2-person

a-ba-a-v-il-w-a-ku a-ba-ntu ba-a-ibwe

AUG-SM₂-PST-lose-APPL-PASS-FV-ENCL₁₇ AUG-2-person CP₂-of-POSS_{2PL}

modulo oti ni e-e-o

9.model like COP AUG-23-DEM_B

'We sympathise a bit with the persons **who have lost** their people in a manner like that one.'

[File ID: StarEC2 | O • Radio talk shows • Health • 2010]

6. to stop

ARVs **waadhivaaku?**

ARVs **o-a-dhi-v-a-ku**

10.ARV **SM_{2SG}-PST-OM₁₀-stop-FV-ENCL₁₇**

'**Did you stop** (taking) the ARVs?'⁸

[File ID: StarEC4 | O • Radio talk shows • Health • 2010]

7. to not disturb, to leave alone

***Nvaaku** iwe akavubuka ye ggu lwaki onneesimbamu engeli eyo?*

n-v-a-ku iwe a-ka-vubuka ye ggu

OM_{1SG}-leave_alone-FV-ENCL₁₇ PERS_{2SG} AUG-12-youth INTERJ INTERJ

lwaki o-n-e-simb-a-mu e-n-geli e-yi-o

INTER SM_{2SG}-OM_{1SG}-RFL-stand-FV-ENCL₁₈ AUG-9-way AUG-PP₉-DEM_B

'**Leave me alone** you young boy; but really why do you stand against me in that way?'

[File ID: Abantub | O • Songs - Traditional • Rehabilitation • 2000s]

8. to finish

*Bamaama bwe banaaba **baviileku** twidha kuba n'ebigambo okuva eli babbaabba.*

ba-maama bwe ba-naa-b-a **ba-v-ile-ku**

2-mother MOD SM₂-FUT₁-be-FV **SM₂-finish-PERF-ENCL₁₇**

tu-idh-a ku-b-a ni e-bi-gambo oku-v-a

SM_{1PL}-come-FV 15-be-FV CON AUG-8-word 15-emanate-FV

e-li ba-bbaabba

SM₂₃-be 2-father

'After the mothers **have finished** (giving their speeches), we shall be with the words emanating from the fathers.'

[File ID: BuwaabGr | O • Celebrations • Inspirational • 2010]

9. to come a (little) bit

Nga bwe twabategeeziiza okuva eila nti munange tugya kuba n'abakungu okuvaaku mu Judicial Service Commission.

nga bwe tu-a-ba-tegeez-is-a oku-v-a eila nti
like as SM1PL-PST-OM2PL-inform-CAUS-FV 15-start-FV already that
mu-na-nge tu-gi-a ku-b-a ni a-ba-kungu
1-AFP-POSS1SG SM1PL-go-FV 15-be-FV CON AUG-2-specialist
oku-v-a-ku mu judicial_service_commission
15-come_a_bit-FV-ENCL17 LOC18 9.Judicial_Service_Commission
'As we already informed you, my friends, we shall be with specialists who
come a little bit from the Judicial Service Commission.'
[File ID: Judicial | O • Radio talk shows • Sensitization • 2010]

Combinations

Together with the noun omusolo 'tax', sense 4 acquires a specific use, as shown below.

-vaaku omusolo = to remit tax

Amakolelelo agavaaku omusolo.
a-ma-kol-ilil-o aga-v-a-ku o-mu-solo
AUG-6-do-RPT-DEV SREL6-let_aside-FV-ENCL17 AUG-3-tax
'Factories that remit tax.'
[File ID: AVATVAT | O • Songs - Traditional • Sensitization • 2000s]

3.4.5 The locativised verb -vaamu

When the class 18 locative enclitic -mu is suffixed to the base verb -v-1, numerous new uses that were not seen for the base verb are found (3. to 5. below), together with variations of the two main uses as also seen for the base verb (1. and 2. below).

okuvaamu < okuva1

- 1. to abandon though it is expected
2. to come out, to flow out, to exit
3. to grow well, to turn out well
4. to yield, to generate
5. to not gain

1. to abandon though it is expected

Yaasangulawo bile ebibaile bili kwogelwa nti akalulu akaviilemu [...]
a-a-sangul-a-wo bi-le ebi-b-a-ile bi-li
SM1-PST-rub-FV-ENCL16 PP8-DEMC SREL8-be-FV-PERF SM8-PROG

ku-ogel-w-a nti a-ka-lulu
15-speak-PASS-FV CON AUG-12-vote
a-ka-v-ile-mu

SM₁-OM₁₂-abandon_though_it_is_expected-PERF-ENCL₁₈

'He cancelled all the things that have been said and as to the election **he abandoned it though it was expected of him.**'

[File ID: AEGY3 | O • Radio talk shows • Health • 2010]

2. to come out, to flow out, to exit

*Owundi ku baisilukale yaamufumita eifumo mu lubavu ela mangu ago **mwavaamu** omusaayi n'amaadhi.*

o-wu-ndi-ku ba-isilukale a-a-mu-fumit-a e-i-fumo
AUG-PP₁-EXC-ENCL₁₇ 2-soldier SM₁-PST-OM₁-pierce-FV AUG-5-spear
mu lu-bavu ela mangu a-ga-o **mu-a-v-a-mu**
LOC₁₈ 11-rib CON quickly AUG-PP₆-DEM_B SM₁₈-PST-come-FV-ENCL₁₈
o-mu-saayi ni a-ma-adhi
AUG-3-blood CON AUG-6-water

'One of the soldiers pierced him with a spear in the rib and quickly (thereafter) blood and water **flew out.**'

[File ID: Missa1 | W • Biblical documents • Religion • 2012]

3. to grow well, to turn out well

*Bw'ozala n'abaawo **ekivaamu** agasa bwa iwanga.*

bwe o-zaal-a ni a-b-a-wo
if SM_{25G}-give_birth-FV CON SM₁-be-FV-ENCL₁₆
eki-v-a-mu a-gas-a bwa i-wanga
SREL₇-turn_out_well-FV-ENCL₁₈ SM₁-benefit-FV just 5-country

'If you give birth and **s/he turns out well**, s/he just benefits the nation.'

[File ID: Bwozaal | O • Songs - Traditional • Inspirational • 2000s]

4. to yield, to generate

*Ekiighulo kya Kyabazinga kyatundibwa aghalala ni Kodh'eyo **kyavaamu** emitwalo kumpi ikumi nga empiiya dhino dha kugheeleza Busoga mu bitundu ebili n'eby'etaago.*

e-ki-ighul-o ki-a kyabazinga ki-a-tund-ibw-a aghalala
AUG-7-food-DEV CP₇-of 1a.King SM₇-PST-sell-PASS-FV together
ni kodh'eyo **ki-a-v-a-mu** e-mi-twalo kumpi
CON 9.Kodh'eyo SM₇-PST-yield-FV-ENCL₁₈ AUG-4-ten_thousand near
ikumi nga e-n-piiya dhi-no dhi-a ku-gheel-is-a
ten CON AUG-10-money PP₁₀-DEM_A CP₁₀-of 15-serve-CAUS-FV
bu-soga mu bi-tundu ebi-li ni e-bi-etaag-o
14-soga LOC₁₈ 8-part SREL₈-be CON AUG-8-need-DEV

'(Tickets to attend) the 'dinner' (in honour) of the King were sold together with Kodh'eyo⁹ and **this yielded** roughly one hundred thousand Shillings which serves Busoga in the parts which have needs.'

[File ID: Kodh'eyo | W • Journalism • Networking • 1997–1998]

5. to not gain

Muzeeyi iwe oidha **kuvoilamu** awo.

mu-zeeyi iwe o-idh-a **ku-v-il-a-mu**
1-old PRON_{2SG} SM_{2SG}-will-FV **15-not_gain-APPL-FV-ENCL₁₈**
a-wa-o
AUG-PP₁₆-DEM_B

'Mzee you, you will **not gain** anything at all.'

[File ID: PFExtaud | O • Interviews • Language • 2012]

Combinations

Together with the noun *enda* 'stomach', sense 2 acquires a specific use, as shown below.

-vaamu enda = to miscarry

[...] mukazi wo lw' **avaamu enda**. Onaagya waalima?

mu-kazi a-o lwe **a-v-a-mu** **en-da**
1-wife PP₁-DEM_B OREL₁₁ **SM₁-come_out-FV-ENCL₁₈** **10-stomach**
o-naa-gi-a o-a-lim-a
SM_{2SG}-FUT₁-go-FV SM_{2SG}-PROG-dig-FV

'[...] (the day) when your wife **miscarries**. Will you go and dig?'

[File ID: Esaalmk1 | O • Radio talk shows • Religion • 2010]

Other word classes

One particular frequent construction has lexicalised and is used as a connective — namely the subject relative of cl. 7, with the past tense marker, and sense 2 of *-vaamu* — as shown below.

ekyavaamu (connective) = what came out of it, what resulted from it

Ekyavaamu, Wampala yaakoogha okutambula nga bwayagala; Wakayima, yaasalawo okugya okwekweka mu ndhu.

eki-a-v-a-mu wa-mpala a-a-koogh-a oku-tambul-a
SREL₇-PST-come_out-FV-ENCL₁₈ 16-lion SM₁-PST-tire-FV 15-walk-FV
nga bwe a-a-gal-a wa-kayima a-a-sal-a-wo
ADV ADV SM₁-PROG-search-FV 16-monkey SM₁-PST-decide-FV-ENCL₁₆
oku-gi-a oku-e-kwek-a mu n-dhu
15-go-FV 15-RFL-hide-FV LOC₁₈ 9-house

'**What resulted from it**, (is that) Mr. Lion got tired of walking as he searched; Mr. Monkey decided to go and hide himself in the house.'

[File ID: MwidTufm | W • Literature • Fables • 1999]

3.4.6 The locativised verb *-vaayo*

When the class 23 locative enclitic *-yo* is suffixed to the base verb *-v⁻¹*, either a variation of sense 5 of the base verb is seen, or a new one.

okuvaayo < *okuva*¹

1. to come out
2. to give way

1. to come out

Abawala muviileyo? *Mwanguyeeke mwanguyeeke.*

a-ba-wala mu-v-ile-yo

AUG-2-girl SM_{2PL}-come_out-PERF-ENCL₂₃

mu-angu-y-e-ku

mu-angu-y-e-ku

SM_{2PL}-hurry_up-CAUS-SUBJ-ENCL₁₇ SM_{2PL}-hurry_up-CAUS-SUBJ-ENCL₁₇

'Girls **have you come out**? Hurry up, hurry up.'

[File ID: IntHadij | O • Celebrations • Marriage • 2008]

2. to give way

Eee! Oooh! Vaayo baidhaakuniina.

eee oooh v-a-yo

ba-idh-a

Ø-ku-niin-a

INTERJ INTERJ give_way-FV-ENCL₂₃ SM₂-will-FV 15-OM_{2SG}-step-FV

'Eee! Oooh! **Give way**, they will step on you.'

[File ID: PFExtaud | O • Interviews • Language • 2012]

3.4.7 The cl. 14 deverbative noun *-vo*

While all previous derivations (§§3.4.3–3.4.6) were derived from *-v⁻¹*, one frequent derivation, the cl. 14 deverbative noun *-vo*, is derived from *-v⁻²*, sense 1, as shown below.

obuvo = *the beginning* < *okuva*²

Inhonhola obuvo n'obwiko bwe ensonga yaali kwogelaku.

inhonhol-a o-bu-v-o

ni o-bu-ik-o

bu-a

explain-FV AUG-14-start-DEV and AUG-14-end-DEV CP₁₄-of

e-n-songa ye a-li ku-yogel-a-ku

AUG-9-issue OREL₉ SM₁-be 15-speak-FV-ENCL₁₇

'Explain the **beginning** and the end of the issue that s/he is talking about.'

[File ID: Omugole | W • Literature - Plays • Marriage • 2007]

3.4.8 Summary of the corpus evidence for the Lusoga verb -v-

The corpus evidence as analysed and illustrated in §3.4.2 through §3.4.7 can now be synthesised as presented in Table 3. The three steps of Hanks's procedure may be recognised, but for a Bantu language the approach is not as linear as suggested in §3.4.1 for English. Part of Step 1, the division 'on broad syntactic grounds', is the outcome of the lemmatisation, which resulted in the distinction between verbal, locativised verbal and nominal uses (column 1 in Table 3). The other half, with connectives and an adverbial use, was only revealed during analysis (column 4 in Table 3). When it comes to Step 2, the division 'into more delicate semantic and syntactic frames' is what we termed *combinations* (column 3 in Table 3). In our case study, these may be combinations of verb + noun, verb + verb, verb + preposition, and verb + locative + noun. Those that include a preposition also turn into prepositional uses. Due to the structure of Bantu languages, some of these lemmas and combinations include codes for entire paradigms (here LOC = any locative, SM = any subject marker). Lastly, Step 3, 'making more subtle distinctions on semantic grounds', goes to the heart of the splitting vs. lumping decisions that every lexicographer must contend with (column 2 in Table 3).

Table 3: Synthesis of the verb -v- in the 1.7m Lusoga corpus, with columns 1 and 4 for Step 1, column 3 for Step 2, and column 2 for Step 3 of the procedure to map meaning onto use. (Manual effort between [1].)

Lemma signs derived from the lemmatised frequency list	Meaning potentials	Combinations, + meaning potentials following '='	Lemma signs for other word classes, + meaning potentials following '='
okuva ¹ [move senses]	1. to leave, to depart, to go away [1] 2. to hail (from) [2] 3. to abandon [3] 4. to make way, to move away [4] 5. to result, to come out 6. to spend (time)	-v- + LOC + <i>mais</i> = to die	<i>kya</i> -SM- <i>va</i> (connective) = that is why
okuva ² [projection and direction senses]	1. to start (and continue onwards) [1]	-v- + <i>-tuuk-</i> = from ... up to [measuring space] -v- + <i>paka</i> = from ... up to, from ... until [measuring time]	<i>kuva</i> (adverb) = since

	2. to be the source (of), to emanate (from)		
okuvaawo < okuva¹	1. to stop existing, to die (out) 2. to leave, to depart, to go away		
okuvaaku < okuva¹	1. to go off, to turn off [1] 2. to abandon 3. to trigger, to cause 4. to let aside, to give up [2] <i>-vaaku omusolo = to remit tax</i> 5. to lose 6. to stop 7. to not disturb, to leave alone 8. to finish 9. to come a (little) bit		
okuvaamu < okuva¹	1. to abandon though it is expected [1] 2. to come out, to flow out, to exit [5] <i>-vaamu enda = to miscarry</i> 3. to grow well, to turn out well [3] 4. to yield, to generate 5. to not gain [4]	<i>ekyavaamu</i> (connective) = what came out of it, what resulted from it	
okuvaayo < okuva¹	1. to come out 2. to give way		
obuvo < okuva²	the beginning		

3.5 Comparison of the manual effort vs. the corpus evidence for the Lusoga verb *-v-*

Any comparison between a manual effort and a corpus-driven one is always unfair, as the corpus tends to 'win'. In doing so, one often forgets about the heroic efforts that went into the manual effort in the first place (Nabirye 2008, 2009a, Nabirye and de Schryver 2010, 2011, 2013). The following, therefore, is only for illustrative purposes.

While the lemmatisation had already revealed that two of the four localised verbs had accidentally been overlooked, including a very frequent one,

as well as a deverbative noun (probably because it was assumed to belong to the grammar rather than the lexicon), all of the trickier derived word classes as well as the truly frequent combinations were also absent from the manual effort. (The one combination offered in the monolingual dictionary, viz. *okuva ku luguudo*, was not found in the 1.7m Lusoga corpus.) With regard to the various meaning potentials: while one notices a few overlaps, one especially notices a good number of additions and more fine-grained descriptions as a result of the corpus analysis. The order of the meaning potentials that do overlap is not always the same either (cf. [1] in Table 3).

What the manual effort does include, and what the corpus does not reveal in the same way, is the long list of 19 proverbs seen in Table 2. This is only partly the result of the fact that proverbs are known to be far less canonical in their use than dictionary-makers try to make you believe (Moon 1998). The proverb *Akaviile mu igi tikatya ikoli* 'The one that has just come from an egg does not fear an eagle' from the monolingual dictionary is for instance found in the corpus as *Akazaalibwa tikatya ikoli* 'The one which has just been born does not fear an eagle', hence without what one would assume to be a core term, 'egg'. Or, more Bantuish in nature, the monolingual-dictionary proverb *Omusaadha kikele kiva kyonka mu bwina* 'A man is a frog, which comes out of the hole by itself' is found in the corpus as *Omusaadha ikere: liva lyonka mu bwina* 'A man is a frog, which comes out of the hole by itself', which appears to be the same in translation, but in Lusoga the canonical form uses the noun in gender 7/8, while it is found in gender 5/6 in the corpus. Given this variation, proverbs have to be spotted mostly manually in a corpus. As to the reverse, a dedicated search does reveal proverbs not included into the otherwise pretty exhaustive manual list, such as *Awava omugulu waila mwigo* 'The stick takes the place of the leg that has left', *Awava omwosi wava omulilo* 'Fire comes from where smoke comes from', etc. Even so, their frequency of use is simply too low to merit inclusion when reasonable corpus frequencies and a nice spread across sources are used as an inclusion criterion.

3.6 Constructing corpus-driven microstructures for the Lusoga verb *-v-*

The data synthesised in Table 3 is the starting point for constructing the various dictionary articles that revolve around the verb *-v-* in Lusoga. In a desk or school dictionary, one may select from that data by taking, say, only the top *n* (frequent) lemmata and for these the top *n* (frequent) meaning potentials. At the other extreme, in a comprehensive dictionary, one will also want to exemplify all possible senses. To do so, reusing the examples that were studied during the analysis is an option, so the sentences and phrases from §§3.4.2–3.4.7 are prime candidates.¹⁰ In doing so, however, it is good to recall that 'giving equal prominence to all senses, when they are not equally common, is a distortion' (Hanks 2002: 157). So, the most frequent meaning potentials could be illustrated with multiple examples, while the lesser-frequent ones could do with

just one or even no examples. Likewise with the combinations: whether or not to include some or all of them will depend on the target. For an unabridged paper dictionary, however, or for a digital dictionary in which the information is layered and where it may be 'peeled off' (Geeraerts 2000: 78-79), one can as well prepare and *optionally* present as much as possible. Adding the sources of the various examples also becomes a worthwhile addition at that point, as is the tradition in dictionaries based on historical principles. While such information on each source could be synthesised in the dictionary itself, a link to the full information, as seen in Addendum 1 of Part 1, could furthermore easily be added. In a digital dictionary actual hyperlinks to the corpus material itself could even be envisaged, thereby handing dictionary users the 'raw data' on which the lexicographers based their decisions, and/or allowing such users to explore the (corpus) data further (cf. de Schryver 2003: 167, 169, i.e. 'Dream # 31'). In short, a maximally populated dictionary writing system is best viewed as a *single* database from which *any number* of dictionaries may be generated, a concept that has been termed 'one database, many dictionaries' (de Schryver and Joffe 2005).

4. Discussion

In this article we have made a strong case for the analysis of corpora to discover word meanings. After two decades of querying corpora for Bantu lexicography in general, and about one decade of corpus-building for Lusoga in particular, we are pretty much convinced that a careful study of the natural production of language that was produced by a multitude of speakers and writers indeed offers the best perspective on how language is truly used, from which meanings may be mapped (as explained and illustrated in the present article), and with which detailed studies of language may be undertaken. Some colleagues remain sceptical however, as voiced by Michael Marlo two years ago:

A criticism that can be levelled at corpus-based approaches is that because they lump together data by individual speakers, it is extremely difficult if not impossible in a corpus-based approach to make sense of variation across individuals which is the result of the speakers having different internal grammars. The present approach seems to reject the idea that grammar is in the heads of individual speakers. It focuses on 'e-language' vs. 'i-language'. That is fine, but the approach has some limitations — such as the ability to state with precision what is a 'language'. (Marlo 2016, personal communication)

By using a corpus in the way we do, one ends up compromising, and indeed focusing on many e-languages (with *e* for 'external/externalised'), rather than on a single or a limited number of i-languages (with *i* for 'internal/internalised'). That said, even though the corpus analyst likes lots and lots of data and ditto examples, it is also true that: "Overwhelming evidence", be it noted, may

consist of no more than a handful of textually well-formed and convincing modern uses' (Hanks 2002: 174). Michael Marlo goes on to suggest:

Moreover, most linguists consider negative evidence to be essential for understanding the rules of language — not just what is common vs. uncommon but determining what is possible vs. impossible. There is considerable discussion of this within the generativist community under the notion of 'poverty of the stimulus' — the idea that speakers of a language know much about the language, even if they have never heard the expressions in question before. (Marlo 2016, personal communication)

In our strand of corpus linguistics, the focus is on the norms, not the exploitations, and the focus is consequently also not on what does not occur or on what occurs infrequently. Of course, this is a choice, but for a language like Lusoga which needs 'first descriptions', focusing on the speech community and their general needs first, and attempting to bring back their own words to them, in this case in the form of corpus-driven dictionary-making, seems like a worthwhile venture.

With this, we have come to the end of our three-part study of corpus-driven Bantu lexicography as applied to Lusoga. To conclude, it is now fitting to point out that our effort is not the first trilogy of articles on the application of corpora in modern dictionary-making. As a matter of fact, Michael Rundell and Penny Stock initiated this trend a quarter of a century ago, with a three-part report on what was then called 'The corpus revolution' (as applied to English lexicography). Compared to our effort, the sequence of their articles is organised differently, however. In their first part, Rundell and Stock (1992a) looked at the relative merits of large-scale text corpora compared to traditional citation banks. In the light of Hanks's theoretical framework of **mapping meaning onto use**, their most important observation in favour of the use of computerised corpora over manual reading and marking is that:

It is astonishingly difficult for even the most experienced person to collect material for ordinary everyday usages since human beings tend to notice the unusual. [...] When using corpus evidence, therefore, the lexicographer works with whatever comes up in the corpus rather than with individually or specially selected examples. (Rundell and Stock 1992a: 13, 10)

The other advantages they list in favour of a corpus remain valid to this day, and have also all been illustrated for Lusoga lexicography: (i) 'it can provide evidence for the comparative frequency of word occurrence and behaviour', (ii) 'It can be of immense help in enabling the lexicographer to give examples to show the word in its most typically or frequently used contexts', (iii) 'It allows the lexicographer to structure an entry in such a way as to reflect how a word is normally used', and (iv) 'It can enable the dictionary maker to give an accurate account of grammatical behaviour at the level of individual senses' (Rundell and Stock 1992a: 14).

In their second part, Rundell and Stock (1992b) looked at the ways in which corpus evidence informs the actual writing of dictionary articles. With de Schryver and Joffe's practical concept of **one database, many dictionaries** in mind, the following observations on what to put in a certain dictionary ring true:

In fact the task of omitting or not including known meanings which are nonetheless inappropriate to a particular dictionary is a very hard one. It is so much easier to play safe and let such meanings in [...] Again the evidence of many millions of examples of usage can be of enormous assistance in strengthening the lexicographer's nerve in such cases [...] (Rundell and Stock 1992b: 25)

On a more generic level, their closing statement has proven to be as valid for Bantu as it is for English:

It is perhaps fairly rare to find all one's preconceptions about a word being overturned on consulting a corpus, but it is equally rare to come away from analysing a given word or use without having learned a great deal that is new, illuminating, and sometimes unnerving. (Rundell and Stock 1992b: 28-29)

In their third part, Rundell and Stock (1992c) mainly deal with corpus building, and try to predict some of the automated tools and procedures that will be developed. These are, using the terms that have come to be adopted since Rundell and Stock's predictions from the early 1990s: (i) lemmatisers, (ii) sampling techniques, (iii) POS-taggers, (iv) parsers, and (v) word-sense disambiguators. Over the past 25 years these have indeed all been created for the world's major languages. More in particular, in Part 2 of our series we have indicated how the lemmatisation and POS-tagging for lexicographic purposes may be achieved for the Bantu languages. Unlike for English, these macrostructural aspects are hugely complex for the Bantu languages, which led Prinsloo and de Schryver to develop instruments known as **part-of-speech rulers and alphabetical (or multidimensional lexicographic) rulers** in order to measure, evaluate, predict and manage Bantu-language dictionary projects. We therefore trust that thanks to corpora, and just as is the case for English, we are now indeed 'emancipated from the role of harmless drudge and empowered to make new insights into every area of language' (Rundell and Stock 1992c: 51).

Abbreviations

#	noun class number	AP _x	adjectival prefix (of cl. x)
ADV	adverb	AUG	augment
AFP	affiliation prefix	CAUS	causative
APERF	after-perfect	cl.	class
APPL	applicative	CON	connective

COP	copulative	OM _x	object marker (of cl. or ps. x)
CP _x	connective prefix (of cl. x)	OREL _x	object relative (of cl. x)
DEM _{A,B,C}	demonstrative (of position A, B, C)	PASS	passive
DEV	deverbative	PERF	perfect
ENCL _x	locative enclitic (x = cl. 16, 17, 18, 23)	PERS _x	personal pronoun (of ps. x)
EXC	exclusive pronoun	PL	plural
ext.	extension	POSS _x	possessive (of ps. x)
FUT ₁	future tense <i>-naa-</i>	PP _x	pronominal pronoun (of cl. x)
FUT ₂	future tense <i>-li-</i>	PROG	progressive
FV	final vowel	PRON _x	pronoun (of ps. x)
HAB	habitual	ps.	person
INC	inclusive pronoun	PST	past tense
INTENS	intensifier	REC	reciprocal
INTER	interrogative	RFL	reflexive
INTERJ	interjection	RPT	repetitive
loc.	locativised	SG	singular
LOC _x	locative (x = cl. 16, 17, 18, 23)	SM _x	subject marker (of cl. or ps. x)
MOD	modality	SREL _x	subject relative (of cl. x)
MODF	modifier	SUBJ	subjunctive
NEG _{A,B}	negative (of type A, B)	UNS	unspecified noun

Acknowledgements

The research for this article was funded by the Special Research Fund of Ghent University. Thanks are due to the two anonymous referees.

Endnotes

1. Parts of this theoretical discussion are based on sections from Nabirye (2016).
2. The reference to 'real' language is taken from the first-ever corpus-based dictionary, the *Collins COBUILD English Language Dictionary* (Sinclair 1987a), which was advertised as such.
3. In corpus-linguistic circles, Sinclair may be best known as the founder of the COBUILD (i.e., the Collins Birmingham University International Language Database) project in lexical computing (Sinclair 1987b), and the chief editor of the *Collins COBUILD English Language Dictionary* (Sinclair 1987a). The managing editor of the latter dictionary was Patrick Hanks.
4. Conversely, typologists may be interested in simply knowing what a language is capable of, and want answers to questions like: 'What is the longest possible verb form in this or that Bantu language?'
5. Also present in the TLex database, but not frequent enough to have been lemmatised, are *-v̄w-*, a spoken variant of *-v-*, and *-evaamu* 'dare; be brave', the reflexive form of *-vaamu*.
6. 2 Timothy 2:21

7. The locativised verb *-vaawo* also has a variant, namely *-vaagho*, but its frequency is too low to have made it into the lemmatised frequency list.
8. ARVs = antiretrovirals (i.e., drugs to treat HIV)
9. *Kodh'eyo* was a short-lived newspaper (1997–1998) written in Lusoga.
10. For the value of corpus examples over 'invented' (but more didactic) ones see Fox (1987).

References

- Atkins, B.T.S. 1996. Bilingual Dictionaries: Past, Present and Future. Gellerstam, M., J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström and C.R. Pappmehl (Eds). 1996. *Euralex '96 Proceedings I-II, Papers Submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden*: 515-546. Gothenburg: Department of Swedish, Göteborg University.
- Bosch, S.E. and D.J. Prinsloo. 2002. 'Abbreviated Nouns' in African Languages: A Morphological, Semantic and Lexicographic Perspective. *South African Journal of African Languages* 22(1): 92-104.
- Bostoën, K. and G.-M. de Schryver. 2015. Linguistic Innovation, Political Centralization and Economic Integration in the Kongo Kingdom: Reconstructing the Spread of Prefix Reduction. *Diachronica* 32(2): 139-185 + 13 pages of supplementary material online.
- Bostoën, K., F. Mberamihigo and G.-M. de Schryver. 2012. Grammaticalization and Subjectification in the Semantic Domain of Possibility in Kirundi (Bantu, JD62). *Africana Linguistica* 18: 5-40.
- Chabata, E. and D. Nkomo. 2010. The Utilisation of Outer Texts in the Practical Lexicography of African Languages. *Lexikos* 20: 73-91.
- De Kind, J. and K. Bostoën. 2012. The Applicative in ciLubà Grammar and Discourse: A Semantic Goal Analysis. *Southern African Linguistics and Applied Language Studies* 30(1): 101-124.
- De Kind, J., M. Devos, G.-M. de Schryver and K. Bostoën. 2013. Negation Markers, Focus Markers and Jespersen Cycles in Kikongo (Bantu, H16): A Comparative and Diachronic Corpus-based Approach. Available online at: https://www2.hu-berlin.de/predicate_focus_africa/data/2013-12-10_deKind_Negation.in.Kikongo.pdf.
- De Kind, J., S. Dom, G.-M. de Schryver and K. Bostoën. 2015. Event-centrality and the Pragmatics-Semantics Interface in Kikongo: From Predication Focus to Progressive Aspect and Vice Versa. *Folia Linguistica Historica* 36: 113-163.
- de Schryver, G.-M. 1999. *Cilubà Phonetics, Proposals for a 'Corpus-based Phonetics from Below'-Approach* (Recall Linguistics Series 14). Ghent: Recall.
- de Schryver, G.-M. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies* 11(2): 266-282.
- de Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2): 143-199.
- de Schryver, G.-M. 2005. Book Review: M.-H. Corréard, ed. 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. *Lexicographica: International Annual for Lexicography* 21: 420-425.
- de Schryver, G.-M. 2006. Compiling Modern Bilingual Dictionaries for Bantu Languages: Case Studies for Northern Sotho and Zulu. Corino, E., C. Marelllo and C. Onesti (Eds). 2006. *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 / Proceedings XII Euralex International Congress, Torino, Italia, September 6th-9th, 2006*: 515-525. Alessandria: Edizioni dell'Orso.

- de Schryver, G.-M.** 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English / Pukuntšuu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford*. Cape Town: Oxford University Press Southern Africa.
- de Schryver, G.-M.** 2008. Why does Africa need Sinclair? *International Journal of Lexicography* 21(3): 267-291.
- de Schryver, G.-M. and R. Gauton.** 2002. The Zulu Locative Prefix ku- Revisited: A Corpus-based Approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201-220.
- de Schryver, G.-M. and D. Joffe.** 2005. One Database, Many Dictionaries — Varying Co(n)text with the Dictionary Application TshwaneLex. Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan and Y.Y. Tan (Eds). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1–3 June 2005, M Hotel, Singapore*: 54-59. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.
- de Schryver, G.-M. and B. Lepota.** 2001. The Lexicographic Treatment of Days in Sepedi, or When Mother-Tongue Intuition Fails. *Lexikos* 11: 1-37.
- de Schryver, G.-M. and M. Nabirye.** 2010. A Quantitative Analysis of the Morphology, Morphophonology and Semantic Import of the Lusoga Noun. *Africana Linguistica* 16: 97-153.
- de Schryver, G.-M. and D.J. Prinsloo.** 2000. 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.** 2001. Towards a Sound Lemmatisation Strategy for the Bantu Verb through the Use of Frequency-based Tail Slots — with Special Reference to Cilubà, Sepedi and Kiswahili. Mdee, J.S. and H.J.M. Mwansoko (Eds). 2001. *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: 216-242, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.
- de Schryver, G.-M. and E. Taljard.** 2006. Locative Trigrams in Northern Sotho, Preceded by Analyses of Formative Bigrams. *Linguistics, An Interdisciplinary Journal of the Language Sciences* 44(1): 135-193.
- de Schryver, G.-M. and E. Taljard.** 2007. Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho. *Lexikos* 17: 37-55.
- de Schryver, G.-M., E. Taljard, M.P. Mogodi and S. Maepa.** 2004. The Lexicographic Treatment of the Demonstrative Copulative in Sesotho sa Leboa — An Exercise in Multiple Cross-referencing. *Lexikos* 14: 35-66.
- Devos, M. and G.-M. de Schryver.** 2013. From 'habitually going' to 'maybe': Grammaticalization and Lexicalization of an Epistemic Sentence Adverb in Swahili. *Abstracts of The 21st International Conference on Historical Linguistics*: 29. Oslo: University of Oslo.
- Devos, M. and G.-M. de Schryver.** 2016. From Usually Going to Epistemic Possibility. Origin and Development of an Epistemic Sentence Adverb in Swahili. *6th International Conference on Bantu Languages, Workshop on the Expression of Mood and Modality in Bantu Languages*: 11. Helsinki: University of Helsinki.
- Devos, M., M.-J. Misago and K. Bosto.** 2017. A Corpus-based Description of Locative and Non-locative Reference in Kirundi Locative Enclitics. *Africana Linguistica* 23: 47-83.
- Dom, S., G. Segerer and K. Bosto.** 2015. Antipassive/Associative Polysemy in Cilubà (Bantu, L31a): A Plurality of Relations Analysis. *Studies in Language* 39(2): 354-385.

- Faaf, G.** 2010. *A Morphosyntactic Description of Northern Sotho as a Basis for an Automated Translation from Northern Sotho into English*. Unpublished Ph.D. dissertation. Pretoria: University of Pretoria.
- Fox, G.** 1987. The Case for Examples. Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*: 137-149. London: Collins ELT.
- Gauton, R., G.-M. de Schryver and L. Mohlala.** 2004. A Corpus-based Investigation of the Zulu Nominal Suffix -kazi: A Preliminary Study. Akinlabi, A. and O. Adesola (Eds). 2004. *Proceedings of the 4th World Congress of African Linguistics, New Brunswick 2003*: 373-380. Cologne: Rüdiger Köppe Verlag.
- Geeraerts, D.** 2000. Adding Electronic Value. The Electronic Version of the *Grote Van Dale*. 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*: 75-84. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Geeraerts, D.** 2009. Currents and Undercurrents in Lexical Semantics, Twenty Years After. Beijck, E., L. Colman, M. Göbel, F. Heyvaert, T. Schoonheim, R. Tempelaars and V. Waszink (Eds). 2009. *Fons Verborum. Feestbundel voor prof. dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie*: 421-430. Amsterdam: Gopher BV.
- Geeraerts, D.** 2010. *Theories of Lexical Semantics*. New York: Oxford University Press.
- Gouws, R.H. and D.J. Prinsloo.** 1997. Lemmatisation of Adjectives in Sepedi. *Lexikos* 7: 45-57.
- Gouws, R.H. and D.J. Prinsloo.** 2005. Left-expanded Article Structures in Bantu with Special Reference to isiZulu and Sepedi. *International Journal of Lexicography* 18(1): 25-46.
- Hanks, P.** 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*: 156-198. s.l.: Euralex.
- Hanks, P.** 2004. Corpus Pattern Analysis. Williams, G. and S. Vessier (Eds). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 87-97. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.
- Hanks, P.** 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hanks, P.** 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: The MIT Press.
- Hanks, P.** 2015. Cognitive Semantics and the Lexicon. *International Journal of Lexicography* 28(1): 86-106.
- Joffe, D. and G.-M. de Schryver.** 2002–18. TLex Suite — Dictionary Compilation Software. Available online at: <http://tshwanedje.com/tshwanelex/>.
- Kawalya, D.** 2017. *A Corpus-driven Study of the Expression of Modality in Luganda (Bantu, JE15)*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Kawalya, D., K. Bostoën and G.-M. de Schryver.** 2014. Diachronic Semantics of the Modal Verb -sóból- in Luganda: A Corpus-driven Approach. *International Journal of Corpus Linguistics* 19(1): 60-93.
- Kawalya, D., G.-M. de Schryver and K. Bostoën.** 2018. From Conditionality to Modality in Luganda (Bantu, JE15): A Synchronic and Diachronic Corpus Analysis of the Verbal Prefix -andi-. *Journal of Pragmatics* 127: 84-106.
- Kilgarriff, A.** 2003–18. Sketch Engine. Available online at: <https://www.sketchengine.co.uk>.

- Klein, J.** 2010a. Can the New African Language Dictionaries Empower the African Language Speakers of South Africa or Are They Just a Half-hearted Implementation of Language Policies? Dykstra, A. and T. Schoonheim (Eds). 2010. *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010)*: 1485-1496. Leeuwarden/Ljouwert: Fryske Akademy.
- Klein, J.** 2010b. Nord Sotho Wörterbücher als Implementierungsstrategien der *South African Languages Bill* und der *National Lexicographic Units Bill*. Buchmann, L., L. Fuhrmann, N. Nassenstein, C. Vogel, M. Weinle and A. Wolvers (Eds). 2010. *Beiträge zur 3. Kölner Afrikawissenschaftlichen Nachwuchstagung (KANT III)*: 1-11. Cologne: University of Cologne.
- Kosch, I.** 2013. An Analysis of the *Oxford Bilingual School Dictionary: Northern Sotho and English* (De Schryver 2007). *Lexikos* 23: 611-627.
- Lafkioui, M., E. Nshemezimana and K. Bostoën.** 2016. Cleft Constructions and Focus in Kirundi. *Africana Linguistica* 22: 71-106.
- Madiba, M. and D. Nkomo.** 2010. The *Tshivenda-English Tahalusamaipfi/Dictionary* as a Product of South African Lexicographic Processes. *Lexikos* 20: 307-325.
- Mberamihigo, F.** 2014. *L'expression de la modalité en kirundi : Exploitation d'un corpus électronique*. Unpublished Ph.D. dissertation. Brussels; Ghent: Université libre de Bruxelles; Ghent University.
- Mberamihigo, F., G.-M. de Schryver and K. Bostoën.** 2016. Entre verbe et adverbe : Grammaticalisation et dégrammaticalisation du marqueur épistémique *umeengo/umeenga* en kirundi (bantou, JD62). *Journal of African Languages and Linguistics* 37(2): 247-286.
- Misago, M.-J.** 2018. *Les verbes de mouvement et l'expression du lieu en kirundi (bantou, JD62) : Une étude linguistique basée sur un corpus*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Moon, R.** 1998. *Fixed Expressions and Idioms in English. A Corpus-based Approach*. Oxford: Oxford University Press.
- Nabirye, M.** 2008. *Compilation of the Monolingual Lusoga Dictionary*. Unpublished M.A. dissertation. Kampala: Makerere University.
- Nabirye, M.** 2009a. Compiling the First Monolingual Lusoga Dictionary. *Lexikos* 19: 177-196.
- Nabirye, M.** 2009b. *Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwegwa Olusoga [A Dictionary of Lusoga. For speakers of Lusoga, and for those who would like to learn Lusoga]*. Kampala: Menha Publishers.
- Nabirye, M.** 2016. *A Corpus-based Grammar of Lusoga*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Nabirye, M. and G.-M. de Schryver.** 2010. The Monolingual Lusoga Dictionary Faced with Demands from a New User Category. *Lexikos* 20: 326-350.
- Nabirye, M. and G.-M. de Schryver.** 2011. From Corpus to Dictionary: A Hybrid Prescriptive, Descriptive and Proscriptive Undertaking. *Lexikos* 21: 120-143.
- Nabirye, M. and G.-M. de Schryver.** 2013. Digitizing the Monolingual Lusoga Dictionary: Challenges and Prospects. *Lexikos* 23: 297-322.
- Nong, S., G.-M. de Schryver and D.J. Prinsloo.** 2002. Loan Words versus Indigenous Words in Northern Sotho — A Lexicographic Perspective. *Lexikos* 12: 1-20.
- Nshemezimana, E.** 2016. *Morphosyntaxe et structure informationnelle en kirundi : Focus et stratégies de focalisation*. Unpublished Ph.D. dissertation. Ghent: Ghent University.
- Nshemezimana, E. and K. Bostoën.** 2016. The Conjoint/Disjoint Alternation in Kirundi (JD62): A Case for its Abolition. van der Wal, J. and L.M. Hyman (Eds). 2016. *The Conjoint/Disjoint*

- Alternation in Bantu* (Trends in Linguistics. Studies and Monographs 301): 390-425. Berlin: Mouton de Gruyter.
- Prinsloo, D.J.** 1992. Lemmatization of Reflexives in Northern Sotho. *Lexikos* 2: 178-191.
- Prinsloo, D.J.** 1994. Lemmatization of Verbs in Northern Sotho. *South African Journal of African Languages* 14(2): 93-102.
- Prinsloo, D.J.** 2002. The Lemmatization of Copulatives in Northern Sotho. *Lexikos* 12: 21-43.
- Prinsloo, D.J.** 2003. The Lemmatization of Adverbs in Northern Sotho. *Lexikos* 13: 21-37.
- Prinsloo, D.J.** 2005. Electronic Dictionaries Viewed from South Africa. *Hermes, Journal of Linguistics* 34: 11-35.
- Prinsloo, D.J.** 2006. Compiling a Bidirectional Dictionary Bridging English and the Sotho Languages: A Viability Study. *Lexikos* 16: 193-204.
- Prinsloo, D.J.** 2009. Current Lexicography Practice in Bantu with Specific Reference to the *Oxford Northern Sotho School Dictionary*. *International Journal of Lexicography* 22(2): 151-178.
- Prinsloo, D.J.** 2012. Die leksikografiese bewerking van verwantskapsterme in Sepedi. *Lexikos* 22: 272-289.
- Prinsloo, D.J.** 2014a. A Critical Evaluation of the Paradigm Approach in Sepedi Lemmatization — The *Groot Noord-Sotho Woordeboek* as a Case in Point. *Lexikos* 24: 251-271.
- Prinsloo, D.J.** 2014b. Lexicographic Treatment of Kinship Terms in an English / Sepedi–Setswana–Sesotho Dictionary with an Amalgamated Lemmatist. *Lexikos* 24: 272-290.
- Prinsloo, D.J. and S.E. Bosch.** 2012. Kinship Terminology in English–Zulu / Northern Sotho Dictionaries — A Challenge for the Bantu Lexicographer. Fjeld, R.V. and J.M. 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., T.J.D. Bothma and U. Heid.** 2014. User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages. Abel, A., C. Vettori and N. Ralli (Eds). 2014. *Proceedings of the XVI EURALEX International Congress: The User in Focus, 15–19 July 2014, Bolzano/Bozen*: 819-827. Bolzano/Bozen: EURAC Research.
- Prinsloo, D.J., T.J.D. Bothma, U. Heid and D.J. Prinsloo.** 2017. Direct User Guidance in e-Dictionaries for Text Production and Text Reception — The Verbal Relative in Sepedi as a Case Study. *Lexikos* 27: 403-426.
- Prinsloo, D.J. and G.-M. de Schryver.** 1999. The Lemmatization of Nouns in African Languages with Special Reference to Sepedi and Cilubà. *South African Journal of African Languages* 19(4): 258-275.
- Prinsloo, D.J. and G.-M. de Schryver.** 2001. Taking Dictionaries for Bantu Languages into the New Millennium — with Special Reference to Kiswahili, Sepedi and isiZulu. Mdee, J.S. and H.J.M. Mwanasoko (Eds). 2001. *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: 188-215. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.
- Prinsloo, D.J. and G.-M. de Schryver.** 2002. Reversing an African-language Lexicon: The *Northern Sotho Terminology and Orthography No. 4* as a Case in Point. *South African Journal of African Languages* 22(2): 161-185.
- Prinsloo, D.J. and R.H. Gouws.** 1996. Formulating a New Dictionary Convention for the Lemmatization of Verbs in Northern Sotho. *South African Journal of African Languages* 16(3): 100-107.
- Prinsloo, D.J. and R.H. Gouws.** 2006. Lexicographic Presentation of Grammatical Divergence in Sesotho sa Leboa. *South African Journal of African Languages* 26(4): 184-197.

- Prinsloo, D.J., U. Heid, T.J.D. Bothma and G. Faaß.** 2012. Devices for Information Presentation in Electronic Dictionaries. *Lexikos* 22: 290-320.
- Rundell, M. and P. Stock.** 1992a. The Corpus Revolution 1. The first in a series of three reports on the development and use of electronic language corpora and their impact on dictionaries. *English Today, The International Review of the English Language* 8(2): 9-14.
- Rundell, M. and P. Stock.** 1992b. The Corpus Revolution 2. A consideration of the practical benefits to English-language lexicographers of the evidence derived from computer corpora (second article of three). *English Today, The International Review of the English Language* 8(3): 21-29.
- Rundell, M. and P. Stock.** 1992c. The Corpus Revolution 3. A consideration of the prospects and potential of corpus-and-concordance lexicography (third article of three). *English Today, The International Review of the English Language* 8(4): 45-51.
- Scannell, K.P.** 2003–18. An Crúbadán — Corpus Building for Minority Languages. Available online at: <http://crubadan.org/>.
- Scott, M.** 1996–2018. WordSmith Tools. Available online at: <http://www.lexically.net/wordsmith/>.
- Sewangi, S.S.** 2000. Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-specific Terms from Corpus. *Nordic Journal of African Studies* 9(2): 60-84.
- Sewangi, S.S.** 2001. *Computer-assisted Extraction of Terms in Specific Domains: The Case of Swahili*. Unpublished PhD dissertation. Helsinki: University of Helsinki.
- Sinclair, J.M.** 1966. Beginning the Study of Lexis. Bazell, C.E., J.C. Catford, M.A.K. Halliday and R.H. Robins (Eds). 1966. *In Memory of J.R. Firth*: 410-430. London: Longmans.
- Sinclair, J.M.** 1987a. *Collins COBUILD English Language Dictionary*. London: William Collins Sons & Co.
- Sinclair, J.M. (Ed.)**. 1987b. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins ELT.
- Taljard, E.** 2006. Corpus-based Linguistic Investigation for the South African Bantu Languages: A Northern Sotho Case Study. *South African Journal of African Languages* 26(4): 165-183.
- Taljard, E.** 2012. Corpus-based Language Teaching: An African Language Perspective. *Southern African Linguistics and Applied Language Studies* 30(3): 377-393.
- 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.
- Taljard, E. and G.-M. de Schryver.** 2016. A Corpus-driven Account of the Noun Classes and Genders in Northern Sotho. *Southern African Linguistics and Applied Language Studies* 34(2): 169-185.
- Tognini-Bonelli, E.** 2001. *Corpus Linguistics at Work* (Studies in Corpus Linguistics 6). Amsterdam: John Benjamins.
- Toscano, M. and S.S. Sewangi.** 2005. Discovering Usage Patterns for the Swahili *amba-* Relative Forms cl. 16, 17, 18: Using Corpus Data to Support Autonomous Learning of Kiswahili by Italian Speakers. *Nordic Journal of African Studies* 14(3): 274-317.
- Tummers, J., K. Heylen and D. Geeraerts.** 2005. Usage-based Approaches in Cognitive Linguistics: A Technical State of the Art. *Corpus Linguistics and Linguistic Theory* 1(2): 225-261.

Addendum 1: *Eiwanika ly'Olusoga* (Nabirye 2009b), page 379

(o)ku.v.a

(o)ku.v.a¹ [(o) kúvá] *kt.* [L] [-viile] [nviile] **bl:** [Lg: okuva] 1) Okusimbuka mu kifo ekilala waayolekela ekindi. *gez:* *Nva Mayuge.* 2) Okusibuka. *gez:* *Nva Iganga.* 3) Okulekelela ekintu ky'obaile okola. *gez:* *Ebyo nabiviileku.* 4) Okuseguka mu ngila oba mu kifo. *gez:* *Leka nkaviile ofine eidembe.* 'sk: *Okuva ku luguudo: Okwonooneka / Okuva ku mulembe.*' ggl: Awava akwita n'awava akukobela; Awava ennume waila nnume. Awava mwino tiwaila mwino: awava eliino waila ilibu; Awava mwino tiwaila mwino: Awava eliiso waila itulu; Akaviile mu igi tikatya ikoli; Edhiva okulala n'embilo; Empambo eva ku kiwalo; Ennhonhi eva ewala temala mutonto; Ensanafo eva ku mugendelo telwa kufuuka kabasa; Atava ku mulungi afa t'awoza; Ka nduviile ku ntobo oti n'omuyala atuuse we bafumba; Olusubi olulala we luva ku ndhu tetoonha; Omukazi omulungi nimumo ya ngila buli avayoy agyegwaniza; Omukwano guva mu ngabo; Omukwano guva mu ngila gwatuuka eka; Omusaadha kikele kiva kyonka mu bwina; Va we ndi takulwanya; Va ku ntebe ya lata awulilila ku ise; W'ova tosoile w'otela okwila.

bbgz: *Okuviila, Okuviisa.*

(o)ku.v.a² [(o) kúvá] *kt.* [L] [-viile] [nviile] **bl:** [Lg: okuva] Okutandiikila mu kifo ekilala okutuuka ku kifo ekindi. *gez:* *Ennhandha ya Nalubaale eva Indhindha.*

(o)ku.vaabiil.a [(o) kúváábíílá] *kt.* [L] [-vaabiile] [nvaabiile] **bl:** [Lg: okuvaabiila] 1) Okuvuukila ebyokulya. 2) Okuliisa amailu.

bbgz: *Okuvaabiilila, Okuvaabiiza.*

(o)ku.vaal.a [(o) kúvaalá] *kt.* [L] [-vaile] [nvaile] **bl:** [Lsw: vaa, Lg: okuvaala] 1) Okubisa omutwe, emikono, amagulu oba ebigele mu kintu ng'olugoye oba engaito. *gez:* *Abantu bavaala engoye baleke kubita madu.*

bbgz: *Okuvaalibwa, Okuvaalika, Okuvaalila, Okuvaaza.*

(o)ku.vool.a

(o)ku.v.a.a.ku [(o) kúvááku] *kt.* [T] [-viileku] [nviileku] **bl:** [Lg: okuvaako] 1) Okuzima. *gez:* *Amasaanhalaze gaviileku.* 2) Okuleka. *gez:* *Omwenge nguviileku. / Oyo namuvaaku naafina owundi.*

bbgz: *Okuviilwaku, Okuviisaaku.*

(o)ku.v.a.a.mu [(o) kúváámu] *kt.* [T] [-vilemu] [nviilemu] **bl:** [Lg: okuvaamu] 1) Obutatuukiliza kye wasuubiza omuntu. *gez:* *Tubaile tusuubiila nti agya kutuyamba aye atuvilemu.* 2) Okulyamu olukwe. *gez:* *Gwetwateesa naye mwene neeyatuvaamu.* 3) Okuwa oba okumaliliza. *gez:* *Bw'obifumba bivaamu bulungi.* 4) Okubula ky'ofuna oba ky'ogobolola mu kintu. *gez:* *Nze nviililemu awo.* 5) Okusagala. *gez:* *Eno engaito evaamu.*

bbgz: *Okuviilamu, Okuviisamu.*

(o)ku.vangan.a [(o) kúvangáná] *kt.* [T] [-vangaine] [nvangaine] 1) Okwagaanana okw'abantu abava mu bifo eby'endhawulo baawaya. 2) Okusonhiwagana mwailagana nga mubaile muyombye mwayawukana. 3) Okugaitagaita.

bbgz: *Okuvangania, Okuvanganila.*

(o)ku.vook.a [(o) kúvooká] *tbk:* Okúvootóká *kt.* [L] [-voose] Omubuli gw'ekintu okwenhiga gwaila munda oba gwefunha. *gez:* *Enthupa evoose.*

bbgz: *Okuvookela, Okuvoosa.*

(o)ku.vool.a¹ [(o) kúvóólá] *kt.* [L] [-voile] [nvoile] Okunhigiliza omubili gw'ekintu gwaingila munda ave nga tiki-menheike. *gez:* *Oniinhie ku mukebe waaguvoola.*

bbgz: *Okuvoolabwa, Okuvooloka, Okuvooza, Okwevoola.*



(o)ku.vool.a² [(o) kúvóólá] *kt.* **bl:** [Lg: okuvvoola] (*stt*) Okunhooma oba okugaya omuwendo gw'ekintu. *gez:* *Ekigambo ekyo kivoola embeela y'omuntu.*

Addendum 2: Top orthographic corpus types in the 1.7m Lusoga corpus underlying the verbal lemmas *-v-*, *-vaawo* (cl. 16), *-vaaku* (cl. 17), *-vaamu* (cl. 18) and *-vaayo* (cl. 23) as well as the nominal *-vo* in cl. 14

Types in the 1.7m Lusoga corpus considered for the lemma *-v-*, verb:

abaava [-va] *verb, transitive cl. 2* 1 came from; 2 diverted / *freq.70* rank 2606 # texts 22;
abava [-va] *verb, transitive cl. 2* come from / *freq.81* rank 2289 # texts 32;
abaviile [-va] *verb, transitive cl. 2* 1 have come from; 2 have left / *freq.13* rank 10420 # texts 6;
agava [-va] *verb, transitive cl. 6* come from / *freq.43* rank 3953 # texts 21;
aghava [-va] *verb, transitive cl. 16* 1 comes from; 2 leaves / *freq.22* rank 6788 # texts 13;
ava [-va] *verb, transitive cl. 1* from / *freq.389* rank 509 # texts 107;
ave [-va] *verb, intransitive cl. 1* 1 comes from; 2 leaves / *freq.37* rank 4478 # texts 17;
aviile [-va] *verb, transitive cl. 1* 1 came from; 2 diverted; 3 left / *freq.35* rank 4655 # texts 25;
aviire [-va] *verb, transitive cl. 1* 1 has come from; 2 has left / *freq.13* rank 10467 # texts 8;
awava [-va] *verb, transitive cl. 16* 1 comes from; 2 departs / *freq.13* rank 10468 # texts 5;
baava [-va] *verb, transitive cl. 2* 1 came from; 2 diverted / *freq.118* rank 1598 # texts 40;
bava [-va] *verb, transitive cl. 2* from / *freq.120* rank 1572 # texts 47;
baviile [-va] *verb, transitive cl. 2* 1 have come from; 2 have left / *freq.13* rank 10514 # texts 10;
baviire [-va] / *freq.17* rank 8404 # texts 10 {Notes: see *baviile*};
biva [-va] *verb, transitive cl. 8* 1 from; 2 come from / *freq.63* rank 2851 # texts 35;
biviire [-va] *verb, transitive cl. 8* 1 came from; 2 diverted / *freq.104* rank 1787 # texts 7;
buva [-va] *verb, transitive cl. 14* come from / *freq.53* rank 3296 # texts 35;
byava [-va] *verb, transitive cl. 8* came from / *freq.14* rank 9926 # texts 8;
dhaava [-va] *verb, transitive cl. 10* came from / *freq.15* rank 9399 # texts 10;
dhiva [-va] *verb, transitive cl. 9* come from / *freq.26* rank 5968 # texts 19;
ebiva [-va] *verb, transitive cl. 8* come from / *freq.106* rank 1753 # texts 35;
ebyava [-va] *verb, transitive cl. 2* 1 came from; 2 resulted from / *freq.30* rank 5300 # texts 18;
edhiva [-va] *verb, transitive cl. 10* come from / *freq.29* rank 5457 # texts 18;
ekiva [-va] *verb, transitive cl. 7* comes from / *freq.38* rank 4408 # texts 16;
ekyava [-va] *verb, transitive cl. 7* 1 came from; 2 left / *freq.18* rank 8060 # texts 7;
eva [-va] *verb, transitive cl. 9* from / *freq.118* rank 1602 # texts 44;
eyava [-va] *verb, transitive cl. 1* came from / *freq.46* rank 3748 # texts 21;
gava [-va] *verb, transitive cl. 6* come from / *freq.33* rank 4920 # texts 25;
guva [-va] *verb, transitive cl. 3* comes from / *freq.27* rank 5809 # texts 16;
gwava [-va] *verb, transitive cl. 3* 1 came from; 2 left / *freq.13* rank 10656 # texts 11;
kava [-va] *verb, transitive cl. 12* 1 comes from; 2 departs / *freq.13* rank 10693 # texts 9;
kikuyiire [-va] *verb, transitive cl. 7 + 15* 1 has escaped; 2 has come from / *freq.17* rank 8539 # texts 1;
kiva [-va] *verb, transitive cl. 7* from / *freq.94* rank 1984 # texts 39;
kuva [-va] *verb, transitive 1* leave; 2 let alone / *freq.267* rank 749 # texts 96;
kwava [-va] *verb, transitive 1* came from; 2 resulted from / *freq.20* rank 7437 # texts 4;
kyava [-va] *verb, transitive cl. 7* 1 came from; 2 diverted / *freq.51* rank 3434 # texts 24 {Notes: see also *kye* & *ava* > *ky'ava*};
kyebaava [-va] / *freq.19* rank 7772 # texts 5 {Notes: see *kye* & *baava*};
kyenva [-va] / *freq.36* rank 4597 # texts 17 {Notes: see *kye* & *nva*};
kyetuva [-va] / *freq.13* rank 10766 # texts 10 {Notes: see *kye* & *tuva*};
kyeyava [-va] *verb, transitive 7 + 1* that is why / *freq.143* rank 1329 # texts 16 {Notes: see *kye* & *yava*};
liva [-va] *verb, transitive cl. 11* comes from / *freq.32* rank 5052 # texts 15;
luva [-va] *verb, transitive cl. 11* 1 comes from; 2 results from / *freq.20* rank 7443 # texts 12;
lyava [-va] *verb, transitive* came from / *freq.17* rank 8589 # texts 6;
muva [-va] *verb, transitive 2pl.* come from / *freq.30* rank 5351 # texts 15;
naava [-va] *verb, transitive 1sg.* 1 left; 2 diverted; 3 parted with / *freq.20* rank 7476 # texts 13;
nava [-va] *verb, transitive 1sg.* 1 came from; 2 left / *freq.69* rank 2654 # texts 27;

nva [-va] *verb, transitive 1sg.* from / *freq.162* rank 1190 # texts 71;
nviile [-va] *verb, transitive 1sg.* 1 have left; 2 have come from / *freq.18* rank 8200 # texts 16;
nviire [-va] / *freq.17* rank 8657 # texts 8 (Notes: see **nviile**);
obuva [-va] *verb, transitive* come from / *freq.32* rank 5086 # texts 17;
oguva [-va] *verb, transitive cl. 3* comes from / *freq.26* rank 6067 # texts 11;
okuva [-va] *verb, transitive* 1 since; 2 from; 3 from ... to; 4 to leave or be from / *freq.2668* rank 58 # texts 207;
okuviila [-va] *verb, transitive* 1 leave; 2 come from / *freq.27* rank 5887 # texts 18;
okuviira [-va] *verb, transitive* 1 come from; 2 move away from / *freq.58* rank 3075 # texts 18;
oluva [-va] *verb, transitive cl. 11* 1 from which; 2 comes from / *freq.40* rank 4260 # texts 28;
olwava [-va] *verb, auxiliary cl. 11* came from / *freq.30* rank 5382 # texts 12;
omutuviira [-va] *cl. 1 + 1pl.* leave / *freq.93* rank 2015 # texts 5;
omuva [-va] *verb, transitive cl. 3* 1 from which; 2 comes from / *freq.42* rank 4090 # texts 28;
ova [-va] *verb, transitive 2sg.* 1 return; 2 reason / *freq.325* rank 619 # texts 103;
tava [-va] *verb, transitive 2sg.* does not come from / *freq.19* rank 7919 # texts 16;
tova [-va] *verb, transitive* 1 leave not; 2 divert not / *freq.15* rank 9729 # texts 13;
tunaava [-va] *verb, transitive 1pl. + 1sg.* 1 will leave; 2 will come from / *freq.14* rank 10340 # texts 9;
tuviile [-va] *verb, transitive 1pl.* 1 came from; 2 left / *freq.44* rank 3941 # texts 21;
twava [-va] *verb, transitive 1pl.* 1 came from; 2 diverted; 3 left / *freq.38* rank 4465 # texts 23;
va [-va] *verb, transitive* 1 leave; 2 let alone / *freq.81* rank 2312 # texts 42;
yaava [-va] *verb, transitive cl. 1, cl. 9* 1 come from; 2 diverted / *freq.96* rank 1955 # texts 45;
yava [-va] *verb, transitive cl. 1* from / *freq.188* rank 1041 # texts 55

Types in the 1.7m Lusoga corpus considered for the lemma -vaawo, loc. verb:

okuvaawo [-va-wo] *verb, intransitive* 1 leave; 2 become extinct / *freq.15* rank 9671 # texts 10

Types in the 1.7m Lusoga corpus considered for the lemma -vaaku, loc. verb:

kuvaaku [-va-ku] *verb, transitive* 1 cause; 2 trigger / *freq.14* rank 10102 # texts 14

Types in the 1.7m Lusoga corpus considered for the lemma -vaamu, loc. verb:

avaamu [-va-mu] *verb, transitive cl. 1.* 1 departs; 2 results into; 3 releases / *freq.18* rank 7990 # texts 15
bituviiremu [-va-mu] *verb, intransitive cl. 8 + 1pl.* 1 result into; 2 have come from / *freq.13* rank 10527 # texts 5
ebivaamu [-va-mu] *verb, transitive cl. 8* 1 come out; 2 become visible / *freq.12* rank 11353 # texts 12
ekivaamu [-va-mu] *verb, transitive cl. 7* comes out finally / *freq.37* rank 4494 # texts 24
ekyavaamu [-va-mu] *verb, transitive cl. 7* 1 came out; 2 resulted into / *freq.78* rank 2386 # texts 21
kuvaamu [-va-mu] 1 get out; 2 come forward; 3 betray; 4 end up / *freq.24* rank 6442 # texts 17
mwavaamu [-va-mu] *verb, transitive cl. 18* 1 resulted into; 2 came out / *freq.14* rank 10180 # texts 9
okuvaamu [-va-mu] *verb, transitive* 1 leave; 2 get out; 3 end up / *freq.46* rank 3775 # texts 30

Types in the 1.7m Lusoga corpus considered for the lemma -vaayo, loc. verb:

avaayo [-va-yo] *verb, intransitive cl. 1.* 1 comes out; 2 surfaces / *freq.35* rank 4654 # texts 25
aveeyo [-va-yo] *verb, intransitive cl. 1.* 1 move away; 2 become visible / *freq.14* rank 9851 # texts 13
baveeyo [-va-yo] *verb, intransitive cl. 2* 1 become visible; 2 come out; 3 move away / *freq.13* rank 10513 # texts 12
kuvaayo [-va-yo] *verb, intransitive* 1 come out; 2 become visible; 3 move away / *freq.38* rank 4424 # texts 27
muveeyo [-va-yo] *verb, intransitive 2pl.* 1 clear the way; 2 become visible; 3 come out / *freq.38* rank 4440 # texts 11

navaayo [-va-yo] *verb, intransitive 1sg.* 1 returned; 2 came back; 3 left / *freq.13 rank 10874* # texts 8
okuvaayo [-va-yo] *verb, intransitive* 1 come out; 2 become visible / *freq.45 rank 3858* # texts 30
twavaayo [-va-yo] *verb, transitive 1pl.* 1 left; 2 came out / *freq.15 rank 9746* # texts 9
vaayo [-va-yo] *verb, intransitive 2sg.* 1 move away; 2 come out; 3 become visible / *freq.17 rank 8765* # texts 15
waavaayo [-va-yo] *verb, intransitive cl. 1* 1 come out; 2 become visible; 3 move away / *freq.13 rank 11077* # texts 10
yaavaayo [-va-yo] *verb, transitive cl. 1* came out / *freq.22 rank 7029* # texts 15
yavaayo [-va-yo] *verb, transitive cl. 1* 1 left; 2 came out; 3 became visible / *freq.18 rank 8311* # texts 13

Types in the 1.7m Lusoga corpus considered for the lemma -vo in cl. 14, dev. noun:

buvo [-vo] *noun -/14* place of origin / *freq.21 rank 7085* # texts 14
obuvo [-vo] *noun -/14* origin / *freq.19 rank 7858* # texts 9

New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the Diccionarios Valladolid-UVa

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)*

Sven Tarp, *International Centre for Lexicography, University of Valladolid,
Spain; Department of Afrikaans and Dutch, University of Stellenbosch,
South Africa; and Centre for Lexicography, University of Aarhus, Denmark
(st@cc.au.dk)*

and

Peter Sepstrup, *Ordbogen A/S, Odense, Denmark (pse@ordbogen.com)*

Abstract: This contribution deals with a new digital English–Spanish–English lexicographical project that started as an assignment from the Danish high-tech company Ordbogen A/S which signed a contract with the University of Valladolid (Spain) for designing and compiling a digital lexicographical product that is economically and commercially feasible and can be used for various purposes in connection with its expansion into new markets and the launching of new tools and services which make use of lexicographical data. The article presents the philosophy underpinning the project, highlights some of the innovations introduced, e.g. the use of logfiles for compiling the initial lemma list and the order of compilation, and illustrates a compilation methodology which starts by assuming the relevance of new concepts, i.e. object and auxiliary languages instead of target and source languages. The contribution also defends the premise that the future of e-lexicography basically rests on a close cooperation between research centers and high-tech companies which assures the adequate use of disruptive technologies and innovations.

Keywords: DICTIONARY CONCEPT, EMPIRICAL RESOURCES, LOGFILES, NGRAM VIEWER, INTERNET AS A CORPUS, COMPILATION METHODS, LEXICOGRAPHICAL DATA, ONLINE DICTIONARIES, INTEGRATED DICTIONARIES, WRITING ASSISTANTS, L2-RECEPTION DICTIONARIES, L2-PRODUCTION DICTIONARIES, TRANSLATION DICTIONARIES

Opsomming: Nuwe insig in die ontwerp en samestelling van digitale tweetaalige leksikografiese produkte: Die geval van die Diccionarios Valladolid-UVa. In hierdie bydrae word aandag geskenk aan 'n nuwe digitale Engels–Spaans–Engelse leksikogra-

fiese projek wat begin is in opdrag van die Deense hoëtegnologiemaatskappy Ordbogen A/S. 'n Ooreenkoms is gesluit met die Universiteit van Valladolid (Spanje) vir die ontwerp en vervaardiging van 'n digitale leksikografiese produk wat ekonomies en kommersieel uitvoerbaar is en wat gebruik kan word vir verskillende doeleindes wat verband hou met die uitbreiding daarvan na nuwe markte en die bekendstelling van nuwe hulpmiddels en dienste wat leksikografiese data benut. Die artikel bespreek die filosofie onderliggend aan die projek, belig sommige van die vernuwend elemente wat bekendgestel is, soos die gebruik van log-lêers vir die samestelling van die aanvanklike lemmalys en die volgorde van die samestelling. Die samestellingsmetodologie wat begin by die aanname dat vernuwend konsepte toepaslik is, word ook geïllustreer, d.w.s. primêre en sekondêre tale in plaas van doel- en brontale. In hierdie bydrae word die aanname dat die toekoms van e-leksikografie fundamenteel berus op die noue samewerking tussen navorsingsentrums en hoëtegnologiemaatskappye wat die voldoende gebruik van ontwrigtende tegnologieë en vernuwend elemente verseker, verdedig.

Sleutelwoorde: WOORDEBOEKKONSEP, EMPIRIESE HULPBRONNE, LOG-LÊERS, NGRAM VIEWER, DIE INTERNET AS 'N KORPUS, SAMESTELLINGSMETODES, LEKSIKOGRAFIESE DATA, AANLYN WOORDEBOEKE, GEÏNTEGREERDE WOORDEBOEKE, SKRYFHULPMIDDELS, L2-RESEPSIE-WOORDEBOEKE, L2-PRODUKSIE-WOORDEBOEKE, VERTALENDE WOORDEBOEKE

"Q: Did you do consumer research on the iMac when you were developing it?

A: No. We have a lot of customers, and we have a lot of research into our installed base. We also watch industry trends pretty carefully. But in the end, for something this complicated, it's really hard to design products by focus groups. A lot of times, people don't know what they want until you show it to them. That's why a lot of people at Apple get paid a lot of money, because they're supposed to be on top of these things."

(Interview with Steve Jobs in *Business Week*, Reinhardt (1998))

1. Introduction

In this contribution, we will first briefly discuss the history and philosophy behind a new digital English–Spanish–English lexicographical project which started in 2017 and is expected to see its first practical results launched from 2020 onwards. We will then go into details about the experience to date in the design and compilation of the product which, in various aspects, is innovative and based on cutting-edge technology with the use of completely new lexicographical methods guided by the Function Theory of Lexicography; cf. Fuertes-Olivera and Tarp (2014).

The project started as an assignment from the Danish high-tech company Ordbogen A/S, an international provider of online dictionary portals (ordbogen.com, lemma.com) as well as language services 24/7. Due to its technological innovation and unique business model, both of which have received

several national and international prizes, this company has since 2000 completely surpassed the traditional publishing houses and is now the dominant provider of online dictionaries in Denmark with a clear intention to increase its market share also in the neighboring countries. It is therefore an interesting partner for any lexicographer with a novel idea to be implemented or a dictionary to be distributed on a commercial basis.

As to the project discussed in this contribution, Ordbogen A/S wanted a digital lexicographical product that is economically and commercially viable and can be used for various purposes which are in line with its expansion into new markets and the launching of new tools and services which make use of lexicographical data. The Danish company therefore made contact with the International Centre for Lexicography at the University of Valladolid (Spain), with which it was already collaborating. The collaboration was on two other major online projects, the English–Spanish Accounting Dictionaries (available since 2012) and a set of monolingual Spanish dictionaries (under construction), both of which are to be commercialized under the brand name *Diccionarios Valladolid-UVa*; see Fuertes-Olivera (forthcoming).

The contract signed between the two partners stipulates, among other things, that the Danish company provides the technological support for the project, including the Dictionary Writing System (DWS) with lexicographical database, interfaces, search engines and grammar, as well as part of the empirical basis. The Spanish counterpart is in charge of the practical production of the required lexicographical data by means of highly specialized human resources, and project management. In addition, Ordbogen A/S finances part of the production costs at the International Centre for Lexicography with the Centre and the University of Valladolid providing the remaining funds. The project develops on a contractual basis as an international cooperation between two independent partners, each with their specific know-how and experience. To our knowledge, this represents a rather atypical lexicographical project model as projects of a similar scope in most cases are carried out either directly in the publishing houses or by independent entities and lexicographers who subsequently offer their products to the former. So far, the experience has been highly positive. For instance, Ordbogen A/S has agreed to transfer 50,000 euros a year to the International Centre for Lexicography for paying contracted and freelance lexicographers working on this project. These lexicographers are expected to compile around 25,000 senses a year, including definitions, collocations, synonyms, antonyms, examples and other data. This means that this bilingual project is expected to use approximately 2 euros per sense.

2. Lexicography, technology and current trends

During its more than four thousand years of existence as a cultural practice, lexicography has always depended strongly on the available technology in order to compile and present its products which, until now, have mainly taken

the form of dictionaries, although history has also known other forms of lexicographic endeavour. Hanks (2013: 507), for instance, reports how, at the dawn of European lexicography (500 B.C.), "it was customary for Greek scribes to insert glosses into manuscript copies of the works of Homer and other earlier writers" in order to explain "obsolete and unusual words".

These early *context-adapted* lexicographical glosses, which later developed into separate glossaries, allow for two important conclusions which we think are undervalued in the scholarly literature: 1) that lexicographers, as a matter of fact, do not compile dictionaries but lexicographical data which subsequently can be used for different purposes, among them, and notably, to edit dictionaries; and 2) that the standardized dictionary which was totally dominant in the printed environment is not the only type of lexicographical product known to history. Both conclusions are highly relevant for the correct interpretation of the current tendencies in lexicography where new disruptive technologies are turning the discipline upside down.

Prior to the advent of computer and information technologies, the introduction of the printing press more than 500 years ago had, in many respects, a similar impact on the discipline. A lot has been written about this phenomenon and some of its consequences (see, for instance, Hanks 2010), whereas other consequences have been less adequately dealt with, although they may not be less important in the long run.

In conclusion, it can be established that the introduction of printing technology implied big changes in:

- the production and presentation of the lexicographical product;
- the empirical basis with the increased use of index cards based on written texts;
- the design of dictionary articles with the incorporation of new data categories;
- the distribution and use of dictionaries;
- the number of users;
- the topics treated in dictionaries; and
- the research areas of scholarly interest.

To this can be added the growing social prestige of lexicographers, some of whom became nationally and internationally famous personalities, as well as the fact that lexicography turned into an increasingly successful business. Over a few hundred years, printing technology led to an almost total revolution of the discipline.

A similar thing is happening today where the technological innovations affect lexicography in its four main dimensions, i.e. the production, presentation, usage and financing of the lexicographical product. Fuertes-Olivera (2016) refers to the current situation as a "Cambrian explosion" where new forms constantly appear and disappear. This indicates that the adaptation to the new technological environment is a complex process that is far from one-dimen-

sional. Of special concern is the fact that the new technologies, especially the use of the Internet to make dictionaries available to their users, has undermined the existing business model and thrown lexicography into a sort of identity crisis where many publishing houses have reduced or even closed down their lexicographical sections due to dramatically reduced sales. Consequently, the continuous production of high-quality products is under attack. A new business model is therefore necessary but this is, obviously, the publishers' task — although nothing prevents lexicographers from contributing new ideas.

It is important to understand that the roots of the current crisis for lexicography are not only objective (disruptive technologies and an obsolete business model), but also subjective (ingrained habits and a frequently conservative approach to the new challenges). In this regard, lexicographers also have a big responsibility to the future of the discipline. They are above all challenged with the task of engaging in interdisciplinary collaboration with programmers and designers in order to guarantee still higher productivity without compromising quality and exploring new ways of presentation of the lexicographical product as the old static dictionary article is becoming increasingly obsolete. This presupposes a good dose of technological sensibility and understanding of the lexicographical potential of the continuous innovations, development of new compilation methods, and visionary thinking that offers new solutions to both new and old problems. In this perspective, Rundell and Kilgarriff (2011) have treated some of the technological and methodological advances in terms of the automation of the compilation process but the very title of their contribution leaves, understandably, an important question to be answered: "Where will it all end?"

In the following, we will take Rundell and Kilgarriff's (2011) reflections a little further and look into new methods of lexicographical data compilation. However, we are firmly convinced that this is not possible without knowing, or at least having a qualified idea of, the destination of these data and how they will eventually be presented to the users. A careful observation of the current trends related to this aspect of lexicography unveils four big transformations that are taking place simultaneously:

1. The first big transformation is from the *printed dictionary* to the *digital dictionary*. This process is still ongoing and characterized by a large number of dictionaries, especially online ones, that are either digitalized editions of already printed dictionaries or designed from scratch without taking into proper consideration the new options provided by the digital media.
2. The second transformation is from the traditional *stand-alone dictionary*, either printed or online, to the *integrated dictionary*, i.e. a dictionary integrated into other information tools such as e-readers, writing assistants or learning tools.
3. The third transformation is from the *standardized dictionary*, which is a typical result of the printed book format, to a more *personalized dictionary*

that adapts to the user's specific needs in each situation.

4. Finally, it is also possible to observe an inevitable move away from the *dictionary as such to lexicographical data for different uses*. Today, many publishing houses are increasingly receiving their revenue from selling lexicographical data. Many integrated information tools do not present dictionaries as such to their users but only a selection of data types taken from a lexicographical database.

The growing tendency to work with separate lexicographical data is also the reason why this contribution mostly refers to *lexicographical products* instead of dictionaries. In some cases, the data may even be taken from different sources. An example of this can be seen in Figure 1, which is a screenshot from a Danish–English writing assistant where the English equivalent *donation* and its Danish definition have been taken from two different digital dictionaries in order to get the best result for the user. This has become necessary because neither of the two dictionaries was originally planned and designed to be used as a support for a writing assistant.

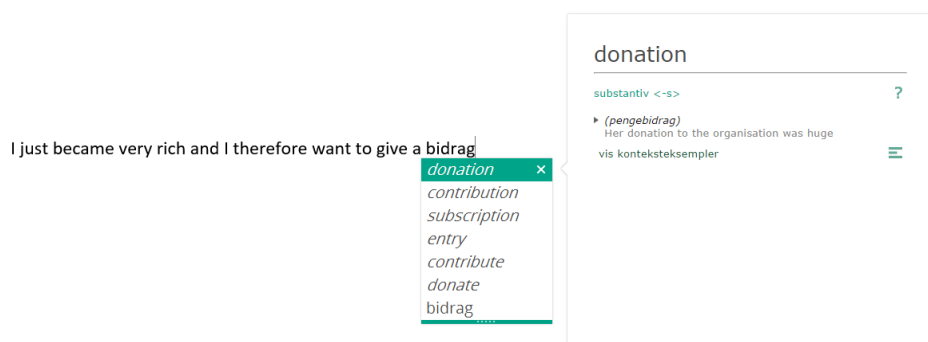


Figure 1: Screenshot from a Danish–English Write Assistant (Fisker 2018)

These four transformations are interwoven and herald a near future of integrated information tools that are based on digital platforms and provide *personalized service* by making use of lexicographical data. Personalized service is a general consumer demand in modern society and has therefore been a dream of many lexicographers in recent years; see, for instance, Rundell (2010).

In a lexicographical perspective, personalized service can be defined as the provision of the exact amount and types of data required to meet the user's needs in each concrete consultation, neither more nor less (Fuertes-Olivera and Tarp 2014: 64). This requires that the lexicographical data are of a high quality, that there are enough data, and that data and information overload is avoided as requested by Gouws and Tarp (2017).

Until a few years ago, it was conceived by many lexicographers, among

them Tarp (2011), that a personalized solution meeting these requirements would be something like "a set of components which customers can mix and match according to their needs" (Rundell 2007: 50). However, subsequent technological development has shown that such a solution, although innovative and useful in many aspects, is not the final word in this respect as it entails clear structural limitations, especially in terms of the resulting information costs, i.e. the time required to find and retrieve the needed information; see Nielsen (2008). It is now clear that a completely personalized service is only possible in an integrated information tool which, like a GPS, is designed to "observe" its users' behavior and prescribes the exact amount of data that is likely to meet their needs in each concrete case.

3. Presentation of the project

The combination of data from two different sources, which was shown in Figure 1, is a typical example of lexicographical databases that were prepared without knowing the exact use of the included data. In this case, the databases were designed several years before the work with the writing assistant started, and the problem was inevitable, at least to a certain extent. It nevertheless shows that meticulous work is required when starting a lexicographical project as small inadvertent "mistakes" could have big consequences at a later stage.

In the project we are discussing, the assignment from Ordbogen A/S was clear. The company wanted a bilingual lexicographical database that could feed two new products, namely a "traditional" online English–Spanish–English dictionary (to be included into the portal *Diccionarios Valladolid-Uva*) as well as a Spanish–English Write Assistant like the one described by Tarp et al. (2017). Both products are intended for native Spanish-speaking users.

After signing the contract, the lexicographers at the International Centre for Lexicography in Valladolid were tasked with 1) establishing the respective lexicographical functions, 2) framing a project concept including the required data categories, 3) preparing a compilation methodology that guarantees productivity and quality, and 4) engaging with programmers and designers from Ordbogen A/S in order to design a Dictionary Writing System suited for this particular project; cf. Fuertes-Olivera (forthcoming).

As to the functions, it was evident that the Write Assistant in the first instance had to assist Spanish users when writing texts in English and, secondarily, when translating from Spanish into English. These two functions are also applicable to the dictionary section of the project to which should be added two other functions that were exclusively relevant to the dictionary, i.e. assistance with reception of English texts and English–Spanish translation.

Based on these prioritized functions, a list of data categories to be included in the lexicographical database was drawn up. These categories included formal grammar, definitions, synonyms, antonyms, collocations, example sentences, etc. Apart from that, and due to the design and functionality of both the

Write Assistant and the dictionary, the data categories were divided into their smallest relevant parts so that they could be presented separately to the users who are expected to work on devices with varying screen sizes (laptops, tablets and smartphones).

As to the presentation of the product, it is predicted that the dictionary will be made available on the Internet with different function buttons which allow the users to get more specialized and individualized service such as:

- English definitions
- English grammar
- English synonyms and antonyms
- English collocations and examples
- Spanish–English translation
- Spanish–English translation of collocations and examples
- Spanish–English translation of fixed expressions
- English–Spanish translation
- English–Spanish translation of collocations and examples
- English–Spanish translation of fixed expressions
- Etc.

As to the Write Assistant, its design and functionality is still being improved. However, at this point it is clear that the tool demands English equivalents to Spanish words, Spanish definitions of English words (to be used as meaning discrimination), English inflectional forms, English synonyms and antonyms, English collocations and example sentences, etc. This suggests that the data categories already envisaged for the dictionary project are sufficient in order to meet the requirement of the Write Assistant.

As mentioned, both the dictionary and the writing assistant are intended for native Spanish-speaking users. This means that English in both cases is the language where the users need to have special assistance, whereas Spanish is used both to access and explain English, and as a lexicographical metalanguage. We therefore call these two languages *object language* (English) and *auxiliary language* (Spanish), respectively. These two terms are used with a different meaning than the one defined by Hartmann and James 1998 in their *Dictionary of Lexicography*. In this respect, they break with the terminology which is traditionally used to describe bilingual dictionaries (source language, target language, both of them treated as object languages) and which was coined in a period when practical solutions to users' needs were influenced and also limited by the existing technology, especially due to the restraints of the printed book format. We do not find this old terminology to be the most adequate and helpful if lexicography is expected to make the most out of modern computer and information technology as it may constitute a mental barrier that stands in the way of developing new solutions.

The new terminology makes us focus on the object language, i.e. English. It is English that has to be described and explained to the Spanish users. It is in

English where they need instructions on how to write and produce texts. Spanish is "only" used to access the English words and expressions as well as to explain these and give indications on how to use them in context. This means that Spanish is not going to be treated at the same level as English.

This approach has direct consequences for the methodology used in the project. Whereas traditional mono-directional, biscopal dictionary projects usually take their point of departure in the users' native language, the Valladolid project does the opposite. It starts with a selection and description of English lemmata including separation in senses, definitions, Spanish equivalents, grammar, etc. An automatic and simultaneous inversion is then made where the Spanish equivalents to one or more English lemmata become new lemmata whereas the English lemmata become equivalents with the brief Spanish definitions used as meaning discrimination. This inversion is, of course, revised by the lexicographers who also rely on an independent list of Spanish lemma candidates as will be explained in paragraphs 4 and 5.

The described compilation methodology is then in close collaboration with programmers and designers from Ordbogen A/S, incorporated into the Dictionary Writing System which, so far, has proved to be very user-friendly, efficient and economical in terms of productivity and quality.

4. Empirical basis

In the scholarly literature, there is a long and rich reflection on the most adequate empirical basis of the different types of dictionary. As it was briefly mentioned in the historical overview in paragraph 2, as is the case with the compilation and presentation of the lexicographical product, its empirical sources also change over time as the result of the continuous technological development. Since the 1960s, and especially since the disruptive publication of the *Collins Cobuild English Language Dictionary* in 1987, there has been a strong reliance on still bigger corpora as the main empirical source of dictionaries (see Sinclair 1987, Bergenholtz 1996, Kilgarriff 1997, Atkins and Rundell 2008, and Hanks 2012, among many others). The positive results of this development are indisputable and excellent dictionaries have been produced with this point of departure. However, the generalized use of corpora also gives rise to new questions and challenges, especially in the light of new technological innovations such as digital dictionaries, the Internet and logfiles. This is the case with the selection of lemmata to dictionaries with a limited lemma stock. Some lexicographers, like Kilgarriff (2013), advocate, at least until recently, the use of corpora in these cases:

Building a headword list is the most obvious way to use a corpus for making a dictionary. *Ceteris paribus*, if a dictionary is to have N words in it, they should be the N words from the top of the corpus frequency list. (Kilgarriff 2013: 79)

However, there are two questions which in our opinion have not been paid suf-

ficient attention, namely 1) whether users actually consult the lemmata included in dictionaries, and 2) the relationship between corpus frequency and lexicographical frequency, i.e. the frequency with which the users consult the words in a dictionary. As to the first question, Bergenholtz and Norddahl (2012) have reported that the study of logfiles shows that a considerable number of words have never been consulted in an online Danish dictionary after more than 20 million lookups. The dictionary in question is a big general one with more than hundred thousand lemmata and the conclusion may therefore not be representative for dictionaries with a more reduced lemma stock as the ones to which Kilgarriff (2013) refers. However, research into logfiles by other scholars confirms another of Bergenholtz and Norddahl's (2012) conclusions, namely that there is a certain, and therefore lexicographical relevant, discrepancy between the most frequent words in a corpus and the words most frequently looked up in dictionaries; see De Schryver et al. (2006) and Trap-Jensen et al. (2014).

This last conclusion implies that it would be better to start a lexicographical project with a reduced lemma stock with lemmata selected from logfiles instead of a corpus, and then use the method recommended by Bergenholtz and Johnsen (2005) and De Schryver (2013), among others, to supplement the lemma list with additional lemmata that appear in the logfiles once the dictionary has been published online. This is, at least, the method used in the project discussed in this contribution, which uses four main empirical bases: logfiles; Ngram Viewer; the Internet; and existing dictionaries. These are used in the above order and nobody working in the project can change the order, as this could clearly jeopardize the whole project, as we will show in the following paragraphs. This is critical for the project as we believe that someone initially consulting a dictionary will be clearly influenced in their lexicographical work by the data found in the consulted dictionary.

4.1 Logfiles

As already mentioned, our bilingual project started in 2017 with an initial lemma list of around 20,000 English words and 16,000 Spanish words. These were compiled at Ordbogen A/S headquarters in Odense (Denmark) by using big data analytics for around two months.

The process comprises several stages and is based on an analysis of around one million daily searches in several of the company's dictionaries, e.g. an English–Spanish/Spanish–English dictionary, an English–German/German–English dictionary, an English monolingual dictionary, a Spanish monolingual dictionary, and so on. Around 80% of the searches can be matched, i.e. the same search is identified in the logfiles of different dictionaries and can therefore be interpreted with the aim of identifying the most popular dictionary articles in the dictionaries under scrutiny. After two months of work with the logfiles of the searches, which amount to more than 60 million, IT people at Ordbogen A/S were able to produce the above-mentioned lists of 20,000 Eng-

lish words and 16,000 Spanish words. They comprise the words most searched for in the period under analysis and were used by the editor of the project for compiling the initial lemma lists of the bilingual project. Below, we copy some of the searched words, most of which are currently lemmata in this project:

- English words starting with "ang-": *angel, angelic, angelica, anger, angered, angina, angiogenesis, angiogram, angioplasty, angle, angled, angler, Anglican, anglicize, angling, Anglo, Anglo-American, Anglo-Danish, Anglo-Saxon, Anglophile, Anglophobe, Anglophone, Angola, angrily, angry, angst, anguish, anguished, angular, angular momentum, angularity, and angulation.*
- English words starting with "bed-": *bed, bed linen, bed-sitting room, bedazzle, bedbug, bedchamber, bedclothes, bedcover, bedding, bedevil, bedew, bedfellow, bedlam, Bedouin, bedpan, bedplate, bedpost, bedraggled, bedridden, bedrock, bedroom, bedside, bedsit, bedsore, bedspread, bedspring, bedstead, bedstead canopy, bedtime and bedtime story.*
- English words starting with "defe-": *defeasible, defeat, defeated, defeatist, defecate, defecation, defect, defection, defective, defector, defence, defenceless, defend, defendant, defender, defenestrate, defenestration, defensible, defensive, defensively, defensiveness, defer, deference, deferent, deferential, deferment, deferral, and deferred income.*
- English words starting with "equ-": *equable, equal, equalization, equalize, equalizer, equality, equally, equanimity, equanimous, equate, equation, equator, equatorial, equestrian, equidistant, equidistribution, equifinality, equilateral, equilibrate, equilibration, equilibrist, equilibrium, equine, equinox, equip, equipment, equipoise, equipotential, equipped, equitable, equitably, equity, equity capital, equity ratio, equity warrant, equivalence, equivalency, equivalent, equivalently, equivocal, equivocality, equivocally, equivocate, and equivocation.*

A comparison of the above words with the lemma list of the *Oxford English–Spanish Dictionary* indicates three main findings. Firstly, the degree of matching between them is high: 74% of the most searched words are also found in the lemma list of the *Oxford English–Spanish Dictionary*. Secondly, there are 36 searched words (26%) that are not included in the Oxford list and these are basically either formal technical words, e.g. medicine words, or multiword lemmata, i.e. extended units of meaning (Rundell 2018). Thirdly, this second list, composed of the searched words that are not lemmatized in the Oxford dictionary, offers some clues about users' interest for some semantic fields, — these are: medicine, sports, law, sex, and economics — and for multiword lemmata. In addition, they also indicate the adequacy of logfiles for offering more than possible lemmata: they can also help lexicographers to disambiguate meanings and offer interesting data for crafting additional data types, typically sentence examples and collocations, i.e. chunks of words that offer clues on the meaning and use of the lemma and/or equivalent.

4.2 Ngram Viewer

As already mentioned, in this bilingual project we have used three other empirical bases apart from the logfiles: Ngram Viewer; the Internet; and existing dictionaries. Ngram Viewer is being used for four main lexicographical purposes. Firstly, it is used for augmenting the initial lemma list with "extended units of meaning", i.e. strings of recurrent words that adhere to Sinclair's idiom principle which assumes that language users regularly resort to "an inventory of semi-preconstructed phrases that constitute single choices" (Sinclair 1991: 110). In our bilingual project these are lemmatized when they refer to bearers of meaning, e.g. they refer to material things, feelings and emotions, human beings, etc. both in their literal and figurative meanings. For instance, we have searched in Ngram Viewer (English) the following strings: *air* * _ADJ, *air* * _NOUN, * _NOUN *air*, * _VERB *air*, * _ADJ *air*, *air* * _PREP, *air* * _CONJ and *air* * _VERB.

An analysis of the hits as well as the results obtained during the process of compilation, e.g. by means of Google searches and look ups in existing dictionaries (see below), have resulted in around 100 multiword lemmata with "air": *air current, air dry, air offensive, air pollution, air force, air conditioning, air temperature, air pressure, air flow, air transport, air space, air pump, compressed air rifle, air rifle scope, confined air, air letter, air ambulance, air assault, air attack, air ball, air brake, air bubble, air cargo, air cleaner, air conditioner, air conditioning unit, air dam, air filter, air freight, air freshener, air gap, air gauge, air guitar, air hammer, air hockey, air hockey table, air horn, air hostess, air hunger, air intake, air kiss, air leak, air lock, air mail, air map, air marshal, air mass, air mattress, air mile, air out, air piracy, air pocket, air power, air pressure, air rage, air raid, air scoop, air spring, air strike, air taxi, air traffic, air traffic control, air travel, air valve, air vent, air waybill, air-condition, air-conditioned, air-cooled, air-dried, air-filled, air-raid shelter, air-sea rescue, air-to-air, air-to-ground, air-to-surface, air traffic controller, airbag, airbase, airbed, airtight, airway, air lane, on-air, airing, airman, airwoman, air gun, hot air, airfare and airdrop.*

It is interesting to highlight that around 40% of these multiword lemmata are not lemmatized in the *Oxford English-Spanish Dictionary*. This finding indicates that the use of several empirical bases may be needed for compiling online dictionaries, especially because of the disappearance of the space constraints associated with a printed dictionary. For instance, neither *air ball* nor *airball* is lemmatized in the *Oxford English-Spanish Dictionary*, although it is frequently used in Spanish television during the broadcasting of a basketball match where an *airball* is "an unblocked shot which misses the basket, the rim, and the backboard entirely" (*Wikipedia*).

Secondly, Ngram Viewer is used for deciding which word variety is used as lemma and which other varieties are included but not lemmatized. For example, the English varieties "color" and "colour" refer to the same reality and are totally synonymous for Spanish native speakers, who are the main users of these

products. We normally lemmatize the most frequent variety and include the other varieties in several data fields, e.g. as synonyms with their corresponding tag (e.g. UK English or US English) or as not recommended (Figure 2). This decision does not hinder users' searches in a reception situation: the dictionary entry for "color" and "colour" is the same and will be recovered searching "color" or "colour".

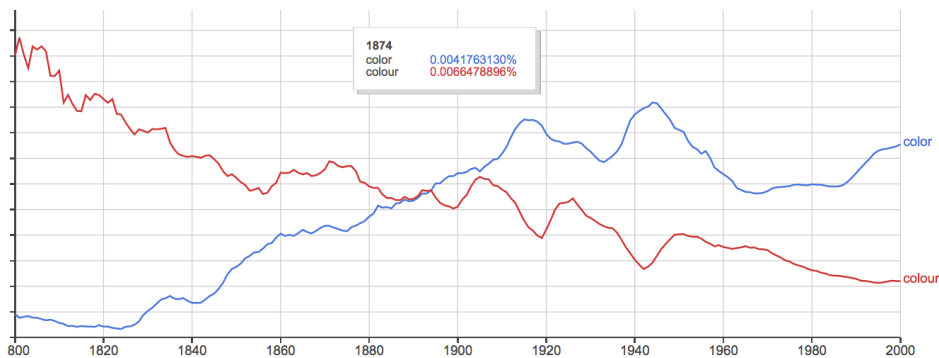


Figure 2: Comparison of **color** and **colour** with Ngram Viewer

Thirdly, we also use Ngram Viewer for checking the inflections and grammar forms of all lemmata, especially those lemmata that can be problematic for users, e.g. countable and uncountable English nouns, masculine and/or feminine Spanish nouns, and so on. For instance **air power** is described as "mass noun" in the *Oxford English Dictionary*. However, Ngram Viewer shows that "air powers" is used in English, especially during the Second World war (Figure 3):



Figure 3: Uses of **air powers** with Ngram Viewer

Hence, in our project, **air power** is lemmatized as uncountable and countable, each with its own grammar, definition, examples, synonyms, and so on (examples 1 and 2):

air power

flexions

air power, air powers

Definition

unidad del ejército de un país encargada de todo lo relacionado con el ejército del aire

Equivalent

fuerza aérea

Example

China had "become one of the major air powers of the world"

China se ha convertido en una de las principales fuerzas aéreas del mundo

Example (1): Extract for **air power** (countable)

air power

flexions

Sin flexion (uncountable)

Definition

1. fortaleza o capacidad del ejército del aire de un nación para defender sus territorios o atacar otros

Equivalent

poder aéreo

Example

Military air power was used to protect relief efforts.

El poder aéreo militar se usó para proteger labores humanitarias.

Definition

2. energía producida por la acción del viento sobre un molino o aerogenerador

Equivalent

energía eólica

Example

In 2008, the U.S. became the world's leading provider of air power.

En 2008, EE.UU. se convirtió en el proveedor líder de energía eólica en el mundo.

Example (2): Extract for **air power** (uncountable)

Finally, we are also using Ngram Viewer for identifying frequent combinations of particular words that have not been lemmatized as multiword lemmata but are included in example sentences or other parts of the dictionary articles, especially when they offer something relevant such as a translation pattern as the different types of "air" recorded in Figure 4. All of them have similar Spanish translations and are therefore adequate for machine learning and neural network software: "aire de la noche", "aire de la mañana", "aire de la habi-

tación", "aire de la montaña", "aire de la tarde", "aire del mar", "aire salado", aire del verano" or "aire veraniego", "aire del desierto", and "aire del país":

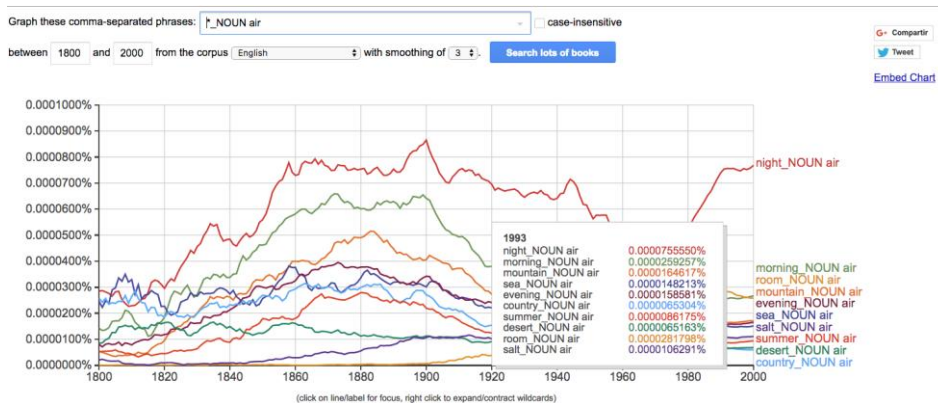


Figure 4: Searching *_NOUN air with Ngram Viewer

4.3 Internet

The Internet is also one of our main empirical bases. As shown in Tarp and Fuertes-Olivera (2016), we use it for crafting definitions, selecting different types of sentence examples, synonyms and antonyms, and so on. In Tarp and Fuertes-Olivera, we have shown that the analysis of Google minitexts, i.e. three-line texts that appear as a result of a Google search, has resulted in dictionary articles that, on average, describe 30% more different senses than existing dictionaries. Examples (1) and (2) above corroborate this reflection. **Air power** in this bilingual project has three senses (In all the dictionaries consulted, it has only one sense. For instance, the *Oxford English Dictionary* describes **air power** as "The ability to defend or attack by means of aircraft."). One of the senses recorded in this bilingual project refers to "wind energy", and this sense is obtained by analyzing texts such as the following, all of them extracted from the Internet, and recorded in the dictionary entry either as collocations or as sentence examples (example 3):

Collocations

- that air power is considered one of the purest energy sources
- the use of air power and solar installations
- the air power, biomass and waste treatment sectors
- air power, solar energy and other renewable energies

Examples

- In 2008, the U.S. became the world's leading provider of air power.

Example (3): Collocations and examples extracted from the Internet for crafting one of the senses of **air power**

The Internet is also used for four more key lexicographical tasks. First, we *always* consult *Wikipedia* for explaining *terms*, i.e. words describing concepts in specialized fields (Humbley 2018). For instance, **angioplasty** (one of the words from the logfiles), is explained in terms of the data reported in *Wikipedia*, especially with regard to following: (a) a definition for our Spanish users; (b) several synonyms, e.g. **balloon angioplasty**; (c) several types of **angioplasty**, all of which are also lemmatized, e.g. **coronary angioplasty**; **peripheral angioplasty**; **carotid angioplasty**; (d) its Spanish equivalent; (e) a usage note indicating that this medical procedure did not exist until 1964 and is therefore not found in texts before such year; (g) collocations and sentence example(s), e.g. "Angioplasty is typically used to treat atherosclerosis." (h) link(s) to the *Wikipedia* entry, images, and so on.

Secondly, we also use *Wikipedia* and other available texts, e.g. maps, lists of cities, rivers, oceans, seas and mountains for completing and describing our lemmata. For instance, the *Wikipedia* entry for **Amazon River** offers reliable data on its left and right tributaries (they are lemmatized when their length is more than 1,000 km) as well as other data for crafting its definition. In addition, several blogs offer interesting data about its flora and fauna, ecosystem, characteristics of the rainforest, etc. All these texts are analyzed and some of their data are included.

Thirdly, we also use the Internet for searching for texts that can be of use for our users such as free online pronouncing dictionaries (e.g. <https://www.howjsay.com/>), images, and so on. We include a link to unprotected texts, especially to texts produced by individuals or companies whose business model is based on the number of clicks, i.e. their revenues come from clicks, no matter where these are done.

Finally, we also use the Internet for finding equivalents. For such a task, we query Google with the lemma, some words related with its meaning(s), and the expression "in Spanish" or "in English" if we are searching for the Spanish or English equivalent. For instance, for finding the English equivalent of Spanish "cobro revertido", we googled "cobro revertido in English", and obtained the English equivalent "reverse charge". Analyzing it was very fruitful: we discovered that **reverse charge** is a synonym of US English **collect call** as well as a term related with the accrual of VAT (lemma **VAT reverse charge**), the charging of batteries, and a trick in pen spinning. In existing dictionaries **reverse charge** is only explained as UK English for "a telephone call paid for by the recipient" (*Oxford English Dictionary*).

4.4 Existing dictionaries

Existing dictionaries are also used as empirical bases. They are consulted once the rest of the empirical bases have been used for compiling the dictionary entries. This consultation has three main purposes.

Firstly, we check for any possible missing sense. In such a situation the lexicographer in charge of a particular dictionary entry must analyze whether the missing sense is still in use, e.g. by googling the lemma with some key parts of the definition or equivalent found.

Secondly, we check for possible grammar discrepancies (e.g. countable or uncountable nouns), lexicographical notes (e.g. a usage note about a particular lemma), formal and informal tags, and so on. If something new is found we double-check it before incorporating it in the dictionary entry.

Finally, we consult existing dictionaries for finding information about geographical varieties, something that is difficult to obtain from the rest of the empirical bases.

5. Phases and steps in the compilation process

As indicated in the previous section, there are three main phases, each with their own sub-phases or steps in the compilation process of this bilingual project.

5.1 First phase

The first phase comprises the work in the English–Spanish section of the project. It starts with the editor of the project analyzing the logfiles submitted by Ordbogen A/S, i.e. the list of around 20,000 English words, with two main aims, deciding which of the most searched words will be lemmatized in the project and establishing the order of compilation of each lemma, as it is expected that the project will go public before it has been completed. This order of compilation is important as we have found out that users do not search randomly but tend to search for specific words. For instance, an analysis of the around 2,300 English words starting with "a-" submitted by Ordbogen A/S to the editor shows that around 25% of them refer to five topics: medicine; sports; law; sex and economics.

The second step comprises the working in the 'Lemma section' of the Dictionary Writing System (Figure 5). This is the editing tool with up to 20 slots, i.e. fields for including lexicographical data and administrative data. The editor of the project enters the basic grammar data of the lemma and decides who will finish the rest of the dictionary entry ('Assigned user' in Figure 5), its status (Review, in progress, finished, start, etc.) and its log history, e.g. who has completed the entry, who has reviewed it, etc. This information is important for the editor, who can, for example, decide to assign the Spanish–English entry to the same lexicographer who has completed the English–Spanish one.

Figure 5: The 'Lemma section' of the Dictionary Writing System

By basic grammar data we mean the information that applies to the lemma in all situations, varieties, registers, etc. For most lemmas, this information comprises the following:

- Number: it differentiates between homonyms, for instance **air power** (countable) and **air power** (uncountable).
- Lemma: it records the dictionary form or canonical form of the lemma. For instance, in the logfiles we have found that users typically search for "clothes" instead of "cloth". In the project, however, we have lemmatized three examples of "cloth": (a) cloth as countable noun; (b) cloth as uncountable noun; and (c) cloth as singular noun in the collective noun "the cloth".
- Word class: it indicates the part of speech of the lemma.
- Inflexions: depending on the word class of the lemma, it stores singular and plural noun forms, comparative and superlative adjectival forms, some regular and irregular verbal forms, and so on.
- Discouraged inflexions: it also stores the above types of inflexions but with an indication that they are not recommended for some reason, e.g. **airball** is less used than **air ball**.

- Grammar remark in Spanish: a grammar comment in Spanish, e.g. **air power** is a countable noun and has singular and plural forms (Lemma 1) and **air power** has no inflexions as it is an uncountable noun (Lemma 2).
- Reference: for internal reference, i.e. cross-references, or for external reference, e.g. a link to a free pronouncing dictionary.
- Valency: it includes syntactic information of the lemma, e.g. "someone plucks something from the air" in the lemma **pluck from the air**.

The third step in phase one comprises work in the 'Meaning section' of the Dictionary Writing System (Figure 6). The assigned lexicographer works in this section, which consists of up to 27 slots with the aim of offering five types of information: (a) meanings; (b) tags, for indicating the style and type of English, if necessary; (c) remarks, e.g. with this meaning it is only used in negative; (d) references, both internal cross-references and external references with links to homepages, e.g. a FLICKR image; and (e) synonyms and antonyms. It is important to highlight that all the information contained in this part is linked to the Arabic number on the left side. This serves for "bundling" all the data of the "meaning part" to each meaning, i.e. associating each meaning to its synonyms, antonyms, production notes, external and internal references, tags and so on. For instance, the meaning "wind energy" of **air power** (example 2 above) is always associated with the data types describing this meaning and use.

The screenshot shows the 'Meaning' section of the Dictionary Writing System. It features a light blue background with a white header area containing the title 'Meaning'. Below the header, there are several input fields and buttons. At the top left, there is a 'Number' field with the value '1'. To its right are 'Meaning in English' and 'Meaning in Spanish' fields. The 'Meaning in Spanish' field contains the text 'unidad del ejército de un país encargada de todo lo relacionado con el ejército del aire'. To the right of these fields are 'Style' and 'Type in' dropdown menus. The 'Style' menu is set to 'neutral' and the 'Type in' menu is set to 'English'. Below these are four input fields for 'Lexical remark Spanish', 'Production remark Spanish', 'Lexical remark English', and 'Production remark English'. A red button labeled 'Delete Meaning' is located on the right side. At the bottom, there are three buttons: 'Reference', 'Antonyms +', and 'Synonyms +'. Below these are several more input fields for 'First reference', 'Second reference', 'See also', 'Internet link: Text', 'Internet link: Image', 'Presentation: Text', and 'Presentation: Image'. The 'Internet link: Text' field contains the URL 'https://es.wikipedia.org/wiki/Fuerza_a%C3%A9' and the 'Presentation: Text' field contains 'wikipedia'.

Figure 6: 'The Meaning section' of the Dictionary Writing System

The fourth step in phase one is working in the 'Translation section' of the Dictionary Writing System (Figure 7). It includes up to 20 slots, all of them concerned with the Spanish equivalent and the meaning and function of lemma and equivalent in context. In this section we also have an administrative button "Create lemma", which will be used in Phase 3 of the compilation process. Regarding the equivalent, lexicographers include the Spanish equivalent, its word class, grammar and contrastive remarks as well as syntactic information, e.g. that an English verb is only used with "something" and not with "someone". On average, we only include one equivalent per meaning, although there are exceptions. For instance, the English lemma **teacher** can refer to a male or female teacher. As the Spanish gender system is different we include the Spanish equivalents **profesor** and **profesora** (male and female teacher). This distinction is important for several reasons and has important consequences in our project. We will not comment on it further for reasons of space. Finally, the buttons "collocations", "examples" and "formation" in Figure 6 record data for contextualizing the meaning of the lemma in different translation situations, in which we can or cannot have the Spanish equivalent. For instance, one of the meanings of **leave up in the air** refers to an unsettled issue or plan. Its Spanish equivalent is "dejar en el aire". One of the collocations in this meaning is "that the whole matter was left up in the air for the whole weekend", which is translated into Spanish as "que todo quedó en el aire durante todo el fin de semana", i.e. the Spanish translation does not use the equivalent but an adaptation, i.e. "quedar en el aire" instead of "dejar en el aire".

Figure 7: The 'Translation section' in the Dictionary Writing System

5.2 Second phase

Phase 2 consists of a single step, i.e. reviewing the dictionary entry completed in Phase 1. This phase is assigned to a member of the International Centre for Lexicography who must check the work done before assigning the status "finished" to the dictionary entry and sending it to the editor of the project for initiating Phase 3. The reviewing phase consists of correcting possible errors, deciding whether the collocations and examples support the meaning and checking possible omissions, e.g. by comparing the dictionary entry with those in existing dictionaries. Should the reviewer find omissions, he or she must analyze them before sending the entry back to the lexicographer with indications about the omissions found. So far, we have found a small number of omissions, less than 2% of the completed entries have been sent back to lexicographers due to omissions.

5.3 Third phase

Phase 3 starts with the editor of the project checking the Spanish equivalent compiled in Phase 1. The editor decides either to convert the equivalent into a Spanish lemma or to leave it only as equivalent. Accepting the equivalent as lemma means reversing the English–Spanish word list. The reversion occurs when the editor clicks "create lemma" (Figure 7) and adds the Wordclass of the equivalent. The "Create lemma" button changes to "Open lemma" (now in green) and clicking on it opens the 'Lemma section' of the Dictionary Writing System corresponding to the Spanish–English side. Figure 8 shows the results of the reversion. An interesting feature is the opening of a drop-down menu on the right side of the 'Lemma section'. This menu refers to the English–Spanish section in which the present lemma was an equivalent and is identified as 'Select to open article'.

Working in this section of the Dictionary Writing System includes all the steps commented on in the above paragraphs, and two more new steps, one for the editor and another one for the lexicographers of the project. The editor has to analyze the presence of the equivalent in the list of 16,000 Spanish words extracted from logfiles and submitted by Ordbogen A/S. If the equivalent is in this list, the equivalent is included in the Spanish lemma list and work continues as shown in the above paragraphs. In most cases, however, the equivalent is not in the list of most searched words. In such a situation, the editor also lemmatizes the equivalent but may postpone the order of compiling, a decision depending on two variables. First, the editor checks whether the equivalent which has now become a lemma is not treated in the monolingual Spanish dictionary, which is also part of the *Diccionarios Valladolid-UVA* and which has more than 50,000 completed lemmata at the time of writing this article (July 2018). In such a situation, the editor usually assigns the equivalent-turned-lemma over to a lexicographer and the work continues as explained in the previous

paragraphs. Second, the equivalent-turned-lemma is not completed in the monolingual Spanish dictionary and is not connected with one of the five topics mostly searched for by users and discovered by lexicographers by analyzing the log files of English words starting with "a-". In such a situation, working on the equivalent-turned-lemma is usually postponed.

The screenshot shows the 'Lemma' section of a dictionary writing system. On the left, there is a sidebar with 'Spanish / English' and a search bar. The main area is titled 'Lemma' and contains several sections: 'Number Lemma' (0, dejar en el aire), 'Wordclass' (expression), and 'Found matching translation(s)' (- Select to open article). Below this are 'Inflexions' and 'Discouraged Inflexions' sections with input fields. There are also 'Grammar remark Spanish' and 'Grammar remark English' fields. At the bottom, there are 'Reference' and 'Valencies +' sections with input fields for 'First reference', 'Second reference', and 'See also'. A 'Delete Lemma' button is located in the top right corner of the main area.

Figure 8: The 'Lemma section' in the Dictionary Writing System after reversion

Lexicographers have an additional step on this side of the bilingual project. They must evaluate the information found when clicking on "Select to open article" (upper right part of Figure 8), which corresponds to English lemmata already described and finished in the English-Spanish section of the bilingual project. The purpose of such an evaluation is to find the best possible English equivalent for the Spanish lemma. For instance, one of the meanings of **abate** is "to put an end to a law, decree, etc." The best Spanish equivalent for such a meaning is *revocar* and therefore *revocar* is lemmatized in the Spanish-English section of the dictionary. When lexicographers study the Spanish word *revocar* and start to explain its different meanings, they must decide that the best English equivalent for the above meaning is **revoke**. There are several reasons for using **revoke** instead of **abate** in such a situation. Three of them are important and illustrate our method of working. The first one is that **abate** is restricted to formal written legal texts, whereas **revoke** is used in a greater variety of texts.

The second one is that **abate** has a lower frequency of use than **revoke** today (Ngram Viewer). And the third one is that **abate** shows a steep downward trend in use from 1800 to 2000 (Ngram Viewer). Hence, for the Spanish lemma **revocar**, its English equivalent is *revoke* whereas *abate* is a synonym that is assigned the tag "formal" and a synonym remark that explains that its use is restricted to formal English written texts.

Once lexicographers finish the steps already commented on they send their entries for reviewing and reviewers check their work before sending them back to the editor of the project who starts the process again. To sum up, the different phases and steps start with the English–Spanish section and continue with the Spanish–English section, which currently has half of the lemmata of the other section. The data will initially, and mainly, be used to feed the bilingual Spanish–English–Spanish dictionary as well as the Spanish–English Write Assistant, both for native Spanish-speaking users. In the long run, Ordbogen A/S is also planning an English–Spanish Write Assistant for English-speaking users. When the latter has to be prepared lexicographically, we will have to change our compilation order and have more or less the same amount of data in both sections of the bilingual project.

6. Conclusions

The bilingual project described in this paper offers four interesting conclusions for the future of e-lexicography. First, lexicographical data have an intrinsic economic value. This value can be realized provided it is prepared in such a way that it can be used for as many projects as possible, e.g. for developing writing assistants and online dictionaries, and for well-defined users and uses. Secondly, as lexicography is in the middle of a Cambrian explosion, the use of disruptive innovations is necessary to be competitive. For instance, this project shows that collaboration between research institutions and technological companies is fruitful as it guarantees funds, cutting-edge technology and knowledge. In addition, the project uses on a regular basis, systems, methods and resources that have not been used before on a large scale, e.g. logfiles for selecting the initial lemma list and the order of compilation, the Internet for searching for senses and the Ngram Viewer for searching for extended units of meaning. Thirdly, the project's point of departure is the idea that lexicography at its most abstract level is no more and no less than the science concerned with the theory and practice of *dictionaries*, i.e. dictionaries, encyclopedias, lexica, glossaries, vocabularies, terminological knowledge bases, and other information tools covering areas of knowledge and its corresponding language. And finally, the project is based on new ideas and concepts that have not been used so far in the scholarly literature related to bilingual projects, e.g. the existence of an object language and an auxiliary one and the interrelationship of the big transformations affecting today's lexicography.

Acknowledgments

Special thanks go to the Ministerio de Economía y Competividad for financial support (grant FFI2014-52462-P). Our special thanks to two anonymous reviewers and to Elsabé Taljard for their comments on a previous draft of this paper.

References

Dictionaries

Oxford English Dictionary: <https://en.oxforddictionaries.com/>.

Oxford Inglés-Español/Español-Inglés: <https://es.oxforddictionaries.com/>.

Wikipedia: <https://www.wikipedia.org/>.

Other literature

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Bergenholtz, H. 1996. Korpusbaseret Leksikografi. *LexicoNordica* 3: 1-15.

Bergenholtz, H. and M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes* 34: 117-141.

Bergenholtz, H. and B. Norddahl. 2012. Ordbogsartikler som ingen læser. *LexicoNordica* 19: 207-223.

De Schryver, G.-M. 2013. The Concept of Simultaneous Feedback. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 548-556. Berlin: Walter de Gruyter.

De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Fisker, K. (Ed.). 2018. *Write Assistant. Danish-English*. Odense: Ordbogen A/S.

Fuertes-Olivera, P.A. 2016. A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. *International Journal of Lexicography* 29(2): 226-247.

Fuertes-Olivera, P.A. Forthcoming. Designing and Making Commercially-driven Dictionary Portals: The *Diccionarios Valladolid-UVa*. *Lexicography*.

Fuertes-Olivera, P.A. and S. Tarp. 2014. Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography. Berlin/Boston: De Gruyter.

Gouws, R.H. and S. Tarp. 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.

- Hanks, P.** 2010. Lexicography, Printing Technology, and the Spread of Renaissance Culture. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6–10 July 2010*: 988-1016. Leeuwarden: Fryske Akademy.
- Hanks, P.** 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hanks, P.** 2013. Lexicography from Earliest Times to the Present. Keith, A. (Ed.). 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University Press.
- Hartmann, R.R.K. and G. James.** 1998. *Dictionary of Lexicography*. London/New York: Routledge.
- Humbley, J.** 2018. Specialised Dictionaries. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 317-334. London/New York: Routledge.
- Kilgarriff, A.** 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.
- Kilgarriff, A.** 2013. Using Corpora as Data Sources for Dictionaries. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 77-96. London: Bloomsbury.
- Ngram Viewer:** <https://books.google.com/ngrams>.
- Nielsen, S.** 2008. The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos* 18: 170-189.
- Reinhardt, A.** 1998. Steve Jobs: 'There's Sanity Returning'. *Business Week*, 25 May 1998. <https://www.bloomberg.com/news/articles/1998-05-25/steve-jobs-theres-sanity-returning>. Accessed on 29 June 2018.
- Rundell, M.** 2007. The Dictionary of the Future. Granger, S. (Ed.). 2007. *Optimizing the Role of Language in Technology-enhanced Learning*: 49-51. <https://hal.archives-ouvertes.fr/hal-00197203/document/>. Accessed on 30 June 2018.
- Rundell, M.** 2010. What Future for the Learner's Dictionary? Kernerman, I. and P. Bogaards (Eds.). 2010. *English Learners' Dictionaries at the DSNA 2009*: 169-175. Jerusalem: Kdictionaries.
- Rundell, M.** 2018. Searching for Extended Units of Meaning — and What To Do When You Find Them. *Lexicography*. <https://doi.org/10.1007/s40607-018-0042-1>. Accessed on 4 July 2018.
- Rundell, M. and A. Kilgarriff.** 2011. Automating the Creation of Dictionaries: Where Will It All End? Meunier, F., S. de Cock, G. Gilquin and M. Paquot (Eds.). 2011. *A Taste for Corpora. In Honour of Sylviane Granger*: 257-282. Amsterdam/Philadelphia: John Benjamins.
- Sinclair, J.M.** 1987. Introduction. *Collins Cobuild English Language Dictionary*: xv-xxi. London: HarperCollins.
- Tarp, S.** 2011. Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 54–70. London/New York: Continuum.
- Tarp, S. and P.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.
- Tarp, S., K. Fisker and P. Sepstrup.** 2017. L2 Writing Assistants and Context-aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.
- Trap-Jensen, L., H. Lorentzen and N.H. Sørensen.** 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 2(2): 94-113. <https://doi.org/10.4312/slo2.0.2014.2.94-113>. Accessed on 4 July 2018.

'n Leksikografiese datatrekkingstruktuur vir aanlyn woordeboeke

Rufus H. Gouws, *Departement Afrikaans en Nederlands,
Universiteit Stellenbosch, Stellenbosch, Suid-Afrika (rhg@sun.ac.za)*

Opsomming: Op die gebied van die leksikografie het die oorgang vanaf gedrukte na aanlyn woordeboeke 'n ingrypende invloed op talle aspekte van sowel die leksikografieteorie as die leksikografiepraktyk. In die voortgesette ontwikkeling van die leksikografieteorie moet hierdie invloed verwoord word sodat daar riglyne gebied kan word vir die optimale benutting van die vooruitspruitende aanpassings in die leksikografiepraktyk. Woordeboekstrukture moet opnuut ondersoek word om vas te stel watter strukture in die nuwe medium behou kan word, watter strukture aan aanpassings onderhewig is en watter nuwe strukture na vore tree.

In hierdie artikel is die fokus op aanpassings in leksikografiese strukture. Daar word verwys na strukture waarvoor aanpassings reeds in die metaleksikografie bespreek is. Die hoofklem is op verskillende vorme van die dataverspreidingstruktuur in aanlyn woordeboeke. Voorsiening word gemaak vir 'n omvattende dataverspreidingstruktuur wat in woordeboekportale gebruik kan word en die gebruiker toegang tot woordeboeksterne bronne gee. Die behoefte aan groter vryheid van die gebruiker om data te kies wat benodig word, lei tot voorstelle vir 'n nuwe struktuur, naamlik die datatrekkingstruktuur. Met behulp van hierdie struktuur kan gebruikers regstreeks vanuit 'n bepaalde posisie in 'n aanlyn woordeboek toegang kry tot die internet as leksikografiese korpus om aan die data daar die inligting te onttrek wat in 'n bepaalde gebruikssituasie verlang word. Die datatrekkingstruktuur bevestig die status van woordeboeke as geïntegreerde inligtingsinstrumente en plaas hulle binne die bestek van 'n oorkoepelende datastruktuur.

Sleutelwoorde: AANLYN WOORDEBOEK, DATASTRUKTUUR, DATATREKKINGSTRUKTUUR, DATAVERSPREIDINGSTRUKTUUR, EENVOUDIGE DATAVERSPREIDINGSTRUKTUUR, LEKSIKOGRAFIESE STRUKTURE, OMVATTENDE DATAVERSPREIDINGSTRUKTUUR, OOR-KOEPELENDE DATAVERSPREIDINGSTRUKTUUR, STOOTMEDIUM, TREKMEDIUM, UITGEBREIDE DATAVERSPREIDINGSTRUKTUUR, WOORDEBOEKPORTAAL, WOORDEBOEK-PORTAALSTRUKTUUR

Abstract: A Lexicographic Data Pulling Structure for Online Dictionaries. In the field of lexicography the transition from printed to online dictionaries has had a significant influence on numerous aspects of both lexicographic theory and the lexicographic practice. In the continued development of lexicographic theory this influence has to be formulated in order to present guidelines for the optimal application of the resulting adaptations in the lexicographic practice. Dictionary structures should be investigated to determine which structures can be maintained in the new medium, which structures need to be adapted and which new structures are coming to the fore.

The focus in this article is on adaptations in lexicographic structures. Reference is made to structures of which the adaptations have already been discussed in metalexigraphy. The main emphasis is on different types of data distribution structures in online dictionaries. Provision is made for a comprehensive data distribution structure that can be employed in dictionary portals to give the user access to dictionary-external sources. The need of users for more freedom to select their required data, leads to proposals for a new structure, namely the data pulling structure. By employing this structure users can access the internet as lexicographic corpus from any point in an online dictionary to retrieve from the data there the information they require in a specific situation of use. The data pulling structure confirms the status of dictionaries as integrated information instruments and puts them within the scope of an over-arching data structure.

Keywords: COMPREHENSIVE DATA DISTRIBUTION STRUCTURE, DATA DISTRIBUTION STRUCTURE, DATA PULLING STRUCTURE, DATA STRUCTURE, DICTIONARY PORTAL, DICTIONARY PORTAL STRUCTURE, EXTENDED DATA DISTRIBUTION STRUCTURE, LEXICOGRAPHIC STRUCTURES, ONLINE DICTIONARY, OVER-ARCHING DATA DISTRIBUTION STRUCTURE, PULL MEDIUM, PUSH MEDIUM, SINGLE DATA DISTRIBUTION STRUCTURE

1. Inleiding

'n Oorsig oor die ontwikkeling van die leksikografiepraktyk lewer bewys van 'n verskeidenheid belangwekkende oorgange, byvoorbeeld die oorgang van kleitablette na perkamentrolle en na papier, die oorgang van 'n tematiese ordening van woordeboeke na 'n alfabetiese ordening, die oorgang na die drukpers en later na die rekenaar as instrumente in die samestelling van woordeboeke en die koms van elektroniese korpora. Die ingrypendste oorgang, sonder twyfel, was die oorgang van die gedrukte na die elektroniese medium en meer spesifiek die leksikografiese toetrede tot die aanlyn omgewing. Dit was nie net 'n oorgang wat die medium waarin die leksikografiese werk gedoen word, betref nie, maar veel meer as dit. Hierdie oorgang was 'n paradigmaskuif wat die tipe leksikografiese werktuie, die inhoud van woordeboeke, die wisselwerking met ander naslaanbronne en die aard en omvang van die verpakking van data verander het (vergelyk Rundell 2012: 72)

Aanlyn woordeboeke gee tans nie altyd blyke van die nodige metaleksikografiese onderbou nie. Dit is nie net die gevolg daarvan dat die samestelling dikwels deur mense met die nodige tegniese kennis maar gebrekkige leksikografiese kennis gedoen word nie, maar ook omdat metaleksikograwe nog nie voldoende ondersteuning aan die nuwe leksikografiepraktyk gebied het nie. Te veel aspekte van die teoretiese leksikografie is nog slegs op gedrukte woordeboeke gerig. In die voortgaande leksikografiese gesprek is dit dringend noodsaaklik dat die moontlikhede wat die aanlyn omgewing aan die leksikografie bied uitvoerig ter sprake moet kom sodat die nodige aanpassings gedoen kan word ter daarstelling van die formulering van 'n algemene leksikografieteorie wat nie medium-spesifiek is nie. Hierdie aanpassing by die aanlyn omgewing

geld alle aspekte van die leksikografie.

Leksikografiese strukture as een onderafdeling in die formulering van 'n algemene leksikografieteorie maar ook in die beplanning en samestelling van woordeboeke, word in hierdie artikel aan die orde gestel. Daar word kortliks verwys na aanpassings in strukture soos die makro-, artikel-, adresserings- en toegangstruktuur. Daarna word meer aandag gegee aan die dataverspreidingsstruktuur en voorstelle word gemaak vir die erkenning van 'n nuwe leksikografiese struktuur, naamlik die datatrekkingstruktuur. Die siening van 'n woordeboek as 'n houer van kennis, sien McArthur (1986), en as 'n draer van tekssoorte, sien Wiegand (1996) word ook kortliks krities onder die loep geneem.

Volgens Wiegand (1989: 251) is leksikografie 'n praktyk wat gerig is op die produksie van woordeboeke sodat 'n verdere praktyk, naamlik die kulturele praktyk van woordeboekgebruik, 'n aanvang kan neem. Woordeboeke is praktiese gebruiksinstrumente en hierdie funksie word nie bepaal deur die medium van 'n woordeboek nie. Gevolglik moet die gesprek oor aanlyn leksikografie steeds die belang van die gebruiker in ag neem. In die hieropvolgende paragraaf kom dit kortliks ter sprake.

2. Die gebruikersperspektief

Die gebruikersperspektief met sy fokus op die vasstelling van 'n spesifieke teikengebruiker en die leksikografiese behoeftes en naslaanvaardighede van daardie gebruiker staan reeds lank sentraal in die metaleksikografiese gesprek, vergelyk onder meer Hartmann (1989; 2001). As deel van die reaksie op die behoeftes van die teikengebruiker moet daar ook duidelikheid wees oor die funksie wat die betrokke woordeboek moet hê om die spesifieke behoeftes van die gebruiker te kan bevredig.

Ter sake in sowel gedrukte as aanlyn woordeboeke is die gebruikersperspektief. Ook in aanlyn woordeboeke mag die gebruiker nie langer die bekende onbekende (Wiegand 1977: 59) wees nie. Die behoeftes maar ook naslaanvaardighede van duidelik gespesifiseerde gebruikers moet bepalend wees in die keuse en aanbieding van data in woordeboeke. Ook leksikografiese strukture moet hierdie gebruikers ter wille wees. Daar moet in die teorie-ontwikkeling deurgaans rekening gehou word met 'n veranderende en veranderde gebruikersgroep.

Een van die reuse-uitdagings van die hedendaagse leksikografie is om woordeboeke te produseer wat steeds relevant is en wat gebruik word. In die formulering van die leksikografieteorie asook in die beplanning van nuwe woordeboeke moet vernuwende denke daarop gerig wees om gebruikers en potensiële gebruikers daarvan bewus te hou dat woordeboeke steeds voorkeurbronne vir die oplossing van bepaalde probleemtypes kan wees. Dit moet veral ook in ag geneem word dat 'n wesenlike deel van die potensiële gebruikersgroepe van die toekoms tot 'n nuwe geslag behoort, onder meer Generasie Z. Dit is lede van die samelewing wat na 2000 gebore is.

Woordeboeke moet ook werktuie wees waarmee lede van Generasie Z inligting kan bekom wat aan hulle leksikografiese behoeftes voldoen. Hierdie generasie is digitale burgers wat die internet as vanselfsprekend aanvaar en alles daar wil vind — vergelyk Parker (2013), Finch (2015) en Gouws (2017). Dit het implikasies vir woordeboektipologie, die strukture van woordeboeke en die leksikografiese daarstelling van data. Finch (2015) sê:

It's critical that we recognize Gen Z's differences and meet them where they are, rather than where we want them to be.

Die ingrypende verskil tussen lede van die digitale maatskappy en die voorafgaande geslagte blyk ook daaruit dat hulle 'n ander benadering tot die naslaan-aktiwiteit en naslaanwerke het. Hulle wil verkieslik 'n enkele instrument hê wat toegang tot alle data gee wat hulle nodig het; ook data wat tradisioneel in woordeboeke aangebied word.

Die aanpassing van die teoretiese leksikografie, die leksikografiepraktyk en daarom ook leksikografiese strukture by die digitale omgewing dwing metaleksikograwe en leksikograwe om steeds bedag te wees op die eise vanuit die gebruikersperspektief.

3. Leksikografiese strukture en woordeboeknavorsing

In verskeie van sy publikasies, onder meer Wiegand (1989: 262; 2010: 250) dui die metaleksikograaf Herbert Ernst Wiegand daarop dat die metaleksikografie, as afdeling van die breë terrein van woordeboeknavorsing, in vier hoofonderafdelings verdeel kan word. Hierdie afdelings is historiese woordeboeknavorsing, navorsing oor woordeboekkritiek, navorsing oor woordeboekgebruik en sistematiese woordeboeknavorsing. Volgens Wiegand (2010: 250) is die opdrag aan die sistematiese woordeboeknavorsing om op die basis van empiriese ondersoek 'n teorie van die leksikografiese proses te behandel en wel as deelteorie van 'n algemene leksikografieteorie. In die leksikografiese proses teorie word drie deelteorieë onderskei, te wete 'n teorie van die behandeling van leksikografiese data, 'n teorie van draers van leksikografiese tekste en 'n teorie van woordeboekindeling (Wiegand 2010: 251).

Wiegand (2010: 251) dui ook aan dat in die teorie van die draers van leksikografiese tekste woordeboeke beskou word as teksdraers, dit wil sê hulle is opgebou uit verskillende soorte tekste. In hierdie verband is gedrukte woordeboeke statiese en digitale woordeboeke dinamiese inligtingstelsels. In die teorie van draers van leksikografiese tekste gaan dit onder meer om 'n deelteorie wat op die woordeboekvorm fokus as deel van 'n teorie oor die strukture van leksikografiese tekste en deelt tekste.

Navorsing oor woordeboekstrukture is deeglik ingebed in 'n algemene leksikografieteorie. Hierdie navorsing, en by name die uitgebreide en baanbrekerswerk wat Wiegand in hierdie verband gedoen het, is grootliks op gedrukte woordeboeke gerig. Wiegand stel dit in verskeie van sy publikasies

eksplisiet dat dit op die strukture van gedrukte woordeboeke gerig is, vergelyk onder meer Wiegand (2005), Wiegand en Beer (2013), Wiegand, Feinauer en Gouws (2013), Wiegand en Gouws (2013, 2013a), Wiegand en Smit (2013; 2013a). Ook hierdie gebied van die metaleksikografie moet opnuut bekyk word met die oog op 'n vasstelling van die strukture wat in aanlyn woordeboeke benut word. In hierdie verband moet aandag gegee word aan daardie strukture wat in sowel gedrukte as aanlynwoordeboeke voorkom maar bepaalde aanpassings in aanlyn woordeboeke vereis, maar ook aan die strukture wat slegs in óf gedrukte óf aanlyn woordeboeke voorkom. Vir aanlyn woordeboeke moet die bespreking van leksikografiese strukture gerig wees op hulle bydrae tot 'n dinamiese inligtingstelsel. Bestaande bydraes in hierdie verband is onder meer Gouws (2014, 2014a, 2014b, 2018, 2018a, 2018b) asook Klosa en Gouws (2015).

Die tempo van die beskikbaarstelling van aanlyn woordeboeke en die aantal woordeboeke wat geproduseer is, staan in skrilte kontras tot die trae ontwikkeling in metaleksikografiese beskouing van en teorievorming oor aanlyn woordeboeke. Dit blyk 'n herhaling te wees van die verhouding tussen teorie en praktyk in die ontwikkeling van gedrukte woordeboeke, vergelyk Gouws (2011), waar die praktyk die teorie voorafgegaan het in stede daarvan om die teorie te volg. Die snelheid waarteen leksikografies minder goeie aanlyn woordeboeke op die mark gekom het, stel tans hoë eise aan die metaleksikografie om modelle daar te stel wat 'n gehalteverbetering van aanlyn woordeboeke kan verseker en die leksikografiese behoeftes van 'n vinnig veranderende gebruikersmark op 'n doeltreffende manier kan bevredig.

Die ontwikkeling van die aanlyn leksikografiese praktyk het op verskillende maniere en in verskillende fases plaasgevind. Die vroeë ontwikkeling het tot produkte gelei wat oorspronklik as gedrukte woordeboeke gepubliseer is en toe, soms via die CD ROM-roete en soms nie, gedigitaliseer is en aanlyn beskikbaar gestel is. Na voorkoms was hierdie woordeboeke dikwels slegs gedigitaliseerde weergawes van gedrukte woordeboeke wat elektronies selfs nog deurgebraai kon word. 'n Verdere ontwikkeling het dieselfde woordeboeke beter aangebied met toegangstrukture wat gebruikers vinniger op die verlangde soekroetes kon plaas en 'n aanbieding waarvan die artikels grootliks nog die vorm van dié van hulle gedrukte voorgangers weerspieël het, maar waar die tradisionele bladsybeeld en -uitleg ontbreek het. 'n Ander ontwikkeling het daartoe gelei dat woordeboeke van meet af as aanlyn woordeboeke saamgestel is met al die voordele van onder meer verbeterde soekmoontlikhede. Dit is die werklike aanlyn woordeboeke wat voortaan in hierdie bydrae ter sprake kom.

Alhoewel dit in veral hierdie laaste ontwikkeling dikwels duidelik was dat die nodige metaleksikografiese onderbou ontbreek, was dit vanuit 'n metaleksikografiese perspektief eweneens duidelik dat daar opnuut gekyk moet word na en leiding verstrekkend moet word oor die benutting van woordeboekstrukture. Vrae moet ook gestel word oor of die tradisionele leksikografiese strukture wat vir gedrukte woordeboeke ontwerp is ook ter sake is vir aanlyn woordeboeke. Daar moet ook voorsiening gemaak word vir nuwe strukture

wat na vore tree. 'n Besinning oor leksikografiese strukture as 'n wesenlike deel van 'n algemene leksikografieteorie moet ook tot 'n besinning lei oor of die geldigheidsbestek van 'n algemene leksikografieteorie vir gedrukte woordeboeke ook aanlyn woordeboeke insluit. Dit benadruk weer die behoefte aan een algemene leksikografieteorie wat nie medium-spesifiek is nie.

4. Aanpassing van leksikografiese strukture

4.1 Strukture wat reeds behandel is

Daar is vroeër in hierdie artikel reeds verwys na bydraes soos Gouws (2014, 2014a, 2014b, 2018, 2018a, 2018b) asook Klosa en Gouws (2015) waarin die aanpassing van sekere leksikografiese strukture vir aanlyn woordeboeke ter sprake gebring is. In hierdie bydraes is die fokus veral op die makro-, artikel-, adresserings- en toegangstruktuur. Telkens word die belang van 'n voortsetting van die spesifieke struktuurtype in aanlyn woordeboeke bevestig, maar telkens word daar ook gewys op aanpassings wat gemaak moet word en vernuwing wat danksy die aanlyn omgewing moontlik is.

In aanlyn woordeboeke bly die makrostruktuur 'n ordeningstruktuur, maar dit bied nie noodwendig meer aan gebruikers dieselfde soort oorsig oor die lemmaversameling van 'n bepaalde woordeboek nie. Deeltrajekte tree sterk op die voorgrond en die hooftoegangstruktuur val nie meer noodwendig saam met die makrostruktuur nie (Gouws 2014a). Aanlyn woordeboeke, Gouws (2014), het 'n dinamiese en veelvlakkige artikelstruktuur wat dit vir gebruikers moontlik maak om doelgerig na spesifieke artikelsones te beweeg. Waar die adresseringstruktuur in gedrukte woordeboeke slegs ter sprake kom tussen individuele aanduiders en nooit in die geval van aanduidertekste nie, word daar voorgestel (Gouws 2015) dat aanduidertekste in aanlyn woordeboeke ook aan lemmatekens geadresseer kan word. Dit beklemtoon 'n wesenlike verskil tussen gedrukte en aanlyn woordeboeke wat spruit die aard van die verhoudings wat deur die adresseringstruktuur blootgelê word en die mindere mate van teksverdigting in aanlyn woordeboeke.

Alhoewel die toegangstruktuur in sowel gedrukte as aanlyn woordeboeke 'n belangrike rol speel, toon Gouws (2018a) aan hoe die aanlyn omgewing nuwe tipes toegangstrukture en toegangsroetes moontlik maak. Die toegangstruktuur in aanlyn woordeboeke is waarskynlik die struktuur wat die grootste en opvallendste verskille toon met die voorkoms van die ooreenstemmende struktuur in gedrukte woordeboeke. Die herwinning van inligting uit leksikografiese data hoef nie meer noodwendig via die lemmateken as gidselement van 'n artikel te wees nie. Die toegangstruktuur bied seekroetes wat kundige gebruikers regstreeks na bepaalde seeksones of spesifieke aanduiders kan lei.

In die res van hierdie artikel is die fokus veral op die dataverspreidingsstruktuur en implikasies van die toepassing daarvan, onder meer die erkenning van 'n nuwe struktuurtype.

4.2 Die dataverspreidingstruktuur

4.2.1 Buitetekste en die raamstruktuur

Die dataverspreidingstruktuur is die struktuur wat die plasing en ordening van data in 'n woordeboek as draer van tekssoorte bepaal. Dit gaan nie net om die plasing van die lemmata en die artikelinterne aanduiders nie, maar ook om die plasing van buitetekste in die voor-, middel- en agtertekste-afdelings. Vir die dataverspreidingstruktuur in gedrukte woordeboeke bied die woordeboek se raamstruktuur aan leksikograwe die moontlikheid om verskillende plasing-omgewings vir die verskillende tekstipes te vind. 'n Raamstruktuur, vergelyk Kammerer en Wiegand (1998), is 'n tipe ordeningstruktuur wat nie ter sake is vir aanlyn woordeboeke nie. Alhoewel aanlyn woordeboeke ook tekste kan bevat wat in 'n gedrukte woordeboek as buitetekste beskou sou word, is hierdie tekste nie meer in vaste liniêre posisies tot die sentrale teks as voortekste, middeltekste en agtertekste geplaas nie. In aanlyn woordeboeke gaan dit ook nie noodwendig oor tekste wat die leksikografiese aanbod van die sentrale teks aanvul nie. Daar kan ook klank- of beeldgrepe wees asook grafika en ander tipes data-aanbod. Gevolglik stel Klosa en Gouws (2015) voor dat daar van *buitekenmerke* eerder as *buitetekste* gepraat moet word. In hierdie artikel sal die term *buitekomponente* eerder gebruik word vir daardie tekste of ander datahouers wat buite die grense van 'n woordeboek se alfabetiese komponent val. Die aard van 'n aanlyn woordeboek oorskry dus die siening wat Wiegand (1996: 136) en Kammerer en Wiegand (1996: 224) ten opsigte van gedrukte woordeboeke voorstel, naamlik dat die woordeboek 'n draer van tekssoorte is. Aanlyn woordeboeke kan veel meer as tekste bevat.

Vir die doel van hierdie artikel is dit belangrik om daarop te let dat die aanlyn omgewing nie 'n raamstruktuur gebruik nie. Die buitetekste van gedrukte woordeboeke word in aanlyn woordeboeke as buitekomponente beskou wat nie 'n vaste posisie met betrekking tot die alfabetiese komponent van die woordeboek het nie. Daar moet steeds onderskei kan word tussen aanlyn woordeboeke met en dié sonder buitekomponente. In hulle bespreking van die dataverspreidingstruktuur onderskei Bergenholtz, Tarp en Wiegand (1999: 1779) tussen 'n eenvoudige dataverspreidingstruktuur en 'n uitgebreide dataverspreidingstruktuur. Die eersgenoemde kom voor waar 'n woordeboek se leksikografiese data slegs in die sentrale teks aangebied word, terwyl die tweede tipe verwys na woordeboeke waar die sentrale teks aangevul word deur buitetekste wat ook as draers van leksikografiese data gebruik word. Hierdie onderskeid tussen die twee tipes dataverspreidingstruktuur is steeds vir die aanlyn omgewing bruikbaar en sal in hierdie artikel gebruik word.

Naas die uitgebreide dataverspreidingstruktuur maak die aanlyn omgewing 'n verdere tipe dataverspreiding moontlik, naamlik die omvattende dataverspreidingstruktuur.

4.2.2 'n Omvattende dataverspreidingstruktuur

Volgens Wiegand, Beer en Gouws (2013: 63) is 'n woordeboek as geheel 'n soekveld met elke artikel wat as soekgebied optree en uit 'n aantal soeksones bestaan waarin die verskillende aanduiders en merkers (vergelyk Wiegand en Smit 2013: 153) as funksionele tekssegmente geplaas word.

Die dataverspreidingstruktuur van aanlyn woordeboeke is uiteraard steeds grootliks daarop gerig om data in die onderskeie woordeboekartikels te plaas. Die dinamiese artikelstruktuur en nuwe toegangstrukture lei die gebruiker regstreeks na die tersaaklike soeksones. Naas die artikels tree bepaalde buitekomponente as plasingskandidate vir leksikografiese data op. Dié plasing word met behulp van die uitgebreide dataverspreidingstruktuur deurgevoer. In die aanlyn omgewing is die dataverspreiding nie tot die posisies in die artikels of dié in die buitekomponente beperk nie. Aanlyn woordeboeke tree nie altyd slegs as geïsoleerde naslaanbronne op nie, maar vorm soms deel van 'n woordeboekportaal. Dit lei tot die daarstelling van 'n nuwe struktuur, te wete die woordeboekportaalstruktuur. Hierdie struktuur het implikasies vir die dataverspreiding, vergelyk Gouws (2018b), en leksikograwe moet nogmaals aandag gee aan die bestek van die dataverspreidingstruktuur.

Engelberg en Müller-Spitzer (2013: 1024) gebruik die term *woordeboekportaal* om na 'n spesifieke tipe datastruktuur te verwys:

- (i) that is presented as a page or set of interlinked pages on a computer screen and (ii) provides access to a set of electronic dictionaries, (iii) where these dictionaries can also be consulted as standalone products.

Engelberg en Müller-Spitzer (2013: 1027) onderskei drie verskillende tipes woordeboekportale, te wete 'n woordeboeknetwerk, 'n woordeboeksoekenjin en 'n woordeboekversameling. Woordeboekversamelings, volgens Engelberg en Müller-Spitzer (2013: 1028) die eenvoudigste tipe woordeboekportaal, is dikwels webbladsye met skakels na aanlyn woordeboeke maar die woordeboeke in 'n woordeboekportaal kan almal ook op die portaal se tuisblad verskyn. Vanuit 'n gebruikersperspektief is 'n woordeboekportaal 'n vertrekpunt vanwaar die gebruiker met behulp van verskillende soekroetes toegang kan kry tot verskillende woordeboeke en die leksikografiese data in daardie woordeboeke. Die gebruiker kan hom of haar tot slegs 'n enkele woordeboek wend of dieselfde data in meerdere woordeboeke in die portaal nagaan.

OWID, die *Online Wortschatz Informationssystem Deutsch* (aanlyn woorde-skatinligtingstelsel van Duits) van die Institut für Deutsche Sprache bied die volgende openingsblad ter aanduiding van hulle woordeboekportaal — met 'n aanduiding van die verskillende woordeboeke in die kolom aan die regterkant:

Afbeelding 1: Skermskoot uit OWID

In die beplanning van 'n woordeboekportaal moet daar vir die tradisionele dataverspreidingstruktuur van die individuele woordeboeke voorsiening gemaak word asook vir die verspreiding van data in die portaal as sodanig. Dit geskied deur middel van 'n *omvattende dataverspreidingstruktuur*. In die beplanning van die portaal moet die leksikograaf vasstel watter woordeboeke in die portaal optree asook watter data in watter woordeboek aangebied moet word. Die toegangstruktuur moet dan so ontwerp word dat die gebruiker die relevante inskrywings in die onderskeie woordeboeke kan bereik. Elk van die woordeboeke in die woordeboekportaal kan ook 'n uitgebreide dataverspreidingstruktuur hê en gebruikers moet leiding kry oor watter data alles binne die portaal beskikbaar is.

5. 'n Oorkoepelende datastruktuur

5.1 Woordeboek-eksterne bronne

Die toetrede van woordeboekportale tot die aanlyn leksikografie noodsaak die inwerkingstelling van omvattende dataverspreidingstrukture aangesien die

leksikografies relevante data oor verskillende bronne heen versprei kan word. Die mediostruktuur moet ook aangepas word want verwysings kan nou ook tussen verskillende woordeboeke in die woordeboekportaal gemaak word. Mediostruktuurtipes wat vir gedrukte woordeboeke ontwerp is, maak reeds voorsiening vir woordeboeksterne kruisverwysingsadresse. Dit is die tipe mediostruktuur wat ook deel vorm van die kruisverwysingstelsel in 'n woordeboekportaal, waar die mediostruktuur aangepas moet word om nou ook 'n woordeboeksterne maar portaalinterne kruisverwysingsadres te kan hê.

In én enkelwoordeboeke met óf 'n eenvoudige óf 'n uitgebreide dataverspreidingstruktuur én woordeboekportale met 'n omvattende dataverspreidingstruktuur geld dieselfde kernbenadering. Die leksikograaf besluit op data wat ter beskikking van die gebruiker gestel moet word en die dataverspreidingstruktuur as 'n ordenings- en plasingstruktuur bepaal die leksikografiese ruimte waar die data waaraan gebruikers inligting moet onttrek, geplaas moet word. Daardie leksikografiese ruimte kan verskillende woordeboeke wees en in hierdie woordeboeke hetsy soeksones binne die bepaalde soekveld se soekgebiede (vergelyk Wiegand, Beer en Gouws 2013: 63), hetsy buitekomponente wat onder meer tekste, grafika, beeld- of oudiodata kan wees.

Die vraag wat in 'n vernuwendende benadering tot aanlyn leksikografie gevra kan word, is of die woordeboeksterne bronne nie ook buite die woordeboekportaal kan val en gebruikers dus inligting kan onttrek aan bronne wat nie noodwendig leksikografiese bronne is nie. Leksikografiese strukture is noodsaaklik om sowel die voorkoms en aanbieding as die opsporing van en toegang tot inhoud te optimeer. 'n Algemene leksikografieteorie behoort dus ook aandag te gee aan die daarstelling van strukture wat op 'n nuwe manier die gebruiker kan help om ook woordeboeksterne inligting te bekom. Aanlyn woordeboeke se inskakeling by 'n oorkoepelende datastruktuur kan tot nuwe vorme van datatoegang lei.

Vir die verstaan van strukture in aanlyn woordeboek is dit volgens Müller-Spitzer (2013: 369) noodsaaklik om nie net na die aanbiedingsvlak te kyk nie, maar ook na die databasisvlak. Atkins en Rundell (2008: 264) beskou 'n databasis as 'n gestruktureerde versameling data op grond waarvan woordeboekartikels geskep kan word. Die leksikografiese databasis word gevul met data wat onder meer uit een of meer leksikografiese korpora kom en die samestelling van hierdie korpora en die vind van die gepaste data vir die databasis val binne die bestek van die woordeboekkonseptualiseringsplan wat dataversameling as een van sy onderafdelings het, vergelyk Wiegand (1998: 151). In gedrukte en tradisionele aanlyn woordeboeke strek die gebruiker se soektog na inhoud nie verder as die data wat vanuit 'n korpus in die databasis van die individuele woordeboek of 'n ander woordeboek in die betrokke woordeboekportaal neerslag gevind het nie. Die bestaande leksikografiese strukture laat nie verdere soektogte toe nie.

In die benutting van leksikografiese korpora, die saamstel van die databasis en die onttrekking van data aan die databasis om as aanduiders in woordeboekartikels of inskrywings in buitekomponente te dien, speel die leksiko-

graaf die bepalende rol. Die leksikograaf maak naamlik die keuse, weliswaar gebaseer op die vasgestelde behoeftes van 'n geïdentifiseerde teikengebruiker. In gedrukte woordeboeke is daar 'n vaste en statiese verhouding tussen gebruiker en leksikograaf met 'n eenrigtingverskaffing van data. Die gebruiker word met die data gekonfronteer wat die leksikograaf aanbied en het geen moontlikheid om alternatiewe data in die betrokke woordeboek te vind nie; al het die gebruiker in sommige aanlyn woordeboeke die geleentheid om regstreeks toegang te kry tot slegs sekere buitetekste of soeksones in 'n bepaalde artikel. Vir woordeboeke in die dinamiese aanlyn omgewing word 'n vergelykbare verhouding gehandhaaf van die leksikograaf wat die verskaffer van data is deur keuses uit 'n korpus te maak en met behulp van 'n spesifieke databasisprogram die data 'n neerslagplek in woordeboekartikels of in buitekomponente te laat vind. Die moontlikheid bestaan wel om aan een databasis verskillende woordeboeke te onttrek — vergelyk Bergenholtz en Bothma (2011: 64).

'n Oorsig oor strukture op sowel die aanbiedings- as die databasisvlak van aanlyn woordeboeke bevestig steeds die eenrigtingverhouding tussen leksikograaf en gebruiker; al het die gebruiker soms die moontlikheid om keuses te maak van wat hy of sy wil sien of nie wil sien nie.

5.2 Stoot- en trekbenaderings in kommunikasie

Müller-Spitzer (2013: 369) verwys na die onderskeid tussen 'n trekmedium en 'n stootmedium met onder meer boeke, radio en televisie wat tot die stootmedium behoort en die internet wat sy as 'n trekmedium beskou. Sy meld dat die internet 'n nuwe vorm van kommunikasie bied wat 'n vernuwende kombinasie van nuwe media is. Hierdie kenmerk van die internet het volgens Müller-Spitzer gevolg vir aanlyn woordeboeke as 'n tipe internetteks.

Dit is ter sake om hier kortliks iets oor die terme *trekmedium* en *stootmedium* te sê. In die inligtingswetenskap en rekenaarwetenskap word die stootmedium gesien as 'n medium waar die sender die identiteit van die ontvanger ken en data voorsien wat deur die ontvanger ontvang en dan aanvaar word al dan nie. Die volledige boodskap word deur die ontvanger ontvang voordat die ontvanger met die verwerking daarvan kan begin — vergelyk Duan et al. (s.j.). Die sender beheer watter boodskap gelewer moet word en ook wanneer dit gelewer moet word. Die ontvanger weet nie watter boodskap ontvang gaan word nie.

Volgens Duan et al. (s.j.) word die oordrag van 'n boodskap in 'n trekmedium deur die ontvanger in werking gestel deurdat die ontvanger dit van die sender vra. Die sender lewer die verlangde inligting wanneer dit gevra word deur die ontvanger. Hulle beweer dat die ontvanger in so 'n medium groter beheer oor die oordrag van die boodskap het en meer vertrouwe stel in die inhoud wat ontvang word in vergelyking met die stootmodel. Belangrik is dat die ontvanger die keuse het om sy/haar vlak van belang in die inhoud (asook die reputasie van die sender) te bepaal alvorens die spesifieke inhoud aangevra word. Dit is die sender se verantwoordelikheid om die data te berg en te beheer totdat die ontvanger gereed is om dit aan te vra.

Duan et al. (s.j.) beweer dat die een nadeel van 'n trekbenadering is dat die sender belas word met meer bestuur van die inhoud, maar die voordeel is dat dit die probleem van ongevraagde data-oordrag verminder.

Vanuit die perspektief van die ontvanger verg die stootmodel 'n passiewe en die trekmodel 'n aktiewe benadering tot dataverkryging.

Volgens Deolasee et al. (s.j.) het die trekgebaseerde benadering nie noodwendig altyd hoë betroubaarheid nie want die data verander vinnig. 'n Stootgebaseerde benadering het groter betroubaarheid ten opsigte van vinnig veranderende data, maar dit gebruik veel meer plek en is makliker aan mislukking blootgestel.

Vir die doel van hierdie artikel is die wesenlike onderskeid tussen die stoot- en die trekmedium dat in die stootbenadering in netwerkkommunikasie die sender besluit watter data voorsien moet word, terwyl in die trekbenadering die versoek om data van die ontvanger kom en dat hierdie data dan deur die sender voorsien word.

Die interdisiplinêre aard van die leksikografie noodsaak dit dat daar in die teorie-ontwerp asook in die praktyk ruim aandag gegee moet word aan ontwikkelinge in verwante dissiplines, onder meer die inligtingswetenskap. Die gerigtheid op die behoeftes van die gebruiker maar ook die teikengebruikersgroep se naslaanvaardighede en naslaanomgewingsvoorkeure moet eweneens bepalend wees in die ontwikkeling van die aanlyn leksikografie.

Alhoewel die stoot- en trekbenaderings nie vir die leksikografie ontwerp is nie en ook nie sonder meer daarop van toepassing gemaak kan word nie, is daar wel sekere grondliggende aspekte van hierdie benaderings wat leksikografies waardevol en prakties bruikbaar kan wees. Na analogie van hierdie benadering in die inligtingswetenskap en rekenaarwetenskap waar verskillende media in terme van 'n trek- al dan stootbenadering geklassifiseer kan word, kan daar in die leksikografie beweer word dat gedrukte woordeboeke asook tradisionele aanlyn woordeboeke grootliks binne die stootmedium val. Die leksikograaf, as sender, voorsien naamlik die data en die gebruiker, as ontvanger, het geen seggenskap oor die aard en omvang van die data wat beskikbaar gestel word nie. Selfs waar verskillende tipes data aan een databasis onttrek word volgens die profiel van die gebruiker, vergelyk Bergenholtz en Bothma (2011: 63), is die inligting wat die gebruiker uiteindelik ontvang 'n ont-trekking aan data waarop die leksikograaf besluit het.

5.3 Ten gunste van 'n datatrekkingstruktuur as deel van 'n oorkoepelende datastruktuur

Rundell (2016: 11) sê:

The Web and social media have created conditions which have overturned the older, top-down media model, where a small number of providers (whether journalists or lexicographers) delivered expertly-curated content to a large number of consumers. Consumers were for the most part passive: a handful of "Let-

ters to the Editor" of a newspaper (or of a dictionary) represented the limits of user-participation. In the new paradigm, ordinary individuals can make a contribution, and increasingly expect to do so.

Hierdie opmerking van Rundell sluit aan by sy betoog oor die groter betrokkenheid van woordeboekgebruikers by die saamstel van woordeboeke en die benutting van skarehulp ("crowdsourcing") in die daarstelling van gebruiker-geskepte inhoud. Groter betrokkenheid van gebruikers kan egter ook bereik word deur aan hulle die geleentheid te bied om inligting aan woordeboeksterne bronne te onttrek, veral waar hierdie bronne nie deel van 'n woordeboekportaal is nie en dus nie deel van data wat die leksikograaf vir 'n spesifieke naslaanprosedure vanuit 'n datastootbenadering gelewer het nie. Teenoor die benadering om 'n woordeboek as 'n stootmedium te beskou met die gepaardgaande situasie dat die leksikograaf as sender volledige beheer het oor die data wat aangebied word, vergelyk Duan et al. (s.j.), kan 'n benadering waar aanlyn woordeboeke aangepas word om 'n trekbenadering te hê tot 'n veel groter en aktiewer deelname van gebruikers aan die materiaalversamelingsfase van die leksikografiese proses en die onttrekking van inligting ter bevrediging van individuele gebruikersbehoefes lei.

Met verwysing na die stoot- en trekbenaderings in netwerkkommunikasie sê Müller-Spitzer (2013: 369) oor woordeboekgebruikers:

Therefore, users are both sender and receiver. They are active in 'pulling' data from the website, saving relevant parts, etc. Thus, the Internet provides a very new form of communication in general. It is communication in an innovative combination with new media.

En dan ook:

The process of pulling and, thus, representing lexicographic data according to a user request is essential for EDs (electronic dictionaries — RHG) and must be considered when the textual structures of EDs are being looked at.

Die skep van 'n leksikografiese datatrekkingstruktuur kan aanlyn woordeboeke binne die bestek van 'n *oorkoepelende datastruktuur* plaas waar gebruikers inligting aan 'n verskeidenheid woordeboeksterne bronne kan onttrek. Die leksikograaf as primêre sender het steeds die besluit oor watter bronne beskikbaar gestel word, maar het nie beheer oor die inligting wat gebruikers aan die data in hierdie bronne onttrek nie.

'n *Leksikografiese datatrekkingstruktuur* kan gesien word as 'n struktuur met 'n reeks geordende elemente om die stappe daar te stel wat 'n woordeboekgebruiker kan volg om vanuit 'n gegewe posisie in 'n bestaande aanlyn woordeboek (waar hierdie posisie óf 'n aanduiding of soeksone in 'n woordeboekartikel óf 'n artikeleksterne posisie, byvoorbeeld 'n inskrywing in 'n buitekomponent, kan wees) toegang te kry tot woordeboeksterne bronne waaraan die gebruiker inligting kan onttrek ter bevrediging van 'n bepaalde leksikografiese behoefte.

6. Bronne wat teikens van 'n datatrekkingstruktuur kan wees

6.1 Bestaande gebruike van die internet as korpus

Gouws en Tarp (2017: 391) voer aan:

Today we are in the middle of a new transition of the material and technological basis of lexicography with the introduction of new production tools and methods as well as new platforms and media for presenting the lexicographic product and the extensive use of corpora for the collection of material. The development and technological innovation are going faster than ever before. (...) We know the point of departure but we still only have a vague idea of where we will eventually arrive.

Die veranderende verhouding tussen leksikografie en tegnologie het 'n invloed op talle aspekte van die leksikografie en die leksikografiese proses; ook op die benutting van korpora. Ook die omvang van die toekomstige gebruik van leksikografiese korpora sal aangepas word. In sy bespreking van korpora noem Fuertes-Olivera (2012: 51) onder meer dat 'n leksikografiese korpus 'n versameling tekste is. Tarp en Fuertes-Olivera (2016: 277) noem dat die internet ook 'n versameling tekste is en ook as 'n tipe leksikografiese korpus beskou kan word. Leksikograwe vul reeds die gebruik van tradisionele korpora aan deur die internet as korpus te gebruik vir die vind van 'n verskeidenheid datatipes, vergelyk onder meer Tarp en Fuertes-Olivera (2016). Die leksikograaf is steeds die voorsiener van tersaaklike data en hierdie data word deur die leksikograaf gekies en as aanduiders in die woordeboek verstrek — dikwels met erkenning aan die internetbron waaraan die data onttrek is. Vergelyk in hierdie verband afbeelding 2 — die aanlyn woordeboek *lexico* se illustrasieprente wat foto's is wat op die internet gevind is. In hierdie geval is dit aan die gratis prentedatabank *pixelio.de* onttrek en hierdie bron word deur die leksikograaf erken.



Afbeelding 2: Skermskoot uit *lexico*

Tarp en Fuertes-Olivera (2016: 277) dui ook aan dat die internet as korpus op twee maniere deur die leksikograaf gebruik kan word. 'n Korpus kan naamlik saamgestel word uit tekste wat op die internet gevind is of die leksikograaf kan die internet regstreeks as korpus gebruik en data daaraan onttrek in 'n spesifieke leksikografiese bewerking. Die gebruiker het egter steeds geen sê in watter data uit die internet gebruik moet word nie; die aanlyn leksikografie bly 'n stootmedium.

Daar is 'n ander tipe interaktiewe verhouding tussen woordeboek en teks waar daar van 'n bepaalde teks na 'n ingeboude woordeboek beweeg kan word, soos in die geval van Amazon se Kindle e-boeke. Deur 'n woord in die Kindle-boek uit te lig, word die gebruiker outomaties na 'n geskakelde e-woordeboek herlei waar die gebruiker die betekenis van die uitgeligte woord kan kry sonder om die boek wat hulle lees te verlaat. Kindle bied tans die moontlikheid om vanuit die opskietvenster wat op enige bladsy van die boek verskyn toegang te kry tot sy eie woordeboek en, indien die leser meer hulp benodig, ook toegang tot Google te kry. Na aanleiding hiervan moet die moontlikheid in meer besonderhede vir die leksikografie ondersoek word oor hoe om die gebruiker van aanlyn woordeboeke regstreeks met 'n Google-soektog te verbind wat dan toegang tot die internet as databron verskaf. Dit bied ruim geleentheid vir verdere navorsing.

Hedendaagse tegnologie maak dit moontlik om met uitgebreide en verspreide soekmetodes spesifiek daardie data in 'n woordeboek te vind wat die gebruiker nodig het, vergelyk Bothma en Prinsloo (2013: 168). Dit sluit aan by Tarp (2009: 29) se siening van dinamiese data in woordeboekartikels wat uniek is vir elke soekprosedure volgens die spesifieke gebruiker en gebruiksituasie. 'n Verdere moontlikheid wat vir die leksikografie nog in meer besonderhede ondersoek moet word, is om die gebruiker regstreeks met 'n Google-soektog te verbind wat toegang tot die internet as databron verskaf.

6.2 Leksikografiese datatrekkingstrukture en 'n nuwe gebruik van die internet as korpus

Woordeboeke lê op die kontinuum van inligtingsbronne en die benutting van 'n aanlyn woordeboek moet 'n sikliese aard hê wat sowel stoot- as trekmoontlikhede aan die gebruikers bied. Die benutting van 'n datatrekkingstruktuur in aanlyn woordeboeke bied aan gebruikers die geleentheid om vanuit 'n bepaalde punt in die woordeboek 'n eksterne bron te raadpleeg om bykomende inligting. Leksikograwe kan byvoorbeeld 'n aanlyn woordeboek koppel aan 'n bron soos Wikipedia en die gebruiker wat op soek is na bykomende hulp ter bevrediging van 'n kognitiewe funksie kan vanuit die aanlyn woordeboek na Wikipedia gelei word en daar die bykomende inligting vind. 'n Klik op enige lemmateken in die aanlyn woordeboek kan die nodige skakel bied om data elders te gaan vind. Dit is 'n gedeeltelike trekbenadering want alhoewel die gebruiker kies waar hy/sy hulp nodig en alhoewel die leksikograaf nie weet

watter data die gebruiker aan die woordeboeksterne bron gaan onttrek nie, is die bron steeds deur die leksikograaf gekies en die data deur 'n stootbenadering tot die gebruiker se beskikking gestel.

'n Ware datatrekkingbenadering kom voor waar die internet as geheel deel van 'n aanlyn woordeboek se datastruktuur en die teiken van 'n datatrekkingstruktuur se toepassing is. Die tegniese aspekte hiervan word nie in hierdie artikel bespreek nie, maar die ideaal van 'n volwaardige datatrekkingstruktuur word voorgehou. Dit vereis 'n wisselwerking tussen die data wat woordeboek-intern aangebied word, die databasis van die woordeboek en die internet as woordeboeksterne bron, 'n buitekomponent in eie reg.

Wanneer enige aanduiding in 'n woordeboekartikel uitgelig word, verskyn daar 'n klein opskietspyskaart met 'n lysie datatipes wat die gebruiker mag nodig hê, byvoorbeeld koteksinskrywings, waaronder voorbeeldsinne en kollokasies, uitvoeriger betekenisparafrases, etimologiese inligting, uitspraak-inligting, konteks-inligting, ensovoorts. Dit is weliswaar datatipes wat die leksikograaf kies, maar oor die inhoud van die kategorieë het die leksikograaf geen seggenskap nie. Deur op 'n spesifieke spyskaartitem te klik, word die gebruiker deur 'n soekroete wat deur die datatrekkingstruktuur bepaal word na 'n teks in die internet gelei waar die nodige databystand verkry kan word. Dit vereis dat die databasis van die woordeboek op so 'n manier saamgestel word dat die gebruiker via die klik van 'n spyskaartitem in die regte rigting na die internet gestuur word. Deur die stel geordende fases van die datatrekkingstruktuur kan 'n woordeboekgebruiker toegang tot 'n veel wyer data-aanbod kry as wat die woordeboekinterne dataverspreidingstruktuur kan voorsien.

Die datatrekkingstruktuur wat die internet as teiken het, verander die aard van die betrokke aanlyn woordeboek as netwerkinstrument en maak daarvan 'n geïntegreerde inligtingswerktuig. Die woordeboekgebruiker word 'n aktiewe deelnemer wat besluitnemingsmag het oor die data waarna hy/sy wil soek as deel van die uitvoering van 'n voortgaande leksikografiese proses. So 'n proses lei tot 'n ander perspektief op die gesag van woordeboeke. Die leksikograaf het geen beheer oor die aard van die data wat die gebruiker aan die internet onttrek nie. Die gebruiker dra die verantwoordelikheid daarvoor om die gehalte van hierdie data te beoordeel.

7. Ten slotte

Metaleksikografiese navorsing moet aandag gee aan voorstelle vir die aanpassing van leksikografiese strukture om die aanlyn omgewing optimaal te kan benut. In hierdie verband is interdisciplinêre samewerking nodig; ook om seker te maak dat die voorstelle van metaleksikografiese tegnies uitvoerbaar is. Die idee van 'n datatrekkingmedium is reeds gevestig in die inligtings- en rekenaarwetenskap. Leksikografiese moet samewerking gee om 'n datatrekkingbenadering in aanlyn woordeboeke moontlik te maak.

Die daarstelling van 'n geordende stel opeenvolgende stappe wat 'n data-

trekkingstruktuur kan skep en die toepassing daarvan kan aan woordeboekgebruikers 'n aktiewe deelname bied om nie net leksikografiese data vir die leksikograaf te voorsien soos in die geval van skarehulp nie, maar om self data uit die internet as leksikografiese data te ontgin en wel op 'n manier wat aan spesifieke behoeftes van spesifieke gebruikers in spesifieke gebruikssituasies voldoen. So 'n trekbenadering sal ook help verseker dat 'n aanlyn woordeboek nie gebuk gaan onder die las van data-oormoed nie (vergelyk Gouws en Tarp 2017).

'n Datatrekkingstruktuur bring ook mee dat McArthur (1986) se siening van woordeboeke as "houers van kennis" minder relevant raak, aangesien die kennis oor verskillende houers — nie net woordeboeke nie — versprei word.

Acknowledgement

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant specific unique reference number (UID) 85434). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author(s), and that the NRF accepts no liability whatsoever in this regard.

Bronnelys

- Atkins, Beryl T.S. en Michael Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Bergenholtz, Henning en Theo J.D. Bothma.** 2011. Needs-adapted Data Presentation in e-Information Tools. *Lexikos* 21: 53-77.
- Bergenholtz, Henning, Sven Tarp en Herbert Ernst Wiegand.** 1999. Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. Hoffmann, Lothar et al. (Reds.). 1999. *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft/Languages for Special Purposes. An International Handbook of Special-Language and Terminology Research, Bd./Vol. 2: 1762-1832*. Berlyn: Walter de Gruyter.
- Bothma, Theo J.D. en Daniel J. Prinsloo.** 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries. *Lexicographica* 29: 165-198.
- Deolasee, Pavan et al.** s.j. *Adaptive Push-Pull: Disseminating Dynamic Web Data*. <http://www-ccs.cs.umass.edu/~krithi/web/WWW10/www10/> (Geraadpleeg op 10 Mei 2018).
- Duan, Zhenhai, Kartik Gopalan en Yingfei Dong.** s.j. *Push vs. Pull: Implications of Protocol Design on Controlling Unwanted Traffic*. <https://pdfs.semanticscholar.org/6f63/d37b4f8dd655e3594185e74daf4689f55aa1.pdf> (Geraadpleeg op 10 Mei 2018).
- ellexico*: Online Wörterbuch zur deutschen Gegenwartssprache. Saamgestel deur die Institut für Deutsche Sprache, Mannheim. <http://www.owid.de/wb/lexiko/start.html>.
- Engelberg, Stefan en Carolin Müller-Spitzer.** 2013. Dictionary Portals. Gouws, Rufus H. et al. (Reds). 2013: 1023-1035.

- Finch, Jeremy.** 2015. *What Is Generation Z, And What Does It Want?* <http://www.fastcoexist.com/3045317/what-is-generation-z-and-what-does-it-want> (Geraadpleeg op 28 Mei 2015).
- Fuertes-Olivera, Pedro A.** 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28: 49-70.
- Gouws, Rufus H.** 2011. Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. Fuertes-Olivera, Pedro A. en Henning Bergenholtz (Reds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 17-29. Londen/New York: Continuum.
- Gouws, Rufus H.** 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.
- Gouws, Rufus H.** 2014a. Makrostruktuuraanpassings vanaf gedrukte na e-woordeboeke. *Tydskrif vir Geesteswetenskappe* 54(3): 481-504.
- Gouws, Rufus H.** 2014b. Expanding the Notion of Addressing Relations. *Lexicography* 1(2): 159-184.
- Gouws, Rufus H.** 2017. La sociedad digital y los diccionarios. Domínguez Vázquez, María José en María Teresa Sanmarco Bande (Reds.). 2017. *Lexicografía y didáctica*: 17-34. Frankfurt: Peter Lang.
- Gouws, Rufus H.** 2018. Dictionaries and Access. Fuertes-Olivera, Pedro A. (Red.). 2018. *The Routledge Handbook of Lexicography*: 43-58. Londen: Routledge.
- Gouws, Rufus H.** 2018a. Accessibility, Access Structures and Access Procedures. Jesenšek, Vida en Milka Enčeva (Reds.). 2018. *Wörterbuchstrukturen zwischen Theorie und Praxis*: 35-56. Berlyn: De Gruyter.
- Gouws, Rufus H.** 2018b. Internet Lexicography in the 21st Century. Engelberg, Stefan et al. (Reds.). *Wortschatz: Theorie, Empirie, Dokumentation*: 215-236. Berlyn: De Gruyter.
- Gouws, Rufus H. et al. (Reds.).** 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlyn/New York: De Gruyter.
- Gouws, Rufus H. en Sven Tarp.** 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.
- Hartmann, Reinhard R.K.** 1989. Sociology of the Dictionary User: Hypotheses and Empirical Studies. Hausmann, Franz J. et al. (Reds.). 1989-1991: 102-111.
- Hartmann, Reinhard R.K.** 2001. *Teaching and Researching Lexicography*. Londen: Pearson Education.
- Hausmann, Franz J. et al. (Reds.).** 1998-1991. *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*. Berlyn: De Gruyter.
- Kammerer, Matthias en Herbert Ernst Wiegand.** 1998: Über die textuelle Rahmenstruktur von Printwörterbüchern. Präzisierungen und weiterführende Überlegungen. *Lexicographica* 14: 224-238.
- Klosa, Annette en Rufus H. Gouws.** 2015. Outer Features in e-Dictionaries. *Lexicographica* 31: 142-172.
- McArthur, Tom.** 1986. *Worlds of Reference: Lexicography, Learning, and Language from the Clay Tablet to the Computer*. Cambridge: Cambridge University Press.
- Müller-Spitzer, Carolin.** 2013. Textual Structures in Electronic Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 367-381.
- Parker, Phil.** 2013. <http://www.sec-ed.co.uk/blog/how-generation-z-is-different> (Geraadpleeg Mei 2015).
- Rundell, Michael.** 2012. 'It Works in Practice but Will it Work in Theory?' The Uneasy Relationship between Lexicography and Matters Theoretical. Vatvedt Fjeld, R. en J. Matilde Torjusen (Reds.).

2012. *Proceedings of the 15th EURALEX International Congress, 7–11 August 2012, Oslo*: 47-92. Oslo: Departement Linguistiek en Skandinawiese Studies, Universiteit van Oslo.
- Rundell, Michael.** 2016. (Ongepubliseer.) *Dictionaries and Crowdsourcing, Wikis and User-generated Content*.
- Tarp, Sven.** 2009. Beyond Lexicography: New Visions and Challenges in the Information Age. Bergenholtz, Henning, Sandro Nielsen en Sven Tarp (Reds.). 2009. *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*: 17-32. Frankfurt: Peter Lang.
- Tarp, Sven en 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.
- Wiegand, Herbert Ernst.** 1977. Nachdenken über Wörterbücher: Aktuelle Probleme. Drosdowski, Günther, Helmut Henne en Herbert Ernst Wiegand (Reds.). 1977. *Nachdenken über Wörterbücher*: 51-102. Mannheim/Wenen/Zürich: Bibliographisches Institut.
- Wiegand, Herbert Ernst.** 1989. Der gegenwärtige Status der Lexikographie und ihr Verhältnis zu anderen Disziplinen. Hausmann, Franz J. et al. (Reds.). 1989-1991: 246-280.
- Wiegand, Herbert Ernst.** 1996. A Theory of Lexicographic Texts. An Overview. *SA Journal of Linguistics* 14(4): 134-149.
- Wiegand, Herbert Ernst.** 1998. *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlyn/New York: De Gruyter.
- Wiegand, Herbert Ernst.** 2005. Über die Datenakzessivität in Printwörterbüchern. Einblicke in neuere Entwicklungen einer Theorie der Wörterbuchform. *Lexikos* 15: 196-230.
- Wiegand, Herbert Ernst.** 2010. Zur Methodologie der Systematischen Wörterbuchforschung: Ausgewählte Untersuchungs- und Darstellungsmethoden für die Wörterbuchform. *Lexicographica* 26: 249-330.
- Wiegand, Herbert Ernst en Sandra Beer.** 2013. Textual Architectures in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 253-273.
- Wiegand, Herbert Ernst, Sandra Beer en Rufus H. Gouws.** 2013. Textual Structures in Printed Dictionaries. An Overview. Gouws, Rufus H. et al. (Reds.). 2013: 31-73.
- Wiegand, Herbert Ernst, Ilse Feinauer en Rufus H. Gouws.** 2013. Types of Dictionary Articles in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 314-366.
- Wiegand, Herbert Ernst en Rufus H. Gouws.** 2013. Macrostructures in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 73-110.
- Wiegand, Herbert Ernst en Rufus H. Gouws.** 2013a. Addressing and Addressing Structures in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 273-314.
- Wiegand, Herbert Ernst en Maria Smit.** 2013. Microstructures in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 149-214.
- Wiegand, Herbert Ernst en Maria Smit.** 2013a. Mediostructures in Printed Dictionaries. Gouws, Rufus H. et al. (Reds.). 2013: 214-253.

A Lexicographic Approach to Teaching the English Article System: Help or Hindrance?

Sugene Kim, *Department of English Studies, Nagoya University of
Commerce & Business, Nisshin, Japan (sugene_kim@nucba.ac.jp)*

Abstract: This article reports on changes in EFL learners' article choice performance before and after receiving lessons on the main rules applicable to article usage combined with dictionary consultation guidance. A sample of 43 Korean college students undertook the same forced-choice elicitation task once as a diagnostic test and again as a post-intervention test at three-month intervals. Unlike the diagnostic test, in which the participants were only asked to choose the correct articles, the post-intervention test asked them to give written accounts of their decision-making procedures as well. The analyses of the diagnostic test results, specifically the items requiring the indefinite article or the zero article, demonstrated EFL learners' struggle with indeterminate nominal numbers, underlining the importance of clear lexicographic treatment of such information. Further, the post-intervention test and the written think-aloud data analyses suggested that although using a bilingualised dictionary for nominal countability is useful in general, dictionary consultation can sometimes impede users from using articles correctly. Specific problem areas are discussed.

Keywords: ENGLISH ARTICLE SYSTEM, NOMINAL COUNTABILITY, ARTICLE USE, BILINGUALISED DICTIONARY, KOREAN EFL LEARNERS

Opsomming: 'n Leksikografiese benadering tot die onderrig van die Engelse lidwoordstelsel: 'n Hulp of 'n hindernis? In hierdie artikel word verslag gedoen oor veranderinge in EVT-leerders se keuse van lidwoorde voor en nadat hulle lesse oor die hoofreëls wat van toepassing is op lidwoordgebruik asook leiding oor die raadpleging van woordeboeke ontvang het. 'n Steekproef van 43 Koreaanse kollegestudente het dieselfde opdrag uitgevoer waartydens gedwonge keuses ontlok is, een keer as 'n diagnostiese toets en weer as 'n postintervensietoets drie maande later. Anders as in die diagnostiese toets, waarin die deelnemers slegs die korrekte lidwoorde moes kies, is hulle in die postintervensietoets ook gevra om 'n geskrewe weergawe te gee van die besluitnemingsprosesse wat hulle gevolg het. Die ontleding van die diagnostiese toetsresultate, spesifiek die items wat die onbepaalde lidwoord of die zero-lidwoord vereis het, het getoon dat EVT-leerders sukkel met onbepaalde naamwoordgetalle, wat die belangrikheid van duidelike leksikografiese hantering van sodanige inligting beklemtoon het. Die postintervensietoets en die ontleding van die geskrewe hardop-dink-data het daarop gedui dat, alhoewel die gebruik van 'n verklarende woordeboek met vertalings oor die algemeen nuttig is vir nominale telbaarheid, die raadpleging van 'n woordeboek soms gebruikers kan verhinder om lid-

woorde korrek te gebruik. Spesifieke probleemareas word bespreek.

Sleutelwoorde: LIDWOORDSTELSEL IN ENGELS, NOMINALE TELBAARHEID, LIDWOORDGEBRUIK, VERKLARENDE WOORDEBOEK MET VERTALINGS, KOREAANSE EVTLEERDERS

1. Introduction

Correct article usage is difficult for learners of English as a second language (ESL) or English as a foreign language (EFL) to master, especially when their mother tongue (L1) does not contain the corresponding function system (García Mayo 2008, Ionin et al. 2008, Mizuno 1999). This has previously been observed in the literature and consistently supported by empirical evidence. Indeed, it is an indisputable fact that the rules governing English article usage are particularly unwieldy, with many exceptions and idiosyncrasies, so that article errors are produced even by highly advanced learners (Lennon 1991, Leroux and Kendall 2018, White 2003). For these reasons, some researchers have even claimed that teaching the article system effectively is an elusive goal (Butler 2002). Working as an EFL writing instructor in Korea for over a decade, the author has also heard students' complaints about written corrective feedback in which their native English-speaking teachers added what seemed to the students to be an unlikely *a(n)* or replaced what the students thought should clearly be *the* with *a(n)*.

A reasonable starting point for correct article usage is to identify the numeral aspects of a noun (Butler 2002, Master 1997), and it is exactly at this point that the problem begins. Of course, some count nouns such as *apple* or *pencil* are physically countable so that we can easily count their number on our fingers. By contrast, the numbers of other nouns such as *atmosphere* or the viral infection *cold* are not so obvious, and these nouns are often paired with deviant article choices in EFL writing. Because they lack a clear understanding that the notion of countability is supposed to be understood in a grammatical — not mathematical — sense, many EFL learners try to determine the countability of the noun in question by visualizing themselves finger-counting the "item," rather than by looking for the information in a dictionary (Xue 2010). Hence, they almost never put *an* before *atmosphere* because *an atmosphere* sounds almost as peculiar as *two atmospheres*. In a sense, the term "count(able) noun" itself can be considered misleading, as there are countless count nouns that we simply do not count. In addition, similar to other classifier languages such as Chinese and Japanese, Korean neither distinguishes between count and noncount nouns nor draws grammatical number distinction. Therefore, correct article usage can be extremely difficult for Korean EFL learners, whose L1 lacks not only an article system but also a singular–plural morphology.

The same holds true for some noncount nouns such as *money*. In theory, it is a mass noun, which is uncountable; in reality, we count money without reservation. Since bank tellers behind a counter can sometimes *miscount* customer

deposits, most banks currently use money *counting* machines that *count* money rapidly. In such circumstances, how can anyone communicate to EFL learners that *money* is, in fact, a noncount noun and thus should not be counted? Who can possibly teach the fact that *cold* is countable, while *flu* is not, when we cannot even confidently identify which of the two illnesses we are suffering from? The unfortunate truth is that students will continue struggling unless they are urged to stop creating a mental image of themselves counting things one by one. Rather, they should be explicitly instructed to turn to dictionaries for nominal countability because grammatical countability cannot be accurately detected by intuition (Butler 2002).

While dictionaries are primary sources of reference for the numeral features of a noun, there are doubts about whether "the present lexicographic practice of indicating ... nominal countability in learner's dictionaries is transparent enough" (Xue 2010: 541) to help learners "acquire one of the hardest grammatical features of the English language" (Miller 2006: 435). Xue (2010), for instance, noted that the absence of indicating articles or quantifiers used before a noun limits the effectiveness of learner's dictionaries for production purposes. She pinpointed "equivocal and discrepant indication in the noun countability features" and "inefficient exemplification" as the main causes of the difficulties that Chinese learners of English face in their use of the numeral inflection of a noun. Similarly, Chan (2017a) contended that learners often misinterpret dictionary information, which consequently leads to article use errors. She identified a few sources of problems with *Oxford Advanced Learner's English-Chinese Dictionary 8* in helping Hong Kong Chinese ESL learners determine the countability of English nouns, such as "L1 translation of the corresponding English phrase with different syntactic requirements" and the "provision of insufficient information about noun countability." Tsang (2017) posits that learners' difficulty in nominal countability has not received enough attention in applied linguistics, although countability and plural marking are among the most challenging topics for both ESL and EFL learners (Celce-Murcia and Larsen-Freeman 1999, Han et al. 2006). Furthermore, because a substantial number of English nouns can be used in both count and noncount contexts, researchers such as Allan (1980) and Wisniewski et al. (2003) argue that the traditional practice of merely labelling nouns as either countable or uncountable is not adequate.

Given that an English sentence (except for an imperative) cannot be constructed without a noun (which can be a gerund), the teaching of correct article usage is urgently needed (Chan 2016). Especially in choosing an article for a noun phrase that "is non-specific" (*Oxford Dictionaries*), the choice between the indefinite article and the zero article is determined by the lexical classification of the target noun as a count or noncount noun (Yoon 1993). Whether learners can successfully extract a noun's numeral features from a dictionary is an "important preliminary to correct use of articles" (Celce-Murcia and Larsen-Freeman 1999: 273), but there is an overall dearth of lexicographic research investigating English learners' countability judgement processes and associated

article use. To bridge this research gap, this study explores how Korean EFL learners use a bilingualised dictionary to retrieve the needed nominal countability information and what difficulties they encounter along the way.

2. The study

2.1 Participants

The participants included 43 Korean college students enrolled at a major research university in Seoul, Republic of Korea. They were from two English courses — one offered for humanities majors and the other for education majors — required for all first-year students. The class met for 75 minutes twice per week over a 15-week semester. The participants were 18–20 years old and had learned English at both elementary and secondary schools and private language institutions for an average of approximately 10.5 years by the time they took the course. English was a foreign language for all the participants, and none of them had lived in English-dominant countries for more than one year. Judging from the scores on the school-administered English proficiency test, the participants could be collectively described as intermediate to upper-intermediate learners of English. At the beginning of the semester, they were informed and consented in writing to the possibility that their assignments and test papers would be analyzed for research and teaching-improvement purposes and part of them might be presented in a published paper, with their personal information protected.

2.2 Instrument

A 23-item forced-choice elicitation task (Gass and Selinker 2001) targeting the use of the English articles — *a(n)*, *the*, and zero (\emptyset) — was created to be used as both a diagnostic test (pre-intervention test) and a post-intervention test. The task contained sentences from various sources such as online newspaper and magazine articles; Ionin et al. (2004); and Yoo (2004) retrieved from MIT OpenCourseWare, a web-based publication of MIT course content. Care was taken to ensure that an approximately even number of items were sought for the indefinite article, the zero article, and the definite article for different reference types — anaphoric¹, associative anaphoric², and cataphoric³. Since revisions were made to the original sentences by shortening the sentence length or changing the sentence structure, four English native-speaking professors — all Ph.D. holders in applied linguistics or in English literature — evaluated the naturalness of the revised sentences. In addition, the professors were asked to choose the most natural-sounding article for each sentence to double-check the correctness of article usage and to determine whether alternate answers were possible. Of the 30 initially prepared test items, seven were removed because there were discrepancies among the professors regarding the use of an article with the target

noun in the given context. The finalized elicitation task is presented in Appendix 1, with the correct answers marked in bold.

2.3 Procedure

This study employed a one-group pre-intervention test–post-intervention test design. Since the same instrument was used for both tests, the post-intervention test was administered approximately three months after the pre-intervention test to minimize practice effects (Bachman 1990). To estimate the participants' current understanding of English article usage, the participants were pre-tested in Week 3 without being allowed to use a dictionary. In Week 14, instruction on the English article system was provided for two consecutive sessions, after which the post-intervention test was given as a take-home task. The course curriculum other than Week 14 was framed with an emphasis on the general features of academic reading and writing, occasionally incorporating narrowly focused mini-lessons on grammar (Ferris and Hedgcock 2005) in cases when the grammar point was directly relevant to the class content of the week (e.g. "parallel structure" for writing stated, or direct, thesis statements).

In Week 14, English article instruction was given using the chapter about the main rules of English article usage in *Top 20: Great Grammar for Great Writing* (Folse et al. 2008) — abbreviated as *Top 20* hereafter — which explains the rules based broadly on "nominal countability" and "definiteness." One week before this instruction, the students were told to read the chapter and work on three (of nine) exercise questions in it — one exercise each for the indefinite, definite, and zero articles — to ensure more class time for instruction and guided practice.

The class in Week 14 took place in a computer lab in which each student could work on a computer independently. During the lesson, the instructor first explained the importance of checking the nominal countability given for each sense of the target noun, as it can easily change according to the meaning in a given context. She then demonstrated how to consult a dictionary for the countability features of a target noun, using exercise questions (other than the assigned ones) from *Top 20*. The online *Naver Dictionary* was adopted for the instruction and subsequent in-class practice because it is by far the most widely used bilingualised dictionary among Korean college students. By default, the *Naver Dictionary* provides information retrieved from the *Oxford Advanced Learner's English–Korean Dictionary*, followed by the English–English definition retrieved from the *Collins COBUILD Advanced Learner's English Dictionary*.

Then, the instructor explained the concept of "(in)definiteness" by employing Master's (1990) binary schema, in which Master reduced the four features required to correctly determine the article — definiteness, specificity, countability, and number — and proposed a simplified dichotomy based on classification ([–definite, ±specific], *a(n)* or \emptyset) and identification ([+definite, ±specific], *the*) as an overarching framework. While the "classification/identification dichotomy

is invoked first, followed by the count/noncount dichotomy" in Master's (1990: 470) schema, the reversed order was adopted in this study because while countability status *can* be checked in a dictionary, the classification vs. identification distinction is not always clear even to English native speakers, let alone EFL learners (Bickerton 1981, Miller 2005). Thus, it was assumed that applying the reverse order would make it easier for the participants to complete the first stage and proceed to the next.

After imparting the lessons that cover the usage rules for each article in relation to nominal countability and definiteness, the instructor had the students form groups of three or four and check their answers for the assigned exercise questions with one another. While the students were engaged in these group discussions, checking the countability status of the target noun if necessary, the instructor circulated around the classroom to answer questions when requested. When the group discussions had been completed, the instructor provided the answer sheet and reviewed the key usage rules for the whole class.

After the second instruction session had been completed, the students were given a take-home post-intervention test, for which they were requested to consult dictionaries unless they were completely certain about the countability feature of a target noun in the given context. Drawing on the view of Ericsson and Simon (1984: 11) that learning is a cognitive process that can be seen as "a sequence of internal states successively transformed by a series of information processes," the students were additionally required to indicate the procedure they followed in choosing the answer in the same manner as they would do the think-aloud protocol, except that they provided written — not verbal — accounts of the thought processes between the introduction of a task to the final product. To assist the students in developing the ability to perform think-alouds independently, the instructor gave demonstrations using one exercise set from *Top 20* consisting of four questions. The demonstrations were given in both Korean and English, and the students were informed that they could choose either language. The assigned task was collected one week later.

2.4 Data analysis

To measure whether giving lessons on the main rules for article usage combined with dictionary consultation guidance facilitated Korean EFL learners' ability to use the English articles correctly, the participants' pre- and post-intervention tests were scored by checking whether the answers given were correct. Then, the post-intervention test scores were compared with the pre-intervention test scores by means of a paired-samples *t*-test. The statistical analysis was performed at a significance level of .05. In addition, to examine what difficulties the participants encountered in the use of the English articles and what specifically caused them to make correct or incorrect article choices, their written think-aloud data were analyzed. Since all participants used Korean in describing their decision-making procedures, the author translated the data into Eng-

lish verbatim. The comments were categorized using thematic analysis and then ranked by frequency.

3. Results

The paired-samples *t*-test result showed that the mean correct answer rate increased from 65.2% on the pre-intervention test to 82.6% on the post-intervention test. Unsurprisingly, the *p*-value was far lower than the pre-selected alpha ($p < .001$), confirming that the students had made significant improvements in using the articles correctly after receiving the instruction. Although the overall mean score improved meaningfully on the post-intervention test, participants' performance level differed sharply depending on "what purpose the noun is used for (i.e. classification vs. identification)" and "whether required countability information (RCI) is provided for the target noun." Table 1 summarizes the participants' performance according to the nature of the target noun defined by the purpose, the lexicographic treatment of RCI for the nouns used for classification purposes, and the reference types of the nouns used for identification purposes.

Table 1: Mean correct answer rates for article use purposes and reference types

Purpose (Definiteness)	Lexicographic treatment of RCI / Reference type	Item number	Test	Mean correct answer rate	
classification (-definite)	RCI provided	4, 5, 20	pre-intervention	65.8%	
			post-intervention	95.3%	
	RCI not provided	2, 3, 7, 8, 16, 18, 19, 21, 23	pre-intervention	32.5%	
			post-intervention	62.0%	
			subtotal	pre-intervention	40.8%
			post-intervention	70.3%	
identification (+definite)	anaphoric	9, 11, 13	pre-intervention	100.0%	
			post-intervention	100.0%	
	associative anaphoric	1, 10, 12, 14	pre-intervention	96.5%	
			post-intervention	100.0%	
	cataphoric	6, 15, 17, 22	pre-intervention	80.2%	
			post-intervention	89.0%	
	subtotal	pre-intervention	91.5%		
	post-intervention	96.0%			
total	pre-intervention	65.2%			
post-intervention	82.6%				

As is apparent from the mean correct answer rates shown in Table 1, the correct article choice for the nouns used for identification purposes seemed quite straightforward, as the mean correct answer rates for both the pre- and post-intervention tests were as high as 91.5% and 96.0%, respectively. Specifically, for the nouns used for anaphoric or associative anaphoric reference, almost all the participants chose the correct answers on both tests. The mean correct answer rates for the nouns used for cataphoric reference were slightly lower —

80.3% on the pre-intervention test and 89.0% on the post-intervention test. Since the participants' overall post-intervention test performance on definite article use was fairly high, the items in this category are not discussed further, except when the participants misunderstood the given discourse context as [-definite] and the lexicographic presentation of the target noun countability caused an article selection error.

For the nouns used for classification purposes, the mean correct answer rate for the pre-intervention test was only slightly over 40%, suggesting EFL learners' difficulties with the indeterminate, variable numeral features of a noun (Butler 2002, Wisniewski et al. 2003, Xue 2010). Although the post-intervention test mean score improved significantly by almost 30% (from 40.8% to 70.3%), the results revealed a substantial post-intervention test performance gap depending on whether the required countability status of the target noun is provided in the dictionary (refer to the discussion section for details). As is illustrated in Table 1, while the mean correct answer rate for the items with the RCI provided was as high as 95.3%, the mean of those without its proper lexicographic treatment averaged only 62.0%.

Overall, the findings of the study suggest that teaching lessons on English article usage combined with dictionary consultation guidance can facilitate EFL learners' ability to use English articles correctly. Nonetheless, the participants continued struggling with correct article use in certain contexts. Causes of weak performance are discussed in the next section.

4. Discussion

The analyses of the participants' written think-aloud data shed light on possible sources of the difficulties that Korean EFL learners encounter when choosing the right article for a noun used for classification purposes, and the findings reveal five main factors relating to current lexicographic practice with nominal countability presentation. A detailed account of each is given in the sub-sections.

4.1 Equivocal criterion for dividing senses with opposite countability features

A vast majority of the participants reported experiencing difficulty distinguishing between at least two senses with opposite countability for Items 2 (*exercise*), 3 (*business*), 16 (*improvement*), and 21 (*distinction*). The participants' task performance on these items is delineated in two sub-categories below.

4.1.1 Provision of identical English synonyms for multiple senses

For Item 2 (*Any exercise is a / the / Ø good **exercise**, but when it comes to losing weight, nothing can beat running*), the entry for the target noun *exercise* provides

identical English synonyms (*activity/movements*) for both senses 1 and 2, with one being uncountable and the other being countable (see Figure 1). Furthermore, the Korean translations are identical except for the modifying information placed in the parentheses — "exercise (for physical, mental health)" (sense 1) and "exercise [physical exercise] (comprising a series of movements); practice [training] (for sharpening skills)" (sense 2), to translate them into English.

1. ACTIVITY/MOVEMENTS | [U] (신체적·정신적 건강을 위한) 운동

Swimming is good exercise. 

수영은 좋은 운동이다.

I don't get much exercise sitting in the office all day. 

나는 하루 종일 사무실에 앉아 있어서 운동을 별로 하지 않는다.

The mind needs exercise as well as the body. 

몸뿐만 아니라 마음도 운동이 필요하다.

vigorous/gentle exercise 

활발한/부드러운 운동


to take exercise 

운동을 하다

2. ACTIVITY/MOVEMENTS | [C] (일련의 동작들로 이뤄진) 운동[체조]; (기량을 닦기 위한) 연습[훈련]

breathing/relaxation/stretching exercises 

숨쉬기 운동/정리 운동/맨손 체조

exercises for the piano 

피아노 연습

Repeat the exercise ten times on each leg. 

그 운동을 각 다리에 열 번씩 하라.

Figure 1: Senses 1 and 2 of *exercise*, noun, from the *Naver Dictionary*

Nevertheless, 81.3% of the participants successfully chose the correct article Ø. Most of them commented in their written protocols that although the examples following the sense differentiation were by no means distinguishable from each other in terms of a syntactic structure and semantic meaning, they could decide which sense to choose thanks to the similarly constructed example following the first sense. One participant commented:

I can't tell the difference between senses 1 and 2. However, while one is [U], the other is [C]. Embarrassing. I check examples, looking for a hint. The construction of the first example under the first sense, *Swimming is good exercise*, is almost identical to the question sentence. I pick Ø as in the example.

Considering that EFL learners tend to have a fixed notion that abstract nouns are invariably uncountable (Butler 2002, Master 1994), it seems necessary to direct their attention to the noncount-to-count shift that many abstract nouns undergo (Celce-Murcia and Larsen-Freeman 1999, Greenbaum and Nelson 2009, Master 1988).

Such two-way nouns are generally countable in cases where they denote an instantiated concept (Huddleston and Pullum 2002), as in *You don't meet a courage like hers every day* or *You'll need a good knowledge of English for that job*. Hence, it might be beneficial to include usage notes of "when abstract nouns can be used as countable or uncountable" so that learners can make an informed decision about which article to use.

4.1.2 Provision of identical or interchangeable Korean translations for multiple senses

For Item 16 (*I believe there is room for an / the / Ø improvement in every sportsman*), most participants seem to have encountered a similar type of difficulty: The Korean translations of senses 1 and 2 are interchangeable but have an opposite countability status (see Figure 2). While "few equivalent words in two languages have precisely the same meaning" (Chan 2017a: 201), the corresponding Korean translations — "향상" ("improvement") for the first sense and "개선, 호전" ("improvement, improvement") for the second — are provided without explicit guidance, which inevitably constituted a source of trouble.

1. [U] ~ (in/on/to sth) 향상

Sales figures continue to show signs of improvement. 

판매 수치가 계속 향상되는 흔적을 보이고 있다.

There is still room for improvement in your work. 

당신 작품[작업]은 아직 향상 될 여지가 있다.

We expect to see further improvement over the coming year. 

다가오는 해에는 더 많은 향상을 보게 되기를 기대합니다.

2. [C] ~ (in/on sth) 개선, 호전

a significant/substantial/dramatic improvement 

중대한/실질적인/극적인 개선

a slight/steady improvement 


약간의/꾸준한 호전

an improvement in Anglo-German relations 

영-독 관계 개선

This is a great improvement on your previous work. 

이것은 당신의 이전 작품[작업]보다 크게 개선된 것이다.

improvements to the bus service 

버스 운행의 질 개선

Figure 2: Senses 1 and 2 of *improvement*, noun, from the *Naver Dictionary*

Amid the indistinguishable Korean translations, however, 79.0% chose the correct article \emptyset , with the majority commenting that they took advantage of the similarly phrased example (*There is still room for improvement in your work*) under the first sense. To quote one participant who correctly chose \emptyset :

I check both senses carefully. They look the same, but they are divided into two separate senses, not one with a [U, C] code. I read examples carefully. There's a sentence under the first [sense] including the same phrase "room for improvement." I choose " \emptyset ," not "a," solely because of this example.

While translations in a bilingualised dictionary are usually regarded as preferably insertable (Gauton 2008) and highly useful for decoding purposes (Cowie 1999), the results of this study suggest that they are "not equally useful for encoding" (Chan 2017a: 201) due to possible syntactic discrepancies between the learners' L1 and the target language. As is shown in Figure 2, the provision of the syntactic specifications — "~ in/on/to sth" for the first sense and "~ in/on sth" for the second — is not very useful, not only because they overlap for the most part but also because the provided specifications are not comprehensive. The last example under the second sense (*improvements to the bus service*) shows that, just like the uncountable *improvement* (sense 1), the countable counterpart (sense 2) can also be followed by *to*, although it is not specified in the sub-entry.

Similarly, for Item 21 (*It is important to draw a distinction between what you want and what you need*), the entry for the target noun *distinction* provides two senses — senses 1 (차이[대조]) ("difference[contrast]") and 4 (구분, 차별) ("distinction, discrimination") — that are immediate synonyms in Korean but are specified with an opposite countability status (see Figure 3).

1. [C] ~ (between A and B) (특히 비슷하거나 관련이 있는 것들 사이의 뚜렷한) 차이[대조]

distinctions between traditional and modern societies

전통 사회와 현대 사회 사이의 차이

Philosophers did not use to make a distinction between arts and science.

예전에는 철학자들이 예술과 과학 사이에 차이를 두지 [예술과 과학을 구별하지] 않았다.

We need to draw a distinction between the two events.

우리는 그 두 사건 사이에 차이를 둘 [그 두 사건을 구별할] 필요가 있다.

4. [U] 구분, 차별

The new law makes no distinction between adults and children.

그 새 법률은 어른과 아이를 구분하지 않는다.

All groups are entitled to this money without distinction.

모든 집단이 아무런 차별 없이 이 돈을 받을 권리가 있다.

Figure 3: Senses 1 and 4 of *distinction*, noun, from the *Naver Dictionary*

Although the first sense provides the syntactic structure "~ between A and B," the very first example under the fourth sense takes the same construction (*The new law makes no distinction between adults and children*). Examples are generally considered "an effective way to demonstrate syntactic behaviour [of a noun] in context" (Xue 2010: 549), but perusing the examples following the fourth sense added to the confusion in this case. However, 62.7% of the participants managed to choose the correct article *a*, thanks to one of the examples under the first sense that includes the phrase "draw a distinction" (*We need to draw a distinction between the two events*). Compared with the rate for the other items for which the dictionary provides a similar or identical phrase as in the given question, the correct answer rate was relatively lower — the fourth lowest of all 23 item mean scores — because approximately one-third of all participants mistook the given discourse context as [+definite] and incorrectly chose *the*. In line with Chan (2017b), the participants in this study frequently used the term "specific" in their written protocols to explain the [+definite] status of target nouns. One respondent explained her choice as follows:

Regardless of its countability, the correct answer is *the* because "distinction" in this sentence means specific "distinction" *between what you want and what you need*, not just any "distinction."

By contrast, participants sometimes benefited from the provision of distinctive English synonyms for senses with an identical Korean translation. For Item 3 (*A / The / Ø business always has some teams that are hotspots for creativity*), for instance, the entry for the target noun *business* provides indistinguishable — senses 1 (사업, 상업, 장사) ("business, commerce/business, business") and 4 (사업체) ("business/company") — or identical — senses 1 (사업, 상업, 장사) ("business, commerce/business, business") and 3 (사업) ("business") — Korean translations. Despite the ambiguity, 74.4% of the participants correctly chose *a* on the post-intervention test — a 32.2% increase from the pre-intervention test mean score — thanks to the English synonyms provided for each sense (*trade*, *work*, and *company* for senses 1, 3, and 4, respectively) (see Figure 4). To quote one participant's written comments:

It's difficult to pick the right sense because all of the first four senses make sense in Korean. Examples under each sense are unhelpful. [I] can't understand why the same meaning is divided into three senses [senses 1–3]. Luckily, there are English definitions that are different from one another. I choose the fourth sense because [the target noun] "business" here means "company" so that there can be "teams" in it. Thus, [the answer is] *a*.

1. TRADE | [U] 사업, 상업, 장사 참고 agribusiness, big business, show business

business contacts / affairs / interests ▶▶

사업상의 인맥/일/사업자들

a business investment ▶▶

사업 투자

It's been a pleasure to do **business** with you. ▶▶

(당신과) 함께 사업을 하게 되어 기쁩습니다.

She has set up in **business** as a hairdresser. ▶▶

그녀는 미용사로 사업을 시작했다.

When he left school, he went into **business** with his brother. ▶▶

그는 학교를 마치고 형과 함께 장사를 시작했다.

She works in the computer **business**. ▶▶

그녀는 컴퓨터 업계에서 일한다.

2. WORK | [U] (직장의) 일, 업무

Is the trip to Rome **business** or pleasure? ▶▶

로마 여행은 출장이세요 아니면 관광차 가세요?

a business lunch ▶▶

업무상 (하는) 점심

He's away on **business**. ▶▶

그 분은 업무차 자리를 비우셨어요.

3. WORK | [U] (회사 등의) 사업 [명명] (실적)

Business was bad. ▶▶

사업이 잘 안 되고 있었다.

Business was booming. ▶▶

사업이 활기를 띠고 있었다.

Her job was to drum up **business**. ▶▶

그녀의 업무는 영업을 신장시키는 것이었다.

How's **business**? ▶▶

사업 잘 되세요?

4. COMPANY | [C] 사업체(회사-가게-공장 등)

to have/start/run a **business** ▶▶

사업체를 갖고 있다/시작하다/운영하다

business premises ▶▶

사업체 구내

Figure 4: Senses 1, 2, 3, and 4 of *business*, noun, from the *Naver Dictionary*


4.2 Absence of nominal countability information

For Item 6 (*It is hard enough to get a / the / Ø job of your dreams, no matter what it may be*), approximately 44% of the participants made an incorrect article choice on the post-intervention test. They all made a similar comment that

because the numeral features of the target noun *job* are not provided in the dictionary, they chose to check examples, in which *job* was mostly preceded by the indefinite article (see Figure 5).

명사

1. PAID WORK | (정기적으로 보수를 받고 하는) 일, 직장, 일자리

He's trying to get a job. 

그는 취직을 하려고 애쓰는 중이다.

She took a job as a waitress. 

그녀는 웨이트리스로 취직했다.

His brother's just lost his job. 

그의 형 [동생]이 얼마 전에 실직을 했다.

a summer/holiday/Saturday/vacation job 


하계/방학[휴가]/토요일/방학[휴가] 중에 하는 일 [일자리]

a temporary/permanent job 

임시직/영구직

I'm thinking of applying for a new job. 

난 새로운 직장에 지원을 해 볼 생각이다.

The takeover of the company is bound to mean more job losses. 


그 회사의 기업 인수는 틀림없이 더 많은 실직 사례가 생길 것을 의미한다.

Many women are in part-time jobs. 

많은 여성들이 파트타임직으로 일하고 있다.

Did they offer you the job? 

그들이 당신에게 그 (일)자리를 제의했나요?

He certainly knows his job. 

그는 자기 일에 대해 확실히 안다.

I'm only doing my job. 

난내가 (보수를 받기 때문에 해야) 할 일을 하고 있을 뿐이다.

He's been out of a job for six months now. 

그는 이제 실업자가 된 지 6개월째이다.

She's never had a steady job. 

그녀는 한번도 안정된 직장을 가져 본 적이 없다.

Figure 5: The entry for *job*, noun, from the *Naver Dictionary*

The analysis of the protocol data suggested two possibilities: Either those who wrongly chose *a* misunderstood the given discourse context as being [-definite] because they memorized the phrase "get a job" as a fixed collocation; or, in the absence of the required lexicographic information, they became preoccupied

with the nominal countability search to the point where they became oblivious to the fact that they had another decision to make — whether the noun is used for identifying or classifying purposes. Either way, all the participants who answered this item incorrectly chose *a* in place of *the* in the post-intervention test, and the correct answer rate for the post-intervention test remained unchanged from that for the pre-intervention test (55.8%), yielding the third lowest of all the mean scores. The following quotations from two participants who wrongly selected the indefinite article outline their reasons for such a decision:

There is no countability symbol [for this word]. However, fortunately, I know for sure that *job* is countable because I've heard of the phrase "get a job" countless times. The correct article is *a*.

While we *must* check the countability status for each definition, there is no such information! I read the examples under the first [correct] sense carefully to check which one [article] is most common. I count the [occurrence] number [of each article], and [the one for] *a* is the largest. It's either "~ a job" or "~ possessive + job." Over 90 percent. No "Ø job." Therefore, I choose *a*.

4.3 Provision of both countable and uncountable features without explicit usage notes

For Items 7 (*crisis*), 18 (*food*), and 19 (*shortage*), the *Naver Dictionary* labels the countability of their target nouns as [C, U], meaning that the noun is used mostly as countable but can be used as uncountable as well. The analyses of the students' written think-aloud data revealed that almost all the participants relied heavily on checking examples to decide which countability status to apply.

For Items 7 (*However, it [getting your dream job] will get even harder for anyone if a / the / Ø worldwide financial crisis occurs*) and 19 (*The United Nations estimates that the world will face a / the / Ø severe water shortage by 2025*) — whose target nouns are preceded by *a*, which is consistent with the countability label — the mean correct answer rates for the post-intervention test were 74.4% and 93.0%, respectively.

In the case of Item 18 (*A / The / Ø Korean food is known for being spicy*), by contrast, the target noun *food* takes the zero article, the use of which is defined by its lexicographic label [C, U] as less frequent than that of the indefinite article. Possibly due to the incongruity between the article to be used (Ø) and the lexicographic label suggesting which countability status takes priority, the post-intervention test mean score dropped by 14% from the pre-intervention test and averaged out at 79.0%. Almost 30% of the 34 respondents who correctly chose Ø were found to have wrongly chosen the first sense (labelled as [U]), which happened to lead to the correct article choice. Most of the remaining students (who correctly chose the second sense) commented that although the noun is labelled as countable first and uncountable second, it

would be "safer" to follow the similarly phrased example (*Do you like Italian food?*) rather than merely to rely on the [C, U] abbreviation (see Figure 6).

1. [U] 식량, 음식, 식품; 먹이

a shortage of food/food shortages

식량 부족

food and drink

음식물

the food industry

식품 산업

2. [C, U] (특정한 유형의) 음식[식품/먹이] **참고** convenience food, fast food, functional food, health food, junk

food, seafood, soul food, wholefood

Do you like Italian food?

이탈리아 음식 좋아하세요?

frozen foods

냉동식품

a can of dog food

개 먹이 한 강통

He's off his food.

그는 식음을 전폐했다.

Figure 6: The entry for *food*, noun, from the *Naver Dictionary*

Given that almost all the participants commented that they had to examine the examples exhaustively for further specifications about the use of a determiner, it is posited that presenting a noun as both countable and uncountable using [C, U] or [U, C] specifications without any usage notes can result in confusion rather than assurance (Xue 2010). As Chan (2017a: 203) has pointed out, most learners cannot possibly discern the "subtle differences between the countable and uncountable uses of the target noun." In the case of Item 18 (*food*), it was obvious that the absence of usage notes adversely affected the participants' determination of the numeral features of the target noun and the associated article selection. Therefore, it seems essential to supplement the marking of countability for two-way nouns with adequate contextual usage examples so that learners can correctly apply the concept in production activities (Hausmann and Gorbahn 1989).

4.4 Inadequate labelling of nominal countability

Of all items, Item 23 (*Scholarships can ease the costs of a / the / Ø college **edu-***cation*) had by far the lowest mean correct answer rates on the pre- and post-*

intervention tests — 0.0% and 4.6%, respectively. While the target noun *education* can be preceded by both \emptyset and *a*, its countability in the corresponding sense (sense 1) is simply marked as [U, sing.] (see Figure 7).

명사

1. [U, sing.] 교육

primary / elementary education

초등 교육

secondary education

중등 교육

further / higher / post-secondary education

고등 교육

students in full-time education

풀타임[전 시간] 교육을 받는 학생들

adult education classes

성인 교육 수업들

a college / university education

대학 교육

the state education system

국가 교육제도

a man of little education

교육을 거의 못 받은 남자

She completed her formal education in 1995.

그녀는 1995년에 정규 교육(과정)을 마쳤다.

2. [U, sing.] (특정한 종류의) 교육[지도/훈련]

health education

건강 교육

Figure 7: The entry for *education*, noun, from the *Naver Dictionary*

Although the *Naver Dictionary* shows a phrase (*a college/university education*) under the first sense showing the target noun being used as countable, embedding the phrase in such a manner without any guiding notes seems to have done more harm than good to its users. Undoubtedly, the participants' disappointing performance on this item may be attributed to the lexicographic failure to mark "the different uses of nouns associated with any differences in their countability status" (Lock 1996: 24) and the related use of determiners in a user-friendly format.

Given the circumstances, it was rather unexpected that two students who correctly chose the indefinite article on the post-intervention test opted to check the examples when the entry information for the target noun seems quite straightforward in terms of countability. To quote one of their protocol data:

It's weird that the dictionary shows [U, sing.] for *education* — an abstract noun. The [U] code already implies that *education* is used exclusively as singular, but then why the redundant [sing.] code? I happened to spot the phrase *a college/university education*, which was weirder as it contradicts the dictionary specification. I am very weak in English grammar, so generally using the articles is tricky, but this one is insane.


4.5 Incongruent countability presentation of the English–English definition with that of the bilingualised version

For Item 8 (*There is a significant difference between an interview and an / the / \emptyset interrogation*), the mean correct answer rate for the post-intervention test was the second lowest (9.3%). Three of four respondents who correctly chose *an* commented that they checked the English–English definition for further clarification of the countability symbol [UC], not [U, C] with a comma in between "U" and "C" (see Figure 8).

명사

1. [UC] 질문, 심문; 의문
2. 의문 부호
3. 질문장

영영사전

[NOUN] An interrogation is the act of interrogating someone.
the right to silence in police interrogations. 

출처: Collins Cobuild Advanced Learner's English Dictionary

속어 (12건)

note[point, mark] of **interrogation** 미국식  영국식 

물음표 ((?))

conduct an **interrogation** 미국식  영국식 

심문을 행하다.

an **interrogation** mark 미국식  영국식 

의문부호

quit an **interrogation** 미국식  영국식 

질문을 그만두다.

Figure 8: The entry for *interrogation*, noun, from the *Naver Dictionary*

Despite the incomprehensible countability notation, the three respondents who took the time to scroll down to the English–English definition and the following Phrasal Expression ("숙어") sections managed to choose the correct article. Interestingly, they almost unanimously wrote in their written think-aloud comments that it was deemed safer to go with the English–English definition, which provides an example (in which *interrogation* is used in its plural form, denoting its countable status), than with the Korean version, which offers the puzzling [UC] code only. The majority of the participants also similarly related in their protocols that they wondered whether [UC] is a typing/printing mistake in the dictionary for [U, C] or for [U] because this notation is not used in the *Naver Dictionary* for any other (more than 100) target nouns they consulted for the exercise questions in *Top 20*. (There is no user's guide available on the *Naver Dictionary* explaining why such a code is used.) The following written think-aloud data vividly depict the struggle that English learners can encounter in such situations:

Definitely, more information is needed. [UC] — I wonder what that means. Possibility (1): UnCountable; possibility (2): Uncountable, but Countable [is] okay too. I check the usage example [section] and count the instances of each [article usage] shown on the first page. "Uncountable" seems to stand a fairer chance. I choose \emptyset . Why on earth do I have to calculate the probability even when using a dictionary?

1. [U] ~ (in/on/to sth) 향상

Sales figures continue to show signs of improvement. 

판매 수치가 계속 향상되는 흔적을 보이고 있다.

There is still room for improvement in your work. 

당신 작품[작업]은 아직 향상 될 여지가 있다.

We expect to see further improvement over the coming year. 

다가오는 해에는 더 많은 향상을 보게 되기를 기대합니다.

2. [C] ~ (in/on sth) 개선, 호전

a significant/substantial/dramatic improvement 

중대한/실질적인/극적인 개선

영영사전

[NOUN] If there is an improvement in something, it becomes better. If you make improvements to something, you make it better.

the dramatic improvements in organ transplantation in recent years 

[NOUN] [usu sing, oft N on n] If you say that something is an improvement on a previous thing or situation, you mean that it is better than that thing.

The new Prime Minister is an improvement on his predecessor 

출처: Collins Cobuild Advanced Learner's English Dictionary

Figure 9: The entry for *improvement*, noun, from the English–English dictionary section of the *Naver Dictionary*

In a similar vein, for Item 16 (*I believe there is room for an / the / \emptyset improvement in every sportsman*), approximately one-quarter of the participants commented that they additionally referred to the English–English definition for further clarification (as discussed earlier in Section 4.1.2, the entry for the target noun *improvement* provides two senses whose Korean definitions are interchangeable). As one student explained,

Improvement in the English–English dictionary is defined as a countable noun — “~ an improvement” and “~ improvements” (see Figure 9). Initially, my choice leaned toward [U] after checking the examples. However, since the English–English dictionary says otherwise, I am confused. I choose *an* according to the English–English definitions, but I feel somewhat uncomfortable [with the choice].

5. Conclusion and implications

Although the findings of this study have important pedagogical and lexicographical implications, several limitations should be noted. Obviously, one major limitation concerns the data collection setting. Due to the limited class time available (the instructors teaching the course from which the participants were drawn had to complete the syllabus written by the school, which outlines specific parts of the required textbook that have to be covered), valid pre-intervention test–post-intervention test designs could not be implemented. As described earlier, the pre-intervention test was conducted in class, whereas the post-intervention test was administered as a take-home task, which must have affected the participants' performance. In addition, since this study adopted a quasi-experimental design with no control group, but with the pre-intervention test results acting as a set of control data, it cannot be attested whether the improvement in article choice performance in the post-intervention test resulted solely from the experimental intervention. Arguably, previous exposure to the same task (the pre-intervention test) could have primed the participants for the post-intervention test, or they could simply have become familiar with the types of test items by the time they took the post-intervention test, which led to better performance.

Notwithstanding these limitations, this study has several strengths. It explained some of the perennial problems encountered with article use, suggesting that giving EFL learners explicit instruction on the main rules for article usage combined with dictionary consultation guidance can foster their ability to use the English articles more correctly. In particular, the results indicate that the use of the indefinite article and the zero article can be a straightforward task for most learners (Miller 2005) if they take the time to check the countability status of the target noun in a dictionary. As one participant's comment “never in my wildest dreams did I expect a change in countability status in relation to a sense” well illustrates, most EFL learners tend to wrongly assume that countability is a static property that is not affected by the sentence context

(Butler 2002). However, the participants' newly — albeit not necessarily voluntarily — formed habit of consulting a dictionary for nominal countability after receiving the instruction seems to have contributed positively to their improved post-intervention test performance.

For lexicographic practice, the findings are valuable because a number of problems have been identified with regard to the present lexicographic practice of presenting the solicited nominal countability information, and the dictionary users' authentic voices reported in this study would be useful to lexicographers in improving their products. The identified problem areas include, but are probably not limited to, applying equivocal criteria for dividing senses with opposite countability status; failing to provide nominal countability features; presenting both countable and uncountable features without their distinct usage information; inadequately labelling countability features, resulting in some examples with conflicting countability status; and supplementing an English–English definition that does *not* accord with the countability status labelled in the bilingualised version. Such observations accentuate the importance of clear lexicographic indications of the numeral features of a noun in a bilingualised dictionary according to semantic differences as well as syntactic requirements. As Kirkness (2004: 78) has rightly maintained, dictionaries should consistently serve their role as "the single most valuable source of linguistic information ... of the target language," actively accommodating "lexicographic needs arising in concrete situations" (Xue 2010: 550).

In addition, since a number of noncount nouns (e.g. abstract nouns such as *beauty, truth, crime, law, or education*; and mass nouns such as *cheese, wine, tea, chocolate, or aspirin*) can also have a countable form without substantially changing the meaning (Celce-Murcia and Larsen-Freeman 1999), a learner's dictionary should supplement usage information for two-way nouns so that its users can decide whether the indefinite or the zero article is appropriate in a given discourse context. Given the general tendency for such nouns to be used as countable when referring to a particular type or instance — as opposed to referring to the abstract concept — it might be also useful for ESL/EFL teachers to design instructional materials that present a set of sentences containing the same noun in different contexts to alert students to "the variability of noun countability and related article use" (Chan 2017a: 202).

Although the English article system has been seen by some linguists as strangely immune to instruction and acquirable only through exposure (e.g. Doughty and Williams 1998, Lightbown and Spada 2013), a growing body of research indicates the contrary, presenting empirical data that many aspects of the English article system are in effect teachable because of the clearly defined rules associated with it (e.g. Ferris 2011, Master 1994). The findings of this study provide additional support for the lexicographic approach to teaching article usage for nouns used for classification purposes under the condition that learners are clearly provided with their countability status. Since the problems investigated are relevant to almost anyone using a dictionary, and particularly

second language learners, it is suggested that this study is replicated with other language combinations. Meanwhile, teachers need to acquaint their students with the fact that nominal countability is a variable, context-sensitive feature that should be checked by consulting a dictionary.

Acknowledgements

I am grateful to the two anonymous reviewers for their incisive, expert commentaries, which have led to significant improvements in the presentation of this paper. I would also like to thank Robert Yeates for helping me finalize the instrument. As always, all remaining errors are my own.

Endnotes

1. Anaphoric reference means that a word in a text refers back to other ideas in the text for its meaning, as in *An elegant, dark-haired woman entered the compartment, and I immediately recognized the woman* (Lyons 1999).
2. Associative anaphoric reference means that first mentions of new referents within a discourse can be identified via another, already present referent, as in *I have a bicycle, but the gears are out of order* (Allan 2009).
3. Cataphoric reference means that a word refers to ideas later in the text, as in *I remember the beginning of the war very well* (Chesterman 1991).

References

- Allan, K. 1980. Nouns and Countability. *Language* 56(3): 541-567.
- Allan, K. (Ed.). 2009. *Concise Encyclopedia of Semantics*. Oxford: Elsevier.
- Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bickerton, D. 1981. *Roots of Language*. Ann Arbor, MI: Karoma.
- Butler, Y.G. 2002. Second Language Learners' Theories on the Use of English Articles: An Analysis of the Metalinguistic Knowledge Used by Japanese Students in Acquiring the English Article System. *Studies in Second Language Acquisition* 24(3): 451-480.
- Celce-Murcia, M. and D. Larsen-Freeman. 1999. *The Grammar Book: An ESL/EFL Teacher's Course*. Second edition. Boston, MA: Heinle & Heinle.
- Chan, A.Y.W. 2016. How Much Do Cantonese ESL Learners Know about the English Article System? *System* 56: 66-77.
- Chan, A.Y.W. 2017a. The Effectiveness of Using a Bilingualized Dictionary for Determining Noun Countability and Article Selection. *Lexikos* 27: 183-213.
- Chan, A.Y.W. 2017b. Why Do Hong Kong Cantonese ESL Learners Choose a Certain English Article for Use? *The Asian Journal of Applied Linguistics* 4(1): 16-29.
- Chesterman, A. 1991. *On Definiteness: A Study with Special Reference to English and Finnish*. Cambridge Studies in Linguistics 56. Cambridge: Cambridge University Press.

- Cowie, A.P.** 1999. *English Dictionaries for Foreign Learners: A History*. Oxford: Clarendon Press.
- Doughty, C. and J. Williams.** 1998. *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press.
- Ericsson, K.A. and H.A. Simon.** 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.
- Ferris, D.** 2011. *Treatment of Error in Second Language Student Writing*. Second edition. Ann Arbor, MI: University of Michigan Press.
- Ferris, D.R. and J.S. Hedgcock.** 2005. *Teaching ESL Composition: Purpose, Process, and Practice*. Mahwah, NJ: Lawrence Erlbaum.
- Folse, K.S., E.V. Solomon and B. Smith-Palinkas.** 2008. *Top 20: Great Grammar for Great Writing*. Second edition. Boston, MA: Heinle.
- García Mayo, M.P.** 2008. The Acquisition of Four Nongeneric Uses of the Article *the* by Spanish EFL Learners. *System* 36(4): 550-565.
- Gass, S.M. and L. Selinker.** 2001. *Second Language Acquisition: An Introductory Course*. Second edition. Mahwah, NJ: Lawrence Erlbaum.
- Gauton, R.** 2008. Bilingual Dictionaries, the Lexicographer and the Translator. *Lexikos* 18: 106-118.
- Greenbaum, S. and G. Nelson.** 2009. *An Introduction to English Grammar*. Third edition. Harlow, UK: Pearson Longman.
- Han, N.-R., M. Chodorow and C. Leacock.** 2006. Detecting Errors in English Article Usage by Non-native Speakers. *Natural Language Engineering* 12(2): 115-129.
- Hausmann, F.J. and A. Gorbahn.** 1989. COBUILD and LDOCE II: A Comparative Review. *International Journal of Lexicography* 2(1): 44-56.
- Hornby, A.S.** 2008. *Oxford Advanced Learner's English-Korean Dictionary*. Oxford: Oxford University Press.
- Huddleston, R. and G. Pullum.** 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Ionin, T., H. Ko and K. Wexler.** 2004. Article Semantics in L2 Acquisition: The Role of Specificity. *Language Acquisition* 12(1): 3-69.
- Ionin, T., M.L. Zubizarreta and S.B. Maldonado.** 2008. Sources of Linguistic Knowledge in the Second Language Acquisition of English Articles. *Lingua* 118(4): 554-576.
- Kirkness, A.** 2004. Lexicography. Davies, A. and C. Elder (Eds.). 2004. *The Handbook of Applied Linguistics*: 54-81. Oxford: Blackwell.
- Lennon, P.** 1991. Error and the Very Advanced Learner. *International Review of Applied Linguistics in Language Teaching* 29(1): 31-44.
- Leroux, W. and T. Kendall.** 2018. English Article Acquisition by Chinese Learners of English: An Analysis of Two Corpora. *System* 76: 13-24.
- Lightbown, P.M. and N. Spada.** 2013. *How Languages Are Learned*. Fourth edition. Oxford: Oxford University Press.
- Lock, G.** 1996. *Functional English Grammar: An Introduction for Second Language Teachers*. Cambridge: Cambridge University Press.
- Lyons, C.** 1999. *Definiteness*. Cambridge: Cambridge University Press.
- Master, P.** 1988. Teaching the English Article System. *English Teaching Forum* 26(3): 2-7, 18-21, 25.
- Master, P.** 1990. Teaching the English Articles as a Binary System. *TESOL Quarterly* 24(3): 461-478.

- Master, P.** 1994. The Effect of Systematic Instruction on Learning the English Article System. Odlin, T. (Ed.). 1994. *Perspectives on Pedagogical Grammar*: 229-252. Cambridge: Cambridge University Press.
- Master, P.** 1997. The English Article System: Acquisition, Function, and Pedagogy. *System* 25(2): 215-232.
- Miller, J.** 2005. Most of ESL Students Have Trouble with the Articles. *International Education Journal* 5(5): 80-88.
- Miller, J.** 2006. An Investigation into the Effect of English Learners' Dictionaries on International Students' Acquisition of the English Article System. *International Education Journal* 7(4): 435-445.
- Mizuno, M.** 1999. Interlanguage Analysis of the English Article System: Some Cognitive Constraints Facing the Japanese Adult Learners. *International Review of Applied Linguistics in Language Teaching* 37(2): 127-152.
- Naver Bilingual (English–Korean/Korean–English) Dictionary.* Available at: <http://endic.naver.com/>.
- Oxford Dictionaries.* Available at: <http://en.oxforddictionaries.com/>.
- Sinclair, J. (Ed.).** 2008. *Collins COBUILD Advanced Learner's English Dictionary*. Sixth edition. Glasgow, UK: Harper Collins.
- Tsang, A.** 2017. Judgement of Countability and Plural Marking in English by Native and Non-native English Speakers. *Language Awareness* 26(4): 343-359.
- White, L.** 2003. Fossilization in Steady State L2 Grammars: Persistent Problems with Inflectional Morphology. *Bilingualism: Language and Cognition* 6(2): 129-141.
- Wisniewski, E.J., C.A. Lamb and E.L. Middleton.** 2003. On the Conceptual Basis for the Count and Mass Noun Distinction. *Language and Cognitive Processes* 18(5/6): 583-624.
- Xue, M.** 2010. Countable or Uncountable? That Is the Question — Lexicographic Solutions to Nominal Countability in Learner's Dictionaries for Production Purposes. *Lexikos* 20: 540-558.
- Yoo, I.W.** 2004. *The English Articles and Nouns*. Accessed at: http://ocw.mit.edu/courses/global-studies-and-languages/21g-213-high-intermediate-academic-communication-spring-2004/readings/MIT21G_213S04_articles.pdf (22 October 2016).
- Yoon, K.K.** 1993. Challenging Prototype Descriptions: Perception of Noun Countability and Indefinite vs. Zero Article Use. *International Review of Applied Linguistics in Language Teaching* 31(4): 269-290.

Appendix 1: Forced-choice elicitation task

Circle the correct answer for each question.

1. We went to a wedding yesterday. A / The / Ø bride was wearing a lovely dress.
2. Any exercise is a / the / Ø good exercise, but when it comes to losing weight, nothing can beat running.
- 3.–5. A / The / Ø business always has some teams that are hotspots for creativity, and a / the / Ø creative ideas need a / the / Ø special climate to grow.
- 6.–7. It is hard enough to get a / the / Ø job of your dreams, no matter what it may be. However, it will get even harder for anyone if a / the / Ø worldwide financial crisis occurs.
8. There is a significant difference between an interview and an / the / Ø interrogation.
9. Julian ordered a cup of coffee and a dessert, but he didn't touch a / the / Ø dessert.
- 10.–12. At a gallery, I saw a beautiful landscape painting. I really wanted to meet an / the / Ø painter of a / the / Ø painting, but a / the / Ø gallery owner said he didn't have her contact information.
13. Robert was discussing an interesting book in his class. I went to discuss a / the / Ø book with him afterwards.
14. We have just arrived from New York. A / The / Ø plane was five hours late.
15. A / The / Ø happiness that I felt when Charlene became pregnant was beyond description.
16. I believe there is room for an / the / Ø improvement in every sportsman.
17. A / The / Ø tea that I received for my birthday is high-quality.
18. A / The / Ø Korean food is known for being spicy.
- 19.–20. The United Nations estimates that the world will face a / the / Ø severe water shortage by 2025 if we continue to use a / the / Ø water at today's rates.
21. It is important to draw a / the / Ø distinction between what you want and what you need.
22. An / The / Ø anger he felt after the accident nearly ended his career.
23. Scholarships can ease the costs of a / the / Ø college education.

An Empirical Study of EFL Learners' Dictionary Use in Chinese–English Translation

Pengcheng Liang, *School of Foreign Languages and Cultures,
Nanjing Normal University and Bilingual Dictionary Research Center,
Nanjing University, Jiangsu, China (richardl@126.com)*

and

Dan Xu, *School of Foreign Languages and Cultures,
Nanjing Normal University, Jiangsu, China (sophy3230@126.com)*

Abstract: This article reports on the results of a study which investigated English as Foreign Language (EFL) learners' use of an electronic dictionary in a L1–L2 translation task. Forty-seven university graduate students from a Chinese university were asked to translate a Chinese passage into English on computers with the support of an embedded dictionary. Screen recorders were used to record their dictionary use behavior and a follow-up interview was conducted to tap into the thinking processes behind their behavior. The results of the study show that when translating, EFL learners demonstrate preferences for L2 equivalents and content words in their lookups, and reveal specific problems such as a preoccupation with L2 equivalents and lack of awareness of other lexical information, which may hinder correct application of dictionary information. This study suggests that dictionary use behavior may affect the development of students' ability to translate and requires attention from both EFL learners and teachers. It is further suggested that translation teachers should alert learners to the importance of checking other lexical information in a dictionary in their translation practice.

Keywords: DICTIONARY USE PREFERENCES, DICTIONARY USE PROCESSES, EFL LEARNERS, TRANSLATION TASK, INTERVIEW, SCREEN RECORDING, LOG FILES

Opsomming: 'n Empiriese studie van EVT-leerders se woordeboekgebruik in Chinees-Engelse vertaling. In hierdie artikel word verslag gelewer oor die resultate van 'n studie waarin leerders van Engels as Vreemde Taal (EVT) se gebruik van 'n elektroniese woordeboek in 'n L1–L2-vertalingsopdrag ondersoek is. Sewe en veertig nagraadse studente van 'n Chinese universiteit is versoek om op die rekenaar 'n Chinese stuk in Engels te vertaal met behulp van 'n ingeboude woordeboek. Skermopnemers is gebruik om hul gedrag rakende woordeboekgebruik vas te lê, en 'n opvolgonderhoud is gevoer om die denkprosesse wat hul gedrag rig, te probeer bepaal. Die resultate van die studie dui daarop dat EVT-leerders in die naslaanproses 'n voorkeur vir L2-ekwivalente en inhoudswoorde toon, en dit lê spesifieke probleme soos 'n behepthed met L2-ekwivalente en 'n onkunde oor ander leksikale inligting bloot, wat kan verhinder dat die woordeboekinligting korrek toegepas word. Hierdie studie suggereer dat woordeboekgebruiksgedrag die ontwikkeling van studente se vertaalvermoëns mag affekteer en dat sowel EVT-leerders as -onderwysers aandag hieraan moet skenk. Daar word ook voorgestel dat vertaalonderwysers leer-

ders se aandag moet vestig op hoe belangrik die kontrolering van ander leksikale inligting in 'n woordeboek in hul vertaalpraktyk is.

Sleutelwoorde: WOORDEBOEKGEBRUIKERSVOORKEURE, WOORDEBOEKGEBRUIKS-PROSESSE, EVT-LEERDERS, VERTALINGSOPDRAG, ONDERHOUD, SKERMOPNAME, LOG-LÊERS

1. Introduction

Traditionally, the dictionary is considered an important tool in language learning. A number of studies have demonstrated that dictionary use contributes to the acquisition of a foreign language (Lew and Doroszewska 2009; Chen 2011; Dziemianko 2014; Chen 2017; Liang and Xu 2017). However, the importance of applying dictionary information correctly has not been thoroughly examined. Researchers often fail to acknowledge that for most language learners, the purpose of looking up words in a dictionary is not to memorize vocabulary or acquire language, but to solve problems in various language tasks such as reading, writing and translation. In other words, vocabulary acquisition is incidental in dictionary use, while the availability, accessibility and application of lexical information are the immediate needs of most dictionary users. In addition, it is the correct use of retrieved information that forms the basis of vocabulary (language) acquisition. In this sense, studies of dictionary use should not only focus on the incidental acquisition of words but also the application of lexical information. After all, the incidental acquisition of vocabulary in dictionary use depends on the correct application of the target words on repeated occasions. On the other hand, the majority of research on dictionary use has employed elicitation tasks to collect data, either through various forms of production questionnaires (Barnhart 1962; Tomaszczyk 1979; Hartmann 1983; Atkins and Varantola 1997; Sánchez Ramos 2005) or log files (Laufer and Hill 2000; Lew and Doroszewska 2009; Chen 2011; Liang and Xu 2017). Research employing more naturalistic data is needed to explore how learners use dictionaries in real situations, particularly in language learning contexts such as reading, writing and translation. Tarp (2009: 293) argues that various methods should be combined to obtain more knowledge about real user needs. Lew (2011b) believes that there is room for engaging both positivistic and naturalistic approaches, as in fact they do not exclude, but rather complement one another. Possibly due to the difficulty of collecting naturalistic data, and the relative recency of electronic dictionaries, these kind of studies of dictionary use seem to be under-represented in the field of lexicography.

The present exploratory study aims to contribute to the literature on user research in lexicography by employing mixed research methods (observation, a test and interviews) to collect naturalistic data, exploring what EFL learners look up in electronic dictionaries and how they use the lexical information in L1–L2 translation.

The article is structured as follows: first, we provide a brief summary of research on dictionary use by English as a foreign language (EFL) learners and research on the application of dictionary information. In the second part of the article, we present our study, starting with the research questions, participants and a description of the experimental dictionary. We then describe the methods used to observe user behavior and collect data, present and discuss the results obtained, and conclude the article with a summary and suggestions for future studies.

2. Literature review

The literature review consists of two parts. The first part is concerned with studies of dictionary use in general and the second part focuses on studies of the application of dictionary information.

2.1 Studies of dictionary use

Studies of dictionary use have a long history as lexicographers have learned to recognize the importance of this research field. According to Welker (2010: 531), about 70 empirical studies were published from 1962 to 1989 and there have been more than 250 investigations since 1990. Some studies have focused on assessing the dictionary skills of learners (Frankenberg-Garcia 2011; Chan 2012), discovering where students look up multi-word expressions (Tono 1989; Bogaards 1998, 2003; Frankenberg-Garcia 2011; Gromann and Schnitzer 2016), as well as which type of dictionary — bilingual, monolingual or semi-bilingual — is easiest to use and gives students the most reliable results (Laufer and Melamed 1994; Laufer and Hadar 1997; Kaneta 2011; Chen 2011; Chan 2014). According to Nesi (2014), dictionary use research covers five themes: learners' preference and attitudes, the influence of dictionaries on text comprehension, the influence of dictionaries on text production, the role of dictionaries as an aid to English language learning and English language learners' dictionary consultation behavior. Lew (2011a: 1) notes that interest in the empirical study of dictionary use is on the rise.

As electronic dictionaries replace print dictionaries (Lew 2012: 243), research into dictionary use is increasingly focusing on the former. In the digital age, the status of the dictionary is changing, and so are the patterns of user behavior. As such, we need to know more about user behavior in the digital environment (Lew and De Schryver 2014). Carolin Müller-Spitzer (2014: 46) found that a majority of studies had been concerned with bilingual dictionaries and the comparison in students' use of bilingual and monolingual dictionaries. This is connected to the fact that some of the studies concentrate in particular on vocabulary learning (Laufer and Hill 2000; Lew and Doroszewska 2009; Chen 2010; Dziemianko 2010; Chen 2017). Laufer and Hill (2000) and Chen (2010),

for example, investigated the relationship between which low-frequency words students looked up while reading and how well those words were remembered. In their studies, the relevant lexical information was incorporated into a CALL program comprised of a text, highlighted low-frequency words, and access to different lexical information about these words (with explanations in English, translations into the L1, sound and "extra" information). These studies reveal the role of electronic dictionaries in vocabulary learning but they are not without problems. One concern is that they do not reflect actual dictionary use, because users looked up both low-frequency words and high-frequency words in reading and translation, especially when high-frequency words have many different senses (Bogaards 1998; Frankenberg-Garcia 2011; Koplenig, Meyer and Müller-Spitzer 2014). Another problem is that most users consulted a dictionary to solve the problems arising during the linguistic activities rather than to memorize words. When learners have difficulty understanding a word or expressing an idea in linguistic activities, they turn to a dictionary for help. Then they try to understand and use the word. In addition, these studies only investigated incidental vocabulary acquisition while reading. As we know, receptive tasks such as reading are less demanding than productive activities such as writing and L1-L2 translation because they do not require learners to know lexical information in great depth. In decoding tasks, users "will be 'blind' to the grammatical contexts in which a target word appears"(Chan 2012: 134).

2.2 Information application study

At present, only a few studies have focused on the application of dictionary information, but they have not provided a complete picture of dictionary use, likely because they fail to combine positivistic and naturalistic research methods. For instance, Atkins and Varantola (1992, 1993, 1997) carried out a series of studies monitoring dictionary use in translation. They performed a detailed examination of the words looked up by users and the motivations for their look-ups. They aimed to monitor the dictionary look-up process in as natural a situation as possible. To that end, the researchers asked students to note down what their partners looked up, using forms designed for the research. This method was unobtrusive but it only recorded the information that the form required. In addition, participants in their study were not from the same language background, so the user group was not representative of any particular language community. Also, the researchers did not rate how successful the look-ups were. Harvey and Yuill (1997) also studied the use of monolingual dictionaries by EFL learners while writing. They asked students to recall what they had looked up. This method may well help to produce easily quantifiable results from natural settings but sacrifices the crucial criterion of reliability in data collection by relying solely on students' memory. Bogaards (2003: 26-33) concludes that 'uses and users of dictionaries remain for the moment relatively unknown'.

In recent years, researchers (Dziemianko 2010, 2014; Chen 2011, 2012, 2017; Chan 2012, 2014; Hu and Zhang 2013; Frankenberg-Garcia 2015) have focused on dictionary use and language acquisition. However, they tend to report on factors that affect learners' use of information while not providing a comprehensive description of the application of this information in natural settings. In addition, some studies only investigated the effects of dictionary features without gathering information from the users about their behavior. For instance, Frankenberg-Garcia (2015) investigated the effect of the type and number of examples in dictionary entries by asking 161 students to perform partial sentence translations. Students were made aware that their production might be problematic and they were encouraged to make revisions. She found that the number of examples did affect users' production. Although experiments of this kind are indispensable, we need to know both the effects and the causes underlying the performance of users by combining positivistic and naturalistic research methods.

These studies highlight the need for an in-depth analysis of dictionary use in a more natural setting. In response, this study offers a detailed account of how students applied dictionary information in a production assignment, analyzes the possible relationship between application and behavior and explores underlying causes, and identifies some implications for the presentation of information in electronic dictionaries.

3. Research design

3.1 Research questions

Our specific research questions were as follows:

1. What is the status quo of EFL learners' use of the electronic dictionary during a translation task?
 - (1) What do users look up in an electronic dictionary during a production activity?
 - (2) Are there any look-up preferences? If yes, what are they?
 - (3) What are the underlying causes of users' different lookup preferences?
2. How well do EFL learners apply retrieved information in the translation task?
 - (1) What types of errors did learners make in the application of dictionary information? And what are the causes?
 - (2) What contributed to the learners' successful application of dictionary information?

3.2 Participants

The study included 50 students from two translation classes in a course that was taught in a CALL classroom, wherein students listened to the teacher and practised translation on computers. Participants' ages ranged from 22 to 25. 28 of the students were female and 22 were male. One class of students (27) were majoring in computer science at a Chinese College, and the other class of students were psychology majors. All of the study's participants had passed College English Test Band 4 (CET4 \approx 5 in IELTS) and 30 of them had passed College English Test Band 6 (CET6 \approx 5.5 in IELTS).

3.3 Instruments

The research instruments included a self-designed CALL program with an embedded dictionary which was used to present the translation task, provide dictionary help and record students translation products, AntConc 3.2.1 (Laurence 2007), ICTCLAS 2014 (Zhang 2014), a screen recorder PMLX (Pan 2012), and an outline of our interview questions.

ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) was used to compute how many words a Chinese text (in our study, the translation task) contains and to tag parts of speech onto the words. The accuracy of lexical analysis is 97.58%. AntConc 3.2.1 was first used to analyze the tokens and types of Chinese words in the translation task. The result is a factor we considered when deciding how many entries the embedded dictionary should include. AntConc 3.2.1 was also used to analyze the application of dictionary information (spelling, collocational and syntactical features) in students' translation products. PMLX, a screen recorder software, was used to record students' choice of lexical information category and retrieval behavior.

The researchers designed a CALL program similar to that used in the studies of Laufer and Hill (2000) and Chen (2013). The program in our study consisted of a task box, an embedded dictionary with a search box and a display box, a click counter, and a writing box. The translation task consisted of a text of 328 Chinese characters. It contains about 198 word tokens (according to an analysis using ICTCLAS 2014) and 134 word types. Similar to the programs in Laufer and Hill (2000) and Chen (2013), this program also had a task box, an embedded dictionary with information category labels, and a dictionary information presentation box. Different from their programs, this program had a search box and a writing box because in productive tasks (such as translation and writing) learners usually have more lexical needs than they would in receptive tasks (such as reading and listening). The search box gave users more freedom to look up words (both low-frequency and high-frequency) when needed than previous studies (Laufer and Hill 2000; Chen 2013). In Laufer and Hill's study (2000), the target words included 12 low-frequency words and in Chen's study (2013), the target words were 10 unknown words. In addition,

since the translation activity involved the change of word forms and different ways of expressing meanings, we also added three information categories to help students. They were derivative, collocation and phrase, synonyms and antonyms.

The interface of the CALL program is shown in Figure 1. The top box in the left colored yellow is the translation task. The bottom box in the left colored blue is the writing box. Between them lies the display box colored green. The top box in the right colored green is the search box. The six buttons under this search box are the labels of information categories. When users input a word into the search box and clicked on one information category label, the corresponding information for the word appeared in the display box. Like other electronic dictionaries, this search box also carried a function of association, that is, when users input letter "a" into the search box, a group of words beginning with "a" appeared in the pull-down list. This helped users locate the target word entries immediately. When the translation task was finished, users clicked on the save button below the writing box and the program saved the work.

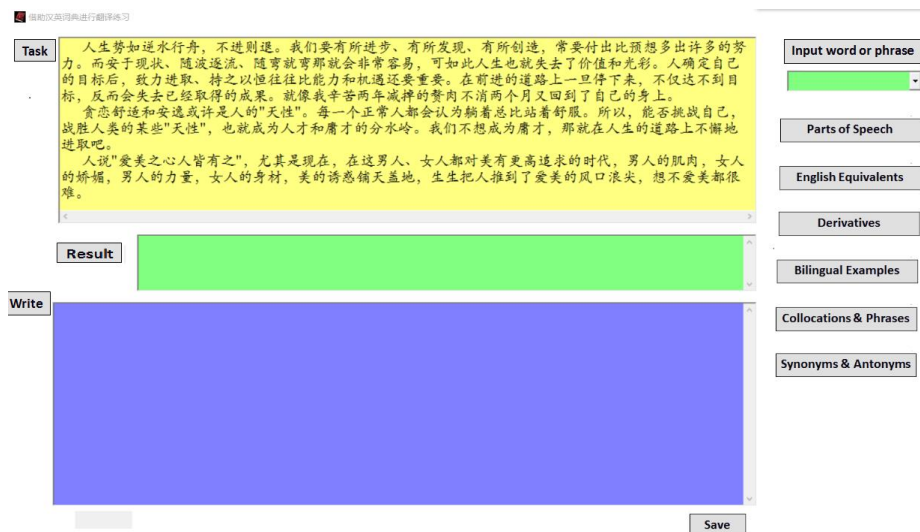


Figure 1: Screenshot of the CALL program

Several factors were taken into consideration for the selection of entries in the dictionary. First, in classroom practice of translation, we found that students did not look up all the words of the translation task in the dictionary. In tasks similar to the one in the test, they usually looked up about one-third of the words in the texts. In the pilot study, the number of words one student looked up in the dictionary was 35 and the other student looked up 30 words. Second,

as all the students passed CET4 (CET4 \approx 5 in IELTS), the two researchers in this study decided that there were about 20 words unfamiliar to students. Third, considering some low-frequency words students might look up, we decided to provide 43 words in the dictionary, which included all the words the two students in the pilot study looked up.

The lexical information of these entries in the dictionary was collected from two print Chinese–English dictionaries (*Chinese–English Dictionary*, 2010 and *New Century Chinese–English Dictionary*, 2012) and two electronic dictionaries (*Youdao Dictionary* and *Kingsoft PowerWord*). Compiled after the founding of the People's Republic of China, the *Chinese–English Dictionary* (1978) is the first of its kind and is regarded as the most authoritative. *Chinese–English Dictionary* (2010) is currently in a third edition. *New Century Chinese–English Dictionary* (2012) is ranked number one of its kind in terms of sales volume. The two print dictionaries both cover over 100 thousand entries and were published in recent years. *Youdao Dictionary* (7.0) and *Kingsoft PowerWord* (2017) are the two most used electronic dictionaries by college students in China (Xie 2014; Yang 2017). *Youdao Dictionary* (7.0), with over 500 million users, is a digitalized collection of many print dictionaries, like LDOCE (*Longman Dictionary of Contemporary English*, 5th edition) and *Collins Learners' English–Chinese Dictionary* (2012). It includes over 37 million entries and 23 million examples. Like *Youdao Dictionary*, *Kingsoft PowerWord* (2017) is a digitalized collection of many print dictionaries, such as the *Collins COBUILD Advanced Learner's English–Chinese Dictionary* (2012). It has 30 million users and is famous for its over 5 million bilingual examples. At present, there is no digitalized form of the two print Chinese–English dictionaries in China. However, the lexical information in the two print dictionaries is too limited for translation learners. Take the word "yù xiǎng" (which literally means "expect") for example.

预想 yù xiǎng <动> anticipate; expect; prefigure; preconceive: ~未来 prefigure the future || 符合~ satisfy sb's preconceptions of sth
|| 这比~的要复杂得多。 It is more complicated than first thought.
(*New Century Chinese English Dictionary* 2012)

This entry only provides pronunciation, one word class of the word, four English equivalents, two phrases and one bilingual example. It does not demonstrate the usage of all the equivalents.

预想 yù xiǎng I 动 anticipate; expect; preconceive ~未来 prefigure(or envisage) the future/~不到的后果 unexpected consequences/得到~的结果 obtain the anticipated results II 名 preconception
(*Chinese English Dictionary* 2010)

This entry only provides pronunciation, two classes of the word, three English equivalents, and three phrases. It does not demonstrate the usage of all the equivalents and it does not even provide one sentence example.

In the pilot study, students reported that they wanted to read more phrases and examples in the dictionary, so we decided to take some information (chiefly bilingual examples and phrases) from these two electronic English–Chinese dictionaries because they support Chinese–English translation and have more phrases and examples. As a result, the information categories of the embedded dictionary included the part of speech (POS), English equivalents, derivatives of the English equivalents, bilingual examples, collocations and phrases, and synonyms and antonyms. Information about the part of speech for the 43 word entries was taken from *Chinese–English Dictionary* (2010), collocations and phrases from *New Century Chinese–English dictionary* (2012), and equivalents from *Chinese–English Dictionary* (2010). Derivatives, Synonyms and antonyms were taken from *Youdao Dictionary* (7.0), and bilingual examples were taken from *Kingsoft PowerWord* (2017). For example, the entry of "yù xiǎng" (again which literally means "expect") in the dictionary is as follows:

预想 (yù xiǎng)

Part of speech: Verb;

English equivalents: anticipate; expect; speculate;

Derivatives: (N.) anticipation; expectation; (ADJ) anticipated; expected

Bilingual examples:

1. 这比预想的要复杂得多。

It is more complicated than first thought.

2. 这次度假的花销超出了我的预想。

The costs of the vacation surpassed my expectation.

3. 西湖的春景要比他的预想更加美丽。

The beauties of the West Lake in spring were beyond his expectation.

Collocations and phrases: 预想未来 Prefigure the future;

符合预想 satisfy one's preconceptions of sth.

Synonyms and antonyms:

近义词 (synonym): foresee; expect; hope;

反义词 (antonym): recall; review; recollect

3.4 Research methods

The data collection methods employed for this research were observation by means of a screen recorder, a translation task, and follow-up interviews used to explore the underlying reasons for users' behavior.

Before the experiment, the researchers conducted a pilot study. Participants in the pilot study were two students from another class in the same grade as participants from two classes in our study. Based on their feedback, some information in the dictionary was revised. After completing the translation task with the embedded dictionary, both students suggested that the bilingual examples in the dictionary should be numbered in the display box to make them more readable. We numbered all the examples in the dictionary. One stu-

dent indicated that some examples were too long. We replaced the examples with shorter ones. They also reported some spelling mistakes in the entry. We corrected them. The pilot study found that both students preferred to click on English equivalents. Li (1998) found that Chinese students often choose the first equivalents in bilingual dictionaries when translating from English into Chinese. This is one reason why they often cannot produce correct translations. Therefore, we decided to investigate their preference in choosing these equivalents when translating from Chinese into English. In designing the embedded dictionary, we chose ten words which have at least three English equivalents and placed the most familiar equivalent in second position. The familiarity with the words was rated by the two students in the pilot study. It needs to be explained that "unremitting" and "vanquish" appeared in the essay *If I Rest, I Rust* written by Orison Marden, the first unit in the textbook students used. That is why the two students rated them as the most familiar words among the equivalents.

The experiment was performed in a computer center where the learners took translation classes. At the beginning, participants were shown a demo about how to use this program without being told the purpose of the study, although they were told that it was part of an innovation program about computer-aided translation training. The experiment was carried out over two classes (90 minutes). Since two students used online dictionaries and one failed to finish the translation task, the data of their performance was excluded. Therefore, the effective number of participants was 47. The experiment produced the following data: video records (entries retrieved and information category click counts), and the products of the translation task.

Based on a preliminary observation of video recordings, interviews with five participants were carried out the second day of the study to learn about the reasons for their consulting behavior. The interviews were guided by the following topical questions: (1) Does the dictionary provide sufficient help in translation? (2) Why did they click on the equivalents most often or click only on the equivalents? (3) Why did they click on the examples, or why not? (4) What are the criteria for their choice of an equivalent?

To minimize the difficulty of expressing their ideas, interviews with students were conducted in Chinese. Interviews were recorded and then transcribed.

4. Results and Discussion

In the pilot study, informants indicated that the embedded dictionary provided sufficient information for the translation task. Video recordings confirmed this. By comparing the number of entries participants input into the search box and the results they obtained, we concluded that the embedded dictionary helped them address most of their lexicographic needs. In the experiment, the consultation could not solve all the problems of students who lacked dictionary use

skills, nor could the dictionary solve all of the problems users met in the translation process. Therefore, we believe that this experiment can represent actual use of a dictionary for translation in a natural setting.

	N	Min	Max	Mean	SD	F
Retrieval success	47	.34	1.00	.8445	.15406	.024
Effective	47					

Table 1: Results of lexical information availability test in the e-dictionary

Table 1 shows that the average retrieval success was 84%. This means that 84% of the words users searched were available in the embedded dictionary. We believe this result reflects the authentic situation of dictionary use in linguistic activities. Firstly, as new words or new usages of existing words emerge almost every day, the available Chinese–English dictionaries cannot immediately include all the words in use. Secondly, since a Chinese character could be part of a word, a word or a phrase, some users do not know the lexical unit they should look up in the dictionary. This is evidenced by some students failing to find the target words in the dictionary because they looked up phrases, clauses or even sentences rather than words. Thirdly, some students lacked the instrumental ability of translation competence. They read only the English equivalents while ignoring other information categories which might be helpful to their translation.

4.1 Translation learners' dictionary use

4.1.1 The words looked up

Drawing on ICTCLAS2014, the original text was found to contain 198 Chinese running words, and our results showed that users consulted high-frequency words most often. This finding is consistent with that of Varantola (1998) and Koplein, Meyer and Muller-Spitzer (2014). The reason could be that users consult dictionaries not only to learn about new words but also to check whether their understanding or use of high-frequency words is accurate. The learners looked up a small number of function words. This is possibly because translation learners felt they were more familiar with these words than other high-frequency words and function words usually were not the barrier to understanding. With regard to the word classes users retrieved, Table 2 reveals that the users consulted content words most often; 89% of the words looked up were verbs, nouns, adjectives and adverbs. This is quite understandable. Firstly, content words are related to both meaning comprehension and pro-

duction. Secondly, the number of content words was high in the original text, about 134. Thirdly, many content words are in wide use and carry multiple meanings which usually cause trouble for students' meaning understanding and production, while function words, users assumed, were relatively familiar to them. The average number of words looked up by each user was 45 while the average number of content words looked up by users was 40. The high proportion of content words in the words that were looked up reflects the fact that users relied heavily on dictionaries to express their meaning.

In addition, the majority of words looked up were basic words such as "effort" (looked up 31 times), "progress" (looked up 35 times), "road" (looked up 31 times), "ability" (looked up 26 times). Some students even looked for the equivalents of such words as "easy" (looked up twice), "important" (looked up 4 times) and "now" (looked up twice). Interviews revealed that some students looked up these words to check whether what they remembered about them was correct.

We also found that users treated multi-word expressions as retrieval units. Most of the items that were looked up were actually phrases and expressions. For instance, *zhì lì jìn qǐ*, (i.e. make great efforts) was looked up 9 times; *què dìng mù biāo* (i.e. set the target) was looked up 4 times and *yóusuo jìn bù* (i.e. make some progress) was looked up 5 times. This revealed that translation students understand texts in terms of semantic units rather than lexical units. Therefore, to help users' retrieval efficiency, we believe more phrases should be included in the dictionary for translation learners. This would be easy to tackle in electronic dictionaries. To improve learners' understanding of these expressions, dictionaries should provide more contextual information within entries for this group of users.

Part of Speech	V.	N.	Adj.	Adv.	Total
Number of Looked-up content words	20	13	5	2	40
Average number of looked-up words					45
Percentage					89%

Table 2: The percentage of content words in the words looked up by users

4.1.2 The information categories users clicked on

Since a small number of the words users searched for were not found in the dictionary, the relevant clicks were not included in the results of this study. Repeated clicks were included, however, as this reflected the users' actual dictionary use behavior and needs.

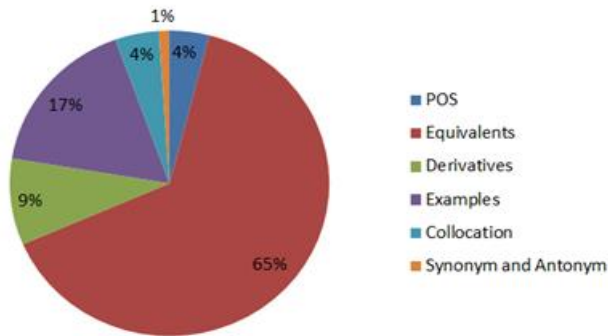


Figure 2: Pie Chart of click counts of information categories

Figure 2 shows that there are big differences between the click counts for different information categories. *Equivalents* are ranked first and account for 65% of the total click counts whereas *Examples* account for 17%, *Derivatives* 9% and *Collocation* and *Part-of-Speech* only 4%. *English equivalents* were the most consulted parts of the dictionary entries while synonyms and antonyms were the least looked-up elements. This might be because equivalents are usually the first step towards transferring an idea into English, but it could also be attributed to the learners' lack of translation skills. Frankenberg-Garcia (2011) found that users do not know which information to look up beyond L1–L2 equivalents. This finding is consistent with that of Atkins and Varantola (1997), who found that in L1–L2 translation, consulting and checking foreign language equivalents accounted for 77% of dictionary look-ups. Our follow-up interviews revealed that the users regarded the provision of English equivalents as a basic requirement for a dictionary and some even went so far as to claim that the provision of English equivalent was sufficient for translation most of the time. If necessary, the learners also browsed other information categories. For example, when they did not know the usage of the equivalent, they looked at other information such as "examples". Interviewees responded that examples could enhance their understanding of the equivalents and could serve as models in translation. When asked about their preference, those who did not browse examples said that examples could be very useful. They did not look at the examples just because they forgot to do so, or because they thought that the equivalents were sufficient for their purposes. If they had looked at examples, their expression would have been more natural and idiomatic. This finding is different from that of Chan (2014: 34), who found that in determining the meaning of words and making sentences, Chinese monolingual English dictionary users relied most on examples (90%), and then definitions (63.6%). Examples could help users learn about the detailed usage of words because they demonstrate the specific use of words in context, but some translation learners in our study lacked the skills to use dictionaries in translation. As

translation learners, they missed the opportunity to learn new words and expressions that could have been used in their translations later.

4.1.3 Selection of equivalents

To understand users' preference in choosing equivalents, we chose ten words with three equivalents and placed the most familiar ones in the second positions in the dictionary. The frequency of the users' choice among the three positions is as follows:

Equivalents (Frequency of choice)	Position		
	1	2	3
Chinese words			
发现(faxian)	discover (22)	find (3)	identify (0)
创造(Chuangzao)	create(28)	produce(1)	bring about (1)
付出(fuchu)	pay(27)	devote(22)	commit(3)
进步(jinbu)	advance(5)	progress(15)	improve (5)
预想(yuxiang)	anticipate(0)	expect(22)	speculate(1)
确定(queding)	determine(9)	define(1)	fix(1)
目标(mubiao)	objective(1)	goal(11)	aim(3)
成果(chengguo)	achievement(13)	gain(1)	harvest(8)
不懈(buxie)	untiring(5)	unremitting(29)	unrelenting(1)
战胜(zhansheng)	defeat(17)	vanquish(2)	conquer(4)
Total	127	107	27

Table 3: English equivalents in three positions and respective selection frequency

Table 3 shows that selection of the first equivalent was most frequent (48.7%). Selection of the second equivalent did not fall far behind (41%). Selection of the third equivalent was the smallest (10.3%). These results confirm the finding of other researchers (Tono 1984; Li 1998) about users' strategy in using a dictionary: they tended to utilize the beginning of an entry. Tono (1984) found that dictionary users tended to choose the first definition unless clear information to reject it was indicated. Li (1998) found that one factor responsible for mistranslation was that dictionary users tended to choose the first equivalent in the dictionary entry. We also wondered why the second equivalents were nearly as popular as the first equivalents. In follow-up interviews, respondents indicated

that they preferred to choose the equivalents they were most familiar with. In the first place, these choices could give them assurance. In the second place, these choices could facilitate their expression because they are more familiar with the usage of these words than other words. For those words they encountered before, this strategy offered learners opportunities to use them again and ultimately could contribute to the acquisition of these words.

4.2 Application of retrieved information

Previous studies (Varantola 1994, Atkins and Varantola 1997) highlight the need for a more in-depth analysis of dictionary use during a translation task. We hold that analysis of dictionary information application would be a step toward that end. Varantola (1998) has argued that it is difficult to evaluate the use of words in the translation product because translators use different standards for their choices. However, we believe that this analysis is significant and feasible. Although there are different ways to evaluate a translation product, we can judge whether the use of words is grammatically correct or not. The analysis of the content can inform us of the application ability of translator trainees, that is, whether a user can adapt the information from a dictionary to the context of a translation text. In other words, this analysis can reveal the particular linguistic and transfer needs of users in translation. To be more effective, bilingual dictionaries should gear their information toward the needs of translation learners as most of them claim that translators and translation learners are their target users. For example, if users have difficulty in choosing the correct part-of-speech form of a word, dictionaries can provide more instruction in presentation of definition, senses or examples.

As for the operation of this analysis, we focus on how well users applied the dictionary information to the translation task by conducting errors analysis, correctness analysis, and by examining possible causes for errors.

4.2.1 Error analysis

From the products of students, we determined that students' errors in using the dictionary information fall into two categories: parts of speech and collocation. We offer a focused case study that illustrates our wider findings in these categories.

For errors in tense, we take the verb "预想 yù xiǎng" (which literally means "expect").

The original sentence:

我们要有所进步、有所发现、有所创造，常要付出比预想多出许多的努力。

Suggested translation: *To make some progress, discoveries or creations, we must make more efforts than expected.*

In the original text, "预想 yù xiǎng" (which literally means "expect") is used as a noun. In the translation, it could be used as a noun. If students want to express it as a verb in English, they must shift the word class of the equivalents.

As "预想 yù xiǎng" (which literally means "expect") is labeled as a verb in the four dictionaries, the dictionary in the study provides three verbs in the *Equivalents* (*anticipate; expect; speculate*) and noun forms (*anticipation; expectation; speculation*) in the *Derivative* category. In the categories such as *Collocation* and *Examples*, it also provides the verb form.

Among the forty-seven students, thirty-three students used "expect". Video recordings informed us that only twenty-two students looked up "yù xiǎng" in the dictionary and eighteen users chose "expect" in the equivalents. That is, most of students chose the most familiar equivalent. A closer observation found that in the use of the word "expect", seven instances of incorrect usage were found. The following sentence fragments were taken from students' products of translation and students' IDs are in the parentheses.

We should often pay out much more effort than expecting if ... (Student 1091)

..., we will make more effort to make it than expecting before. (Student 1128)

We need to devote more than we had expected i f ... (Student 1006)

..., you must pay more than you have expected. (Student 1002)

We should pay much more efforts than our expect so that ... (Student 1133)

..., we usually need to pay out more effort than expect. (Student 1140)

So we should make great efforts which beyond our expectation ... (Student 1132)

From the translation products of students, we can see that two learners (Student 1091 and Student 1128) used it as "expecting" in the context. They used the gerund form of "expect" incorrectly, as there is no objective. The video informed us that these students just read the information category of *English equivalents*. If they had clicked on the other information categories such as *Derivative*, *Collocations* or *Bilingual examples*, they would probably have known more about this word and chosen its form appropriate for this context. Two learners (Student 1006 and Student 1002) used "expect" in the perfect tense. One (Student 1006) used it in the past perfect tense and the other (Student 1002) used it in the present perfect tense. Four students who did not consult the dictionary made similar mistakes. They used it in the past tense as "we expected ...". One user (Student 1133) took it as a noun. If the student had clicked on the *Past of Speech* or *Derivatives*, he or she would have found the noun form of "expect". One learner (Student 1132) transferred the phrase "beyond one's expectation" from the bilingual example "The beauties of the West Lake in spring were beyond his expectation." to his or her translation but the phrase was not used correctly. Such errors indicate that these learners lacked knowledge of the general grammatical rules. It would be an advantage if the description and explanation of some general rules could be incorporated into the dictionary as a separate section and individual dictionary articles could refer to them (Tarp 2008: 234).

Study Pages in the OALD8 (*Oxford Advanced Learner's Dictionary*, 8th Edition) could serve as a good example.

When we asked about errors in tense in follow-up interviews, some students responded that they forgot to find information about the different forms of the equivalents in the dictionary. When they were engaged in translation, they focused on the meaning transfer rather than on the form of the words they used. Others said that they failed to find enough tense information about verbs in the examples. This has implications for both teachers and dictionary compilers. Since the Chinese language does not have as many tense markers as the English language, it would be helpful to Chinese translation learners if the general rules of tense could be incorporated into Chinese–English dictionaries as a separate section. At the same time, teachers should draw learners' attention to this difference between two languages in their instruction.

With regard to collocation, we take "*mù biāo*" (which literally means "goal") as an example.

The original clause: 人确定自己的目标后...

Suggested translation: *After setting a goal, ...*

In the original Chinese text, *mù biāo* (which literally means "goal") collocated with the verb *què dìng* (which literally means "define"). For the verb *què dìng*, the dictionary provided three equivalents, namely, *determine*, *define*, *fix*. In the category of *Collocation*, the dictionary provides two phrases 确定日期 (*què dìng rì qī*) *fix a date*; 确定目标 (*què dìng mù biāo*) *set a goal/an aim*. Video recordings showed that twenty-one students did not look up this word. Twelve students looked up *què dìng* while six students looked up *mù biāo*. Eight students looked up both *què dìng* and *mù biāo*. The following sentence fragments were taken from students' products of translation and students' IDs are in the parentheses.

..., after people fixing on the goals, (Student 1091)

Once people determine their goals, ... (Student 1096)

After people determine their goals, ... (Student 5010)

Once you determine a goal, ... (Student 1003)

Once we set up our goal, ... (Student 1009)

When you set an ambitious goal, ... (Student 1010)

Once we make a clear goal, ... (Student 1132)

Once you have defined your goal, ... (Student 5002)

In the translation products of students, we found that collocate words used with *goal* were as follows: *determine* 9 times, *set* 6 times, *set up* 1 time, *make* 3 times, *define* 1 times, *fix on* 1 time. We checked these collocation choices with *goal* in BNC (*British National Corpus*) and found that in the first 100 collocates; *set* is ranked the second, *make* 13rd, *fix* 93rd. We regard these collocations as acceptable. However, for the collocations with *determine*, *set up*, *define*, and *fix on*, which are not found in the corpus, we regard these collocations as unacceptable.

In follow-up interviews, some students responded that they did not give much attention to collocation. They just picked the first equivalent or the one they were familiar with and then applied it in the translation. Sometimes, even when they looked for the information, they could not find it in the dictionary. In the translation task, the verb *què dìng* collocates with *mù biāo*. So most students who translated word by word felt it unnecessary to think about collocation. That is to say, these translation learners were not aware that the collocation of a word in two languages might be different. This also has implications for Chinese–English dictionary compilers. For example, when they provide equivalents for a word, they should also give more information such as definition, style and collocation, which can help users to identify the distinctions between equivalents and then make informed choices. This problem exists in almost all Chinese–English dictionaries available in China and had already been pointed out by researchers (Wei 2000; Hu and Zhang 2011; Xu 2012).

4.2.2 Correctness analysis

The success of word application in learners' translations was decided by the negotiation between the two researchers. When students looked up the same word but chose different expressions, their application of dictionary information would be regarded as successful if the words or expressions were used correctly. For instance, twenty-two students looked up "yù xiǎng" (which literally means "expect") in the dictionary. Eighteen of them chose "expect" and eleven of them used it correctly. At the same time, one student chose "speculate" and used it as "make more efforts than we speculate", and it was coded as correct. Another student who did not choose any of the equivalents but read the bilingual examples produced the following translation, "make more efforts than our first thought". It was also judged as successful application of dictionary information. Correctness ratio refers to the comparison between the number of words or expressions a student looked up and the words she or he applied correctly in their translation. For example, a student looked up twenty-nine words and found the equivalents of twenty-four words. If fifteen equivalents were used correctly in the translation, his or her correctness ratio would be 63%. The following is an overview of the correctness ratio.

	N	Mini	Maxi	Mean	SD
Correctness	47	.60	1.0	.82	.1188
Number	47				

Table 4: Correctness of dictionary information application

Li (1998) found that 73% of the lookups in English–Chinese translation is successful. Table 4 shows that in our study, 82% of consultation was successful and dictionary use contributed to Chinese–English (L1–L2) translation.

To investigate whether consultation preference has an impact on the correctness of their dictionary information application in translation, we divided students into three groups on the basis of their consulting preferences. Group one consisted of students who only looked up *Equivalents*; Group two consisted of students who consulted both *Equivalents* and *Examples*; Group three consisted of students who consulted *Equivalents*, *Examples* and *Collocation*. A one-way ANOVA test was carried out to explore whether there was a significant difference between the groups. The statistical results indicated that there is a significant difference between these groups ($F=6.968$, $P=0.002<0.01$). The following table shows the result.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.156	2	.078	6.968	.002
Within Groups	.494	44	.011		
Total	.650	46			

Table 5: Results of group difference test

For more detailed information about the difference, we made a further analysis.

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.
1	2.00	-.13281*	.03251	.001
	3.00	-.10193*	.03750	.027
2	1.00	.13281*	.03251	.001
	3.00	.03087	.03328	.628
3	1.00	.10193*	.03750	.027
	2.00	-.03087	.03328	.628

Table 6: Results of group difference identification test

The results show that there is a significant difference between group one and group two (Mean Difference=0.13281, $p=0.001<0.01$). A significant difference also exists between Group two and Group three (Mean Difference=0.10193, $p=0.03<0.05$). This implies that when student translators know more information about a word, the correctness of translation also increases. This is consis-

tent with Laufer (1993) who found that the combination of *definition* and *examples* contributes more to translation than *definition* or *examples* alone. The latter has no significant influence on translation. This could be explained by the fact that examples and collocation provide more detailed usage of words. The information can either demonstrate the usage of words in context or provide exemplary use of the word in the task. As no student only looked up the information category *Examples*, we cannot find out the relationship between examples and correctness of word use. The findings suggest that dictionaries should provide more information for learners and more importantly, translation trainers should encourage students to read more information in the dictionary.

In previous studies, researchers (Peters 2007; Lew and Doroszevska 2009; Chen 2011) also investigated the relationship between click counts and vocabulary retention, with various conclusions. Our study showed that there is no correlation between correctness and click counts. As learners cannot use them correctly in the first place, it can be predicted that there is no correlation between click counts and vocabulary retention. This could be attributed to the fact that the majority of clicks were on the *Equivalents* and this information category did not provide detailed information about the usage of words. In addition, the number of clicks does not necessarily equate to a deepening of understanding. Therefore, it can be concluded that information category rather than click counts has more influence on correctness of lexical information application.

5. Conclusion

This study investigated the use of an electronic dictionary (digitalized print dictionaries) by students in a natural setting. It provides a more complete picture of dictionary use by EFL learners as it utilized both positivistic and naturalistic research methods. It contributes to the literature of dictionary use study by providing a detailed description and analysis of users' dictionary information application during a translation task. The study has five findings: 1) EFL learners' consulting preferences include *Equivalents* and *Examples*; 2) EFL learners preferred to choose the most familiar equivalents; 3) EFL learners looked up content words and phrases more than other words; 4) EFL learners' errors in dictionary information application lie in collocation and parts of speech; 5) EFL learners' correctness of dictionary information application increases as students consult, or click on, additional information categories. These findings have implications for Chinese–English dictionary compilers, who are tasked with providing high-quality equivalents and examples as users relied heavily on them. For example, dictionary compilers could provide more information about equivalents so that users know the difference between them and make the informed choices. In the bilingual examples, compilers could demonstrate the usage of the equivalents so that users could learn how to use these words in

context. To enhance users' retrieval success, dictionaries could provide more content words and phrases. Results also confirm that translation teachers should encourage students to read more information categories in dictionary use.

This study focused on the looking up preferences of translation trainees and their application of dictionary information. It has some limitations: the number of participants is not very large, the number of dictionary entries is small and findings are based chiefly on an observation. To improve its reliability, further studies with mixed research methods should be conducted.

Acknowledgement

The research was funded by the College Philosophy and Social Science Foundation of Jiangsu Provincial Department of Education (Grant No. 2016SJB740007) and by The Third Phase of the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD: Phase iii) (Project No. 20180101). We are especially grateful to two anonymous reviewers and Dr Wei Xiangqing for their insightful comments and constructive suggestions. Our thanks also go to Roy Stamper from NC State University for proof-reading the manuscript and all the participants in the study.

References

A. Dictionaries

- Hornby, A.S. (Ed.).** 2010. *Oxford Advanced Learner's Dictionary* (8th edition). Oxford: Oxford University Press.
- Hui, Y. (Ed.).** 2012. *New Century Chinese–English Dictionary*. Beijing: Foreign Language Teaching and Research Press.
- Kingsoft PowerWord 2017*. Accessed on March 10, 2017. <http://www.iciba.com>.
- Yao, X.P. (Ed.).** 2010. *A Chinese–English Dictionary*. Third edition. Beijing: Foreign Language Teaching and Research Press.
- Youdao Dictionaries 7.0*. Accessed on March 10, 2017. <http://dict.youdao.com>.

B. Other literature

- Anthony, L.** 2010. AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software>.
- Atkins, B.T.S. and K. Varantola.** 1997. Monitoring Dictionary Use. *International Journal of Lexicography* 10(1): 1-45.
- Barnhart, C.L.** 1962. Problems in Editing Commercial Monolingual Dictionaries. Householder, Fred W. and Sol Saporta (Eds.). 1962. *Problems in Lexicography*: 161-181. Bloomington: Indiana University/The Hague: Mouton.

- Bogaards, P.** 1998. What Type of Words do Language Learners Look Up? Atkins, B.T. Sue (Ed.). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*: 151-158. Tübingen: Max Niemeyer.
- Bogaards, P.** 2003. Uses and Users of Dictionaries. Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography*: 26-33. Amsterdam/Philadelphia: John Benjamins.
- Chan, A.Y.W.** 2012. The Use of a Monolingual Dictionary for Meaning Determination by Advanced Cantonese ESL Learners in Hong Kong. *Applied Linguistics* 33(2): 115-140.
- Chan, A.Y.W.** 2014. Using LDOCE5 and COBUILD6 for Meaning Determination and Sentence Construction: What Do Learners Prefer? *International Journal of Lexicography*: 27(1): 25-53.
- Chen, Y.** 2010. Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. *International Journal of Lexicography* 23(3): 275-306.
- Chen, Y.** 2013. A Correlational Study between Dictionary Lookup Behavior and Vocabulary Acquisition under CALL Context. *Foreign Languages and Their Teaching* 5: 46-51.
- Chen, Y.** 2017. Dictionary Use for Collocation Production and Retention: A CALL-based Study. *International Journal of Lexicography* 30(2): 225-251.
- Chen, Y.Z.** 2011. The Use of Bilingualized English–Chinese Learner's Dictionaries: A Survey and An Experiment. *Lexicographical Studies* 2: 141-158.
- Dziemianko, A.** 2010. Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography* 23(3): 257-273.
- Dziemianko, A.** 2014. On the Presentation and Placement of Collocations in Monolingual English Learners' Dictionaries: Insights into Encoding and Retention. *International Journal of Lexicography* 27(3): 259-279.
- Frankenberg-Garcia, A.** 2011. Beyond L1–L2 Equivalents: Where Do Users of English as a Foreign Language Turn for Help? *International Journal of Lexicography* 24(1): 97-123.
- Frankenberg-Garcia, A.** 2015. Dictionaries and Encoding Examples to Support Language Production. *International Journal of Lexicography* 28(4): 490-512.
- Gromann, D. and J. Schnitzer.** 2016. Where Do Business Students Turn for Help? An Empirical Study on Dictionary Use in Foreign-language Learning. *International Journal of Lexicography* 29(1): 55-99.
- Hartmann, R.R.K.** 1983. The Bilingual Learner's Dictionary and Its Uses. *Multilingua* 2(4): 195-201.
- Harvey, K. and D. Yuill.** 1997. A Study of the Use of a Monolingual Pedagogical Dictionary by Learners of English Engaged in Writing. *Applied Linguistics* 18(3): 253-278.
- Hu, W.F. and Y.H. Zhang.** 2011. A Survey of Sense Representation in Chinese–English Dictionaries from the Perspective of Users. *Foreign Languages Research* 3: 78-84.
- Hu, W.F. and Y.H. Zhang.** 2013. The Effect of Definition Model in C–E Dictionaries on Chinese EFL Learner's English Productive Ability. *Journal of Foreign Languages* 34(5): 54-62.
- Kaneta, T.** 2011. Folded or Unfolded: Eye-tracking Analysis of L2 Learners' Reference Behavior with Different Types of Dictionary Interfaces. Akasu, K. and U. Satoru (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives, 22–24 August 2011*: 219-224. Kyoto: Asian Association for Lexicography.
- Koplenig, A., P. Meyer and C. Müller-Spitzer.** 2014. Dictionary Users Do Look Up Frequent Words. A Log File Analysis. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 229-250. Berlin/Boston: De Gruyter.

- Laufer, B.** 1993. The Effect of Dictionary Definitions and Examples on the Use and Comprehension of New L2 Words. *Cahiers de Lexicologie* 63(2): 131-142.
- Laufer, B. and L. Hadar.** 1997. Assessing the Effectiveness of Monolingual, Bilingual, and "Bilingualised" Dictionaries in the Comprehension and Production of New Words. *The Modern Language Journal* 81(2): 189-196.
- Laufer, B. and M. Hill.** 2000. What Lexical Information Do L2 Learners Select in a CALL Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3(2): 58-76.
- Laufer, B. and L. Melamed.** 1994. Monolingual, Bilingual and "Bilingualised" Dictionaries: Which are More Effective, for What and for Whom? Martin, W. et al. (Eds.). 1994. *Euralex 1994 Proceedings, Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands*: 565-576. Amsterdam: Vrije Universiteit.
- Lew, R.** 2011a. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4.
- Lew, R.** 2011b. User Studies: Opportunities and Limitations. Akasu, K. and U. Satoru (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives, 22–24 August 2011*: 7-16. Kyoto: Asian Association for Lexicography.
- Lew, R.** 2012. How Can We Make Electronic Dictionaries More Effective? Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 343-361. Oxford: Oxford University Press.
- Lew, R. and G.-M. de Schryver.** 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.
- Lew, R. and J. Doroszewska.** 2009. Electronic Dictionary Entries with Animated Pictures: Lookup Preferences and Word Retention. *International Journal of Lexicography* 22(3): 239-257.
- Li, L.** 1998. *A Study of Dictionary Use by Chinese University Learners of English for Specific Purposes*. Ph.D. dissertation. Exeter: University of Exeter.
- Liang, P. and D. Xu.** 2017. The Contribution of Dictionary Use to the Production and Retention of the Middle Construction for Chinese EFL Learners. *International Journal of Lexicography* 30(1): 85-107.
- Nesi, H.** 2014. Dictionary Use by English Language Learners. *Language Teaching* 47(1): 38-55.
- Pan, X.S.** 2012. *PMLX (Version V2012)* [Computer Software]. Wenzhou, China: Tianlangxing Software Studio. Available from: <http://www.tlxsoft.com>.
- Peters, E.** 2007. Manipulating L2 Learners' Online Dictionary Use and Its Effect on L2 Word Retention. *Language Learning & Technology* 11(2): 36-58.
- Sánchez Ramos, M.M.** 2005. Research on Dictionary Use by Trainee Translators. *Translation Journal* 12(10): 25-35.
- Tarp, S.** 2009. Reflections on Lexicographical User Research. *Lexikos* 19: 275-296.
- Tomaszczyk, J.** 1979. Dictionaries: Users and Uses. *Glottodidactica* 12: 103-119.
- Tono, Y.** 1984. *On the Dictionary User's Reference Skills*. Unpublished B.Ed. thesis. Tokyo: Gakugei University.
- Tono, Y.** 1989. Can a Dictionary Help One Read Better? On the Relationship Between EFL Learners' Dictionary Reference Skills and Reading Comprehension. James, G. (Ed.). 1989. *Lexicographers and Their Works*: 192-200. Exeter: University of Exeter Press.
- Varantola, K.** 1994. The Dictionary User as Decision Maker. Martin, Willy et al. (Eds.). 1994. *EURALEX 1994 Proceedings*: 606-611. Amsterdam: Vrije Universiteit.

- Varantola, K.** 1998. Translators and Their Use of Dictionaries. Atkins, B.T. Sue (Ed.). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*: 179-192. Tübingen: Max Niemeyer.
- Wei, X.Q.** 2000. Multilevel Exemplification in Active Bilingual Learner's Dictionaries. *Lexicographical Studies* 6: 68-75.
- Welker, H.A.** 2010. Dictionary Use: A General Survey of Empirical Studies. Brasília: Eigenverlag.
- Xie, X.Y.** 2014. *The Evaluation of Online English-Chinese Dictionaries*. Unpublished M.A. thesis. Shanghai, China: Fudan University.
- Xu, H.** 2012. Decoding and Encoding Functions of Examples in English Learners' Dictionaries: A Case Study of the Exemplification of the Word "Monopoly". *Lexicographical Studies* 2: 33-39.
- Yang, S.M.** 2017. Effects on English Learning of Online Dictionaries---Taking *Youdao Dictionary* and *Kingsoft Powerword* for Example. *Journal of Hainan Radio and TV University* 2017(1): 114-117.
- Zhang, H.P.** 2014. *ICTCLAS2014* (Institute of Computing Technology, Chinese Lexical Analysis System) [Computer Software]. Beijing, China: Chinese Academy of Science. Available from <http://ictclas.nlpir.org/downloads>.

Once Again Why Lexicography Is Science*

Tinatin Margalitadze, *Lexicographic Centre, Ivane Javakishvili Tbilisi State University, Tbilisi, Georgia* (tinatin.margalitadze@tsu.ge)

*Lexicography is a scientific practice aiming to bring dictionaries into existence*¹
Franz Josef Hausmann

Abstract: The article addresses some issues connected with the disciplinary status of lexicography. Drawing on the views of scholars such as L. Zgusta, R. Ilson, H. Wiegand, R. Gouws, H. Bergenholtz, S. Tarp, R. Lew and others, the author argues in favour of the viewpoint that lexicography is a science and that working on a dictionary is a scientific activity. The main issues tackled in the paper include understanding the complex nature of word meaning, the role of dictionaries in the description of word meaning and the development of lexical semantics. Attention is also paid to the definitional method of the study of word meaning, which is based on the analysis of dictionary definitions, components of the theory of lexicography, the relation between lexicographic theory and practice, and the teaching of lexicography as an academic discipline at universities.

The author argues that the right approach to lexicography and its disciplinary status is particularly important in our era of globalisation. Only state-of-the-art lexicographic and corpus resources will secure the future of many languages, particularly lesser-used languages, and such resources will not be created until lexicography receives proper recognition as a science with "big interdisciplinary vocation" (Tarp 2017); until lexicography is turned into an academic discipline through advanced theory of lexicography, through teaching lexicography at universities, etc.

Keywords: DISCIPLINARY STATUS OF LEXICOGRAPHY, MEANING OF WORDS, COMPONENTIAL ANALYSIS OF MEANING, DEFINITIONAL METHOD OF ANALYSIS, OED, THEORY OF LEXICOGRAPHY, LEXICOGRAPHIC PRACTICE, TEACHING LEXICOGRAPHY, ACADEMIC DISCIPLINE, MA IN LEXICOGRAPHY

Opsomming: Nog eens waarom leksikografie 'n wetenskap is. In hierdie artikel word 'n paar kwessies met betrekking tot die vakstatus van leksikografie aangespreek. Gebaseer op die sienings van vakkundiges soos L. Zgusta, R. Ilson, H. Wiegand, R. Gouws, H. Bergenholtz, S. Tarp, R. Lew en ander, argumenteer die outeur ten gunste van die siening dat die leksikografie 'n wetenskap is en dat die samestelling van 'n woordeboek 'n wetenskaplike aktiwiteit is. Die hoofkwessies wat in hierdie artikel aangespreek word, sluit die komplekse aard van woordbetekenis, die rol van woordeboeke in die beskrywing van woordbetekenis en die ontwikkeling van die leksi-

* This article is a revised version of a paper presented as keynote address at the Twenty-second Annual International Conference of the African Association for Lexicography (AFRILEX), hosted by the School of Languages and Literatures: African Language Studies Section, Rhodes University, Grahamstown, South Africa, 26–29 June 2017.

kale semantiek in. Daar word ook aandag geskenk aan die definisiële studiemetode van woordbetekenis, wat gebaseer is op die ontleding van woordeboekdefinisies, komponente van die leksikografieteorie, die verband tussen die leksikografiese teorie en -praktyk, en die onderrig van die leksikografie as 'n akademiese dissipline by universiteite.

Die outeur argumenteer dat die korrekte benadering tot die leksikografie en die vakstatus daarvan besonder belangrik in hierdie era van globalisering is. Slegs die heel nuutste leksikografiese en korpushulpbronne sal die toekoms van baie tale, spesifiek minder gebruikte tale, verseker, en voor hierdie hulpbronne geskep kan word, moet die leksikografie behoorlike erkenning as 'n wetenskap met "(n) groot interdisiplinêre taak" (Tarp 2017) geniet; moet gevorderde leksikografieteorie in 'n akademiese vakrigting verander word, moet leksikografie aan universiteite onderrig word, ens.

Sleutelwoorde: VAKSTATUS VAN LEKSIKOGRAFIE, BETEKENIS VAN WOORDE, KOMPONENSIËLE BETEKENISANALISE, DEFINISIËLE ANALISEMETODE, OED, LEKSIKOGRAFIETEORIE, LEKSIKOGRAFIESE PRAKTYK, DIE ONDERRIG VAN LEKSIKOGRAFIE, AKADEMIESE DISSIPLINE, MA IN LEKSIKOGRAFIE

1. Introduction

In 1747 Samuel Johnson writes in his famous work *The Plan of a Dictionary of the English Language*:

WHEN first I undertook to write an English Dictionary ... I knew that the work in which I engaged is generally considered as drudgery for the blind, as the proper toil of artless industry; a task that requires neither the light of learning, nor the activity of genius, but may be successfully performed without any higher quality than that of bearing burdens with dull patience, and beating the track of the alphabet with sluggish resolution. Whether this opinion, so long transmitted, and so widely propagated, had its beginning from truth and nature, or from accident and prejudice; whether it be decreed by the authority of reason or the tyranny of ignorance, that, of all the candidates for literary praise, the unhappy lexicographer holds the lowest place. (Johnson, in *Practical Lexicography*, 2008)

In the 21st century, some lexicographers in some countries still experience the same underappreciation of their work. Below I quote from an *Appeal of Georgian Lexicographers to the Georgian Government and the Academic Community*, adopted at the *First International Symposium in Lexicography* in Batumi (May 2010):

The present status of Georgian lexicography, which has a long history and rich heritage of tradition and experience, gives ground for serious concern. Regrettably, the colossal toil of lexicographers remains almost totally unappreciated in present-day Georgia, namely:

- The result of lexicographic work is not classed among scientific categories in general and in process of present-day contests and rating assessments in particular;

- Lexicographic work and its product are not yet entitled to the right of being competitive participants of modern grant competitions;
- Salaries of lexicographers are inadequate, compelling them to earn livelihood by means of other activities;
- Lexicographers are not awarded academic (scientific) degrees for the lexicographic products they create".²

It is probably worth mentioning here that Oxford University rewarded Samuel Johnson with a Master of Arts degree after the publication of his *Dictionary* in 1755 (he had studied only one year at Oxford, which he had to leave for financial reasons).

In July 2010, the text of the above-mentioned Appeal was forwarded to the Organising Committee of the XIV International Symposium of EURALEX (European Association for Lexicography), held in Leeuwarden, the Netherlands. Georgian lexicographers requested their European colleagues to discuss the Appeal of Georgian lexicographers and express their viewpoints concerning the issues raised in it. The Board of EURALEX agreed to add their voices to the Appeal.

"Within academia, lexicography is frequently overlooked, relegated to being a mere craft rather than an academic discipline. Such a notion is misguided and dangerous. Lexicographers not only study language for what it is, the central tool for communication, but also provide the means by which a language, and its underlying cultural values, may be taught and given full value within a society", wrote then president of EURALEX, Professor Geoffrey Williams in his letter addressed to the Georgian Government and the Academic Community³ (Williams 2016).

During the last couple of years the board of EURALEX has sent several such letters to colleagues from different countries to support their lexicographic projects or their campaigns for the rights of lexicographers.

The XVII EURALEX International Congress, held in Tbilisi, Georgia in September 2016 (<http://euralex2016.tsu.ge>) adopted a resolution addressed to UNESCO, national governments throughout the world, research funding agencies, and universities to acknowledge the status of lexicography as an academic discipline and promote the study of words and languages. 'Our multilingual world needs novel types of dictionaries, which requires proper recognition and support', states the resolution.⁴

Prior to the adoption of the resolution, a round-table discussion was organised within the framework of the congress which was dedicated to the status of lexicography. 'One of the hot topics today is whether lexicography should be seen merely as a "craft", or as a scientific academic discipline whose theory should be taught in universities, like mainstream linguistics', stated the synopsis of the discussion.⁵

These statements reveal that in the 21st century we may still come across

opinions that working on a dictionary is not a scientific activity. Such views are very damaging to lexicography and hinder its proper development.

Lexicography, which has a centuries-old history, has undergone significant evolution. Glosses, glossaries and dictionaries of hard words were replaced by dictionaries which incorporated the whole vocabulary of each particular language. Methods of description and study of word meaning also underwent drastic changes. Corpora of thousands of illustrative phrases and sentences from the works of literature emerged as the main tool of the study of meaning, paving the way for the development of scholarly lexicography. Lexicography has always kept abreast of the newest developments in linguistics and related sciences, frequently even being ahead of these developments. The advent of comparative-historical linguistics was reflected in the entries of the *Oxford English Dictionary on Historical Principles* (OED). The development of electronic corpora and corpus linguistics in the 1980s was also immediately reflected in lexicography, as the study of word meaning since then has been entirely based on the analysis of vast corpus data. The appearance of electronic dictionaries has opened completely new prospects for lexicography turning it into one of the most dynamic and rapidly developing fields of knowledge. Modern lexicography is a complex, multidisciplinary field incorporating multiple components, viz. semantic theories, corpus-based methods, methods and techniques for natural language processing, e-lexicography, research on dictionary use, dictionary criticism, dictionary didactics, terminology, etymology and so on. Consequently, claims that working on a dictionary does not constitute a scientific activity seem to be an unbelievable misunderstanding.

Some scholars such as Ladislav Zgusta (1971, 1992/93), Herbert Wiegand (1984), Robert Ilson (2012), Rufus Gouws (2012), Henning Bergenholtz (2012), Sven Tarp (2017), Robert Lew (2007) and others have published interesting articles on the status of lexicography. This is how Robert Ilson explains the lack of understanding of what lexicography really is:

Between them, the academics, professional lexicographers, and computerniks provided a round view of lexicography as a whole. The problem was, however, that each group had on its own a limited view of the subject. The academics had their Ideas; the computerniks, their Algorithms. But too often, alas, they seemed to lack detailed knowledge of what dictionaries are actually like and how dictionaries are actually produced. On the other hand, the professional lexicographers seemed often to lack detailed knowledge of linguistics; and their superbly detailed knowledge of Really Existing Dictionaries seemed often to be limited to those they had actually worked on ... but lexicographers have scant time or incentive to contribute to learned journals: after all, they have dictionary deadlines to meet. (Ilson 2012)

In his article "Lexicography as an Independent Science", Sven Tarp (2017) gives an interesting classification of different viewpoints on the disciplinary status of

lexicography, himself advocating the view that lexicography is "a science with its own independent core and a big interdisciplinary vocation" and that it should be treated as "an independent discipline with its own theory, own tasks and own methods". The independent disciplinary status of lexicography is also supported by H. Bergenholtz, R. Gouws (Bergenholtz and Gouws 2012), T. Bothma and D. Prinsloo (Bothma et al. 2016).

In this article I also want to formulate my viewpoints on this issue as a practical lexicographer with the experience of working on general and terminological dictionaries (*Comprehensive English–Georgian Dictionary* (CEGD), *English–Georgian Military Online Dictionary* (EGMD), *English–Georgian Biology Online Dictionary* (EGBD), *English–Russian–Georgian Technical Online Dictionary* (ERGTD)), as a scholar who has studied different theoretical aspects of lexicography and as a lecturer who teaches lexicography at all three university levels.

2. Understanding the Complex Nature of Word Meaning

From my personal observation, one of the reasons for the above-mentioned simplistic attitude towards lexicography, stating that it is not a science, stems from the superficial approach to the intricate phenomenon of meaning and related issues.

"As you surely know, one of the many surprising facts about the discipline of linguistics in the 20th century was that the study of lexis and meaning was largely neglected in America, Britain, and their spheres of influence. Honourable exceptions were in the European Saussurean tradition — notably German semantic field theorists such as Trier, Porzig, and Weisgerber and the Romanian Eugene Coseriu; British Firthians such as Halliday and Sinclair, Russians such as Mel'cuk and Apresjan, and others. But these past researchers were hampered by, among other things, lack of evidence and the political crises of their time", writes Patrick Hanks in the new proposal of the University of Wolverhampton "Studying meaning in the 21st century".

One of the reasons may be traced back to descriptive linguistics, which treated the lexical level of language as peripheral and non-structural for decades, concentrating on the description of phonological and morphological systems of language.

After being the philologists' prime object of investigation in the nineteenth century, the lexicon had been neglected in favour of syntax and phonology, as it was more difficult to describe and encapsulate it in rules. Vocabulary was deemed the least significant part of a language by the structuralists. Some of them even doubted that vocabulary was a part of a language. Ullmann in his *Semantics* also confirms that semantics was mostly formal the first three-quarters of the twentieth century and that lexicology was hardly regarded as a branch of linguistics.⁶

This approach to the phenomenon of meaning was also reflected in the methodology of componential analysis which drew its inspiration from structural phonology and like distinctive phonological features, the combination of which describes each phoneme, tried to describe meaning on the basis of a restricted set of semantic components (Geeraerts 2010: 70-80). It is genuinely surprising for me how one could believe that it was possible to describe meaning the same way as a phoneme with a finite set of features.

The complex nature of meaning is determined by the complexity of the cognition of the world with which it is closely connected. Cognition of the world is a multi-step, multifaceted process of perception, generalisation, formation of concepts, etc. A word is not only the main nominative but also the main cognitive unit of a language and its lexical meaning is determined by the reflection of some segment of extralinguistic reality, i.e. a class of things, events, etc. (denotatum) in our minds, in the mind of a language community. Meaning is a concept (designatum) attached to a word. Lexical meaning reflects not a segment of reality (denotatum) but the concept (designatum) that a language community has about it. The world around us is infinite, therefore describing meaning with a finite number of features and formalising it the same way as phonology or syntax was doomed to failure, but such views discouraged its study. As a consequence, if the scientific study of meaning was impossible, then lexicography, which was primarily involved in the study of words and their meanings, could not be a science. Later, this disregard for the content plane of language changed, and nowadays different theories of lexical semantics study meaning from many different angles (Geeraerts 2010), but it has left its mark on the understanding of the essence of lexicography.

The above-mentioned approach to the study of meaning is even more surprising as the dictionaries which emerged in the 19th and 20th centuries provided excellent scientific studies of meaning reflected in their word entries. The proof of this is one of the methods of componential analysis of meaning applied by Georgian linguists (following the tradition of Soviet linguistics), the so-called definitional method of analysis (Margalitadze 2014). The school of linguistics at Tbilisi State University (mostly English philologists) following theories of some Russian (e.g. V.G. Gak) and foreign linguists (e.g. American E. Nida) viewed meaning as a structure consisting of semantic components arranged in a hierarchical order. The Georgian linguist Mary Iankoshvili (1972) regarded the meaning of a word as a structure consisting of a core and peripheral potential semes. According to her theory, the core consists of a grammatical categorial semantic component (form which expresses meaning), a lexical categorial (hyponymic) semantic component and a differential seme or semes. Potential semes are arranged around the core; they reflect different features of denotatum described by the meaning of a word which is characteristic of denotatum or is ascribed to it by a language community. In other words, the core corresponds to the archisemes and differential semes of V.G. Gak; archilexemes and distinctive semes in the terminology of Pottier; or the common and diagnostic

semes of E. Nida. In general, these theories of word meaning distinguished archisemes or hyponymic semantic components, differential semantic components and potential or supplementary semantic components. Traditional lexicographic practice of the second half of the 19th century (OED and its European counterparts) regarded word meaning the same way and defined words in an analytical way by splitting them up into more basic semantic components, Distinctive-Feature Semantics, in other words. This methodology of defining meaning in the 19th century dictionaries follows the tradition of Aristotelian and Thomistic philosophy, which is known as a definition 'per genus proximum et differentias specificas'. The above-mentioned dictionaries described not only hyponymic and differential features of meaning. They also paid a lot of attention to the description of supplementary features of meaning, different potential semes which served as the basis for the development of transferred meanings of polysemous words, and were the basis of metaphor, metonymy and other mechanisms of semantic change.

To illustrate: the OED, while defining the word *father*, alongside lexical categorial (hyponymic) and differential semantic components (*a kinship term, nearest male ancestor*), provides numerous supplementary components: *a male ancestor more remote than a parent, esp. the founder of a race or family, a forefather, progenitor* (definition 2); *one who institutes, originates, calls into being* (definition 3.a); *one who exercises protecting care like that of a father; one who shows paternal kindness; one to whom filial reverence and obedience are due* (definition 4.a); *applied to God, expressing His relation to Jesus, to mankind in general (considered either as His offspring, as the objects of His loving care, or as owing Him obedience and reverence), or to Christians (as His children by regeneration or adoption)* (definition 5.a), etc. (see Figure 1).

In the entry for *heart*, the OED describes not only the hyponymic component of its meaning — the *bodily organ*, or the differential semantic component *The hollow muscular or otherwise contractile organ which, by its dilatation and contraction, keeps up the circulation of the blood in the vascular system of an animal* (definition 1.a) — but various definitions of the entry reveal different supplementary semantic components ascribed to the concept of *heart* by the English language community: *the seat of life* (definition 2); *the seat of one's inmost thoughts and secret feelings* (definition 6.a); *the seat of emotions* (definition 9.a); *the seat of love or affection* (definition 10.a); *the seat of the mental or intellectual faculties* (definition 12), *the seat of courage* (definition 11.a), etc. As reported by South African colleagues, *the seat of courage* in some African languages is the liver and not the heart. Interestingly, heart surgeons would argue that the heart is not the *seat of anything*, but just a pump. It is exactly the existence of these potential semantic components, different features associated with the same object of reality in different languages, that makes the study of meaning worthwhile and interesting.

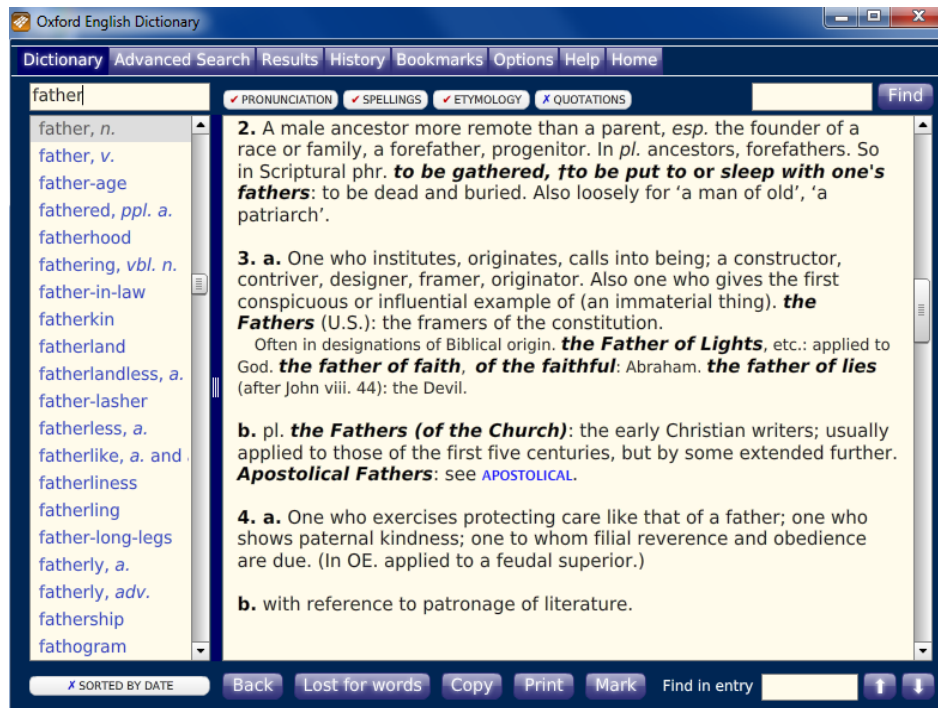


Figure 1: Entry of *father* from the OED

Hanks (2000) argues that the meaning potential of each word is made up of a number of components. These components may be activated cognitively by other words in the context in which they are used and are linked in a network which forms the semantic base of the language. This holds enormous dynamic potential for saying new things and relating the unknown to the known.

Thus the meaning of each word is unique, it consists of a unique combination of semantic components, therefore the meaning of each word is to be analysed individually. As Zgusta justly stresses in his *Manual of Lexicography*, what lexicographers have at their disposal is utterances, concrete instances of the usage of a word in a particular context. On the basis of the study of utterances, lexicographers deduce meaning or meanings of a word. Lexicographers of the 17th–18th centuries knew this quite well. Samuel Johnson collected 250 000 quotations from 500 sources for his dictionary. The 19th century lexicographers developed this method further and the OED team was able to collect 10 million quotation slips to be analysed for their dictionary.

Did lexicographers know what meaning was or how to describe it? Undoubtedly they knew it very well, they created and used corpora for their

research and they described word meaning in a way that transformed dictionaries into the main tools of study of meaning in the following decades.

Dictionaries of the 18th and 19th centuries did not use semantic theories to describe meaning, as there were none, but they created these theories through each word-entry and gave impetus to the development of lexical semantics.

As mentioned above, the method which was developed to study the semantic structure of a word and its semantic components was called the definitional method of analysis, which is based on the comparison and analysis of definitions of comprehensive explanatory dictionaries. Especially noteworthy in this regard are the *Oxford English Dictionary on Historical Principles* (OED) and *Webster's Third New International Dictionary*, whose definitions had become the basic source for the semantic study of English words before the advent of corpus linguistics and its methods.

The growth of the Internet in recent decades, the introduction of corpora as well as corpus linguistics have provided unprecedented opportunities for more objective studying of and research into language and meaning; however, it is not the case that meaning was not studied in previous decades.

3. Theory of Lexicography and its Components

Another reason for not regarding lexicography as a science is the view that lexicography has no theory. I fully agree with Gouws (Gouws 2012) that the authority of some European scholars who voice these claims is partly responsible for such views.

In 1983, at the founding congress of EURALEX, the German linguist Herbert Wiegand (Wiegand 1984) formulated the structure and components of metalexicography: 1. History of lexicography; 2. General theory of lexicography; 3. Research on dictionary use; 4. Criticism of dictionaries.

The general theory of lexicography is subdivided into 4 constituent theories:

- A. General Section;
- B. Theory of organisation;
- C. Theory of lexicographical research on language;
- D. Theory of the lexicographical description of language.

In the general section, Wiegand singles out three components: 1. Purposes of Dictionaries; 2. Relationship to other theories; 3. Principles from the history of lexicography.

Theory B is concerned with the organisation of labour in the three fields of activity.

Theory C comprises three components: 1. Data collection; 2. Data processing; 3. Computer assistance. In theory D two components are distin-

guished: 1. Dictionary typology; 2. Textual theory for lexicographical texts (i.e. the structure of lexicographical texts).

This was an excellent starting point for the development of a unified theory of lexicography and a unified understanding of its components, which has not happened. Defining the scope of lexicographic theory is important, otherwise many theoretical issues will not be sufficiently researched and treated in scientific literature. The study of the theoretical issues is important in the transitional period from printed to online media and particularly at present, when lexicography is at the crossroads of new developments in the era of the Internet and modern technologies.

From my point of view, a theory of lexicography accumulates and develops the knowledge necessary for lexicographers in dictionary production and is made up of the following components:

1. General lexicographic theory

This part of the lexicographic theory comprises the essence and functions of lexicography, dictionary typology, different theories necessary for dictionary production, i.e. theories of lexical semantics, methods of semantic research, including methods of corpus linguistics, theory and methods of natural language processing, etc.

2. History of Lexicography

3. Genres of Lexicography

This part of lexicographic theory includes a description of the lexicographic principles underlying different genres of lexicography: comprehensive monolingual dictionaries; comprehensive bilingual dictionaries; monolingual, bilingual and multilingual learner's dictionaries; historical dictionaries; terminological dictionaries; specialised dictionaries and so on. This section also comprises electronic lexicography and the changes it has brought about in the actual production of dictionaries. Genres of lexicography study the methodology of planning different stages of dictionary production, selection of sources, data collection and processing, producing entries for different types of dictionaries and modern technologies used in the production of different types of dictionaries. I view criticism of dictionaries in this section, as criticism of different types of dictionaries should be based on the knowledge of the genres.

4. Research on Dictionary Use

What is practical lexicography? How is the production of dictionaries connected to the theory?

From our experience, the actual production of dictionaries is not simply the application of theory to practice. The knowledge of the theory of lexicography and its components described above is the basis for lexicographers while planning and implementing their dictionary project. Practical work on a dictionary starts with:

1. The plan of a dictionary, detailed description of the principles that the dictionary will be based on, principles of selection of lemmas, treatment of homonyms, multiword units, etc.; description of the sources, principles of data collection and data processing, etc. For this work lexicographers need the theoretical knowledge mentioned above, knowledge of the target group and their needs and preferences, research on dictionary use and knowledge of other studies and experiments in the field. While planning a dictionary, lexicographers need knowledge of the history of the development of the same type of dictionaries, specificities of the genre, etc. Thus, at the very start, while planning their dictionaries, lexicographers need knowledge of the theory: of the history, of the genre specificities and so on.

2. The second stage is data collection for the dictionary. At this stage lexicographers need the knowledge of general lexicographic theory, theories of lexical semantics, methods of the study of meaning, etc. They also need knowledge of data collection and data processing experience in the genre and so on.

3. The actual compilation of entries is by no means an activity where lexicographers do not need theory. Each entry is unique with a unique meaning which a lexicographer needs to investigate on the basis of the sources and data collected. For each entry, a lexicographer goes through the stages of data collection, data processing, checking sources, deducing meanings, selecting illustrative material, studying connotation of the meaning and range of application and so on and so on (Zgusta 1971).

Theory and practice of lexicography do not exist independently of each other; it is not a ready theory which is uncritically applied in practice. Knowledge of theory is necessary for practical work and practical work is not simply compilation but work based on sound theoretical knowledge and the study of each unique meaning, undertaken by knowledgeable lexicographer-scholars. Each lexicographic project enriches the theory of lexicography with new solutions discovered by lexicographers working on different projects. Lexicographers may need to develop completely new principles for the creation of some dictionaries, but they still need to know the existing best practices to find better solutions for their projects.

"What is called the theory of lexicography is not something opposed to lexicographic practice, nor is it an endeavor that largely coincides with linguistics (theoretical or otherwise)" (Zgusta 1992/93: 137). Robert Lew (2007: 212) understands metalexigraphy as the "theoretical foundation to lexicographic practice".

What can be deduced from the above? Is working on a dictionary a "craft"? We strongly believe that it is a scientific activity rather than a "craft". We believe that only the highly competent, broadly educated lexicographers can work on the creation of dictionaries. The work of such a scholar is creative and intellectual and in its process it is impossible to make a distinction between its

general theoretical and current applied aspects. Consequently, we find the interpretation of lexicography expressed in the following phrase by Franz Josef Hausmann, a prominent German lexicographer and lexicographic theoretician more acceptable: "Lexicography is a scientific practice aiming to bring dictionaries into existence". We think that such an approach is more correct and adequate, giving a better idea of the essence of the subject (Meladze 2016).

4. Should the Theory of Lexicography Be Taught?

"One of the hot topics today is whether lexicography should be seen merely as a 'craft', or as a scientific academic discipline whose theory should be taught in universities, like mainstream linguistics". This is a statement from the synopsis of the Round Table Discussion at Tbilisi Congress in September 2016.

At the founding congress of EURALEX in 1983, mentioned earlier, the British scholar John Sinclair (Sinclair 1984) raised the issue of setting up a master's course in lexicography which would contribute to transforming lexicography from a practical activity into an academic discipline and would develop lexicography in close relation with information technologies, computer linguistics, general linguistics and lexicographic practice.

While developing the curriculum for the MA programme in lexicography at Tbilisi State University, we took into consideration the above-mentioned views, as well as our understanding of the theory of lexicography and its components. The programme comprises the following courses: word meaning and methods of its research; main genres of lexicography; history of lexicography; introduction to corpus linguistics and corpus-based lexicography; theories of lexical semantics; practical courses in general and specialised lexicography and so on.

We fully agree with John Sinclair that it is the unity of theory and practice that turns lexicography into an academic discipline, and with Sven Tarp (2017) that "lexicographical practice can be transformed into a 'scientific activity' when it is guided by an advanced theory (provided this theory is lexicographical)".

5. The Georgian Case

As mentioned in the introduction, views that working on a dictionary is not a scientific activity are very damaging to lexicography and hinder its proper development. The adverse results of underappreciation of lexicography can be well seen by the observation of processes taking place in my native language, Georgian.

Lexicography was a well-developed field of knowledge in Georgia and people involved in lexicographic work were respected by the academic community, as well as by Georgian society. The *Comprehensive English–Georgian Dictionary* was a research project of the Department of English Philology for

more than 30 years and when we started the publication of the Dictionary on a letter-by-letter basis in the 1990s, the presentation of the first fascicle, the letter A, was attended by the intellectual elite of Tbilisi of that time.

This attitude started to deteriorate after the break-up of the Soviet Union and the consequent period of political turmoil. Within ten to fifteen years, interest in lexicography started to decline and this short period proved to be enough to have grave consequences for Georgian lexicography. As referred to earlier, the *Appeal of Georgian Lexicographers to the Georgian Government and the Academic Community* (May, 2010) stated, the status of Georgian lexicography gave ground for serious concern. The appeal expressed their regret that the colossal toil of lexicographers remained almost totally unappreciated in Georgia.

Such circumstances eventually led to a shortage of qualified lexicographers working in the field, a shortage of academic dictionaries, the cessation of terminological work, deterioration of knowledge of foreign languages and quality of translations, etc. One more consequence was the decline of interest in published dictionaries and their application in teaching foreign languages. These processes were further aggravated by new methods for teaching foreign languages. These methods spread to the schools and higher-education institutions of Georgia greatly diminished the role of translation and reduced the practice of using the native language in the process of teaching foreign languages. This naturally led to the elimination of the use of bilingual/translation dictionaries, with an accompanying shift toward the use of explanatory, i.e. monolingual dictionaries (Margalitadze and Meladze 2016). From the same period of decline the Georgian language has been exposed to the comprehensive influence of the English language: the Internet and modern information and communications technologies; growing international contacts as a result of the years of regained independence; the free market economy, new entrepreneurial and legal relations; revolutionary advancements almost in every field of science and technology were linked with the formation of new concepts, with new terms which have naturally inundated Georgian directly via English. Unnecessary loans from English started to flood the vocabulary of Georgian, gradually taking the form of an avalanche, engulfing dozens of Georgian words on a daily basis: დისემინაცია – *diseminatsia* (Eng. dissemination), აროგანტული – *arogantuli* (Eng. arrogant), დაქენსელება – *dakenseleba* (Eng. to cancel), პრირეკვიზიტი – *prirekviziti* (Eng. prerequisite), ალარმირება – *alarmireba* (Eng. to alarm), ოვერლაპი – *overlapi* (Eng. to overlap), ბულით ფონთები – *bulit pointebi* (Eng. bullet points), პატერნი – *paterni* (Eng. pattern), and so on. All the above-cited loans have equivalents in Georgian, sometimes even several equivalents. These tendencies were even more alarming in terminology. New terms, even multi-word terms, were introduced into the Georgian language mainly by means of transliteration: პრეციპიტაცია – *pretsipitatsia* (Eng. precipitation), შაპერონი – *shaperoni* (Eng. chaperone), ქემოატრაქტანტი – *kemoatraktanti* (Eng. chemoattractant), ტრანზიციული მუტაცია – *tranzitsiuli mutatsia* (Eng. transitional mutation), რეზიდუალური სტრესის პატერნი – *rezidualuri*

stresis paterni (Eng. residual stress pattern) and so on. Our recent study has revealed that 90% of terms are introduced into Georgian as transliterated forms of the corresponding English terms. The number of such loans is so extensive that it already hinders communication in society and is a constant source of irritation to the Georgian public. Georgian lexicography should have served as a filter for this situation; the dictionaries should have protected Georgian from chaotic processes and professional lexicographers should have investigated different strategies for introducing emerging new concepts into the lexis of Georgian. During this period it was necessary to compose and publish new English–Georgian terminological dictionaries, to compose new European–Georgian type dictionaries and to intensify the work on the new edition of the *Explanatory Dictionary of the Georgian Language*. It was necessary to revise existing dictionaries, to compose Georgian corpora and terminological databases, to develop language technologies for the Georgian language and so on. Instead, lexicography in Georgia was in a critical state.

The struggle for saving Georgian lexicography started in 2010, with the first symposium in lexicography. The appeal of Georgian lexicographers and the support letter of EURALEX helped to develop a more positive attitude towards lexicography in Georgia. The most important achievement was the setting up of a committee for the enhancement of lexicography in Georgia at the Ministry of Education and Science of Georgia. The Committee is working on a National Programme in Lexicography. MA and PhD programmes in Lexicography were launched at Tbilisi State University, but the damage done to the language and terminology is so great that it will take years of hard work and dedication to mitigate these consequences and to produce a new generation of dictionaries for the Georgian language.

6. Conclusion

From our observation, the viewpoint that working on a dictionary is not a scientific activity is determined by a lack of understanding of the complex nature of word meaning as well as the complexity of its description. The complex nature of meaning is determined by the complexity of the cognition of the world with which it is closely connected. Another reason for such an approach to lexicography is the opinion that lexicography has no theory. Such views hinder the proper understanding of lexicography as a complex, multidisciplinary field incorporating multiple components. From our point of view, the theory of lexicography accumulates and develops the knowledge necessary for lexicographers in dictionary production and is made up of the following components: general lexicographic theory, history of lexicography, genres of lexicography and research on dictionary use. A dictionary is created according to a well-prepared model which is based on a sound theoretical approach. It is the unity of theory and practice that turns lexicography into an academic discipline.

This paper also discussed the Georgian case in order to highlight how the neglect of lexicography in Georgia during the last 10–15 years has led to the deterioration of the State language of Georgia.

The right approach to lexicography and its disciplinary status is particularly important in our era of globalisation. Only state-of-the-art lexicographic and corpus resources will secure the future of many languages, particularly lesser-used languages. Such resources will not be created until lexicography receives proper recognition as a science with "big interdisciplinary vocation" (Tarp 2017). These resources will not be created until the realisation that dictionaries are "great cultural vehicles", repositories of our languages, so vital for the preservation of our national identities. The creation of such resources is not cheap, but governments and societies should realise that this is an investment in the preservation of our languages and cultures, an investment in the democracy of our multilingual world.

Endnotes

1. Hausmann, F.J. (1985)
2. The full text of the appeal is available at the following URL: <http://blog.dictionary.ge/en/archives/114>.
3. The full text of the EURALEX letter is available at the following URL: <http://blog.dictionary.ge/en/archives/134>.
4. The full text of the Resolution of the XVII EURALEX International Congress (September 2016) is available at the following URL: <http://euralex.org/resolution2016/>.
5. The recording of the Round Table discussion is available at the following URL: <http://euralex2016.tsu.ge/media.html>.
6. Quoted from: Henri Béjoint's *The Lexicography of English*, p. 264.

References

- Béjoint, H.** 2010. *The Lexicography of English*. Oxford: Oxford University Press.
- Bergenholtz, H. and R.H. Gouws.** 2012. What is Lexicography? *Lexikos* 22: 31-42.
- Bothma, T.J.D., R.H. Gouws and D.J. Prinsloo.** 2016. The Role of e-Lexicography in the Confirmation of Lexicography as an Independent and Multidisciplinary Field. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity, Ivane Javakishvili Tbilisi State University, Tbilisi, Georgia, 6–10 September, 2016*: 109-116. Tbilisi: Ivane Javakishvili Tbilisi State University. Available at: <http://euralex.org/category/publications/euralex-2016/>.
- (CEGD)** Margalitadze, T. (Editor-in-chief), G. Meladze, A. Chanturia et al. 1995–2012. *Comprehensive English–Georgian Dictionary* (Vol. I–XIV); the Online Version www.dict.ge. Tbilisi: Ivane Javakishvili Tbilisi State University, Lexicographic Centre.
- (EGBD)** Khundadze, G., G. Meladze, T. Margalitadze et al. (Eds.). 2014. *English–Georgian Biology Online Dictionary*. <http://bio.dict.ge>. Tbilisi: Ivane Javakishvili Tbilisi State University, Lexicographic Centre.

- (EGMD) Khundadze, G., G. Meladze, T. Margalitadze et al. (Eds.). 2010. *English–Georgian Military Online Dictionary*. <http://mil.dict.ge>. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.
- (ERGTD) Khundadze, G., G. Meladze, T. Margalitadze et al. (Eds.). 2016. *English–Russian–Georgian Technical Online Dictionary*. <http://techdict.ge>. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.
- Gak, V.G. 1977. *Comparative Lexicology*. Moscow: International Relations (In Russian).
- Geeraerts, D. 2010. *Theories of Lexical Semantics*. New York: Oxford University Press.
- Gouws, R.H. 2012. Theoretical Lexicography and the *International Journal of Lexicography*. *International Journal of Lexicography* 25(4): 450-463.
- Hanks, P. 2000. Do Word Meanings Exist? *Computers and the Humanities* 34: 205-215.
- Hausmann, F.J. 1985. Lexikographie. Schwarze, Christoph and Dieter Wunderlich (Eds.). 1985. *Handbuch der Lexikologie*: 367-411. Königstein/Ts.: Athenäum.
- Iankoshvili, M. 1972. Core Structure of Word Meaning. *Foreign Languages at School* 3: 3-11 (In Georgian).
- Illson, R. 2012. IJL: The First Ten Years — And Beyond. *International Journal of Lexicography* 25(4): 381-385.
- Johnson, S. 1747. The Plan of a Dictionary of the English Language. Fontenelle, Thierry (Ed.). 2008. *Practical Lexicography: A Reader*: 19-30. Oxford/New York: Oxford University Press.
- Lew, R. 2007. Linguistic Semantics and Lexicography: A Troubled Relationship. Fabiszak, M. (Ed.). 2007. *Language and Meaning. Cognitive and Functional Perspectives*: 217-224. Frankfurt am Main: Peter Lang.
- Margalitadze, T. 2014. Polysemous Models of Words and Their Representation in a Dictionary Entry. Abel, A., C. Vettori and N. Ralli. (Eds.). 2014. *Proceedings of the XVI EURALEX International Congress: The User in Focus, EURALEX 2014, Bolzano/Bozen, Italy, July 15–19, 2014*: 1025-1037. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism. Available at: <http://euralex.org/category/publications/euralex-2014/>.
- 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/>.
- Meladze, G. 2016. Towards the Scientific Status of Lexicography. *Spekali* 10. Electronic Bilingual Scholarly Journal. Tbilisi: Ivane javakhishvili Tbilisi State University. Available at <http://www.spekali.tsu.ge/index.php/en/article/viewArticle/10/99/>.
- Nida, E.A. 1975. *Componential Analysis of Meaning*. The Hague/Paris: Mouton.
- (OED) *Oxford English Dictionary on Historical Principles*. (Vol. I–XIII; suppl. I–IV). Oxford: Oxford University Press.
- Sinclair, J. 1984. Lexicography as an Academic Subject. Hartmann, R.R.K. (Ed.). 1984. *LEXeter '83 Proceedings. Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983*: 13-30. Tübingen: Max Niemeyer.
- Tarp, S. 2017. Lexicography as an Independent Science. Fuertes-Olivera, Pedro A. (Ed.). 2017. *The Routledge Handbook of Lexicography*: 19-33. London: Routledge.

- Wiegand, H.E.** 1984. On the Structure and Contents of a General Theory of Lexicography. Hartmann, R.R.K. (Ed.). 1984. *LEXeter '83 Proceedings. Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983*: 13-30. Tübingen: Max Niemeyer.
- Williams, G.** 2016. In Praise of Lexicography, and Lexicographers. 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*: 77-88. Tbilisi: Ivane Javakhishvili Tbilisi State University. Available at <http://euralex.org/category/publications/euralex-2016/>.
- Zgusta, L.** 1971. *Manual of Lexicography*. Prague: Academia / The Hague: Mouton.
- Zgusta, L.** 1992/93. Lexicography, Its Theory, and Linguistics. *Dictionaries* 14: 130-138.

The Effectiveness of Using Dictionaries as an Aid for Teaching Standardization of English-based Sports Terms in Serbian

Mira Milić, *Faculty of Sports and Physical Education,
University of Novi Sad, Serbia (mmilic@uns.ac.rs)*

Tatjana Glušac, *Faculty of Law and Business Studies Dr. Lazar Vrkatić,
Novi Sad, Union University, Belgrade (tatjana.glusac@gmail.com)*
and

Aleksandra Kardoš, *doctoral student at Faculty of Philosophy,
University of Novi Sad, Serbia (sandra.kardosh@gmail.com)*

Abstract: This paper reports on the effectiveness of a new teaching method employing dictionaries as an aid for teaching the standardization of English-based sports terms in Serbian. The research was conducted among the students of a sports faculty in 2017 by means of a questionnaire distributed to the students both at the beginning of the second half of an ESP course and again at its end. Its aim was to measure the students' progress related to the acquisition of standardized sports terms in Serbian as an indicator of the effectiveness of the new teaching method. The findings generally indicate a certain degree of improvement of the students' knowledge of standardized sports terminology, though a less than satisfactory amount of progress regarding their linguistic competence. Even though the outcomes did not fully meet the goals set in advance, they do provide solid arguments for further efforts in developing and monitoring dictionary use in teaching the standardization of English-based sports terms in Serbian within the ESP curriculum and, even more importantly, for the systematic education of dictionary usage as part of the mother tongue curriculum.

Keywords: DICTIONARY USE, ENGLISH, SERBIAN, ESP TEACHING, SPORTS TERMINOLOGY, STANDARDIZATION

Opsomming: Die effektiwiteit van die gebruik van woordeboeke as hulpmiddels in die onderrig van die standaardisering van Engelsgebaseerde sportterme in Serwies. In hierdie artikel word verslag gedoen oor die effektiwiteit van 'n nuwe onderrigmetode waarin woordeboeke benut word as hulpmiddels in die onderrig van die standaardisering van Engelsgebaseerde sportterme in Serwies. Hierdie navorsing is in 2017 uitgevoer onder die studente van 'n sportfakulteit deur middel van 'n vraelys wat aan die begin van die tweede helfte van 'n ESD-kursus en weer aan die einde daarvan aan die studente uitgedeel is. Dit

het die evaluering van die studente se vordering ten opsigte van die aanleer van gestandaardiseerde sportterme in Serwies ten doel gehad wat 'n aanduiding sou wees van die effektiwiteit van die nuwe onderrigmetode. Die bevindings dui oor die algemeen op 'n mate van verbetering van die studente se kennis van gestandaardiseerde sportterminologie, maar dui ook op minder bevredigende vordering ten opsigte van hul taalkundige vaardigheid. Alhoewel die resultate nie die doelwitte wat aanvanklik gestel is ten volle bevredig het nie, verskaf dit steeds grondige argumente vir verdere pogings in die ontwikkeling en monitering van woordeboekgebruik in die onderrig van die standaardisering van Engelsgebaseerde sportterme in Serwies in die ESD-kurrikulum en, selfs belangriker nog, vir die sistematiese onderrig van woordeboekgebruik as deel van die moedertaalkurrikulum.

Sleutelwoorde: WOORDEBOEKGEBRUIK, ENGELS, SERWIES, ESD-ONDERRIG, SPORTTERMINOLOGIE, STANDAARDISERING

1. Introduction

This paper reports on the effectiveness of an innovative course of English for Specific Purposes (henceforward referred to as ESP) focused on dictionary use in teaching the standardization of sports terms in Serbian. The course was taught at the Faculty of Sport and Physical Education in Novi Sad in 2017. Given the fact that the literature (Chun 2004; Lew 2011) confirms that dictionaries are not used in language teaching as much as might be necessary, the course was innovative for three reasons: (1) it required the use of dictionaries for teaching standard English-based sports terms in Serbian as a teaching resource, not as a reference book, (2) it emphasized the teaching of the standardization of sports terms, which is not a common practice in ESP teaching, although it is highly desirable, and (3) it promoted the development of contact linguistic competence, which is stressed in literature as an essential component of an ESP course (Prčić 2014). Special attention is paid to the effectiveness of an English–Serbian Dictionary of Sports Terms (*Englesko–srpski rečnik sportskih termina*) (Milić 2006), to be referred to henceforward as ESDST, since it is the first bilingual sports dictionary whose Serbian equivalents are subjected to the process of standardization. Building on the first author's previous research into dictionary use in teaching ESP (Milić 2016), it is assumed that such a dictionary can significantly contribute to developing a proper approach towards the increasing influx of lexical and other borrowings from English into Serbian. For this reason, the dictionary should be given the status of one of the compulsory ESP teaching resources for building contact linguistic competence (henceforward referred to as CLC), which is "a type of linguistic knowledge related to the use of elements, i.e., words and names, from English as the nativized foreign language in a non-English language that regularly comes into contact with it" (Prčić 2014: 147). The paper is divided into six sections. Following the introduction, Section 2 outlines the theoretical background, Section 3 deals with research methodology, Section 4 presents the research method, Section 5 elabo-

rates on the research results, while the last, Section 6, summarizes the conclusions. The paper also contains an Appendix, which is an English translation of the Final questionnaire.

2. Theoretical framework

This research belongs to the field of teaching ESP and is focused on building ESP students' English–Serbian CLC. Broadly speaking, the requirement to acquire this special type of knowledge related to contact and contrastive aspects of English and a non-English language has occurred as a result of the current global domination of English (cf. Prčić 2011), which has given rise to an incessant influx of lexical and other borrowings from English into other languages that come into contact with it (cf. Furiassi, Pulcini and Rodríguez González 2012). Under such circumstances, non-English language users are increasingly faced with the need to acquire a new type of linguistic knowledge that has only recently been recognized as CLC (Prčić 2014). According to Prčić (2014: 148-150), building CLC comprises three aspects: practical, theoretical, and pedagogical. The practical aspect focuses on the consistent use of standardized English-based elements in Serbian. The theoretical aspect conflates the achievements of three linguistic disciplines: contact linguistics (in terms of different levels of adaptation of English borrowings in Serbian), contrastive linguistics (regarding the principles of establishing correspondence and equivalence between the particular units of two languages), and sociolinguistics (in terms of the principles of language planning and standardization). Lastly, the pedagogical aspect of building CLC concerns the method of building CLC institutionally, more specifically within the EFL/ESP curriculum. Building on the theoretical aspects of this concept, the exposition in this paper is focused on the practical and pedagogical aspects of this knowledge, which involves predominantly institutionalized forms of language planning, lexicography, and language teaching (Prčić 2014: 152). An overview of the past practical endeavors in this field in non-English languages shows that certain efforts have been made in the field of language planning and lexicography, predominantly in specialized terminology (cf. Laurén and Picht 1993; Myking 1997; Gromann and Schnitzer 2015). In this respect, Antia (2000: XIX) states that "investment in local eco-systems by way of creating and planning terminology in less widely used languages is actually very much in tune with globalization." Keeping in mind the practical aspects of CLC, the following sections deal with CLC-related endeavors in the Serbian linguistic community and the use of dictionaries in the language teaching process.

2.1 CLC within the framework of English–Serbian language contact

With respect to the interlingual contacts of English and Serbian, the past few decades have been marked with significant research focused on the linguistic

standardization of English-based elements in the general lexicon of Serbian. A few studies worth mentioning here are: an exhaustive study of English-based lexical and other borrowings in Serbian (Prčić 2011), the first dictionary of recent Anglicisms (Vasić, Prčić and Nejgebauer 2011; originally 2001), a respelling dictionary of personal names from English (Prčić 2008; originally 1998), and an English–Serbian dictionary of geographical names (Prčić 2004). Narrowing the topic down to specialized terminology in Serbian, the common thread of recent findings is the belief of experts in specialized fields that it is only the English term that can convey the meaning of a term accurately (cf. Prčić 2011: Chapters 11 and 12; Milić 2015a; Silaški 2012). Faced with this overwhelming preference of views regarding the high communicative potential of borrowed English terms, a considerable effort has been devoted to the standardization of specialized registers, the most important examples being related to the fields of computers (Prčić 1996), economics (Silaški 2012), medicine (Mičić and Sinadinović 2013), and sport (Milić 2015a). However, in order to foster knowledge of the standardization requirements related to English–Serbian language contact, the latest research findings suggest the need for building CLC through the educational system, as part of the normal curriculum, which is the practical component of building CLC. To do so it is necessary to employ not only relevant language teaching techniques and resources, but also institutionalized forms of language planning and lexicography. Given that terminological standardization requires not only proposing rules and principles, but also monitoring and updating them (cf. Auger 1986, cited in Cabré 1999: 49; Prčić 2011: 247), it is extremely important to educate members of the language community in standardization issues, as well as to carefully monitor feedback from them and update the set standard with new linguistic and specialized requirements. To establish such two-way communication, it is necessary to put more effort into the compilation of lexicographic resources that could be used not only as reference sources but also as teaching resources, which is the primary subject dealt with in this paper.

2.2 The use of dictionaries in language teaching

Even though dictionaries are essential reference books for learning a foreign language, recent research findings indicate that their role in language teaching is often neglected. Empirical interest in the matter, however, has emerged and grown over the past two decades (cf. Lew 2011: 1; Hulstijn and Atkins 1998). Generally, the findings of these studies indicate numerous advantages of dictionary-aided learning (Béjoint 2010; Chi 1998: 575; Hartmann 2001; Hayati and Fattahzadeh 2006; Yamaizumi 2014). One study goes even further, expounding that a dictionary-induced strategy in vocabulary learning is more successful than inferencing from the cognitive science perspective of connectionism (cf. Ellis 2003), since the "rich information of dictionary entries for target words can offer a complexity of connections when multiple aspects of knowledge are con-

structed" (Zou 2016: 382). However, despite the advantages, the use of dictionaries in language teaching has still not received much attention. According to a number of dictionary-use studies (cf. Chun 2004: 20; Lew 2011), the reason for the lack of interest in the pedagogical function of dictionaries is an insufficient knowledge of lexicographic conventions, which confirms an earlier observation that dictionary users need to be trained in how to use the dictionary in order to solve actual typical problems and questions (cf. Scolfield 1982; Vintean and Matiu 2010: 326; Cately 2009; Lew 2013; Akbari 2015). Building on this finding, Frankenberg-Garcia (2011) claims that teaching dictionary use should not start with the dictionary itself, but rather with the problems and activities that prompt dictionary consultation. Moreover, dictionary skills comprise a set of defined activities which users need to be able to execute (cf. Lew 2013; Nesi 1999). They should be mastered and honed both within the mother tongue and a foreign language curriculum alike as they not only raise students' awareness of linguistic matters, but also provide them with an abundance of necessary linguistic information and equip them with skills crucial for autonomous learning later in life (Cately 2009: 501).

Given that good mastery of vocabulary is particularly important for those who learn ESP (cf. Milić 2014: 82; Wu and Wang 2004), a specialized bilingual dictionary is one of the essential means for accomplishing this task in the Anglo-globalized world of today, which is increasingly faced with the requirement of individualization in learning English (cf. Rossner 1985: 98). To this end, Nation (2001) and Nesi (2013) point out that bilingual dictionaries might bring more advantages than monolingual ones, since they offer easily accessible and well thought-out L1-L2 equivalents. With this in mind, two perspectives arise. From the lexicographic perspective, a specialized bilingual dictionary could be used as an ESP teaching resource, which additionally calls for intensive and high quality lexicographic work (Milić 2015b: 184). Viewed from the teaching perspective, this necessitates rethinking and modification of the ESP curriculum, while also monitoring its effects. Employing dictionaries as a teaching resource in ESP courses could lead to desirable learning outcomes. What is more, an attempt should be made to incorporate dictionaries into task-based activities, since some authors (e.g., Sarani and Sahebi 2012) report that these activities are beneficial in teaching technical vocabulary.

Narrowing the topic of dictionary use in ESP teaching to the specialized register of sport in Serbian, research findings indicate that sports terms in Serbian are currently created most often by the adaptation of English terms through transshaping¹ and translation (cf. Milić 2015a). In light of the fact that the Internet offers an abundance of information which forces users to adopt information and linguistic expression in a noncritical and selective manner, new sports terms are often insufficiently adapted to the linguistic system of Serbian, which leaves a strong imprint on the L1 standard. A solution to these problems is not only the standardization of English-based sports terms in Serbian, but also training in terminological standardization involving the education of ESP learners as part of the normal curriculum. To this end, the first

bilingual dictionary of standardized sports terms has been compiled (Milić 2006), and efforts are being made to compile a new English–Serbian dictionary of sports terms in electronic form, as electronic dictionaries are particularly easy to use, being similar to other user-friendly electronic sources and applications (cf. Wang 2012). The model of the standardization of English-based sports terms in Serbian which was applied in the existing dictionary is built on a previous corpus-based study of ball game terms in English and Serbian (Milić 2004), which included six principles arranged in decreasing order of priority: bi-univocity, transparency, systematicity, productivity, concision, and frequency. In order to teach students to apply these principles in an appropriate manner, training in standardization was realized by means of lectures and regular task-based activities focusing on a particular principle of standardized adaptation of borrowed English terms in Serbian. In addition, the students involved in the study were also requested to do three compulsory online tests with multiple-choice answers for the questions posed, which are similar to the questionnaire in the Appendix. The principles are briefly defined and exemplified in the following paragraph².

Bi-univocity is the most important principle, according to which a given term should designate only one concept in a register, e.g., *7m line* > LINIJA SEDMERCA, but not SEDMERAC, which used to be the same translation equivalent of two English terms, *7m line* and *7m shot*. The second most important principle is transparency, which means that the concept a term designates should be inferred without a definition and that it should be motivated etymologically, semantically, or morphologically, e.g., *throwing* > BACANJE ZA LOPTOM, which is given preference over SUVANJE, as this is archaic. The third principle is systematicity, which means that a term must be in accordance with the linguistic standard of Serbian on the level of: orthography, phonology, and morphosyntax, e.g., *playoff* > PLEJOF, but not PLAYOFF, since this is a recently borrowed Anglicism in Serbian, which is adapted according to the acoustic impression. The fourth principle is productivity, which means that the standard term should imply a higher derivational and combining potential than its competitors, e.g., *held ball* > NOŠENA LOPTA, which is given preference over DRUGI KONTAKT S LOPTOM, a term/phrase used previously, since the standard term allows for several derivations of the modifier NOŠEN (NOSITI, NOSILAC, NOŠENJE), whereas the same is not true of the other term. Concision is the fifth principle, which gives preference to a term, justified from the aspect of linguistic economy, e.g., *offending player* > PREKRŠILAC, which is given preference over IGRAC KOJI JE NAPRAVIO PREKRŠAJ, which existed before. Finally, the sixth principle of frequency means that the standard term should be the term with the highest frequency of use, e.g., *corner kick* > KORNER which is given preference over UDARAC SA UGLA, which is used less frequently.

Concerning ESP teaching in the field of sport, practical steps towards the innovation of the ESP curriculum, focused on using dictionaries as an aid in teaching standardization, were taken in 2014, when research was conducted with master students of a sports faculty (cf. Milić 2016). Building on the results

of that research, the aim of this subsequent research conducted in 2017 with bachelor students of the same faculty was to test the effectiveness of the teaching of ESP using dictionaries as an aid for teaching the standardization of English-based sports terms in Serbian, which is the subject of the following section of this paper. Building on the respective findings, this paper will also attempt to justify the usage of an English–Serbian dictionary of standardized terms not only as a reference book, but also as one of the mandatory ESP teaching resources for building students' CLC.

3. ESP course design

The ESP course that served as the grounds for the study generally comprises 60 classes and is taken in the second semester. During the research period, students attended an innovative ESP course focused on the standardization of sports terminology in Serbian and the use of dictionaries, with special emphasis placed on making full use of the English–Serbian dictionary of sports terms³ (Milić 2006). To be eligible for the course, all students were expected to have reached a B1 level of English proficiency (Council of Europe 2001), which means that they have mastered a minimum of 2000 general lexical items (cf. Nation 2001: 15).

3.1 Aims of the innovative ESP course

In order to train students to be capable of dealing with the challenges encountered in the standardization of English-based sports terms in Serbian, learning is understood as a process-oriented activity in which "the individual develops understanding and awareness and creates possibilities for future learning" (Finney 2002: 73). From this standpoint, special emphasis is placed on good mastery of specialized vocabulary, which is essential for ESP learners (cf. Nation 2001: 187). As Nation (2001) advocates, this can be achieved through ESP exercises by means of exploiting a particular context with certain specialized vocabulary of continuing interest to students while helping learners grasp as much information about each new term as possible by providing them with appropriate activities that ensure multiple encounters with the new terminology, meaning-focused input and output, and fluency development. Moreover, students also need to be instructed on how to reach the standard L1 equivalent in case there is not a direct correspondent in their mother tongue. An inevitable teaching and learning resource in such ESP learning is a dictionary and it is of utmost importance that students learn how to use it for several reasons: (1) it encourages autonomous life-long learning, (2) it is a resource for the acquisition of new vocabulary, (3) it improves students' CLC, (4) it ensures students learn standardized terms, and (5) it contributes to the decrease of the influx of non-standardized English terms that permeate the students' mother tongue. Although dictionary-aided activities can be done in pairs or groups as well, for the aim of this research to be achieved, regular class activities had to be heavily dependent on individualized learning, in which the dictionary plays the role of

a teaching resource rather than a reference book. Accordingly, it was extremely important to encourage students not only to build the habit of dictionary use but also to learn how to make full use of dictionary information. The number of the respondents who took part in the newly-designed course allowed for this individualized approach, and the authors believed that such an approach would have more productive and long-lasting learning effects. In light of this scenario, this study aimed to assess how well students increased their CLC and to evaluate their progress achieved through the practical application of the dictionary-assisted learning contents related to the standardization of sports terms in Serbian.

3.2 An innovative ESP course program and its realization

In order to communicate the general idea of the innovative ESP content in brief, this exposition begins with a flowchart of the course content, shown in Figure 1, which is elaborated upon in more detail in the text that follows.

Using a dictionary in teaching the standardization of English-based sports terms in Serbian necessitated an enquiry into the extent to which students were informed about lexicographic resources and what type of information they looked up in these sources. The enquiry was realized through two administered surveys. The fact that the first half of the course dealt with less specialized texts directed the first survey towards determining the role of general dictionary use for fulfilling communication goals in English. Accordingly, Survey 1 was conducted at the beginning of the ESP course (February 26th). With the aim of gathering background information for the further teaching of standardization, the intention of this survey was to get information on the bilingual and monolingual dictionaries the students used in communicative situations of text reception. In order to get written proof of the information, students were instructed to make a list of reference sources they used. For this task, the students were offered three options: a yes/no question related to whether or not they used dictionaries; if the answer was positive, they were instructed to indicate what information they looked up in one or more dictionaries, as well as to make a list of dictionaries they used, while, if the answer was negative, they were instructed to indicate if they would instead apply a keyword search via the Internet. Even though other options are certainly available, the Internet was the only offered alternative to dictionary use since it had served as the predominant method of lexical disambiguation among previous generations. Building on these findings, an effort was made to prompt dictionary consultation in meaning-focused input/output exercises, which was followed by language-focused instruction (cf. Nation 2002: 267-272). Owing to the fact that ESP texts for reading predominantly deal with specialized topics that ESP students are mostly familiar with, the problem with input exercises is the possibility it enables of unknown words being simply learned rather than fully understood. It was thought that this could potentially be solved by providing suitably-

graded input in a number of different contexts, which would then be supported by language-focused instruction. Concerning meaning-focused output, it is possible to influence spoken/written production by careful designing and monitoring what vocabulary could be learned from the given tasks. At the end of the first half of the course (April 14th), students attended a lecture on different types of general dictionaries and the quality and quantity of dictionary information, which was followed by a brief introduction to the concept of language standardization and its impact on lexicographic description. Finally, at the very end of the first half of the course, the initial questionnaire was handed out, which will be commented on in more detail in the following section of this paper.

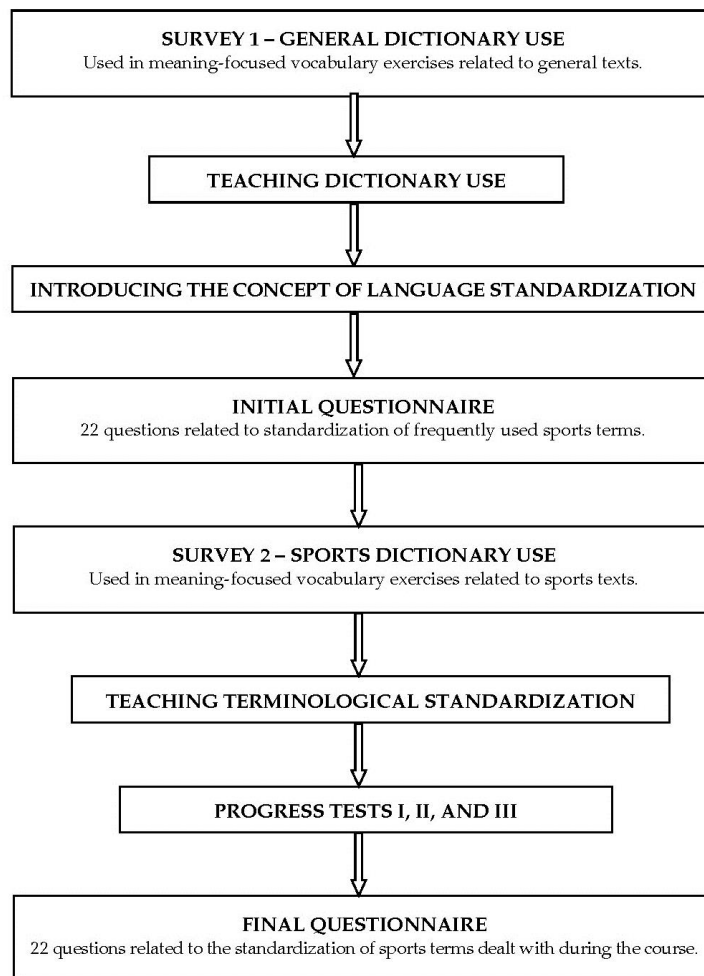


Figure 1: Innovative ESP course design

Survey 2 was conducted at the beginning of the second half of the course (April 18th), which focused on sports texts. The tasks were the same as in Survey 1, except that in place of general dictionaries, bilingual and monolingual dictionaries of sports terms were prioritized. The major finding of both Survey 1 and Survey 2 is that the general practice of the majority of students was to search the Internet via Keywords rather than consult dictionaries. This probably explains why few students managed to compile a list of more than three dictionaries. While even for those few that did regularly consult a dictionary, they were primarily interested in bilingual sources that provide L1–L2 equivalents, which means that they were likely unaware of what other information a dictionary can offer. This indicates a possible situational negligence towards dictionary use in language teaching and learning, perhaps due to the apparent predominance of the teacher-centered method (cf. Müller 2002: 717-8), as well as to more convenient access to the desired information using various Internet applications.

Following the compiling of the lists of dictionaries and a practical demonstration of how students should use them in meaning-focused input/output exercises, the students attended a lecture on the standardization of English-based sports terms in Serbian. This was followed by intermittent in-class discussions of terms of a specific sport from the aspect of standardization, which was complemented by homework assignments related to dictionary consultation aimed at finding information regarding a particular term. During the class, special emphasis was placed on standardization-focused instruction illustrated by examples in the ESDST (Milić 2006). If the dictionary did not provide proper examples, since it includes only terms representative of the five most popular ball games, students were instructed to employ analogies with similar dictionary entries. To motivate students towards higher academic achievement, they were offered an option of compiling an English–Serbian glossary of standardized terms for a sport with which they had dealt, each compilation requiring at least 50 entries, which they were expected to present orally with proper arguments. The main reference source for proposing a standard English-based equivalent in Serbian was the ESDST, which is based on the six principles model of standardization dealt with in Section 2.2. This activity was expected to provide not only an indication of the students' progress in learning, but also give insight into the existing state of sports terminology in Serbian, which is the first stage of standardization.

In order to gain insight into the students' progress in learning the standardization of English-based sports terms in Serbian, three online progress tests were conducted at monthly intervals, which students were expected to solve within a week. Each progress test consisted of 10 English terms with three equivalent terms in Serbian offered as options for each. Students were instructed to choose the one that best fits the Serbian standard, as well as to give an argument for the chosen answer. These tests revealed not only the score of correct answers but also the arguments governing the respondents' answers,

which provided indirect feedback for developing new teaching activities. Though this component is more relevant for constructing further teaching activities rather than the research itself, the findings can also be taken as an indirect source of information regarding the effectiveness of using dictionaries as an aid for teaching standardization as a main component of building English–Serbian CLC. The results of the three tests, which were calculated using IBM SPSS Statistics 20, are presented according to the principles of standardization (Table 1 below).

Progress test	Bi-univocity	Transparency	Systematicity	Productivity	Concision	Frequency
I	43.8%	65.65%	38.55%	85.4%	81.3%	66.7%
II	27.8 %	31.67%	50.75%	28.3 %	36.9%	51.5%
III	42.37%	41.35%	59.19%	18.4%	18.9%	41.3%

Table 1: Progress test results related to the principles of standardization

The most important result of the three tests is a decreasing number of correct answers for all the principles of standardization except for systematicity. This could be due to this principle being given special attention in ESP teaching or due to the students not mastering the specialized registers of most sports dealt with during the ESP classes, since the majority of these sports are not taught during the first year of studies. Regarding the arguments behind choosing answers, they appear to be highly diversified. From the highest frequency of use to the lowest, the results were: professional knowledge, lexical/grammatical knowledge, the process of elimination of incorrect answers, the frequency of term use, concision of a term, and knowledge retention from English classes. It is worth mentioning that the students generally demonstrated heavy reliance on professional vocabulary in use, which is probably the reason why they found it difficult to decide whether to borrow or translate terms from English into Serbian. In case of the latter, they showed an excessive reliance on the existing Serbian terms, most of which are stylistically marked units.

In response to this feedback information, further teaching activities were then focused on the adaptation of English-based terms in Serbian with special emphasis placed on the relevant reference sources of information, especially general and specialized dictionaries. Regarding this approach, the authors would also like to emphasize the need to intensify general and specialized lexicographic efforts in Serbian, corresponding additionally to the findings of Prčić (2016; 2018)

and Milić (2015b). Regarding the subjects of this study, a new homework assignment was also introduced which focused on the adaptation of an English term in Serbian that reflected a particular linguistic aspect of terminological standardization. Each word was carefully chosen from the reading text scheduled for a particular class. Students were expected to look up a particular terminological unit in different reference sources in order to find out its meaning, its L1–L2 translation equivalent, relevant grammatical information, and other details, as well as to note down the consulted sources. The answers were discussed at the beginning of the following English class, and the final solution was reached using the ESRST as the main source of reference. As already mentioned, the linguistic competence of students was additionally exercised through the discussion of an English–Serbian glossary of a particular sport that was presented in the form of an oral presentation by advanced students. Despite the students' heavy reliance on professional knowledge rather than the available lexicographic sources, the scores suggest strongly that this activity yielded positive results, since it sensitized students to recognize the potential danger of Anglo-Serbian pseudo-norms.

At the very end of the course (May 31st), the Final questionnaire was administered, and commentary dedicated to it is found in the next section.

4. Research method

The research for this study is based on a questionnaire that was conducted in 2017 with 255 first year students of the Faculty of Sport and Physical Education in Novi Sad. In order to examine the effectiveness of the dictionary-aided innovations in teaching ESP, it was necessary to make an assessment of the students' learning practices and to evaluate their progress in the practical application of the dictionary-assisted learning content related to standardization. To do so, two questionnaires were handed out after having obtained the Dean's consent. The Initial questionnaire was conducted on a sample of 167 examinees at the beginning of the second half of the ESP course (April 26th), whereas the Final questionnaire was conducted on a sample of 255 examinees at the end of the course (May 31st). It is worth mentioning here that one potential limitation of the study might be different numbers of examinees, since the Initial questionnaire was administered to 167 students, whereas the number of examinees in the Final questionnaire was 255. Even though the Final questionnaire contained a control question asking whether they had filled the Initial questionnaire or not, which further directed them to proceed with answering only if their answer was positive, it was impossible to match the results of the two research instruments due to the fact that the questionnaire was anonymous, so the cumulative research results for the enrolled students in 2017 are presented. The difference in the number of respondents was probably the result of the Initial questionnaire taking place after the first half of the semester (April 26th),

when students were not attending classes of English regularly, whereas the Final questionnaire was conducted immediately before the examination term, i.e., on 31st of May, when students were taking a more active role in their studies. In accordance with previous teaching experience of students of sport and the findings of Survey 1 and Survey 2, the students had demonstrated a preference for keyword search via the Internet and therefore this was included as one of the offered answers to each question. Accordingly, both questionnaires consisted of 22 multiple-choice questions related to the standardization of sports terms with predominantly four offered options: standard term, non-standard term, consult a dictionary, and keyword search via the Internet⁴. The difference between the Initial and Final questionnaire was a difference of exemplification rather than tasks. However, a more important differentiation between the two research instruments concerns the differing interpretation of the last two offered answers in each survey. Namely, *consulting a dictionary* and *keyword search via the Internet* in the Initial questionnaire was an indicator of the students' preferred reference source of information, whereas selecting these options in the Final questionnaire indicated the students' failure in terms of the standardization-related learning content. As shown in the Appendix, all standardization principles included several questions, excepting bi-univocity, which was expected to be easily understood since it reflects the requirements of the sports register. The principles are codified as follows: A (bi-univocity), B (transparency), C (systematicity), D (productivity), E (concision), and F (frequency). Accordingly, A examined the students' awareness of the requirement to have a different term for each sports concept (1); B examined the understanding of the meaning of terms in both languages, over-translation, and the use of archaic words in Serbian (2, 3, 4, 5); C included multiple aspects of English–Serbian linguistic standards: collocations, morphosyntax, nominal modifiers, the adaptation of Anglicisms in oblique cases, choosing between Anglicisms and translation equivalents, compounds and semi-compounds, the adaptation of the decimal point in Serbian, and the phonological and morphosyntactic adaptation of Anglicisms (6, 7, 8, 9, 10, 11, 12, 13); D examined the students' ability to apply derivation in order to get a single word-term in Serbian, or to cut down the number of words in a Serbian translation as much as possible (14, 15); E examined solving the problem of definitional translation,⁵ in which case it is justified to use an Anglicism or give preference to single-word terms or the fewest words of a phrasal term over multi-word polylexical ones (16, 17, 18); F tested the students' preference for Anglicisms over translation equivalents (question 19), as well as their understanding of the conditions in which the principle of frequency should (not) be applied (20, 21, 22).

Correct answers in both questionnaires are shown in Table 2 in percentages initially calculated for each question, and then for a set of questions related to a particular principle of standardization. A comparative analysis of the two questionnaires would be expected to provide information on the students' progress of learning the standardization of sports terms in Serbian, with

a special emphasis on the process of acquiring CLC, i.e., learning linguistic and English–Serbian contact and contrastive linguistic aspects of terminological standardization in Serbian, which would be expected to be the most demanding task for sports professionals. The findings of this analysis are presented in the following section.

5. Research results

The most important finding based on the comparison of the scores of the Initial and Final questionnaire (see Table 2) is a certain extent of improvement related to the six principles of standardization.

	Principles of standardization					
Questionnaire	Bi-univocity	Transparency	Systematicity	Productivity	Concision	Frequency
Initial	29.94%	43.71%	32.63%	19.76%	49.70%	33.53%
Final	51.76%	49.90%	40.15%	77.11%	50.46%	39.12%
Improvement	21.82%	6.19%	7.52%	57.35%	0.76%	5.59%
Percentage point increase	16.53%					

Table 2: Comparative indicators of correct answers in the Initial and Final questionnaire

Additionally, the results of the Initial questionnaire are similar to those obtained in 2014 with master students (Milić 2016: 373), since the lowest scores are for bi-univocity (29.94%), systematicity (32.63%), productivity (19.76%), and frequency (33.53%). Given that bi-univocity essentially concerns the technical aspect of standardization, the low score is probably due to a substantial disregard for, or inconsistency in the use of, terminological units, whereas systematicity and productivity likely reflect a lack of linguistic knowledge of English and Serbian alike. However, the low score of the frequency principle is contrary to the authors' original expectations, since the exemplified terms are believed to be units with a high frequency of use. The higher score of the trans-

parency principle, which reflects the technical aspect of standardization, might be explained by the Initial questionnaire's focus only on frequently used sports terms, whereas the high score of the concision principle is probably due to it reflecting the pragmatic aspects of the standardization process, which essentially concerns the use of terms.

Focusing on the progress in learning standardization, the findings of a comparative analysis of the Initial and Final questionnaires indicate improvement in all six principles, amounting to 16.53% on average. Additionally, the fact that the highest scores in the Final questionnaire were achieved regarding the principles of bi-univocity (51.76%) and productivity (77.11%) is encouraging, since these principles reflect higher-order linguistic and technical aspects of standardization in the model of Milić (2006; 2015a).

Guided by the findings of the three progress tests, as well as by the results of the previous research with master students, according to which the systematicity principle, which reflects the linguistic aspect of standardization, accounted for the lowest score on the test (cf. Milić 2016: 374), it was considered wise to do the assessment of the students' knowledge of the orthographic and grammatical standard of Serbian. To this end, a certain number of the non-standard answers provided in the Final questionnaire (see Appendix) were deliberately entered in grammatically and orthographically incorrect forms, as exemplified in question (12a) (the use of a decimal point in Serbian), questions (13b) and (20b) (the use of nonadapted Anglicisms), questions (8a) and (11a), (13a) and (21a) (nonstandard adaptation of noninflectional nominal modifiers in Serbian) and (20a) (the use of English-spelled terms in Serbian). The outcomes of these tasks show an average percentage of incorrect answers of 38.21% in the Initial questionnaire and 30.03% in the final one. The lower percentage of incorrect responses in the Final questionnaire suggests a certain level of improvement in terms of linguistic competence in Serbian. However, the analysis of the percentages of incorrect answers according to individual principles reveals a slight increase of 11.54% for the principle of systematicity in the Final questionnaire, a finding contrary to the authors' original expectations. Though this might be due to the higher number of examinees in the final testing and/or an increase in the amount of grammatically incorrect options in the Final questionnaire, these results suggest a need for rethinking the methods of teaching linguistic issues of standardization in ESP, and perhaps even more so in teaching English as a foreign language at the elementary and pre-intermediate level. Moreover, consideration should also be given to teaching standardization as part of the mother tongue curriculum in order to raise students' linguistic awareness of the rules of standardization.

Another indicator of the students' progress in the practical application of the learning content related to the standardization of sports terms in Serbian is the percentage of answers of dictionary use and/or keyword search via the Internet (see Table 3).

	Initial questionnaire	Final questionnaire
Looking up a word in a dictionary	12.92%	9.46%
Keyword search via the Internet	8.26%	5.19%

Table 3: Percentage of answers related to dictionary use and keyword search via the Internet

In light of the fact that the standard term served as one of the offered options in each question, the scores in Table 3 have been interpreted as indicators of the extent to which students had mastered the specialized sports terminology. Accordingly, it can be concluded that the students possessed a fairly advanced knowledge of specialized vocabulary, since the need to consult a dictionary or make a keyword search via the Internet accounts for a rather low percentage in both questionnaires. However, the more significant finding at this stage of research is the slight decrease in the students' preferred activity of searching the keywords via the Internet in favor of dictionary use, which is encouraging given that the Internet is not a reliable source of standard terms in Serbian.

6. Conclusions

The study presented in this article is a questionnaire-based investigation into the effectiveness of an innovative curriculum of ESP for undergraduate students of sport, focused on using dictionaries as an aid in teaching the standardization of English-based sports terms in Serbian, as a means of building CLC. To assess the students' learning and evaluate their progress in the practical application of the learning content related to the standardization of sports terms in Serbian, questionnaires were conducted after the second half of the ESP course in 2017 and again at its end (April 26th and May 31st, respectively). The findings indicate an average improvement in student performance of 16.53% in employing the six principles of standardization applied by Milić (2006), a reduction in grammatically and orthographically incorrect answers, and a slight decrease in students' preference for keyword Internet searching as a direct substitute for dictionary use. All things considered, the results suggest that an ESP course aimed at increasing students' awareness of standardization requirements through the use of user-friendly dictionaries would likely lead to positive learning outcomes. Moreover, the findings indicate the need for further research into the ESP dictionary-aided curriculum, as well as the need to pay more attention to educating dictionary users through the educational system, as part of the normal ESP curriculum and the mother tongue curriculum

alike. From a wider perspective, there appears to be a need to intensify lexicographic work and include a specialized English–Serbian dictionary in the basic ESP literature as one of the relevant teaching resources for building English–Serbian CLC. In order to eliminate the complication of the different numbers of examinees taking the Initial and Final questionnaires, which may have impacted the interpretation of the findings, further research should be even more carefully planned so as to motivate all students towards full cooperation. Perhaps more importantly, the effects of further activities related to dictionary use in teaching the standardization of English-based sports terms in Serbian should be the subject of ongoing monitoring aimed at building English–Serbian CLC among students, through assistance from qualified instructors.

Notes

1. According to Prčić (2011: 124), "'transshaping' describes the creation of a new form, whose inherent content is taken from English, but which is adapted to the orthographic and semantic standard of Serbian".
2. For more details, see Milić (2015a).
3. A detailed presentation of the macrostructure of the ESRST and its microstructure can be found in Milić (2015a).
4. A translated version of the final questionnaire is presented in the Appendix.
5. According to Prčić (2005: 177-178), definitional translation involves "a translation in the form of concise definition."

Acknowledgements

A concise version of this paper was presented at the *4th International Conference on English Language and Anglophone Literatures Today* (Novi Sad, 25th March 2017). The paper is a part of the research on Project No. 142-451-3684/2017-01/01, entitled *Using Dictionaries in Teaching English for Specific Purposes in Tertiary Education*, which is financially supported by the Secretariat for Higher Education and Scientific Research of the Autonomous Province of Vojvodina, Serbia.

The authors are indebted to the reviewers for their invaluable comments, suggestions, and insights.

References

- Akbari, Z. 2015. Key Vocabulary Learning Strategies in ESP and EGP Course Books. *International Journal of Applied Linguistics and English Literature* 4(1): 1-7. DOI: 10.7575/aiac.ijalel.v.4n.1p.1.
- Antia, B.E. 2000. *Terminology and Language Planning: An Alternative Framework of Practice and Discourse*. Amsterdam/Philadelphia: John Benjamins.
- Béjoint, H. 2010. *The Lexicography of English: From Origins to Present*. Oxford: Oxford University Press.

- Cabré, M.T.** 1999. *Terminology: Theory, Methods, and Applications*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/tlrp.1.
- Cately, Y-M.** 2009. Using the WordWeb Online Dictionary in an ESP Class. *ANALELE UNIVERSITĂȚII "DUNĂREA DE JOS" DIN GALAȚI FASCICULA XXIV ANUL II* 1(2): 501-507.
- Chi, M.L.A.** 1998. Teaching Dictionary Skills in the Classroom. Fontenelle, T., P. Hiligsmann, A. Michiels, A. Moulin and S. Theissen (Eds.). 1998. *Proceedings of the 8th EURALEX International Congress on Lexicography in Liège, Belgium, Part 2*: 565-577. Liège: Euralex. Available at: <http://euralex.org/publications/teaching-dictionary-skills-in-the-classroom/> [23 January 2017].
- Chun, Y.V.** 2004. EFL Learners' Use of Print and Online Dictionaries in L1 and L2 Writing Processes. *Multimedia-Assisted Language Learning* 7(1): 9-35.
- Council of Europe.** 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Ellis, N.** 2003. Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. Doughty, C. and M.H. Long (Eds.). 2003. *Handbook of Second Language Acquisition*: 33-68. Oxford: Blackwell.
- Finney, D.** 2002. The ELT Curriculum: A Flexible Model for a Changing World. Richards, J.C. and W.A. Renandya (Eds.). 2002. *Methodology in Language Teaching: An Anthology of Current Practice*: 69-79. Cambridge: Cambridge University Press.
- Frankenberg-Garcia, A.** 2011. Beyond L1-L2 Equivalents: Where Do Users of English as a Foreign Language Turn for Help? *International Journal of Lexicography* 24(1): 97-123. DOI: 10.1093/ijl/ecq038.
- Furiassi, C., V. Pulcini and F. Rodríguez González (Eds.).** 2012. *The Anglicization of European Lexis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Gromann, D. and J. Schnitzer.** 2015. Where Do Business Students Turn for Help? An Empirical Study on Dictionary Use in Foreign-language Learning. *International Journal of Lexicography* 29(1): 55-99. DOI: 10.1093/ijl/ecv027.
- Hartmann, R.R.K.** 2001. *Teaching and Researching Lexicography*. Harlow: Pearson Education.
- Hayati, M. and A. Fattahzadeh.** 2006. The Effect of Monolingual and Bilingual Dictionaries on Vocabulary Recall and Retention of EFL Learners. *The Reading Matrix* 6(2): 125-134.
- Hulstijn, J.H. and B.T.S. Atkins.** 1998. Empirical Research on Dictionary Use in Foreign-language: Survey and Discussion. Atkins, B.T.S. (Ed.). 1998. *Using Dictionaries. Studies of Dictionary Use by Language Learners and Translators*. Lexicographica. Series Maior 88: 7-19. Tübingen: Max Niemeyer. Available at: https://pure.uva.nl/ws/files/2239537/165046_Hulstijn_Atkins_1998.pdf [26 November 2016].
- Lew, R.** 2011. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4. DOI: 10.1093/ijl/ecq044.
- Lew, R.** 2013. Online Dictionary Skills. Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (Eds.) 2013. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*: 16-31. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Mićić, S. and D. Sinadinović.** 2013. Anglicizmi u jeziku medicinske nauke i struke [Anglicisms in the Language of Medical, Scientific and Professional Purposes]. Silaški N. and T. Đurović (Eds.) 2013. *Aktuelne teme engleskog jezika i nauke u Srbiji [Contemporary Topics of the English Language and Science in Serbia]*: 93-105. Beograd: CID Ekonomskog fakulteta.

- Milić, M.** 2004. *Termini igara loptom u engleskom jeziku i njihovi prevodni ekvivalenti u srpskom* [Ball Game Terms in English and their Translation Equivalents in Serbian]. Unpublished M.A. Thesis. Novi Sad: Faculty of Philosophy, University of Novi Sad.
- Milić, M.** 2006. *Englesko–srpski rečnik sportskih termina* [English–Serbian Dictionary of Sports Terms]. Novi Sad: Zmaj.
- Milić, M.** 2014. Process-oriented Approach to Translating Sports Research Papers from Serbian into English. Eraković, B. and M. Todorova (Eds.). 2014. *Topics in Translator and Interpreter Training: Proceedings of the Third Regional Workshop on Translator and Interpreter Training*: 71-86. Novi Sad: Faculty of Philosophy.
- Milić, M.** 2015a. Creating English-based Sports Terms in Serbian: Theoretical and Practical Aspects. *Terminology* 21(1): 1-22.
- Milić, M.** 2015b. Principles of Compiling an English–Serbian Dictionary of Sports Terms in the Modern Anglo-globalized World. *ESP Today: Journal of English for Specific Purposes at Tertiary Level* 3(2): 180-195.
- Milić, M.** 2016. An English–Serbian Dictionary of Sports Terms as an Aid in Teaching Standardization of English-based Sports Terminology in Serbian. Eraković, B. and M. Todorova (Eds.). 2016. *English Studies Today: Prospects and Perspectives. Selected Papers from the Third International Conference English Language and Anglophone Literatures Today (ELALT 3)*: 369-381. Novi Sad: Filozofski fakultet.
- Müller, V.** 2002. The Use of Dictionaries as a Pedagogical Resource in the Foreign Language Classroom. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the 10th EURALEX International Congress, EURALEX 2002 Copenhagen, Denmark August 13–17, 2002*: 717-721. Copenhagen: Center for Sprogteknologi. Available at: http://www.euralex.org/proceedings-toc/euralex_2002/ [8 January 2015].
- Myking, J.** 1997. Standardization and Language Planning of Terminology: The Norwegian Experience. *International Congress on Terminology, UZEI and HAEE-IVAP, UZEI; HAEE-IVAP*: 227-248. Donostia: Gasteiz. Available at: <http://www.uzei.com/modulos/usuariosFtp/conexion/archivos194A.pdf> [1 October 2013].
- Nation, I.S.P.** 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nation, P.** 2002. Best Practice in Vocabulary Teaching and Learning. Richards, J.C. and W.A. Renandya (Eds.). 2002. *Methodology in Language Teaching: An Anthology of Current Practice*: 267-272. Cambridge: Cambridge University Press.
- Nesi, H.** 1999. The Specification of Dictionary Reference Skills in Higher Education. Hartmann, R.R.K. (Ed.). 1999. *Dictionaries in Language Learning. Recommendations, National Reports, and Thematic Reports from the Thematic Network Project in the Area of Languages, Sub-Project 9: Dictionaries*: 53-67. Berlin: Freie Universität Berlin.
- Nesi, H.** 2013. Dictionary Use by English Language Learners. *Language Teaching* 47(1): 38-55. DOI: 10.1017/S0261444813000402.
- Laurén, C. and H. Picht.** 1993. Vergleich der terminologischen Schulen. Laurén, C. and H. Picht (Eds.). 1993. *Ausgewählte Texte zur Terminologie*: 493-539. Wien: IITF.
- Prčić, T.** 1996. Adaptacija i standardizacija kompjuterske terminologije iz engleskog jezika kod nas [Adaptation and Standardization of Computer Terminology from English in Serbian]. Šćepanović, Bogić (Ed.). 1996. *Standardizacija terminologije* [Standardization of Terminology]: 203-205. Beograd: Srpska akademija nauka i umetnosti.

- Prčić, T.** 2004. *Englesko–srpski rečnik geografskih imena* [An English–Serbian Dictionary of Geographical Names]. Novi Sad: Zmaj.
- Prčić, T.** 2008. *Novi transkripcioni rečnik engleskih ličnih imena. Drugo izdanje* [A New Respelling Dictionary of Personal Names from English. Second Edition]. Novi Sad: Zmaj.
- Prčić, T.** 2011. *Engleski u srpskom. 2. izdanje* [English within Serbian. Second Edition]. Novi Sad: Filozofski fakultet.
- Prčić, T.** 2014. Building Contact Linguistic Competence Related to English as the Nativized Foreign Language. *System* 42: 143-154. DOI: 10.1016/j.system.2013.11.007.
- Prčić, T.** 2016. Kakav nam opšti rečnik srpskog jezika najviše treba [What Kind of General-purpose Dictionary of Serbian We Need Most]. Ristić, S., Lazić Konjik, I. and N. Ivanović (Eds.) 2016. *Leksikologija i leksikografija u svetlu savremenih pristupa*: 87-117. Beograd: Institut za srpski jezik SANU.
- Prčić, T.** 2018. *Ka savremenim srpskim rečnicima, Prvo, elektronsko, izdanje* [Towards Modern Serbian Dictionaries, The First Digital Edition]. Novi Sad: Faculty of Philosophy. Available at <http://digitalna.ff.uns.ac.rs/sadrzaj/2018/978-86-6065-454-2>.
- Rossner, R.** 1985. The Learner as Lexicographer: Using Dictionaries in Second Language Learning. Ilson, R. (Ed.). 1985. *Dictionaries, Lexicography and Language Learning*: 95-102. Oxford: Pergamon Press/British Council.
- Sarani, A. and L.F. Sahebi.** 2012. The Impact of Task-based Approach on Vocabulary Learning in ESP Courses. *English Language Teaching* 5(10): 118-128.
- Scolfield, P.** 1982. Using the English Dictionary for Comprehension. *TESOL Quarterly* 16(2): 185-194. DOI: 10.2307/3586791.
- Silaški, N.** 2012. *Srpski jezik u tranziciji. O anglicizmima u ekonomskom registru* [Serbian in Transition. Anglicisms in the Economic Register]. Beograd: CID Ekonomskog fakulteta.
- Vasić, V., T. Prčić and G. Nejgebauer.** 2011. *Du yu speak anglosrpski? Rečnik novijih anglicizama. 2. izdanje.* [Do You Speak Anglo-Serbian? A Dictionary of Recent Anglicisms in Serbian. Second Edition]. Novi Sad: Zmaj.
- Vintean, A. and O. Matiu.** 2010. Electronic Dictionaries and ESP Students. *Studies in Business and Economics* 5(3): 324-329.
- Wang, J.** 2012. The Use of e-Dictionary to Read e-Text by Intermediate and Advanced Learners of Chinese. *Computer Assisted Language Learning* 25(5): 475-487. DOI: 10.1080/09588221.2011.631144.
- Wu, J. and B. Wang.** 2004. *The Role of Vocabulary in ESP Teaching and Learning.* Oral Presentation at the Fourth International Conference on ELT in China "New Directions in ELT in China", May 21–25, 2004. Available at: <http://www.celea.org.cn/pastversion/lw/pdf/wujiangwen.pdf> [10 February 2017].
- Yamaizumi, M.** 2014. Teaching English–Japanese Dictionary Use in University Remedial Courses. *Komaba Journal of English Education* 5: 1-28.
- Zou, D.** 2016. Comparing Dictionary-induced Vocabulary Learning and Inferencing in the Context of Reading. *Lexikos* 26: 372-390.

Appendix: The Final questionnaire related to the standardization of English-based sports terms in Serbian

To answer, please circle one of the offered solutions.

A

1. If there is only one translation equivalent for two English terms, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *coach* > TRENER versus *trainer* > KONDICIONI TRENER

- (a) Retranslate the English terms as follows: *coach* > TRENER versus *trainer* > KONDICIONI TRENER;
- (b) Keep the existing translation equivalent (TRENER) for both;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

B

2. If there are two translation equivalents for one English term, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *ball under* > POTOPLJENA LOPTA, TOPLJENA LOPTA

- (a) Use POTOPLJENA LOPTA;
- (b) Use TOPLJENA LOPTA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

3. If an English term is translated to Serbian as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *ear protector* > ZAŠTITNIK ZA UŠI, ŠTITNIK ZA UŠI

- (a) Use ZAŠTITNIK ZA UŠI;
- (b) Use ŠTITNIK ZA UŠI;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

4. If an English term has two translation equivalents in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *goalkeeper's border line* > GOLMANOVA GRANIČNA LINIJA, GRANIČNA LINIJA ZA GOLMANA

- (a) Use GOLMANOVA GRANIČNA LINIJA;
- (b) Use GRANIČNA LINIJA ZA GOLMANA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

5. If an English term has three translation equivalents in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *screw* > OKRET, UDARAC IZ OKRETA, ŠRAUBA

- (a) Use OKRET;
- (b) Use UDARAC IZ OKRETA;
- (c) Use ŠRAUBA;
- (d) Look it up in a dictionary;
- (e) Apply a keyword search via the Internet.

C

6. If an English term is translated to Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *sprint won* > OSVOJENA LOPTA NA CENTRU, OSVOJENA LOPTA SA CENTRA

- (a) Use OSVOJENA LOPTA NA CENTRU;
- (b) Use OSVOJENA LOPTA SA CENTRA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

7. If an English term is translated to Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *FIFA World Cup* > SVETSKI KUP FIFE, FIFA SVETSKI KUP

- (a) Use SVETSKI KUP FIFE;
- (b) Use FIFA SVETSKI KUP;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

8. If an English term comprises a nominal modifier, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *game point* > GEM LOPTA, GEM-LOPTA

- (a) Use GEM LOPTA;
- (b) Use GEM-LOPTA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

9. If an English term comprises a nominal modifier, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *final four tournament* > TURNIR FAJNALFOR-A, TURNIR FAJNALFORA

- (a) Use TURNIR FAJNALFOR-A;
- (b) Use TURNIR FAJNALFORA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

10. If an English term is adapted using an Anglicism in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *feint* > FINTA, VARKA TELOM

- (a) Use FINTA;
- (b) Use VARKA TELOM;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

11. If an English term is adapted using two lexical borrowings, one of which is an Anglicism and the other is a Gallicism, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *match point* > MEČ POEN, MEČ-POEN

- (a) Use MEČ POEN;
- (b) Use MEČ-POEN;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

12. If an English term comprises a decimal point, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *semicircle 6.25 m* > POLUKRUG 6.25 M, POLUKRUG 6,25 M

- (a) Use POLUKRUG 6.25 M;
- (b) Use POLUKRUG 6,25 M;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

13. If an English term cannot be translated to Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *kick serve* > KIK SERVIS, KICK SERVIS, KIK-SERVIS

- (a) Use KIK SERVIS;
- (b) Use KICK SERVIS;
- (c) Use KIK-SERVIS;
- (d) Look it up in a dictionary;
- (e) Apply a keyword search via the Internet.

D

14. If an English poly-lexical term has three translation equivalents in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *receiver* > PRIMAČ, IGRAČ KOJI PRIMA LOPTU and HVATAČ

- (a) Use PRIMAČ;
- (b) Use IGRAČ KOJI PRIMA LOPTU;

- (c) Use HVATAČ
- (d) Look it up in a dictionary;
- (e) Apply a keyword search via the Internet.

15. If an English poly-lexical term has two translation equivalents in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *ineffective side passing* > PASIVNO DODAVANJE, DODAVANJE LOPTE OD IGRAČA DO IGRAČA

- (a) Use PASIVNO DODAVANJE;
- (b) Use DODAVANJE LOPTE OD IGRAČA DO IGRAČA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

E

16. If there are two translation equivalents for one English term, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *dribbler* > DRIBLER, IGRAČ KOJI JE PREVARIO PROTIVNIKA

- (a) Use DRIBLER;
- (b) Use IGRAČ KOJI JE PREVARIO PROTIVNIKA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

17. If there are two translation equivalents for one English term, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *sending out* > IZBACIVANJE (IGRAČA), DISKVALIFIKACIJA

- (a) Use IZBACIVANJE (IGRAČA);
- (b) Use DISKVALIFIKACIJA;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

18. If there an English term has two translation equivalents in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *external influence* > SPOLJNI INCIDENT, INCIDENT VAN IGRE

- (a) Use SPOLJNI INCIDENT;
- (b) Use SPOLJNI INCIDENT;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

F

19. If an English term is translated as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *ironman triathlon* > 1. AJRONMEN, 2. MEGA-TRIJATLON

- (a) Use AJRONMEN;
- (b) Use MEGA-TRIJATLON;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

20. If an English term is adapted using a raw Anglicism in Serbian even though it could have been adapted through translation, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *flex offence* > FLEX OFFENCE, FLEKS-NAPAD

- (a) Use FLEX OFFENCE;
- (b) Use FLEKS-NAPAD;
- (c) Look it up in a dictionary;
- (d) Apply a keyword search via the Internet.

21. If an English term is translated in Serbian, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *straddle support* > STREDL IZDRŽAJ, STREDL-IZDRŽAJ

- (a) Use STREDL IZDRŽAJ;
- (b) Use STREDL-IZDRŽAJ;
- (c) Use IZDRŽAJ U PREDNOSU RAZNOŽNO;
- (d) Look it up in a dictionary;
- (e) Apply a keyword search via the Internet.

22. If an English term has several translation equivalents, as exemplified below, what would you do in order to get the standard term in Serbian?

E.g., *center sport* > CENTAR, BELA TAČKA, CENTRALNA TAČKA, SREDIŠNJA TAČKA

- (a) Use CENTAR;
- (b) Use BELA TAČKA;
- (c) Use CENTRALNA TAČKA;
- (d) Use SREDIŠNJA TAČKA;
- (e) Look it up in a dictionary;
- (f) Apply a keyword search via the Internet.

Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources

Carolin Müller-Spitzer, *Institut für Deutsche Sprache, Germany*
(mueller-spitzer@ids-mannheim.de)

María José Domínguez Vázquez, *Universidade de Santiago de Compostela, Spain* (majo.dominguez@usc.es)

Martina Nied Curcio, *Università degli Studi Roma Tre, Italy*
(martina.nied@uniroma3.it)

Idalete Maria Silva Dias, *Universidade do Minho Braga, Portugal*
(idalete@ilch.uminho.pt)

Sascha Wolfer, *Institut für Deutsche Sprache, Germany*
(wolfer@ids-mannheim.de)

Abstract: In the past two decades, more and more dictionary usage studies have been published, but most of them deal with questions related to what users appreciate about dictionaries, which dictionaries they use and what type of information they need in specific situations — presupposing that users actually consult lexicographic resources. However, language teachers and lecturers in linguistics often have the impression that students do not use enough high-quality dictionaries in their everyday work. With this in mind, we launched an international cooperation project to collect empirical data to evaluate what it is that students actually do while attempting to solve language problems. To this end, we applied a new methodological setting: screen recording in conjunction with a thinking-aloud task. The collected empirical data offers a broad insight into what users really do while they attempt to solve language-related tasks online.

Keywords: DICTIONARY USE, OBSERVATIONAL STUDY, LANGUAGE LEARNERS, ONLINE RESOURCES, SEARCH STRATEGIES, ONLINE DICTIONARIES, AUTOMATIC TRANSLATORS

Opsomming: Akkurate hipoteses en noukeurige lees is noodsaaklik: Resultate van 'n waarnemingstudie uitgevoer op leerders wat aanlyn taalhulpbronne gebruik. In die afgelope twee dekades is al hoe meer woordeboekgebruikstudies gepubliseer, maar die meeste van hierdie studies handel oor vraagstukke wat verband hou met wat gebruikers van woordeboeke waardevol vind, watter woordeboeke hulle gebruik en watter tipe inligting hulle

in spesifieke situasies benodig — met die voorveronderstelling dat gebruikers inderdaad leksikografiese hulpbronne raadpleeg. Taalonderwysers en dosente in die linguistiek kry dikwels die indruk dat studente nie genoeg hoëkwaliteitwoordeboeke in hul daaglikse werk gebruik nie. Met hierdie siening in gedagte het ons 'n internasionale samewerkingsprojek van stapel gestuur om empiriese data te versamel om sodoende te kan evalueer wat dit is wat studente in werklikheid doen wanneer hulle taalprobleme probeer oplos. Om hierdie doel te bereik het ons gebruik gemaak van 'n nuwe metodologiese omgewing: skermopnames saam met 'n opdrag wat uitgevoer moet word terwyl daar hardop gedink word. Die versamelde empiriese data verskaf 'n breë insig in wat gebruikers werklik doen terwyl hulle poog om taalverwante take aanlyn op te los.

Sleutelwoorde: WOORDEBOEKGEBRUIK, WAARNEMINGSTUDIE, TAAL(AAN)LEERDERS, AANLYN HULPBRONNE, SOEKSTRATEGIEË, AANLYN WOORDEBOEKE, OUTOMATIESE VERTALERS

1. Introduction

Research into dictionary use has made substantial progress in the past two decades (cf., e.g., Töpel 2014, Welker 2013, Lew 2011; Lew 2015a), especially with regard to online dictionaries (cf., e.g., Müller-Spitzer 2014, Lew 2015b). However, almost all studies in the field deal with the aspects that users value when using dictionaries (e.g. Domínguez Vázquez et al. 2013, Domínguez Vázquez and Valcárcel Riveiro 2015, Müller-Spitzer and Koplenig 2014), which dictionaries or which items in dictionaries are used or required in which situations (e.g. Koplenig and Müller-Spitzer 2014, Nied Curcio 2013), which methods of presenting data are most user-friendly (e.g. Lew 2010, Lew et al. 2013), or which information is most frequently looked up in online dictionaries (e.g. De Schryver et al. 2006, Hult 2012, Koplenig et al. 2014). Therefore, these studies either assume that lexicographic tools are actually used or put the test subjects into concrete situations in which they are asked to imagine what lexicographic tools they would use. At the same time, many language teachers and lecturers in linguistics are under the impression that students do not use a sufficient amount of (good) dictionaries in their everyday life (see, e.g., Frankenberg-Garcia 2011). Accordingly, there is a gap between empirical research on dictionary use and the reality of learners' or students' actual everyday language challenges. We still have too little empirical data to be able to assess the role dictionaries play in day-to-day work. As Levy and Steel put it:

The study reported here, with data drawn from a large-scale survey, reports on what students *say* they do when using electronic dictionaries. This reportage does not necessarily reflect what students actually *do* [...]. Smaller-scale studies are needed to complement and enrich the findings of the present study. (Levy and Steel 2015: 194)

With this in mind, we launched an international cooperation project to collect empirical data with which to evaluate the suggested discrepancy. Our aim was

to collect comprehensive and reliable data about what it is that students (starting with German language learners from Romance language-speaking European countries) actually do when they deal with language problems. With the help of this accumulated knowledge about students' actual use of lexicographic resources, these data could then constitute an adequate starting point from which to teach students how to use language resources. Ignoring this aspect can be compared to teaching a language without asking at what level the students currently are.

To get a better idea about what students do during their everyday work, we used a new methodological setting for research into dictionary use: we presented sentences on a notebook computer and the participants were asked to improve these sentences using the online resources of their choice. During this process, we recorded the learners' on-screen actions with a screen recorder and prompted them to think aloud. We collected audio and screen capture data of 42 students from Braga (Portugal), Rome (Italy) and Santiago de Compostela (Spain). All participants were at the A2/B1 level according to the 'Common European Framework of Reference for Languages (CEFR)'¹. The collected data include 1,680 minutes of screen recordings and audio material containing more than 2,200 search procedures. The collected empirical data offers a broader insight into what language users today really do when solving language-related tasks. A wide range of questions can be addressed using the data, e.g.: Are our participants aware of the differences between translation systems and dictionaries? Do they adapt the search string to the type of resources used? Does the number or type of resources used have a positive impact on solving the task? How much time do they spend using the various resources? All these questions are addressed in this paper, which is structured as follows: first, we present the study design and our method for collecting the data (Section 2). Then, in the main part of our paper, we describe and explain the results of our study (Section 3). After some general results (Section 3.1), we focus on search strategies (Section 3.2) and on the factors that influence the quality of the corrections (Section 3.3), especially the influence of careful reading and how strongly overall search behavior was influenced by the initial hypotheses. Our article ends with conclusions (Section 4).²

2. Materials and method

We employed a mixed-methods design combining (i) a language correction task, (ii) screen recording of all on-screen actions, and (iii) audio recordings to create the participants' thinking-aloud protocols (Ericsson and Simon 1993). We distributed written instructions and a declaration of consent to the participants before the experiment. Both documents were in the participants' native languages. The instructions described the task and the setup on the computer screen. Also, we highlighted that the participants did not necessarily have to find a solution or correction for each and every stimulus sentence. The instruc-

tions further contained some suggestions for the thinking-aloud task such as "describe what you are doing, why you are doing it, describe your thoughts while solving the task, describe why you are accessing a specific internet site, what you wish to find on the site, tell us why you are choosing a specific correction and whether you are satisfied with the corrections", and so on. Finally, the instructions indicated that the study would only be used for scientific purposes and not to grade the participants³ in any way. After reading the instructions, the participants were given the opportunity to ask questions. There was a native speaker of the local language (Portuguese, Spanish or Italian) present in the room at all times, along with one or two experimenters. The experimenters could not speak or understand the local languages but explained the experimental setup to the local assistants beforehand. All local assistants also understood and spoke German at a native or near-native level.

The setup consisted of a standard desktop environment on a 15-inch Windows 10 Toshiba notebook with German keyboard layout, a cable-based mouse, a screen resolution of 1920 by 1080 pixels with 8 GB of memory and an Intel i5-6200U CPU. The browser cache and history was cleared after each participant. We used the same notebook for all participants in all locations but adapted the browser language to the respective local language.

2.1 Correction task

Each participant was presented with 18 German sentences⁴ containing one error. The errors were constructed in such a way as to satisfy two requirements: (i) the error was typical for early-stage learners of German whose native language was a Romance language; (ii) the error could not be easily resolved by simply searching the web for the stimulus sentence or the part of the sentence containing the error. All sentences were designed by three of the authors of this paper (Idaete Dias, María José Domínguez Vázquez, Martina Nied Curcio) based on their long-term experience as 'German as a Foreign Language' teachers.

For example, one stimulus sentence was "An unserem Forschungsinstitut **ist** Ihnen unsere Bibliothek 24 Stunden **zur Verfügung**" (Eng. "At our research institute, our library is available to you 24 hours"). This stimulus contains an error in the light verb construction "zur Verfügung *sein*". The correct construction is "zur Verfügung *stehen*", hence, one possible correction would be "An unserem Forschungsinstitut **steht** Ihnen unsere Bibliothek 24 Stunden **zur Verfügung**". In Spanish, a correct version of the sentence would be "En nuestro instituto de investigación, nuestra biblioteca está abierta las 24 horas". The German "ist" can be seen as a direct translation of Spanish "está" (accordingly of Portuguese "está" and Italian "è"). The participants had to identify this as an invalid parallelism between Spanish and German and correct the error accordingly. If you search for the original stimulus sentence in Google, you would be faced with several pages of search results related to the libraries of a wide variety of research institutes, but no results dealing with the linguistic

properties of the sentence or the error itself.

It may be possible to argue that a correction task is a rather "unnatural" task for learners of German. A more "natural" task might have been to translate sentences from the participants' respective native language into German. However, we chose the correction task because it gave us the opportunity to use the same sentences for all participants from all countries. This, in turn, should reduce noise induced by stimulus sentences from different languages. All stimulus sentences can be found in the Appendix.

In terms of the technical setup, we used a simple Excel spreadsheet that contained the stimulus sentences in one column titled "Satz" (German for "sentence") and an empty column titled "Korrektur" (German for "correction") where the participants were to type their corrected sentence. The problematic parts of each sentence were highlighted in bold face (as indicated above), which was also explained in the participants' instructions. By using standard office software, we hoped to provide the participants with an environment they are well acquainted with. The participants were allowed to use Google Chrome or Mozilla Firefox whenever they wanted to refer to web content. They were not allowed to use any built-in assistance devices in Windows 10. The participants were not given a time limit before the experiment to avoid time pressure. After 30 minutes, each participant was told that they had 15 minutes left to work on the corrections. After 45 minutes, we told the participants that they should finish the sentence they were currently working on and then ended the experiment.

2.2 Screen recordings

The screen recording software ActivePresenter was started by one of the two experimenters in the room. We made sure beforehand that screen recordings did not interfere with the task in any way (e.g., pop-ups, screen flickering or the like). All actions of the participants were captured in the native display resolution.

2.3 Audio recordings

Since we did not want to rely on the notebook's built-in microphone to capture the voice of the participants, we recorded the thinking-aloud protocols with a high-definition external microphone. The audio recordings were inserted as the screen recordings' audio track after the experiments to allow for a synchronized investigation of both the screen recordings and thinking-aloud data. After we completed the data collection, the verbalizations of the participants were transcribed by native speakers of the respective language. German translations of the verbal protocols are also available.

2.4 Annotations

The corrections that the participants entered were rated by two native German annotators. Five categories were available: "C", correct (all errors have been resolved), "CE", correct with errors (all errors in the stimulus sentences have been resolved but other errors have been introduced into the response), "D", case of doubt (it cannot be determined without a doubt whether the answer is correct or not), "W", wrong (the linguistic problem in the stimulus was not resolved or had been replaced by another), "N", not dealt with (the sentence had not been worked on, no attempt had been made to correct it). One example may illustrate the different categories: The stimulus sentence "Obwohl ich studiere, **wohne** ich noch **mit** meinen Eltern." (English "Although I am a student, I still live with my parents.") contains a wrong preposition. The correct version would be "Obwohl ich studiere, wohne ich noch *bei* meinen Eltern." This solution would accordingly be annotated as correct. An example of a CE-case (corrected with a new error) is the solution of participant R-02: "Obwohl ich studiere, ich wohne noch bei meinen Eltern." Here, the preposition "bei" is correct, but the word order "ich wohne noch bei" is a new error which is not part of the initial stimulus sentence. A wrong solution is, e.g., one made by participant S-09: "Obwohl ich studiere, ich mit noch meinen Eltern whone." Here, the wrong preposition is still there ("mit"), the word order is wrong, and a new spelling error "whone" occurs.

In 712 out of 816 cases (87.3 %), the two annotators labeled the answers of the participants identically. Weighted kappa (Cohen 1968) is $\kappa = .86$, indicating very good agreement between the annotators (we used the weighted kappa value because it penalizes disagreements that are farther apart from each other — e.g., "C" vs. "W" — more than disagreements that are closer to each other — e.g., "C" vs. "CE"). All disagreements were resolved through discussion.

To analyze research behavior, we also annotated the 2,225 search phrases that the participants used during their research. On the top level, three broad categories were distinguished: non-linguistic queries, metalinguistic terms and linguistic queries. (a) Non-linguistic queries are searches for a special dictionary or a general term like "duden wörterbuch", "alemao" or "pons tedesco". Queries were categorized as metalinguistic terms (b) whenever the query contained a linguistic term like "Konjunktiv 2 mit wenn" ("Konjunktiv 2 [a grammatical mood in German] with if"), "coniugazione verbi tedeschi" ("conjugation of German verbs"), "frases com verbos auxiliares em alemao" ("phrases with auxiliary verbs in German"), "deshalb significato" ("sense of 'deshalb'"), "Konzessivsätze mit 'obwohl' und 'trotzdem'" ("concessive clauses with 'obwohl' and 'trotzdem'"). Linguistic queries (c) are searches for words and phrases and are further divided into single-word searches like "beenden" ("to stop") vs. complex queries with multiple words like "ausser Frage" ("out of question") or "Es steht ausser Frage" ("It is without question"). The complex queries in sentence form

are also annotated for whether they are "(near-) verbatim" or "non-verbatim" copy-and-paste versions of the stimulus sentences.

3. Results and Discussion

3.1 General results

As we explained in the method section (2.1), our participants were presented with a maximum of 18 sentences for correction. On average, they edited 16 sentences. This number was nearly equal in all three locations (cf. Figure 1.1). The median (mean) number of edited sentences in Braga was 10.5 (11.4), 13.5 (12.1) in Rome, and 14.0 (13.1) in Santiago de Compostela. However, the number of correctly (category "C") improved sentences differed considerably between the three locations (cf. Figure 1.2). The median (mean) number of improved sentences in Braga was 2.5 (2.6), 7 (7.5) in Rome and 7 (7.1) in Santiago de Compostela. This result already points in a direction that is later supported by other results: although we hoped that our participants would reach the same language level in all three universities, the actual language level of the participants in Rome and Santiago de Compostela was clearly higher than of those in Braga.

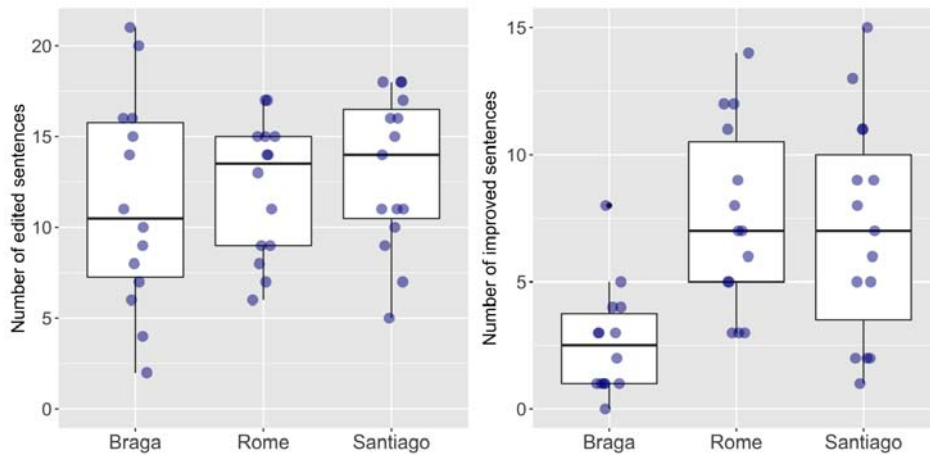


Figure 1: 1.1 (left): Number of corrected sentences in the three locations; 1.2 (right): Number of improved sentences in the three locations. Each dot shows the number of edited/corrected sentences for one participant. A total of 50% of all dots are surrounded by the box. The horizontal line within each box represents the respective median value.⁵

How many participants improved a sentence correctly also depends strongly on the sentence itself (cf. Figure 2). A sentence like "Leider kann ich heute nicht Tennis spielen. Ich bin zu **besetzt**." (in English, correctly: "Unfortunately, I can't play tennis today. I'm too busy.", Sentence-ID 2) with a false friend on the adjective position was improved in 70% of all cases, whereas the error in the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly: "No problem. If the sugar is empty, I'll take honey.", Sentence-ID 4) was obviously very hard to identify and transform into a search. In this case, only 17% of the corrections were annotated as improved. Table 1 shows one short excerpt of the search procedures referring to this sentence, illustrating the difficulties the participant had. The example shows that although the participant had the right idea at the end (looking for an adequate way to say "e'finito" in this context), they did not find an appropriate way to search for it. Another excerpt from a Spanish participant shows similar problems (cf. Table 2). The first idea many other students had concerning this sentence was that the participle, i.e. the grammatical form of "beendet," is wrong, but this is not the problem here. However, this initial idea led the students down the wrong path (for more information on the importance of the initial hypothesis, see Section 3.3.3). The combination of these types of quantitative analyses (here: which problems were difficult to solve?) and the closer qualitative inspection of the data (here: what exactly was difficult here and how did it affect search behavior?) is an advantage of the implemented study design. Thanks to this approach, we are able to evaluate exactly those aspects of dictionary use that cannot be identified on the basis of a log file or in a questionnaire study. Recording these difficulties, which leave the dictionary users at a complete loss, is a very useful insight for research into dictionary use.

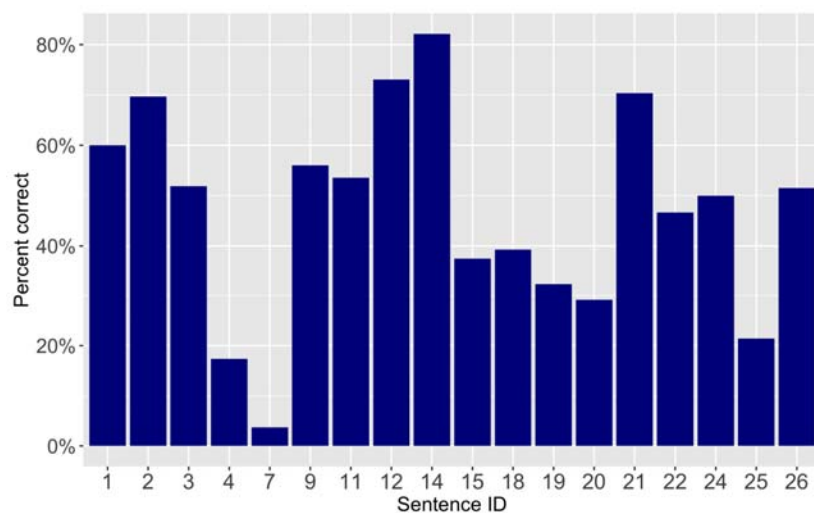


Figure 2: Percentage of improvements per sentence⁶

Action	Think-Aloud-Protocol
Returns to Google results (search string was "beendet")	allora ehm non so cerco esempi perché non mi vengono soluzioni al momento (then ehm do not know I am looking for examples because I don't find solutions at the moment)
opens Deutsches Institut	
opens Bab.la	
returns to Google search results, googles "beendet esempi"	
opens Reverso Context	ehm sto cercando sto leggendo diciamo degli esempi # non ho idea [lacht] (Ehm I'm looking for I'm reading examples # I have no idea [laughing])
opens Excel	
opens Pons Traducaio, searches for "e'finito"	sto cercando # (I'm looking)
opens Leo	sto cercando un modo per dire finito ahm (I'm looking for a way to say finished ahm)
opens Google	
opens Excel, no further corrections	okay non mi viene # non mi viene ahm # passo alla frase dopo perché non mi viene (okay I can't think of anything # I can't think of anything to say # I'm going to turn to the next sentence because I don't know)

Table 1: Excerpt from the study data of participant R-01 concerning the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly "No problem. If the sugar is empty, I'll take honey.")

Action	Think-Aloud-Protocol
opens Leo, searches for "beenden"	vale # verbo beenden # que es acabar ## pero no sé si se puede usar para esto ehhm (ok # the verb 'beenden' # that means acabar # I don't know if it can be used for that ehhm)
opens Linguee, searches for "acabar la comida"	voy a mirar acabar la comida (I'm looking for 'acabar la comida')
searches for "acabar el bocadillo"	no # acabar (no # acabar)
searches for "beenden"	vale # miro en linguee beenden (ok # I'm looking for 'beenden' in Linguee)
opens Excel	
opens Linguee	no sé cómo buscar esto (I don't know how to look for it)
searches for "terminar comida"	
searches for "agotar existencias"	igual agotar (maybe 'agotar')
opens Excel, no correction	voy a dejarlo para después (I'm gonna save it for later)

Table 2: Excerpt of the study data of participant S-11 concerning the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly "No problem. If the sugar is empty, I'll take honey.")

While transferring the screen recordings into analyzable data tables, we also encoded the position of the selected search result on the Google results page. The result is shown in Figure 3: Only the first four hits of the search results list are frequently selected (i.e. almost nobody scrolled because 4 to 5 results were directly visible on the laptop screen, depending on whether the window for the Google Translator was displayed or not). Almost two thirds of all selections (63%) concentrated on the first hit.

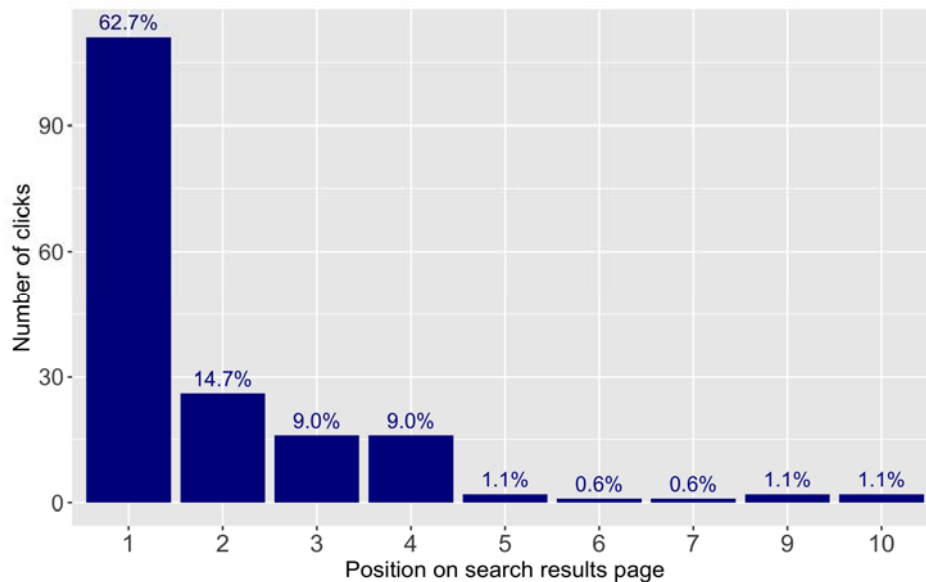


Figure 3: Position of selected search results (only Google was used as a search engine by the participants)

As mentioned at the beginning, we know little about what types of resources students actually use when solving language problems. Some teachers (personal discussion) claim that, nowadays, students use automatic translation programs far more frequently and hardly ever use dictionaries. However, according to questionnaire studies on this topic, online dictionaries are used very frequently (Levy and Steel 2015: 9, Koplenig and Müller-Spitzer 2014: 130). An important question in our study was therefore firstly to find out what types of resources our subjects use, and secondly to see whether they use different search strategies for different resource types or not.

First of all, our study shows that the students used a large number of resources and, above all, many different types (cf. Figure 4): Dictionaries or dictionaries with grammar tables were used the most, followed by search engines (which are, of course, also used to access resources, e.g. by entering "Duden online" in Google). Although 42 subjects is not a large number, the data are valid in the sense that we observed the students directly while working. This means that we do not have to rely on self-reporting, which in the context of language teaching, could be more distorted by some factor of social desirability, since the students usually know that their lecturers like to hear that they do not use automatic translation programs. In this sense, the data collected here may be understood as an encouragement to lexicographic work: indeed, students in our study seem to use dictionaries very often. In the majority (53.8%) of all trials

(i.e., sentence edits), one or more dictionaries were used, and these do not include other types of dictionaries, e.g., dictionaries with grammar tables (used in 35.5% of all trials), dictionaries with parallel texts and grammar tables (16.5%) and dictionaries with just parallel texts (11.5%).

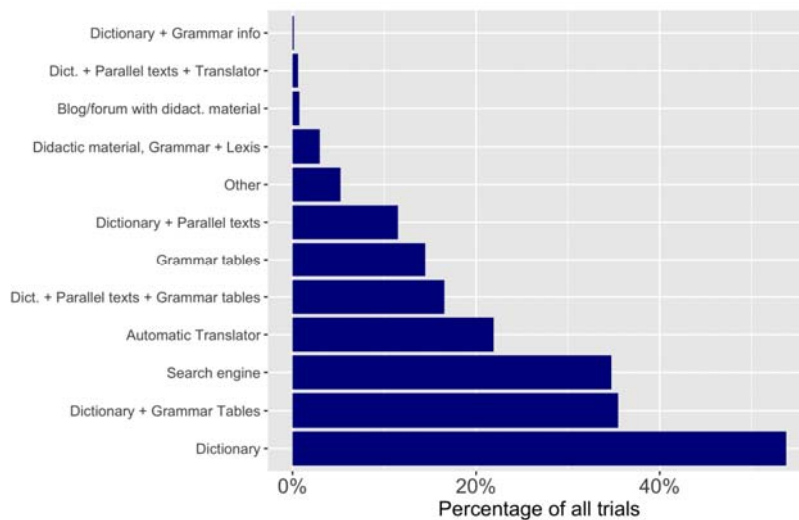


Figure 4: Types of resources used in percentage of all trials. Dict. = Dictionary; Didact. = Didactic

In the following section, we focus on the question whether the students differentiate their search strategies between different types of resources or not.

3.2 Search strategies

The data we gathered contain more than 2,200 search actions. In this section, we focus on the evaluation and analysis of these search actions. Above all, we want to investigate whether students use the various types of resources in different ways.

The language of most search strings is German (aggregated over locations, 69.4% of all search strings are in German), followed by search phrases in the local language (see Figure 5). The use of the local language (aggregated percentage: 22.3%) is remarkable in this study design because students had to conduct an improvement of German sentences, not a translation task. This 'bilingualization' of a monolingual task seems to have to do with the fact that our students want to use their mother tongue as an instance of certainty and/or track down the errors of interference by translating the German stimulus sentence back into their mother tongue and then using bilingual resources. This

strategy works very well in some cases. Interestingly, participants in Braga also rarely (but more often than the others) use English as a relay language. In the screen recordings, one can see that this was mainly done upon realizing that the consulted German–Portuguese bilingual resources achieved poor results (e.g. in an automatic translation program), but a translation from German–English as a first step and then English–Portuguese was more promising (see an example in Section 3.3.3). This use of English as a relay language came as a slight surprise for the language teachers involved in our study, but seems to be a viable strategy in some cases.

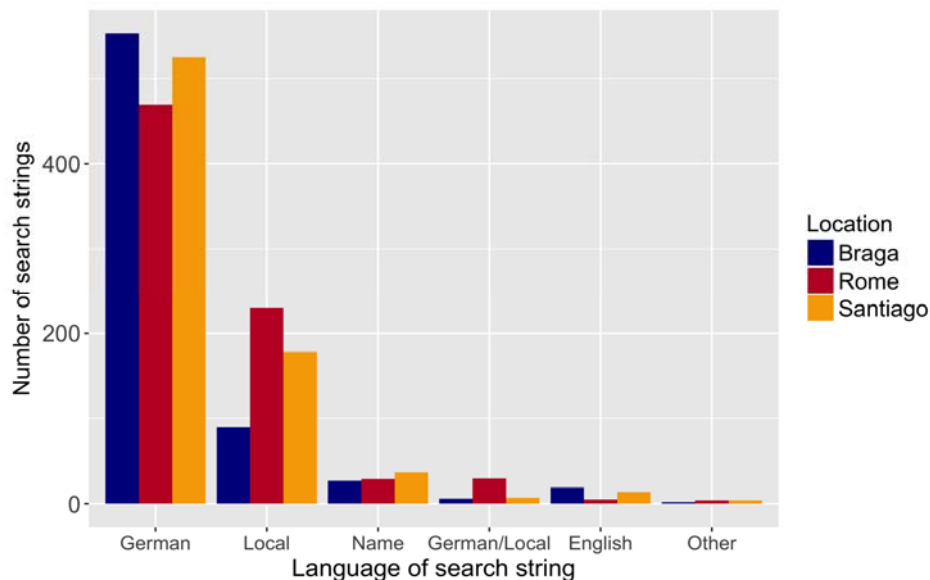


Figure 5: Languages of search strings (Local=Portuguese/Spanish/Italian, Name=Name of a resource)

It is well known that different types of resources are designed for different types of search queries. For example, it is generally not promising to enter entire sentences into the search fields of dictionaries, whereas the more context you have, the better automatic translation programs work. The question is, however, whether students are aware of this and adapt their search strategies accordingly. We wanted to use our data to investigate whether we could prove that our participants are aware of this.

In order to achieve this, we annotated all search strings as explained in the methods section. We see different patterns concerning the complexity of search strings used in different types of resources: although complex queries consti-

tute the minority in all types of resources, the percentage thereof in automatic translation tools is higher than in all the other types (cf. Figure 6.1): In total, 42.5% of all queries in automatic translation tools are complex. These results may indicate that the participants are basically aware of the different functionalities of automatic translation tools vs. other types of resources. This impression is reinforced by the fact that there is an observable difference between use of the different resources from the same publisher or portal. While less than 5.5% of all queries in the Pons dictionary are multiple word items, there are more than 41.9% complex search queries in the Pons Translator even though both resources are presented on the same website (cf. Figure 6.2). Also the distribution of sentential vs. non-sentential queries points in the same direction: while sentential search queries constitute the majority (58.0%) of all queries in automatic translation tools, our participants almost never (1.9%) used them in dictionaries (Figure 6.3).

A further indication that the students use specific resource types depending on the kind of search query comes from the annotation and analysis of "(near) verbatim" and "non-verbatim" search queries (multiple word queries often seem to be verbatim copies of the stimulus sentences, see Figure 7). Google Translate and the Pons translator are clearly preferred if whole stimulus sentences are copied and pasted, i.e. for verbatim queries. In contrast, a resource like Reverso Context is used for non-verbatim queries.

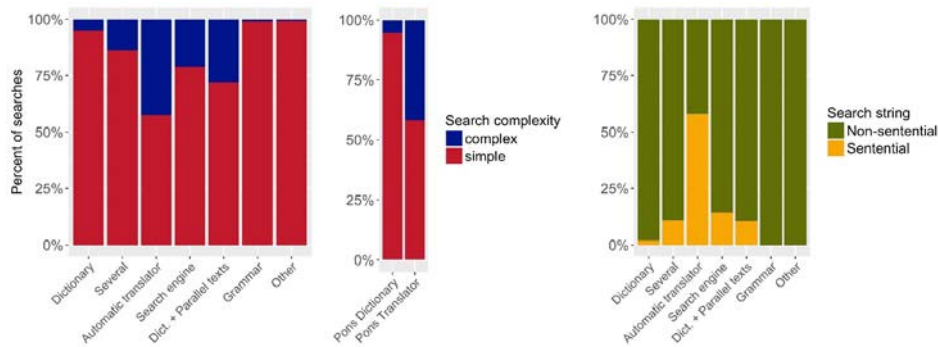


Figure 6: Figure 6.1 (left) Simple (one word) vs. complex (multiple word) queries in different types of resources; Figure 6.2. (middle) Simple vs. complex queries in Pons Dictionary vs. Pons Translator; 6.3 (right) Percentages of non-sentential and sentential search strings in different types of resources

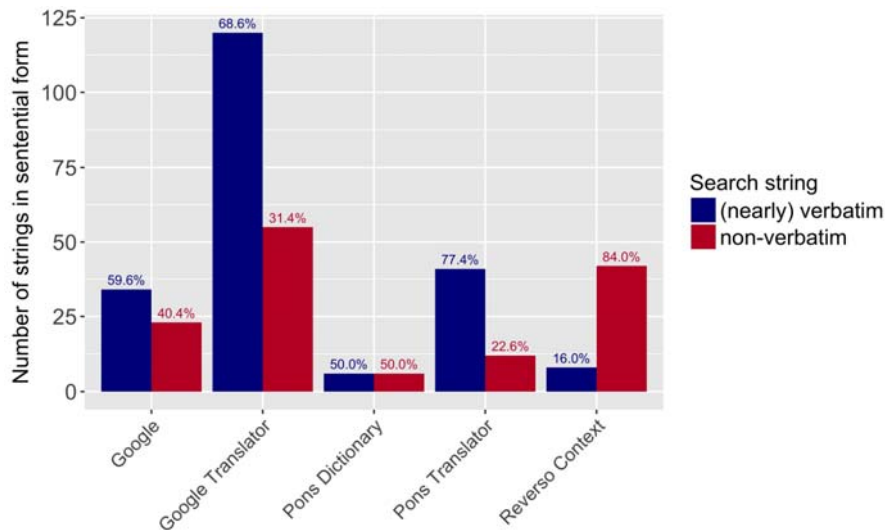


Figure 7: Verbatim vs. non-verbatim queries in sentential form (compared to stimulus sentences) in different types of resources, percentages above bars indicate the distribution within one resource.

Concerning our research question stated at the beginning of this section, we can conclude that the analyses of the search strings have shown that our participants seem to have at least a basic awareness of the different functionalities of the different types of resources used and adapt their search strategy accordingly. In the next section, we will take a closer look at which factors have an impact on the quality of the corrections.

3.3 Which factors influence the quality of corrections?

We now want to investigate whether we can identify systematic factors that influence the correctness of the improvements. We report our results concerning the correlation between types of resources and correction rate (3.3.1), the impact of careful reading (3.3.2) and the importance of initial hypotheses (3.3.3).

3.3.1 Types of resources

One such systematic factor might be the number of different resources that are used to correct a sentence and whether this has a positive impact on the results. However, this is not evident. The main tendency related to the number of resources consulted is very similar in the case of correct improvements (Mean = 1.76, Median = 2), incorrect corrections (Mean = 1.76, Median = 2), cases of doubt (Mean = 1.92, Median = 2) and in the case of not attempting an improvement (Mean = 1.97, Median = 2) at all (cf. Figure 8). Likewise, the pro-

cessing time per sentence has no influence on the improvement (no figure). Similarly, the position of the sentence in the study has no influence on the correctness, i.e. it was not the case that the first sentences were improved correctly more often than the last ones. Rather, it seems that there were sentences that were easy to improve even with few searches and in a short time, but others were not easy to correct even with a long overall processing time and many resources used.

In contrast, the type of resources used has an impact on the correctness rate. Two things in particular are striking. First, those participants who used more dictionary resources were more successful. The relationship is presented in Figure 9.1. This correlation is fed, in particular, by the participants from Braga, who revised only a few sentences correctly. A further subdivision of dictionary resources shows that dictionaries with parallel corpus examples such as Linguee tend to produce even better results. However, we must examine this particular connection in more detail before we can draw reliable conclusions. Second, our analyses show that the participants who rely more on automatic translation programs achieved poorer revision results (cf. Figure 9.2). As shown in Figure 9.2, this correlation is mainly influenced by our Portuguese participants who were less proficient in solving the task in general. Also note that the majority of participants used very few automatic translation programs (or none at all). This means that this correlation is driven by the fact that the better students also used more dictionaries and/or the worse ones use more automatic translation programs.

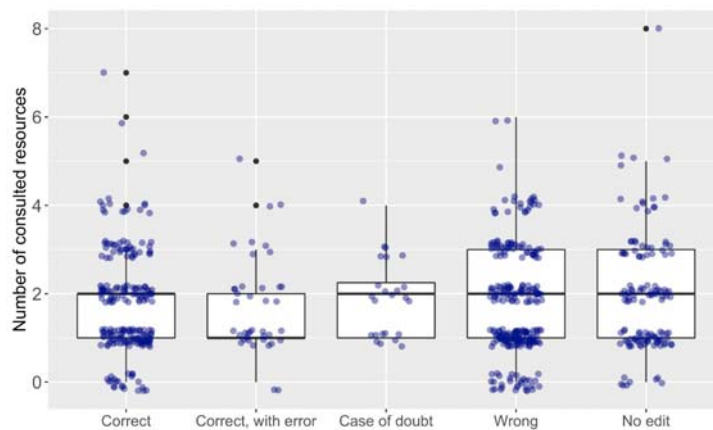


Figure 8: Number of resources used differentiated by correct improvement, correct improvement with new error, case of doubt, wrong corrected or no correction attempt at all (no edit). Each dot represents one sentence of one participant (one 'trial'). The box surrounds 50% of all data points. The horizontal line in each box represents the median value.

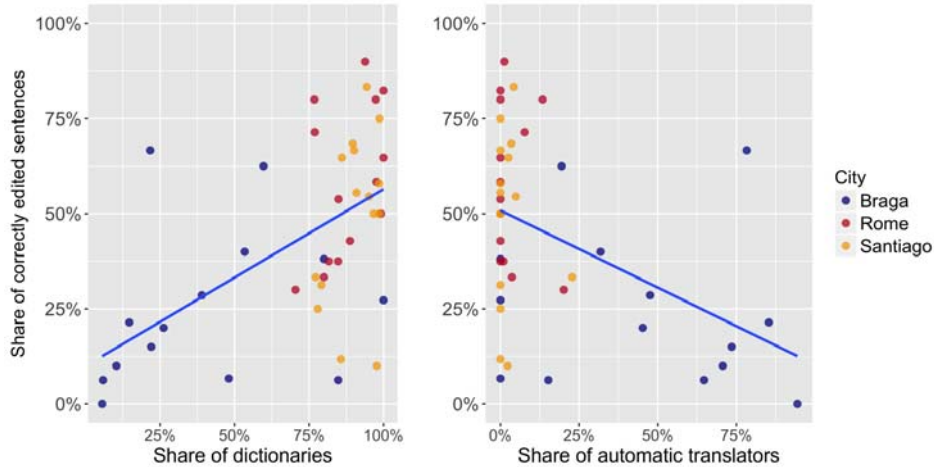


Figure 9: 9.1: Share of dictionaries in all used resources and percentage of improved sentences; 9.2: Share of automatic translation tools in all used resources and percentage of improved sentences. Each dot represents one participant (location is color-coded). The blue line represents the result of a linear regression fitted to the data.

3.3.2 Time spent using the resources and careful reading

Another key factor is time. Looking at the data (Figure 10), we found that there is a relationship between the average time spent using the resources and the correctness of the sentences. The mean difference between wrong and correct outcomes is relatively slight (only 2.4 seconds). However, it should be noted that this difference means that — on average — the time spent on each single resource is 2.4 seconds longer in each sentence edit that results in a correct sentence. During the course of the experiment, this difference may well amount to a much larger overall difference between correct and incorrect sentences. Interestingly, the different performance of the students in the different locations is also reflected in the time spent using the resources (Figure 11): On average, the students from Braga spent less time (Mean = 15.3 sec, Median = 14.8 sec) on the resources than the participants from Rome (Mean = 17.7 sec, Median = 16.1 sec) and Santiago de Compostela (Mean = 18.1 sec, Median = 16.3 sec).

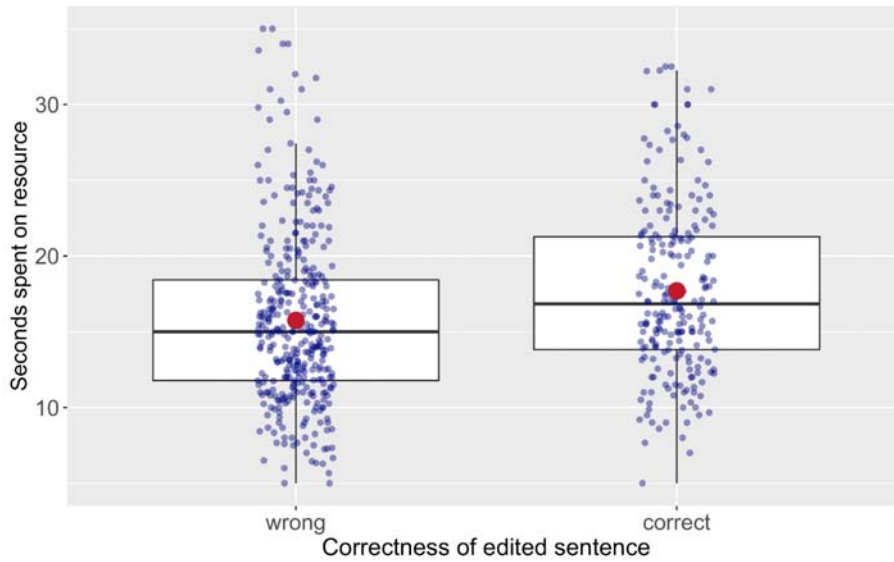


Figure 10: Average time spent using the resources and correctness of the sentences. Each dot represents one sentence edit from one participant (a 'trial'). Boxes are interpreted as in previous figures.

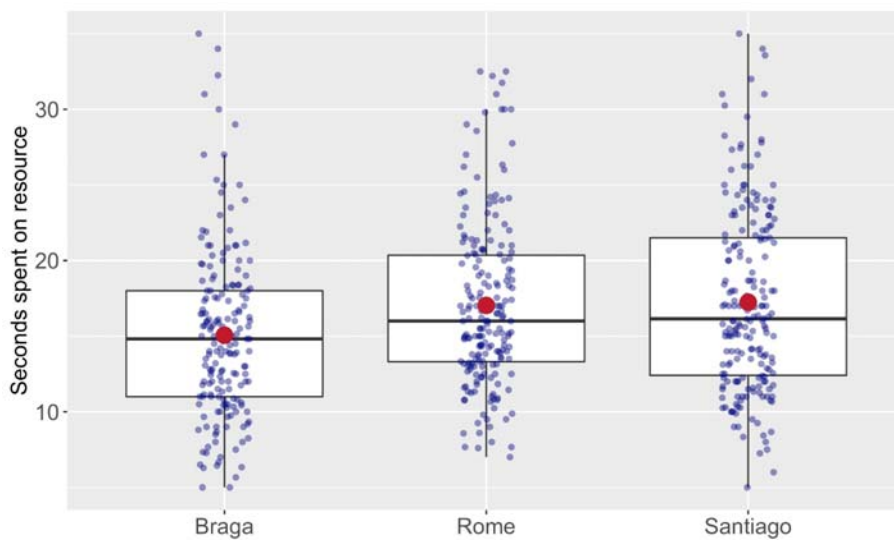


Figure 11: Time spent using the resources at the different locations. Each dot represents one sentence edit from one participant (a 'trial'). Boxes are interpreted as in previous figures.

Obviously, the short time spent using resources implies that students frequently switched between them. One example may illustrate this: Subject R-01 spent a total time of 3.5 minutes on sentence 1, undertaking 25 actions, which means that an average time of 7.65 seconds was spent on a single resource (without correction time). A look at the thinking-aloud-protocols (TAPs) confirms — even on a linguistic level — that subject R-01 takes hardly any time for the individual search queries; they often say "un attimo" or "un attimino" ('a moment'/'a minute'/'just'), e.g., "vado *un attimo* a vedere la costruzione di ehm la coniugazione di enden" ('I will just look at the construction ehm at the conjugation of enden'), "sto vedendo *un attimo* il verbo la coniugazione del verbo per essere sicura" ('I will just look at the verb's conjugation to make sure') or "okay vedo *un attimo* stipendium # okay" ('okay I will just look for stipendium # okay'). It is also very interesting that this student generally gave up very quickly if the solution could not immediately be found and then ascribed blame to the machine by saying "non mi trova niente" ('it doesn't find anything for me') or "cioè non mi sta trovando neanche degli esempi dello stesso verbo" ('so, it doesn't even find examples of the same verb for me') using the 3rd person singular to refer to the computer and/or the resource.

Other participants in contrast spent more time using each individual resource, reflected more upon their actions and achieved better results. These students seemed to solve problems very constructively, were aware of the potential difficulties, had previously developed language awareness, read attentively, and persisted in trying to tackle the same problem from various angles. Another example from Rome (R-07) illustrates this via excerpts from the TAP. Regarding the sentence *Obwohl sich der Junge beeilt hat, hat er die U-Bahn verloren* ('Although the boy hurried, he missed his train') the student was aware of the polysemy of the Italian verb *perdere* ('to lose', 'to miss'), which means that s/he had already developed a certain language awareness; they knew that in combination with a vehicle like *U-Bahn* ('underground train') the German verb *verlieren* (English to lose) was not correct and that a specific verb had to be selected. The student was aware that certain words belong together (collocations) and consequently searched for a specific word in the resources. That is the reason why a word-by-word translation (*perdere* – *verlieren*), which in these selected sentences usually leads to interference errors, could be avoided. In addition, the participant knew about various resources and opened an appropriate resource related to the search query, i.e. in order to find out the meaning of *verloren*, Pons was accessed; for the conjugation of *verpassen*, Reverso Coniugazione (Italian version) was the chosen resource. The student also used linguistic strategies such as searching for synonyms of *U-Bahn*, like *Zug* ('train'), which were considered more prototypical, and synonyms for *verlieren*. It is also interesting that subject R-07 often double-checked, i.e. by changing the search direction and checking the hypothesis, although R-07 was quite sure of the solution. This implementation of multiple strategies was also responsible for the high number of correct sentences. Of course, there is also the willingness to

solve the problem or to investigate it more rigorously and the will not to give up, as we can see in the extract in Table 3.

non posso purtroppo non posso andare allora in die Klasse gehen cerco ehm gehen se mi dà qualche utilizzo con Klasse magari se mi dà una frase simile quindi eh # allora (camminare andare a passeggio # andare in una stanza Zimmer in ein Zimmer) quindi allora se devo andare ehm devo usare in più l'accusativo quindi non è sbagliato l'articolo probabilmente il verbo # cerco Klasse se mi dà un un contesto d'uso per esempio no (viaggiare in prima seconda classe) no ehm okay quindi Klasse potrebbe essere anche una categoria forse ho capito male la frase quindi cerco anche Arzttermin (unfortunately I cannot go in die Klasse gehen so I will look up if ahm gehen somehow is used with Klasse maybe it will give me a similar sentence so ahm # so [camminare andare a passeggio # andare in una stanza Zimmer in ein Zimmer] so when I go ahm I must ahm I must use in plus Accusative so the article is not incorrect maybe the verb is incorrect # I will check if Klasse for example specifies a context of use no [viaggiare in prima seconda classe] no ahm okay so Klasse could also be a category maybe I didn't understand the sentence correctly so I will also search for Arzttermin).

Table 3: Extract of the TAP of student R-07 while working on the sentence "Morgen habe ich einen Arzttermin und kann deshalb nicht **in die Klasse gehen**".

However, it must be mentioned at this point that the time factor should not be considered in isolation. Due to the methodological design (including the TAPs), a longer and therefore more detailed, probably more intentional reflection influences the time spent using each resource. As a consequence, we cannot make a clear statement about the direction of the effect: are the more proficient students better at understanding the information in the resources and therefore spend more time using them, or does careful reading alone really lead to success? In other words, language proficiency, time spent using resources, and careful reading of dictionary entries form a complex inter-connected relationship. To allow for inferences, more experimentally controlled studies are required.

Additionally, a rigorous inspection of individual examples such as the ones presented above, incurs a risk of inferring general trends, which may not be confirmed by the overall data set. So, one has to make sure that the importance of individual examples is not overrated. However, as we have seen from the example of time spent on the sentences, the advantage of the data we collected is that these types of qualitative inspections encourage quantitative analyses which can then verify some data or adjust qualitative impressions. And, vice versa, quantitative results can be more closely examined through quantitative analyses (cf. Wolfer et al. 2018).

3.3.3 Searching guided by hypotheses

While analyzing the TAPs and the screen recordings, there seemed to be evidence that students' search behavior might be influenced by the initial hypotheses they formulate when analyzing the stimulus sentence. We will try to show that students tend to focus their initial hypotheses, thereby ignoring

relevant information in the online resources. In the following, we will describe this behavior in detail in order to make sense of students' search actions and develop a schema based on the observed search behavior patterns.

We begin our analysis with a description of the search actions carried out by a Portuguese student while trying to improve the following stimulus sentence: "Ich möchte ein Stipendium beim DAAD bewerben" ('I would like to apply for a scholarship at the DAAD'). Correcting the sentence involves identifying that: (i) the verb "bewerbem" is a reflexive verb "sich bewerben" and (ii) "sich bewerben" is used with a prepositional phrase introduced by the preposition "um" followed by the object of the preposition in the accusative case: "Ich möchte *mich* beim DAAD *um ein Stipendium* bewerben".

From the TAP it is clear that the student does not know what the verb "bewerbem" means. This leads the participant to look up the meaning of the verb in the Pons German–Portuguese Dictionary. The result provided by entering the search word "bewerbem" is shown in Figure 12.



Figure 12: Result of search query "bewerbem" in the Pons German–Portuguese Dictionary

As can be seen in Figure 13, the entry contains all the necessary information needed for the student to solve the task of correcting the stimulus sentence: (1) "bewerben" is a reflexive verb; (2) it requires the specific preposition "um"; (3) an example sentence is provided; (4) equivalents in the students' native language are provided. In addition, this information appears in the uppermost part of the entry. This means that, in effect, the student does not have to scroll through the entry looking for the answer(s) in order to correct the stimulus sentence. Taking into account studies on patterns of look-up behavior (Tono 1984, Lew et al. 2013), one would expect the student to pay special attention to the central information provided at the beginning of the entry.

Following from the TAP, the student reads the Portuguese equivalent "candidatar-se a", concludes that it is a reflexive verb and all further search actions aim at validating the hypothesis: the verb 'bewerben' in the stimulus sentence is missing the reflexive pronoun. The student focuses on the missing pronoun and does not analyze the entry any further. The information concerning the preposition "um" and the example sentence in the entry go completely unnoticed. To confirm the formulated hypothesis, the student applies the following steps:

- (i) S/he copies the entire stimulus sentence from the Excel file and pastes it in Google Translate (cf. Figure 14.1).
- (ii) Since the Portuguese translation equivalent of the stimulus sentence sounds strange ("Eu quero aplicar uma bolsa do DAAD"), the student changes the target language of the translation to English (cf. Figure 14.2) and keeps using English until the task is over. The result is an incorrect German sentence corresponding to a correct English translation equivalent: "I would like to apply for a scholarship at the DAAD".
- (iii) S/he switches the source and target languages and uses the correct English sentence as the source sentence (cf. Figure 14.3). The result is the correct German translation "Ich möchte mich beim DAAD um ein Stipendium bewerben". So this is an example where including English as a relay language was a promising strategy. Interestingly, based on the TAP and the correction proposal ("Ich möchte mich ein Stipendium beim DAAD bewerben"), the student focuses exclusively on the presence of the reflexive pronoun in the correct German sentence, thereby validating the initial formulated hypothesis, and pays no attention to the preposition "um". This example shows how students use Google Translate and switch between languages to confirm their hypotheses.

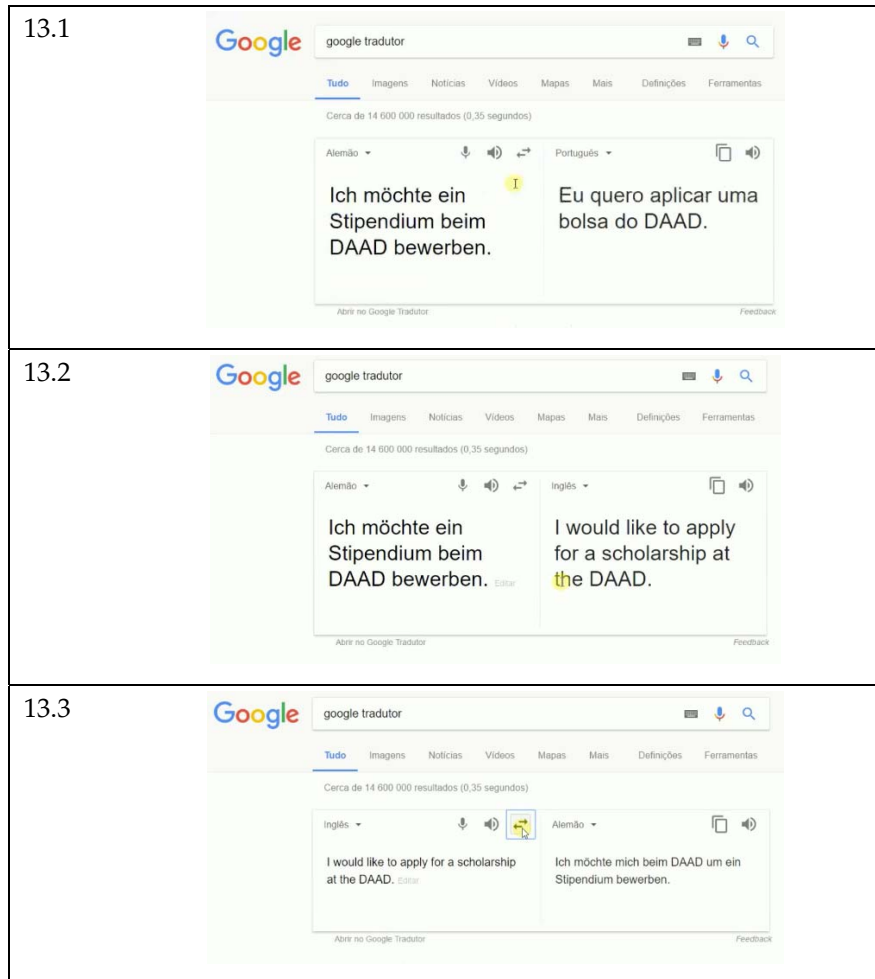


Figure 13: 13.1: Google Translate result for the language pair German–Portuguese; 13.2: Google Translate result for the language pair German–English; 13.3: Google Translate result for the language pair English–German

Based on qualitative observations of the focalization hypothesis in all three participant groups, we arrived at the focalization hypothesis search pattern which can be explained as follows: The students begin by formulating an initial hypothesis (like e.g., in this example: "bewerben" is a reflexive verb), based either on intuition before initiating a search process or on hypotheses formulated on the basis of a specific search action, such as the search for the meaning or translation equivalent of a word. From this point onwards, the whole search process focuses exclusively on the attempt to confirm this hypothesis (see more exam-

ples in the Euralex proceedings paper of this study, Wolfer et al. 2018: 109-111). The observational data seems to indicate that students normally focus their attention on the first result they find in the resources that matches their hypothesis and do not search any further. We also observed that an incorrect initial hypothesis in most cases leads to absurd search actions and results. Furthermore, participants who experience difficulties confirming their hypothesis usually cease to make an effort to correct the stimulus sentence.

The focalization hypothesis described above was identified while conducting a qualitative analysis of participants' search behavior. We aim to complement these qualitative findings with quantitative methods in order to compare the datasets in a more systematic manner and gain a deeper insight into students' search behavior.

4. Conclusions

The combination of quantitative and qualitative methods via the examination of verbal protocols and screen recordings has proven to be an effective approach with which to identify search strategies and patterns common to a specific participant or across participant groups. We received empirical data on important questions related to using online language resources and can draw the following conclusions based on our data.

Although our participants' language proficiency levels are not very high, they use a rather broad range of language resources, most of which are accessed via a Google search and not consulted directly, e.g. by typing in the name of a dictionary in the address bar. Our participants are quite aware of the different functionalities of search engines, translation tools and dictionaries. This can be seen in the fact that they adapt their search strings according to the type of tool. Verbatim or near-verbatim parts of the stimulus sentence are mostly looked up in automatic translation tools and not in dictionaries. We identified three factors that influence the correction rate systematically. (i) Participants who use dictionaries more often than other types of tools are more successful in correcting the stimulus sentences. (ii) Participants who spend more time using the language resources are also more successful in correcting the errors. So, careful reading seems to be one influential factor for solving the task in our study. (iii) Another important factor seems to be whether or not the correct hypotheses are formulated before launching the online search. Participants who had the wrong hypotheses did not see the right solutions although they were presented on the screen. One should keep in mind that we do not know whether the students with a higher level of language proficiency also use dictionaries more frequently (because they have more competence in doing so), spent more time using the language resources (because they can gain a deeper understanding from the presented content) and have better initial hypotheses. Or if two students with the same level of language proficiency really perform differently if they vary in their use of dictionaries vs. translation tools, read

more or less carefully and spend more or less time reflecting on the initial hypotheses. It may also be the case that some students were particularly motivated and therefore read very carefully. So, this is a classical chicken-and-egg question. But what we can see in our data is that these three factors — using dictionaries, careful reading and starting with the right hypotheses — seem to be indicators of successful user behavior.

This leads us to aspects we would change in further studies. Above all, we would do two things differently in future studies: Firstly, we would conduct a short language test prior to the study because this would allow us to identify whether there is a clear connection between language competence and search behavior. We suspect this for our participants in Braga in contrast to those in Santiago de Compostela and Rome, but are not able to prove this assumption. Secondly, we would use a translation task instead of improving sentences in the foreign language. For this study, one central point was to have the same task for all three locations. However, the data we gained show that the task was quite artificial for the students, especially by jumping back and forth between the native and foreign language. On the other hand, as one of the reviewers of this paper argued, the sentence improvement task had the advantage of demanding specific correct vs. incorrect answers and a translation task would not be as clear-cut. For further studies, this issue must be taken into careful consideration. The methodical structure with screen recording and thinking aloud, on the other hand, worked very well. However, in the future we would practice thinking aloud before starting the test, at least briefly with each test person, in order to facilitate speaking during the study.

Empirical studies such as this one are also important because many of the results of our study were unexpected for the language teachers involved: the use of the local language, sometimes even English as a third language in alternation with German, the differentiation between dictionaries and translation programs, the measurable influence of careful reading and the strong influence of the correct starting hypothesis. All this seems almost predictable in retrospect, but was not so beforehand. In our opinion this is exactly where the teaching of language should begin: instead of making general assumptions about what resources are used by students today, our study data could firstly be used as an opportunity to discuss with own students and language learners what resources they use and what strategies they implement. Secondly, at least according to this study, the basic knowledge of different types of language resources should be used to teach even more strategies that support and develop dictionary usage competence. This teaching approach should always be grounded on students' actual use of lexicographic resources. In our opinion, studies such as this one are particularly helpful in this respect. In a further step, it would be important to collect more data in a similar manner in order to investigate whether these results are also confirmed in other countries, for other languages and with other tasks. As Bowker puts it "the key [...] is for lexicographers to listen to users" (Bowker 2012: 396).

Endnotes

1. See <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions> (last accessed 28 June 2018).
2. In *Lexicographica 2018/2019*, there is a German publication on this study with the title: "Recherchepraxis bei der Verbesserung von Interferenzfehlern aus dem Italienischen, Portugiesischen und Spanischen: Eine explorative Beobachtungsstudie mit DaF-Lernenden" (same authors as this article). This year's (2018) Euralex proceedings also include a more methodologically oriented contribution to this study entitled "Combining Quantitative and Qualitative Methods in a Study on Dictionary Use" (Wolfer et al. 2018). — We would like to thank Alexander Koplein for discussing the study results, all assistants and contractors involved in the study, as well as the participants of the IDS Colloquium in fall 2017, the EMLex Colloquium in Stellenbosch and the FaDaF Conference 2018 in Mannheim with whom we discussed the study results. Special thanks go to the participants of the study for their cooperation and to the Institute for the German Language for financing the study. Finally, we would like to thank both reviewers for their very valuable comments.
3. We found this especially important because the participants were recruited by a subset of the authors of this paper who were also their university teachers at that time. By including this section in the instruction and due to the fact that the teachers were not present during the study, we tried to make sure that the participants behaved as "naturally" as possible, i.e. that they did not only consult sites of resources that were taught during their university lessons or avoid specific sites.
4. The first three participants from Braga, Portugal, received 26 sentences instead due to human error on behalf of the experimenters. The 18 sentences that were presented to all 43 participants were also included in the stimuli for these three participants. We will mention the biases and the measures we took to control for them throughout the respective sections.
5. All plots in the present paper were created with the `ggplot2` package (Wickham 2016) for the R environment for statistical computing (R Core Team 2018).
6. The IDs of the sentences go up to number 26, since more sentences were initially meant to be improved, but the pre-tests showed that this was not feasible in the given time.

References

- Bowker, Lynn.** 2012. Meeting the Needs of Translators in the Age of e-Lexicography: Exploring the Possibilities. Granger, Sylviane and Magali Paquot (Eds.). 2012. *Electronic Lexicography*: 379-397. Oxford: Oxford University Press.
- Cohen, Jacob.** 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70(4): 213-220. doi:10.1037/h0026256.
- De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe and Sarah Hillewaert.** 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.
- Domínguez Vázquez, María José, Mónica Mirazo Balsa and Vanessa Vidal Pérez.** 2013. Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch

- lernenden Hispanophonen. Domínguez Vázquez, María José (Ed.). 2013. *Trends in der deutsch-spanischen Lexikographie*: 135-172. Frankfurt a.M.: Peter Lang.
- Domínguez Vázquez, María José and Carlos Valcárcel Riveiro.** 2015. Hábitos de uso de los diccionarios entre los estudiantes universitarios europeos: ¿nuevas tendencias? Domínguez Vázquez, María José, Xavier Gómez Guinovart and Carlos Valcárcel Riveiro (Eds.). 2015. *Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva*: 165-189. Berlin: De Gruyter.
- Ericsson, K. Anders and Herbert A. Simon.** 1993. *Protocol Analysis: Verbal Reports as Data*. A Bradford Book. London: The MIT Press.
- Frankenberg-Garcia, Ana.** 2011. Beyond L1–L2 Equivalents: Where do Users of English as a Foreign Language Turn for Help? *International Journal of Lexicography* 24(1): 97-123.
- Hult, Ann-Kristin.** 2012. Old and New User Study Methods Combined — Linking Web Questionnaires with Log Files from the *Swedish Lexin Dictionary*. Fjeld, Ruth Vatvedt and Julie Matilde Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress 2012, 7–11 August 2012, Oslo*: 922-928. Oslo: University of Oslo, Department of Linguistics and Scandinavian Studies University of Oslo. http://www.euralex.org/elx_proceedings/Euralex2012/pp922-928%20Hult.pdf. (Accessed 11 July 2018.)
- Koplenig, Alexander, Peter Meyer and Carolin Müller-Spitzer.** 2014. Dictionary Users Do Look Up Frequent Words. A Log File Analysis. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 229-250. Berlin/Boston: De Gruyter.
- Koplenig, Alexander and Carolin Müller-Spitzer.** 2014. General Issues of Online Dictionary Use. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 127-142. Berlin/Boston: De Gruyter.
- Levy, Mike and Caroline Steel.** 2015. Language Learner Perspectives on the Functionality and Use of Electronic Language Dictionaries. *ReCALL* 27(2): 177-196. doi:10.1017/S095834401400038X. (Accessed 11 July 2018.)
- Lew, Robert.** 2010. Users Take Shortcuts: Navigating Dictionary Entries. Dykstra, Anne and Tanneke Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6–10 July, 2010*: 1121-1132. Ljouwert: Afûk.
- Lew, Robert.** 2011. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4.
- Lew, Robert.** 2015a. Opportunities and Limitations of User Studies. Tiberius, Carole and Carolin Müller-Spitzer (Eds.). 2015. *Research into Dictionary Use / Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*: 6-16. (OPAL — Online Publierte Arbeiten Zur Linguistik 2015(2)). Mannheim: Institut für Deutsche Sprache. <http://pub.ids-mannheim.de/laufend/opal/pdf/opal15-2.pdf>. (Accessed 11 July 2018.)
- Lew, Robert.** 2015b. Research into the Use of Online Dictionaries. *International Journal of Lexicography* 28(2): 232-253.
- Lew, Robert, Marcin Grzelak and Mateusz Leszkowicz.** 2013. How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. *Lexikos* 23: 228-254.
- Müller-Spitzer, Carolin.** 2014. *Using Online Dictionaries*. (Lexicographica: Series Maior). Berlin/Boston: De Gruyter.

- Müller-Spitzer, Carolin and Alexander Koplenig.** 2014. Online Dictionaries: Expectations and Demands. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 143-188. Berlin/Boston: De Gruyter.
- Nied Curcio, Martina.** 2013. Der Gebrauch zweisprachiger Wörterbücher aus der Sicht italienischer Germanistikstudierender. *Lexicographica* 29: 129-145.
- R Core Team.** 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>. (Accessed 11 July 2018.)
- Tono, Yukio.** 1984. *On the Dictionary User's Reference Skills*. B.Ed. Thesis. Tokyo: Tokyo Gakugei University.
- Töpel, Antje.** 2014. Review of Research into the Use of Electronic Dictionaries. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 13-54. Berlin/Boston: De Gruyter.
- Welker, Herbert Andreas.** 2013. Empirical Research into Dictionary Use since 1990. Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 531-540. Berlin/Boston: De Gruyter.
- Wickham, Hadley.** 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
- Wolfer, Sascha, Martina Nied Curcio, Idalete Maria Silva Dias, Carolin Müller-Spitzer and María José Domínguez Vázquez.** 2018. Combining Quantitative and Qualitative Methods in a Study on Dictionary Use. Čibej, Jaka, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*: 101-112. Ljubljana: Ljubljana University Press, Faculty of Arts. <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2981-1>. (Accessed 11 July 2018.)

Appendix: Stimulus sentences

ID	Satz
1	Meine Nachbarin möchte immer alles wissen. Sie ist sehr kurios .
2	Leider kann ich heute nicht Tennis spielen. Ich bin zu besetzt .
3	Bist du bereit ? Wir müssen jetzt los, wir sind sowieso schon zu spät dran.
4	Kein Problem, wenn der Zucker beendet ist; ich nehme dann Honig.
7	Ich bin einverstanden mit dir .
9	Das erlaube ich dir nicht. Es ist außer Frage .
11	An unserem Forschungsinstitut ist Ihnen unsere Bibliothek 24 Stunden zur Verfügung .
12	Obwohl ich studiere, wohne ich noch mit meinen Eltern.
14	Wenn ich zur Schule ging, habe ich viel Sport gemacht.
15	Morgen habe ich einen Arzttermin und kann deshalb nicht in die Klasse gehen .
18	Ich vorbereite gerade meine letzte Prüfung.
19	Ich möchte ein Stipendium beim DAAD bewerben .
20	Ich habe die Hose viel zu klein gekauft. Jetzt muss ich nochmals ins Geschäft zurück und sie wechseln .
21	Obwohl sich der Junge beeilt hat, hat er die U-Bahn verloren .
22	Er wohnt seit Jahren in Berlin und trotzdem verliert er sich immer noch.
24	Um beim Kartenspielen zu gewinnen, musst du exakt die Regeln folgen .
25	Der Artikel handelt sich um die Migranten in Deutschland.
26	Ich möchte dir heute über einen interessanten Artikel sprechen .

Polish Americans in the History of Bilingual Lexicography: The State of the Art

Mirosława Podhajecka, *Institute of English, University of Opole,
Poland (mpodhajecka@uni.opole.pl)*

Abstract: This paper measures dictionaries made by Polish Americans against the development of the Polish–English and English–Polish lexicographic tradition. Of twenty nine monoscpal and biscopal glossaries and dictionaries published between 1788 and 1947, four may be treated as milestones: Erazm Rykaczewski's (1849–1851), Władysław Kierst and Oskar Callier's (1895), Władysław Kierst's (1926–1928), and Jan Stanisławski's (1929). Unsurprisingly, they came to be widely republished in English-speaking countries, primarily the United States of America, for the sake of Polish-speaking immigrants. One might therefore wonder whether there was any pressing need for new dictionaries. There must have been, assuming that supply follows demand, because as many as eight Polish–English and English–Polish dictionaries were compiled by Polish Americans and published by the mid-twentieth century. The scant attention accorded this topic suggests a chronological approach to these dictionaries is in order, firstly, to blow the dust from the tomes; secondly, to establish their filial relationships; and, lastly, to evaluate their significance for the bilingual dictionary market.

Keywords: HISTORY, BILINGUAL LEXICOGRAPHY, BILINGUAL DICTIONARY, POLISH AMERICANS, SOURCE LANGUAGE (SL), TARGET LANGUAGE (TL), EQUIVALENT, LEXICOGRAPHER, TRADITION

Opsomming: Poolse Amerikaners in die geskiedenis van tweetalige leksikografie: Die jongste stand. In hierdie artikel word woordeboeke wat saamgestel is deur Poolse Amerikaners gemeet aan die ontwikkeling van die Pools–Engelse en Engels–Poolse leksikografiese tradisie. Van die nege en twintig eenrigting- en tweerigtingglossariums en -woordeboeke wat tussen 1788 en 1947 gepubliseer is, kan vier as mylpale beskou word: Dié van Erazm Rykaczewski (1849–1851), Władysław Kierst en Oskar Callier (1895), Władysław Kierst (1926–1928), en Jan Stanisławski (1929). Dit is nie verbasend nie dat hulle wyd in Engelssprekende lande, veral in die Verenigde State van Amerika, ter wille van die Poolssprekende immigrante herpubliseer is. Daar kan dus gewonder word of daar enige dringende behoefte aan nuwe woordeboeke was. Indien daar aangeneem word dat aanbod op aanvraag volg, moes daar wel so 'n behoefte gewees het, aangesien agt Pools–Engelse en Engels–Poolse woordeboeke teen die middel van die twintigste eeu deur Poolse Amerikaners saamgestel is. Die min aandag wat al aan hierdie onderwerp geskenk is, dui daarop dat 'n chronologiese benadering tot hierdie woordeboeke geskik is, eerstens om die woordeboeke te herontdek, tweedens om hul onderlinge verwantskappe te bepaal, en laastens om hul waarde vir die tweetalige woordeboekemark te evalueer.

Sleutelwoorde: GESKIEDENIS, TWEETALIGE LEKSIKOGRAFIE, TWEETALIGE WOORDEBOEK, POOLSE AMERIKANERS, BRONTAAL (BT), DOELTAAL (DT), EKWIVALENT, LEKSIKOGRAAF, TRADISIE

1. Introduction

This paper concentrates on the history of Polish–English and English–Polish bilingual lexicography up until the mid-twentieth century, with particular attention paid to dictionaries compiled by Polish Americans. The dictionary-making tradition has received very little treatment thus far. Suffice it to say that Grzegorzczuk's *Index lexicorum Poloniae* (1967), a bibliography of dictionaries that is now critically out of date,¹ is one of the few sources of information in this respect.

This study, which is a continuation of the research presented in Podhajecka (2016a),² is based on the premise that the bilingual dictionary, despite its ubiquity in the Western world, has been greatly underestimated.³ It was, however, born in response to a real need to understand texts in foreign languages, and it remains a practical tool rather than a book that languishes on the shelf, as Adamska-Salaciak (2014: 1) so simply and disarmingly put it. By bridging gaps between two languages and cultures, moreover, the bilingual dictionary allows for successful cross-linguistic communication and, for this reason alone, its status in the realm of lexicography should be seen as truly unique.

2. The historical background

The history of Polish–English and English–Polish lexicography, in which Polish was either the source language (SL) or the target language (TL), begins with a short glossary in a book of grammar published in 1788. It was compiled by Julian Antonowicz, a Basilian monk and teacher. The glossary included a total of 800 Polish headwords thematically arranged and paired with their English equivalents. It was not, however, devised by Antonowicz. The model was copied from a French–English glossary in Boyer's *The Compleat French-Master ...* (1729), a popular handbook of French aimed at native speakers of English. As Antonowicz spoke French fluently, he translated French headwords into Polish, leaving the English equivalents intact. This is exactly how the glossary came into being, indicating what had become a standard trend in practical lexicography long before the eighteenth century: the use of others' lexical data.

Judging by the number of reprints and new editions,⁴ four dictionaries that came out between 1849 and 1929 may be treated as milestones in the history of Polish–English and English–Polish lexicography. Their potential was soon noticed by publishers in English-speaking countries, primarily the United States of America, and they took steps to republish them for the convenience of the Polish diaspora.⁵ The dictionaries are briefly described below.

[Erazm Rykaczewski]. 1849–1851. *A Complete Dictionary English and Polish and Polish and English Compiled from the Dictionaries of Johnson, Webster, Walker, Fleming and Tibbins, etc., from the Polish Lexicon of Linde and the Polish German Dictionary by Mrongovius. This volume English and Polish ...* (Vol. 1). *Dokładny słownik polsko-angielski i angielsko-polski, czerpany z najlepszych źródeł krajowych i obcych; a mianowicie ze słowników polskich: Lindego, Mrongoviusa i Ropelewskiego; z angielskich: Johnson, Webster, Walker, Fleming-Tibbins i innych. Ten tom polsko-angielski ...* (Vol. 2). Berlin: B. Behr.

This was the first comprehensive dictionary of Polish and English, the English–Polish part including over 26,000 headwords and the Polish–English part nearly 30,000, compiled by Erazm Rykaczewski (1803–1873), a graduate of Vilnius University, a historian, editor and teacher, and an insurgent in the November Uprising forced to seek refuge abroad. Rykaczewski was a novice in the field of practical lexicography, but he was a polyglot and an experienced dictionary user. Unsurprisingly, to compile a dictionary for languages that had not been paired before,⁶ he turned to reference works for other language pairs, primarily English–French and English–German, rendering the TL lexicographic information taken from them into Polish. Although Rykaczewski used a handful of dictionaries, J.G. Flügel's *Complete Dictionary of the English and German ...* (1830) was his main source of data. He is claimed to have compiled both volumes during a ten-year stay in Scotland (Lewandowski 1992: 473). From 1870 onwards, the dictionary appeared under the authorship of Alexander Chodźko (1804–1891); the change in copyright remains a controversy which, despite Chodźko's note appended to the 1870 edition, has not been fully resolved (see Podhajecka 2016a: 100–103). The last edition known to exist came out in Chicago around 1950.⁷

Kierst, W. and O. Callier. 1985. *English–Polish and Polish–English Dictionary / Słownik języka polskiego i angielskiego*. Vols. 1–2. Leipzig: Otto Holtzes Nachfolger.

The dictionary by Władysław Kierst (1868–1945) and Oskar Callier (1846–1929) was a small pocket edition, a mere quarter the size of Rykaczewski's, offering close to 29,000 headwords, i.e. half the scope of Rykaczewski's. It was, in all likelihood, compiled single-handedly by Kierst.⁸ Arranging words in the little space available required the use of specific devices and niching, i.e. the clustering of related words alphabetically in an entry, turned out to be an effective solution. In the long run, it helped the dictionary successfully to challenge Rykaczewski's monopoly, the more so because it was aimed at the mass market. It was republished until 1961.

Kierst, W. (Ed.). 1926–1928. Trzaska, Evert and Michalski *A Dictionary English–Polish and Polish–English. Część pierwsza angielsko-polska* (Vol. 1). *Second Part: Polish–English* (Vol. 2). Warsaw: Trzaska, Evert & Michalski.

In the 1920s, Trzaska, Evert & Michalski (TEiM), the well-known Warsaw publishers, issued another dictionary by Władysław Kierst, who had in the meantime made his name as the translator of Edgar Rice Burroughs's Tarzan stories (e.g. Burroughs 1922). With 32,000 headwords in two handy volumes, the dictionary was only a little more comprehensive than its predecessor.⁹ Nevertheless, Kierst approached his task diligently, working methodically to update the entire text of the dictionary with monolingual works, including *Webster's Collegiate Dictionary* (1919) for English and *Słownik ilustrowany języka polskiego* (1916) for Polish. The last edition appeared in 1984.

Stanisławski, J. 1929. *An English–Polish and Polish–English Dictionary / Słownik angielsko–polski i polsko–angielski*. Warsaw: Skład Główny Księgarnia Wysyłkowa G. Dorn.

This biscopal dictionary was a large pocket edition with nearly 30,000 headwords in both parts.¹⁰ Comparative analyses revealed that it was not compiled from scratch. The extent to which Stanisławski (1893–1973) borrowed from Kierst's TEiM dictionary has been estimated at in excess of 80% (Podhajecka 2016a: 358). One should not underestimate Stanisławski's expertise, however, as he was not a mere imitator, but a genuine innovator. More exactly, he improved three aspects of lexicographic description: firstly, he included phonetic transcription closer to the International Phonetic Alphabet, the European standard at that time, than any other system applied thus far; secondly, he added new headwords and natural-sounding equivalents;¹¹ and, thirdly, he used brief explanatory glosses to distinguish between senses. The dictionary was in use until 1993, and most probably beyond this date, when it was last published by Tormont, a Canadian publisher.

3. Polish American dictionaries

The above dictionaries were widely republished in the United States, so one might wonder whether there was any need to bring onto the market brand-new endeavors. Apparently there was, at least judging by the list of lexicographic works compiled by Polish Americans.¹² Suffice it to say that demand for bilingual dictionaries appears when there is a real or anticipated need for interlingual communication and, hence, works facilitating it.¹³ As Micklethwait (2005: 133) remarks, "A nation of immigrants, especially one with a cultural inferiority complex and an insatiable appetite for self-improvement, provides a ready market for dictionaries".¹⁴ The dictionaries will be presented and briefly evaluated below. If there were any subsequent reprints or editions, this will be signified by means of a vertical arrow. Translations of Polish headwords or equivalents in square brackets are my own and so are translations of quotations accompanied by my initials.

S.Z. and W.B. 1899. *Słownik kieszonkowy polsko–angielski i angielsko–polski*. Chicago: Nakład Spółki Wydawnictwa Polskiego.

Słownik kieszonkowy polsko–angielski i angielsko–polski was the first English–Polish and Polish–English dictionary aimed specifically at Polish immigrants in America, particularly those residing in Chicago, who were soon to become the city's dominant ethnic group (Hargraves 2011: 50). It was compiled by two authors whose identities were concealed behind the initials S.Z. and W.B. Tracing the authorship of anonymous works is fraught with difficulty and this lexicographic work is a case in point. The Polish Publishing Company built on Chicago's Polish Catholic circles is the only clue to the identities of the compilers.

The first of the authors was Szczęsny Zahajkiewicz (1861–1917), who used the initials S.Z. as his cryptonym (Maciejewski and Szweykowski 1982: 253). He was a recognized teacher, editor, poet and playwright, but his lexicographic activity was a new string to his bow. The identity of his co-author is more problematic because there were at least three people in Chicago's Polonia with the initials W.B. with whom Zahajkiewicz collaborated: Wincenty Barzyński (1847–1899), a veteran priest at the St. Stanislaus Kostka Church; Władysław Barwig (1858–?), a parish secretary and the head of a drama group there; and Wiktor Bardoński (1852–1928), the first Polish pharmacist to practise in the state of Illinois.

Still another possible collaborator was Władysław Bełza (1847–1913), one of the most active *literati* in nineteenth-century Lvov, in today's Ukraine (Czartoryski-Sziler n.d.). In fact, Bełza never visited America, but he knew Zahajkiewicz from his home city and corresponded with members of Chicago's Polonia, such as Paweł Sobolewski.¹⁵ Having been regarded as a chronicler of Polishness who "provided many proofs of his merits and unblemished integrity" [M.P.] (Kąsinowski 1913: 83), Bełza was a likely collaborator in the compilation of a dictionary aimed at Polish immigrants. For the time being, however, the involvement of any of the above-mentioned figures cannot be established reliably, so the authorship of the dictionary remains a riddle.¹⁶

S.Z. and W.B.'s dictionary was a simple abridgement of Rykaczewski's endeavor. The compilers took approximately 30% from the wordlist of the parent dictionary in each part, thus omitting a great many headwords. The entry structure consisted of their copying the equivalents, but disregarding most of the dictionary text (e.g. cross-references, grammatical information, SL contextual uses and their TL translations). Sense division was also significantly limited, as is shown below.

S.Z. and W.B.	Rykaczewski
Czas, u, s. m. time; 2) weather; 3) (gram.), tense.	CZAS, U, s. m. time, 2) weather; (gram.), tense. <i>Trawić czas na nauce</i> , to spend one's time in study. <i>Tracić czas</i> , to lose or waste one's time. <i>Powrócić w sam czas</i> , to come back just in time or in the very nick of time. <i>Wolny czas</i> , leisure. <i>Teraz jest właśnie czas po temu</i> , this is a favourable opportunity. <i>Jeszcze nie czas figom</i> , the time of figs is not yet. <i>Za</i>

	<p><i>czasów Abrahama</i>, in the time of Abraham. <i>Za owych czasów</i>, in those times, in those days. <i>Za dawnych czasów</i>, in olden times, in times of old, in times of yore. <i>Od niepamiętnych czasów</i>, in times out of mind. <i>W swoim czasie</i>, in proper time and place. <i>Wszystko dobre w swoim czasie</i>, all is in good time. <i>Ciężkie czasy</i>, troublesome, difficult, hard times. <i>Piękny lub brzydki czas</i>, it is fine or fair, it is fine or bad weather. <i>Od czasu do czasu</i>, from time to time, now and then. <i>Tymczasem</i>, meantime, meanwhile. <i>Wówczas lub podówczas</i>, at that time, then ...</p>
--	--

Table 1: A sample entry in S.Z. and W.B.'s and Rykaczewski's dictionaries

The cuts allowed S.Z. and W.B. to keep their own dictionary compact and concise; the Polish–English part included more than 18,000 entries over approximately 400 pages of pocket format and the English–Polish part more than 18,000 entries over 500 pages. The headwords were printed in the same type as the rest of the entry, at least in volume one, which suggests that user–friendliness was not yet an issue.

Paryski, A.A. 1899. *Wielki ilustrowany angielsko–polski i polsko–angielski słownik, zawierający wszystkie wyrazy, zwroty i przysłowia, używane w mowie i literaturze angielskiej i polskiej, oraz nazwy techniczne i geograficzne, imiona własne, wykazy skrótów, znaków, symbolów i. t.d., z podaniem wymowy, sylabilizacji i form gramatycznych. Część 1: Słownik angielsko–polski [Instalments 1–?]. Toledo, OH.*

Paryski, A.A. 1899. *Kieszonkowy słownik polsko–angielski*. Toledo, OH: Antoni A. Paryski.

Paryski, A.A. 1900. *Kieszonkowy słownik angielsko–polski, zawierający przeszło 40.000 wyrazów używanych w mowie i literaturze angielskiej, z podaniem dokładnej wymowy każdego wyrazu, opracowany na podstawie słowników szkolnych Webster'a, Worcester'a i Standard*. Toledo, OH: Nakład, druk i własność A. A. Paryski.

Paryski, A.A. 1900. *Słownik polsko–angielski i angielsko–polski oraz nauka wymowy angielskiej*. 7th edition. Toledo, OH: Nakład, druk i własność A. A. Paryskiego.

The turn of the twentieth century witnessed robust lexicographic activity in America. This was, in large measure, due to Antoni A. Paryski (1865–1935), another Polish-American entrepreneur to undertake the compilation of bilingual dictionaries. Born in Poland to a peasant family, he must have had a good deal of stamina to pursue a career in America, where he worked hard to acquire English and to learn typesetting, printing, and then journalism to make a living. Several years after arriving in the United States, he set up a highly successful enterprise called the Paryski Publishing Company.

The scale of Paryski's involvement in the American publishing market is astonishing. As well as newspapers, primarily *Ameryka-Echo*, he issued close to eight million books, which earned him the reputation of "the Polish Hearst" (Jaroszyńska-Kirchmann 2015: 2). Majewski (2003: 41) argues that "in conjunction with newspaper publishing, book production could be very cheap, and very

profitable", which suggests that Paryski steered into relatively safe waters. Thousands of the books were apparently dictionaries compiled by himself and, to give users an alternative, by Kierst and Callier. In any case, both were advertised regularly on the pages of *Ameryka-Echo* and sold by Paryski's salesmen criss-crossing the United States.¹⁷

Paryski was the author of the four volumes listed above, of which the first, *Wielki ilustrowany angielsko-polski i polsko-angielski słownik ...* (1899), was hailed as a reference work of unprecedented quality. It was to be an exhaustive dictionary with illustrations and an array of specialist terms, perhaps modeled on encyclopedic works such as Ogilvie's *The Imperial Dictionary ...* (1859–1860). It is unknown what brought the innovative project to a halt. According to Chojnacki (n.d.), there were only a few issues published in 1899. It is to be regretted that no single copy of the dictionary appears to have survived to date.

That the pocket versions offered the same lexical material is confirmed by the number of pages: the English–Polish part comprised 299 pages and the Polish–English part 155 pages.¹⁸ It is interesting that, even though the volumes were parts of one dictionary, they were priced differently: one monoscopal dictionary was sold at \$1, and the other at \$1.5. This may be indicative of Paryski's marketing skill. To compile the dictionary, he borrowed from a range of lexicographic works. His sources included Rykaczewski's (under Chodźko's name) *A Complete Dictionary English and Polish ...* (1890), Kierst and Callier's *English–Polish and Polish–English Dictionary* (1895), Whitney's *The Century Dictionary ...* (1895), and Webster's *A Primary School Dictionary of the English language ...* (1871).

It is unclear whether it was Webster's school dictionary or a description of English phonics regarded as a reading instruction, such as *Practical Phonics ...* (1881),¹⁹ that prompted Paryski's ingenious idea of combining spelling with pronunciation (see also Emans 1968, Barry 2008). Figure 1 provides an example of this.

eöckäde', kokarda.	eöeffiö'ienöy, współdziałanie.
eäckätö', kakadu (papuga).	eöeffiö'ient, współczynnik,
eöck'ätříce, 1. bazyliżek.	współdziałający.
eöck-böat, mały statek, łódź.	eöö'qüäl, współrówny (ity).
eöck'öt, kwit zapłaconego cła.	eöörce', przymusić (cion, cive).
eöck'figłt, walka kogutów.	eöössen'üäl, współlistotny.
eöe'k'le, 1. kół; 2. gatunek	eööstäte', wspólny majątek.
eöck-löft, facyata. [muszli.	eöötär'näl, współwiekuisty.
eöck'nöy, 1. Londyńczyk; 2.	eöötär'nity, współwiekuistość
gap'.	eöö'väl, 1. jednako stary, ró-
eöck'pít, 1. arena do walki	wieśny; 2. rówieśnik.
kogutów; 2. izba na okrę-	eööx'ist', współlistnieć.
cie pod pokładem do ar-	eööx'ist'énöe, współlistnienie.
eöck'röach, karaluch. [mat.	eööxtën'siön, współrozcią-
eöck's'eömö, 1. grzebień ko-	głość. [napój].

Figure 1: A sample of Paryski's *Słownik polsko-angielski i angielsko-polski ...* (1900)

There is no way of knowing what American students consulting Webster's school dictionaries thought of the notation system consisting of a range of diacritical marks and awkward characters,²⁰ but they were evidently drilled on it in the classroom. By contrast, the system must have been extremely confusing for Polish learners of English. Be that as it may, one of the reasons why users consult bilingual dictionaries is for information on spelling. To my knowledge, there is no research data on dictionary use in the past, but it may be safely assumed that bilingual dictionaries in immigrant communities were purchased mainly to be used at home.

Paryski's dictionary, including over 24,000 entries in both parts, was largely a labor of love. The English and Polish wordlists were far from comprehensive, grammatical information was missing, contextual uses were infrequent, phraseology was scarce, and TL equivalents were not always chosen with semantic wisdom, e.g. *cartel* 'dostawa żywności' [food supplier] (> 'kartel'), *chess* 'warcaby' [checkers] (> 'szachy'), *fiance* 'narzeczona' [fiancée] (> 'narzeczony'), and *swum* 'płynął' [(he/it) swam] (> 'past participle of *swim*'). On the other hand, Paryski added new headwords, new senses, and new equivalents (e.g. *szynk* 'saloon; inn') in order to modify the dictionary in accordance with the changing times and, in particular, the American context. For example, he was the first lexicographer to pair *dandruff* with its modern counterpart 'łupież' (cf. Rykaczewski's 'papry na głowie' [dirt on the head]) and *whisky* with 'wódka' (cf. Rykaczewski's 'gorzalka' [booze]).

The E-P part was more extensive than the P-E part. One might assume that Paryski experienced some sort of "alphabet fatigue" (Osselton 2007: 81-91) in compiling it.²¹ To put it differently, this part is unsophisticated content-wise, as Paryski usually paired each Polish headword with only one English equivalent. In doing so, he also borrowed from Rykaczewski's dictionary much more frequently and less critically than he did in the E-P part. All this suggests that he may have been a brilliant entrepreneur with a flair for business, but he was no first-class lexicographer.

Since S. Z. and W. B.'s and Antoni Paryski's dictionaries relied on Rykaczewski's work, their volumes might have been similar. Still, while S.Z. and W.B. took all their lexical material, truncating it severely, from the parent dictionary with no serious modifications, Paryski at least attempted to contribute to the dictionary in his own way.

Słowniczek Polsko-Angielski z wymową fonetyczną. [pre-1905]. Chicago: Smulski Publishing Company.

This tiny booklet, a collection of basic words, was published anonymously by the Chicago-based Smulski Publishing Company. It is a great rarity today inasmuch as the Polish Museum of America in Chicago is the only acknowledged institution that has a copy of it. The publication date of the *Słowniczek* [Little dictionary] is missing from the title page, but there are reasons for claiming that it appeared prior to 1905.

The envisaged readership was unspecified, but the dictionary was probably compiled for the benefit of those Polish immigrants who knew little or no English, but who could at least read and write. After all, thousands of newcomers to America at that time were illiterate. The *Słowniczek* included around 2,500 headwords in three columns (see Figure 2), the majority of which were borrowed from Rykaczewski's dictionary and some from Kierst and Callier's. To the entry structure was added simple phonetic transcription expressed solely in Polish graphemes. The size and low price must have been the book-let's main strengths.

Tekst polski.	Tekst angielski.	Wymawia się.
stary,	old,	old.
statek,	vessel, ship,	wessel, szyp.
stanąć,	to stop, to pause,	tu stap, tu pauz.
stawiać,	to set, to place,	tu set, tu plejs.
stękać,	to groan,	tu gron.
stełmach,	wheel-wright,	hull-rajt.
ster,	rudder, helm,	rodder, neim.
stłuczony,	broken to pieces,	broken tu pises.
stół,	table,	tejbel.
stolarz,	joiner,	dżojner.
stołek,	stool,	stul.
stolica,	capital,	kiapytel.
stopa,	foot,	fut.
stopień,	step,	step.

Figure 2: A sample of the *Słowniczek* (pre-1905)

An analysis of the content suggests that it is likely to have been compiled by Modest Maryański, a brief biographical sketch of whom is given below, albeit there is no concrete evidence for such an accreditation.

Maryański, M. 1906. *Jedyny w swoim rodzaju przewodnik polsko-angielski i słownik polsko-angielski dla wychodźców polskich i przybyszów do Stanów Zjednoczonych Ameryki Północnej i Kanady, ułatwić mający stawianie kroków pierwszych w kraju obcym i naukę języka angielskiego, z podaniem wymowy i brzmienia każdego wyrazu angielskiego według metody fonetycznej, z dołączeniem niektórych uwag, rad i wskazówek.* Chicago: Własnym nakładem.

↓

1907. Warsaw: Gebethner and Wolff.

The volume titled "a unique Polish-English guide-book" for "Polish immigrants and newcomers to the United States of North America and Canada" [M.P.] was compiled by Modest Maryański (1854-1914), a Polish mining engineer. Although the date on the title page is 1905, the book was not published until 1906.²²

The story has it that, in 1887, Maryański went to the United States (his original destination was Australia) in search of "bread", as he himself put it (Paszkowski 2008: 292). This is somewhat difficult to believe since he was previously the founder of a company exploiting the Truskawiec mine in Galicia (Maryański 1882: 3), in the Austrian-Hungarian Empire, and, two years later, its managing director (Chłapowski 1884: 173).²³ Since Maryański's biography is very patchy, we may only rely on hypotheses. One of them assumes that he invested his money in a risky business or unsuccessful speculations, bringing his family (wife and son) to the brink of poverty.²⁴ In America, with no English, he worked as a builder and miner, experiencing various ups and downs on the way. His plight was so difficult at one point that he was considering suicide. It was only working as a mining expert at a gold mine in Colorado that eventually made him a wealthy man; as stated by Paszkowski (1987: 304), the Consolidated Kosciusko Mine was employing twenty to seventy five miners and "the returns were satisfactory", at least at first. Even though Maryański's fortunes soon changed to see him plying his trade as a newspaper editor, he was admittedly one of the best-known Polish-American self-made men of the day.

stary,	old,	öld.
starzec,	an old man,	en öld men.
statek,	vessel, ship,	wessel, szyp.
staw,	joint, pond,	dżojnt, pand.
stanać,	to stop, to pause,	tu stap, tu páz.
stawiać,	to set, to place,	tu set, tu plejs.
stękać,	to groan,	tu grön.
stękanie,	groaning,	gröning.
stelmach,	wheel-wright,	huil-rajt.
step,	prairie,	preri.
stempy,	stamping mill,	stamping myll.
ster,	rudder, helm,	rodder, helm,

Figure 3: A sample of Maryański's dictionary part (1906)

The book was divided into three parts: "Przewodnik Polsko-Angielski" [A Polish-English Guide-book], "Słownik Polsko-Angielski z wymową fonetyczną" [A Polish-English Dictionary with Phonetic Transcription] and "Część trzecia" [Part Three]. The first and the last part included information which, in Maryański's view, was essential to Polish immigrants. As for the dictionary part (see Figure 3), each page was divided into three vertical columns: one for the English headwords, one for the Polish equivalents, and one for the phonetic transcriptions. The similarity between Maryański's wordlist and the *Słowniczek* is not incidental: among the 4,000 Polish headwords that found their way into

Maryański's dictionary, 2,500 were taken directly from the *Słowniczek*. The transcription was also aimed at an inexperienced dictionary user, and the only difference here was a macron over vowel sounds (e.g. ā) to signify length.

Szumkowski, L. 1908. *Dykcjonarz kieszonkowy polsko-angielski i angielsko-polski. Zawiera 12 000 słów polskich, 18 000 słów angielskich*. Chicago: [no publ.].

↓

1909. Chicago [no publ.]; 1912. Chicago: L. Szumkowski (3rd corrected edition).

This pocket dictionary was another reference work by a Polish American targeted at the huge, and steadily growing, Polish community in Chicago. The dictionary, the costs of whose publication were covered by Leonard S. Szumkowski (1885–1954), went into three editions. As an advertisement in *Słowo Polskie* (1912) informed prospective buyers,²⁵ the main attributes of the dictionary were its exhaustive wordlists, its low price (75 cents), its hard-wearing leather cover, and its pocket format. The claims regarding its coverage at least were no exaggeration: the Polish–English part included over 11,000 headwords and the English–Polish part over 14,500, even though the figures fall short of Szumkowski's own estimates.²⁶ The text, as may be seen in Figure 4, was dense in order to save space.

There is no doubt that Szumkowski made an effort to modernize and Americanize the lexical material, but his dictionary was hardly an update. It was compiled solely on the basis of Rykaczewski's and Paryski's works. In addition to the core vocabulary, it included peripheral items useless to the target readership, such as historicisms (e.g. *chaperon* 'kaptur' and *scholiast* 'tłumacz') and dialecticisms (e.g. *bożyć* 'to swear' and *przynuka* 'compulsion'). Although most TL equivalents were satisfactory, some were clearly wrong (e.g. *grzywna* 'reward' [fine, penalty] and *weird* 'czarownik' [wizard]), and the quality of the phonetic transcription was poor.

housemaid, -mejd, służąca	hurdle, hor'dł, płotek
housewife, -wajf, gospodyni	hurl, horł, zgiełk
housing, -ing, czaprak	hurrah, hura', hura!
hovel, how'eł, chata	hurricane, hor'ykiejn, bu-
hover, -er, unosić się	rza
how, hau, jak	hurry, -y, spieszyć
however, hauew'er, jednak	hurt, hort, urazić
howl, hauł, wyć	husband, hoz'band, mąż
hub, hob, piasta	husband, gospodarować
huckster, -ster, szachraj	husbandman, rolnik
huddle, hod'ł, pogmatwać	husbandry, rolnictwo
hue, hiu, barwa	hush, hoś, cyt!
huff, hof, chępliwość	husk, hosk, łupina
huffish, -yś, fukliwy	husking, -ing, łuszczyć
hug, hog, uściśnienie	husky, -y, łupiniasty
huge, hiudź, wielki	hussar', huzar

Figure 4: A sample of Szumkowski's English–Polish part (1908)

Szumkowski arrived in the United States as a boy and attended American schools, so he must have been fully bilingual. He was a doctor of medicine specializing as a surgeon, but his professional career was not limited to his medical practice. Above all else, his technical inventions allowed him, in 1918, to establish the Ursus Motor Company. In his capacity as president of the corporation, he left for Europe in 1920 with the idea of helping the newly-re-established Poland. Having allegedly received "the best possible concessions from the Polish government" [M.P.] (Lokański 1920: 345), he purchased premises in Warsaw in order to begin the production of trucks and tractors. He came back to the United States in May 1921, but there is no information on his later exploits.

Jesień, W. 1925. *Słownik angielsko-polski zawierający 4000 najpospoliej używanych słów*. Warsaw: Michał Arct.

Wacław Jesień's *Słownik angielsko-polski ...* (1925) is a large pocket edition with a collection of American English vocabulary paired with Polish equivalents. It was aimed at immigrants or, more precisely, prospective immigrants aspiring to the status of naturalized Americans. As the title page informed prospective buyers, Wacław Jesień (1886–1937) was "a former specialist in foreign languages at the American Bureau of Education" [M.P.]. Even though little is known about his life, he was indeed employed by the Bureau in the 1910s. The English–Polish dictionary might thus be seen as a continuation of his career in education.

The project is interesting in that the list of 4,000 English words²⁷ was compiled by Alfred E. Rejall, professor of psychology and specialist in adult education for the State of New York. It became the basis for the New York State Regents' literacy test for new Americans, assessing their reading comprehension (Stinchfield-Hawk 1928: 162). As Rejall did not determine the senses that new Americans should be acquainted with, the task of transforming the word-list into a fully-fledged bilingual dictionary was assumed by Jesień. Closer analysis suggests that he consulted both monolingual and bilingual sources for this purpose, but *Słownik angielsko-polski ...* (1925) also reflects his own ideas regarding the appearance of such a "learner-oriented dictionary".²⁸ This includes, among other things, the choice of TL equivalents, a proportion of which were translated literally from English into Polish. One of them is *surprise* '... niespodziewane najście lub atak, zdziwienie, przedmiot wywołujący zdziwienie lub oszołomienie' [an unexpected intrusion or attack, astonishment, an object causing astonishment or bewilderment].²⁹

meal , (mijl) — kasza, posiłek, czas jedzenia.	meeting , (mijt' yng) — spotkanie, zebranie, zgromadzenie
mean , (mijn) — oznaczać, mieć na myśli, zamierzać, zapowiadać, wróżyć; <i>formy</i> : meant (ment) i meant	melody , (mel'odi) — melodja, melodyjność
mean , (mijn) — pospolity, niski, ordynarny, nikczemny; <i>rzecz.</i> średnia, przeciętna; <i>l. mn.</i> środki, majątek, dochody	melon , (mel'au) — melon, arbuz
meantime , (mijn'tajm) — tymczasem, w międzyczasie	melt , (melt) — topić, rozpuszczać, topić się, rozpuszczać się
meanwhile , (mijn'hūajl) — tymczasem, w międzyczasie	member , (mem'ber) — członek
measure , (meź'er) — miara; <i>czas.</i> mierzyć	memorial , (memor'jel) — pamiątkowy, pomnikowy, pamięciowy; <i>rzecz.</i> pomnik, memoriał.
	memory , (mem'ori) — pamięć, wspomnienie

Figure 5: A sample of Jesień's *Słownik angielsko-polski ...* (1925)

On the other hand, Jesień introduced phonetic transcription indicating, fairly effectively, American pronunciation and a significant number of Americanisms (e.g. *Decoration Day* 'dzień wieńczenia grobów (święto amerykańskie)', *drugstore* 'apteka', *eagle* 'orzeł, moneta złota 10-dolarowa', *Fourth of July* 'Czwarty Lipiec (Święto niepodległości w Ameryce)', *gas/gasoline* 'gaz, gazolina, benzyna motorowa ...' and *highway* 'droga, gościniec'). Taking everything into account, the dictionary should have been welcomed by future emigrants, but, for unknown reasons, it never ran into a second edition.

Wilde, T.M. 1928. *Smulski's Dictionary. An English-Polish and Polish-English Pocket Dictionary / Słownik Smulskiego angielsko-polski i polsko-angielski. Słownik kieszonkowy.* Chicago: Polish-American Publishing Co.

↓

1944. Chicago: Polish-American Book Co.

↓

Smulski, J.F. and T.M. Wilde. 1945. *Słownik angielsko-polski polsko-angielski z wymową.* Poznań: Wydawnictwo Polskie R. Wegnera (Oddział w Norymberdze).³⁰

Smulski's Dictionary (1928) was compiled by T.M. Wilde (1858–1943), a sergeant major in the United States Army, a commander in the Kosciuszko Guard, editor of *Kuryer Polski*, and a bank examiner in Wisconsin (Podhajecka 2016a: 319-321). Let me cite the preface to the dictionary in its entirety:

Realizing an urgent need of a modern English-Polish dictionary, Mr. John F. Smulski gave the initiative to this work, and lent it his unstinted support and co-

operation. More than two thousand modern words, not found in any existing English–Polish and Polish–English dictionary, have been included in the pocket edition herewith presented. Owing to the limited space, rare, obsolete, and purely technical words have of necessity been omitted. The dictionary comprises twenty thousand English, and sixteen thousand Polish words, with their proper equivalents. Phonetic pronunciation of each English word is indicated. Polish pronunciation is described and explained. Lessons in conversation for use of beginners will be found in the appendix.

The preparation and printing of the dictionary was financed by John F. Smulski (1867–1928), an American lawyer, millionaire banker, businessman, politician of Polish origin, and a philanthropist (see, e.g. Kantowicz 1975: 64).³¹ The volume was to include as many as 36,000 headwords in both parts, of which more than 2,000 were to be brand new.

Since it was a small pocket edition, the TL equivalents were the main component of the entry structure and other kinds of lexicographic information, such as phonetic transcription and labels, were kept to an absolute minimum (see Figure 6). As my findings reveal, *Smulski's Dictionary* was based on Kierst and Callier's and the TEiM dictionaries. Given the influence of the former, it should come as no surprise that the Polish equivalents (e.g. *aborigines* 'tuziemcy') are sometimes old-fashioned, indicating that Wilde had not succeeded in updating the whole dictionary as consistently as he had planned. Most of the 2,000 new headwords, such as *dziurawka* 'hollow brick', *etażerka* 'what-not' and *lechtaczka* 'clitoris', were borrowed from the so-called *Słownik warszawski* [Warsaw Dictionary], the largest monolingual dictionary of Polish. They were apparently paired with their English counterparts on the basis of Wilde's interlingual competence.

tapioca	telltale
tapioca tapi-ō'ka, tapjoka	teacher -ūr, nauczyciel
tapir tǎ'pūr, tapir	-ka)
tar smoła	teak tik, (Bot.) prze-
tarantula ta-ran'tiu-la,	chownia
krzeczek	teal til, cyranka
tardy tar'dy, późny;	team tim, 1. zaprząg;
opieszaly	sprząg; 2. drużyna
tare ter, 1. wyka; 2.	teamster -stūr, woźnica
źdźbło	teapot ti'pōt, czajnik
tare tara	tear ter, drzeć
target tar'get, tarcza	tear tir, iza
tariff tar'if, taryfa	tease tiz, droczyć
tarnish tar'nisz, 1. śnie-	tease czochrać
dzieć; 2. splamić	teaspoon ti'spun, łyżecz-
tarpaulin -pō'lin, brezent	ka
tarry tar'y, marudzić	teat tit, cycek
tart cierpki	teasel ti'zl, drapacz
tart ciastko owocowe	technical tek'ni-kal,

Figure 6: A sample of Wilde's *Smulski's Dictionary* (1928)

Despite drawing on several existing sources, *Smulski's Dictionary* showed a great deal of lexicographic creativity, providing new equivalents (e.g. Wilde's *tarantula* 'krzeczek' / Kierst's 'tarantula (pająk)') and equivalents for new senses (e.g. Wilde's *angina* 'dusznica' [angina pectoris] / Kierst's 'angina, zapalenie gardła' [tonsillitis]). In this way, it was not only a derivative but also an innovative English–Polish / Polish–English dictionary, even though some equivalents were imperfect (e.g. *szalenie* 'furiously' (> 'madly, exceedingly'), *włosień* 'long hair from horse's tail' (> 'a helminth in mammals') and *wścibiński* 'meddler' (> 'meddlesome')).

Lilien, E.L. 1944. *Lilien's Dictionary. Part 1: English–Polish / Ernesta Liliena słownik. Cz. 1: Angielsko–polski*. Buffalo: Drukiem Dziennika dla Wszystkich [Instalment 1]; 1944–1945. Buffalo: Wydawnictwa Słownika Liliena / Stevens Point: Wydawnictwa Słownika Liliena [Instalments 2–8]; 1947–1951. Stevens Point: Wydawnictwa Słownika Liliena [Instalments 9–19].

This dictionary, compiled by Ernest Lilien (1872–1952), is the only one examined in this paper that is left unfinished. Lilien's death in 1952 thwarted his extraordinary plan to make an entirely new English–Polish and Polish–English dictionary modelled on unabridged American dictionaries, primarily so-called *Webster's Second* (1934), the largest dictionary of English.³² Such a project was an enormous undertaking, particularly as all the editorial duties were assumed by Lilien himself single-handedly (letter to Mizwa of 8 June 1944). The coverage of the English–Polish part, the first and only to have been compiled, was estimated at 115,000 headwords, but only approximately half of it, i.e. 19 out of the 40 planned instalments, appeared in print.

Ernest Lilien was born to a rich Jewish family in Lvov. Before emigrating to America in 1916, he was a businessman.³³ In the United States, by contrast, he was known primarily as a journalist in Stevens Point. Prior to that, he had edited several Polish-language newspapers in Buffalo, Detroit, Toledo, and Chicago, as stated by the author of his obituary in *The Milwaukee Journal* (1952).³⁴ His scope of interests was, however, far wider.³⁵

Lilien worked on his dictionary for the last twenty or so years of his life. The publication, as we learn from the preface, was supported by St. Peter's Foundation in Stevens Point thanks to the Reverend Julius Chylinski.³⁶ Additional backing came from the Kosciuszko Foundation, an organization of American Polonia, which subscribed to "a substantial portion of the printed installments as they appeared" (Mizwa 1961: v).³⁷ The Embassy of the Polish People's Republic in Washington also showed its appreciation for Lilien's effort by subscribing to 200 copies of the dictionary, of which 194 were to be distributed among libraries, universities, and cultural institutions in Poland.³⁸

cat nap (k. nap) króciutka drzemka.

catnip, catnep, s. (-nyp bot. miętna roślina kocimiętka, *NEPETA*, z rodziny wargowych, *LABIATÆ*; N. *CATARIA* Linn. kocimiętka właściwa, bylina szaro-włosista do 4' (120 cm.) w Ang. *catmint*).

cato- przedr. (ke'to; kata' gr. w dół) dolo-, spodo-.

Catonian, adj. (kejto'njen) katoński, dotyczący Katona Starszego albo Katona z Utyki, dwóch starożytnych mężów stanu znanych z prostoty obyczajów i surowości.

Catonian s. hist. kl. stronnik Katona.

cat-o-nine-tails, s. (ket'onajn'tejls') kańczug, przyrząd do smażenia przestępców, z dziewięcioma pletniami, wiązany w guzy, na spójnej rękojeści.

catopter, catoptron, s. (katap'ter, -tron) przest. lustro, zwierciadło.

catoptric, -al, adj. lustrzany, zwierciadlany; c. **light** światło, w którym promienie skupione są przez odbicie, widzialne na odległość znaczna, jak np. w latarni morskiej; c. **telescope** astron. teleskop reflektorowy, luneta mająca zwierciadła.

catoptrically, adv. (-H) jak w lustrze, w odbiciu.

catoptries, s. (-optriks) opt. katoptryka, dział optyki zajmujący się zjawiskami odbijania się światła; nauka pierwszy raz spisana w dziele Euklida "Katoptryka" r. 300 przed Chr.

catopromancy, s. (-tromen'si)

legająca na tworzeniu figur z nitki nawijanej na palce obu rąk.

cat's-ear, s. (kec-ir') 1) med. kocie ucho, chorobowe zniekształcenie ucha ludzkiego na podobieństwo kociego; 2) bot. kocie ucho ob. **capweed**, s. 3) i i.

cat's-eye, s. (kec-aj') 1) med. kocie oko, chorobowy stan oka znamienny szczególnym połyskiem w gromadzie siatkówki, jasna ślepotą, **AMAUROSIS SPLENDENS**; 2) min. kocie oko, zielonawo-szary lub brunatny kwarzec z Indji Wsch., używany jako klejnot; 3) bot. przetańcznik ożankowy **VERONICA CHAMÆDRYS** Linn. i p. bizantyński, V. **BYZANTINA** BSP.

cat's-foot, s. (-fut) bot. kocia stopa, zająca noga, (ziele) owieczki.

Catskill formation (kec'kil form'ejzen) geol. formacja w górnej serii systemu dewońskiego w okolicy Catskill w stanie New York, ob. **geology**.



cat's-paw 1) głupiec, dający się używać za narzędzie do wyciągania komu kasztanów z ognia; 2) bot. kocia stopa ob. **cat's-foot**; 3) mar. a) kocie łapki, wietrzyk poruszający lekko powierzchnią morza miejscami w czas całkowitej ciszy; b) kocie łapki, powrót o dwóch okach na hak, ob. **knot**, także **catspaw**.

GLAUCA z rodziny mydleńcowatych, **SAPINDACEÆ**; w czasie długiej posuchy ścinane na paszę.

c. guard (k. gard) St. Zj. przyrząd automatyczny na krzyżowce wstrzymujący bydło na gościńcu przed torem kolejowym w chwili przejazdu pociągu.

c. louse, (k. lauz) zool. wesz bydła, ogólna nazwa robactwa nawiedzającego rozmaite gatunki zwierząt np. krwiopicia **HÆMATOPTINUS**, zjadający włosy **TROCHODECTES SCALARIS**, wesz kurza **MENOPON PALLIDUM**, wesz kaczka **LIPEURUS SQUALIDUS** Htd.

c. pump, (k. pomp) pompa bydłeca, puszczana w ruch automatycznie przez bydło.

c. range, c. run, (k. rendź, k. ron) obszerny i ogrodzony plac dla bydła.

c. show, (k. szo) wystawa bydła, pokaz bydła.

c. tick, (k. tik) zool. kleszcz rzędu roziczy, nawiedzający bydło w podzw. okolicach Ameryki Płd. i Płn., **BOOPHILUS ANNULATUS**; czeplą się nóg bydła i prawdopodobnie bywa głównym roznośicielem gorączki teksaskiej, **Texas fever**; kleszcz psi, **IXODES RICINUS**, także często w Polsce, szczególnie na psach myśliwskich.

Cattleya, s. (ketil'ie, od nazwiska botanika ang. Williama Cattley'a) bot. rodzaj podzw. amer. epifitów storczyków o kwiatach szczególnej piękności; c. roślina tego rodzaju.

catty, adj. (ke'ti) gw. koci, fałszywy, nieszczerzy, podstępny, złośliwy.

Figure 7: A sample of *Lilien's Dictionary* (instalment 8, 1945)

Lilien's motivations are explained in a two-page preface appended to the first instalment. Briefly, the compilation was triggered by an urgent need for a modern exhaustive English-Polish and Polish-English dictionary for both everyday and specialist uses. Two reference works typical of the American lexicographic tradition, *Webster's Second* (1934)³⁹ and Funk and Wagnalls dictionary (1893), were said to be Lilien's direct models.⁴⁰ A simplified phonetic transcription was employed, which became the subject of harsh criticism in a lexicographic work with such lofty aspirations (e.g. Scherer 1946), and the "most exact translations" and "brief definitions" [M.P.] were regarded as a necessity. It is worthy of note that Lilien not only used American English dictionaries, but he also turned to a number of bilingual and multilingual ventures, including *Słownik morski* [Maritime Dictionary], Matankin's *Słownik wojskowy* [Military Dictionary], and Wlekiński et al.'s *Słownik techniczny* [Technical Dictionary] (Lilien 1944: 3).

With hindsight, any assessment of this dictionary would have to take cognizance of Lilien's intention to describe over 100,000 English words, a great many of which were borrowed from *Webster's Second* and, occasionally, Funk and Wagnalls. Some of the words, including those originating in other world Englishes, were admittedly of little use for Polish Americans. Examples include *billy tea* '(Australia) herbata parzona w kociołku billy' [(Australia) tea brewed in a billy pot], *cachude* 'med. pastylka, używana w Indiach jako odtrutka i jako

środek przeciw zaburzeniom żołądkowym i spazmom' [*med.* a pill used in India as an antidote to gastric problems and convulsions], and *gora, gorah, goura* 'instrument muzyczny Hotentotów i Buszmanów w Afr.' [a musical instrument used by Hottentots and Bushmen in Africa⁴¹]. Other superfluous elements are narrow specialist terms, for which there were no Polish equivalents available (e.g. *frog-eye* 'patol. rośl. nazwa rozmaitych chorób znamienych pierścieniami dokoła schorzałego miejsca ...' [*pathol. veget.* A name of various diseases forming characteristic rings around the infected spot ...]). The TL equivalents were, moreover, not always fitting, particularly from the point of view of translators, since Lilien frequently employed descriptive rather than single-word equivalents (e.g. *gangster* 'St. Zj. członek szajki złodziei, rabusiów, włamywaczy, rzezimieszków, fałszerzy, szantażystów itp ...' [*U.S.* a member of a gang of thieves, robbers, burglars, cutthroats, forgers, blackmailers etc.]). Nevertheless, as Mizwa (1961: v) rightly put it, "the author deserves honorable mention and commendation ... for being the first not only to sense the need of an extensive English–Polish and Polish–English dictionary but also to attempt to do something about it".

The gigantic bilingual material that Lilien collected could not go unnoticed. Indeed, there is some evidence that it was used in the compilation of the English–Polish part of the *Kościuszko Foundation Dictionary* (1959), a renowned lexicographic work compiled in Poland by Kazimierz Bulas, and revised in the United States in collaboration with Francis J. Whitfield and Lawrence L. Thomas, both of them affiliated with the Department of Slavic Languages and Literatures of the California University at Berkeley. Still, neither Bulas nor his American collaborators ever admitted this fact openly.

4. Conclusions

To summarize, the purpose of this paper was to shed some light on Polish–English and English–Polish dictionaries compiled by Polish Americans. Eight such dictionaries, of varying size and targeted at different readership, appeared on the American market between 1899 (S.Z. and W.B.; Antoni A. Paryski) and 1951 (Ernest Lilien). One of them, Waław Jesień's *Słownik angielsko–polski zawierający 4000 najpospoliej używanych słów* (1925), was published in Poland. It is hoped that this historical survey has some important merits as neither the dictionary-makers' biographies (except for Paryski's and Zahajkiewicz's) nor even their names appear in the otherwise comprehensive *Polish American Encyclopedia* (2011) edited by Pula et al.

The attractiveness of a dictionary to the target market is reflected by the number of editions to which it runs. Seen from this perspective, the dictionaries compiled by Polish Americans enjoyed little commercial success. Only four of them went through more than one edition: Maryański's guide-book was republished in Poland by Gebethner & Wolff; Szumkowski published two fur-

ther editions, of which the third is claimed to have been corrected; and *Smulski's Dictionary* had another Chicago edition of 1944, although it was also reprinted, presumably without copyright, in 1945 in Nuremberg by Wydawnictwo Polskie R. Wegnera. Paryski's pocket dictionary alone appeared in new versions; there were at least seven of them, but many more supposedly left the printing presses of the Paryski Publishing Company. It is worth emphasizing that Paryski had the printing infrastructure at his disposal, which he used to his advantage despite the fact that his dictionary was of a low quality. Given his "intelligence, determination, and energy" (Jaroszyńska-Kirchmann 2015: 204), it remains a moot point why he never made an effort to revise and improve it.

The authors described here were people from different walks of life, with different aspirations and different career paths, but none of them was a professional lexicographer. This may explain why their dictionaries — with the single exception of *Lilien's Dictionary* — were unambitious. There is every indication that they turned to practical lexicography to meet the exigencies of the "hyphenated" Polish-American market (see, e.g., Erdmans 2013: 225), cashing in on their knowledge of English which surpassed their fellow countrymen's. In spite of their limited experience in practical lexicography, they tailored their dictionaries, in one way or another, to target users' envisaged needs.

The history of lexicography is indicative of what may be called knowledge development. In a nutshell, establishing TL equivalents for SL headwords requires a network of lexical knowledge that is by no means restricted to a single compiler's cross-linguistic and cross-cultural competence. Polish Americans attempted to modernize and Americanize the network for the sake of Polish immigrants and, as might be expected, some did it more skilfully than others. The limitations of space require that this paper offer merely a sketch, but the history of bilingual lexicography resulting from my research contains fascinating stories about the compilers and the dictionaries they made.

One might ask what remains a desideratum in the field. There is always room for further research. Firstly, many biographical details are still unknown. Evidently the names of Paryski, Maryński, Szumkowski, Wilde, and Lilien may ring a bell for Polish-Americans, but their biographies are patchy and incomplete. Secondly, a few reference works have not been found, Paryski's illustrated dictionary being a case in point. Thirdly, information on how the dictionaries were used would make for a highly informative addition. Lastly, next to nothing is known about the post-war history of Polish–English and English–Polish lexicography, which is another challenge facing a historian of bilingual lexicography.

Endnotes

1. There are a few other bibliographies of dictionaries, including Collison's (1955), Lewański's (1959) and Wojan's (2013), but they are also far from complete.

2. Over the last two years after the publication of the book, a number of additional details enriching the overall picture came to light. They have been diligently collected and shown in this article.
3. According to Hausmann (qtd. in Hartmann 2007: 209), the history of bilingual lexicography has not received much attention. The third volume of *An International Encyclopedia of Lexicography* edited by Hausmann et al. (1991) might suggest otherwise, as it includes a number of chapters dedicated to bilingual dictionaries for different languages pairs. It should be noted, however, that this refers primarily to the world's major languages, such as English, German, French, Italian, Chinese, and Arabic.
4. This distinction requires a word of comment. A new edition is usually seen as different from the original one (cf. Hartmann and James 2001: 47). As my research shows, however, it was not the case with bilingual dictionaries of the past because subsequent reprints were often publicized as "new editions", even though they underwent no revision. This also concerns dictionaries issued by a range of publishers, of which Franciszek Bauer-Czarnomski's Polish-English and English-Polish dictionaries may be a case in point (Podhajecka 2016a: 254).
5. The first wave of Polish political immigration in the nineteenth century was followed by a flood of economic migrants desperate to find social and financial stability for themselves and their families. Most publishers of the Polish-English and English-Polish dictionaries, i.e. Władysław Dyniewicz, Władysław Smulski, the Polish American Publishing Company (Dyniewicz and Smulski, from 1929 Helen and John J. Chrzanowski), the Polish American Book Company (Helen Chrzanowski) and A. A. Paryski, were of Polish descent. This might indicate that providing the Polish community with bilingual dictionaries was some kind of patriotic duty, but it was clearly expected to yield a profit. In the following years, it was not only ethnic publishers (e.g. David McKay) who came to realize that the bilingual dictionary business had the potential to be quite lucrative.
6. Rykaczewski's work had only been preceded by Krystyn Lach-Szyrma's monoscopal English-Polish dictionary (1828).
7. The catalogue of the Wisconsin-Madison Libraries suggests that the copy of the dictionary in the Libraries' possession might have been published between 1950 and 1959, whereas the metadata of the University of California Berkeley Libraries indicates that their copy appeared in 1954. I contacted both libraries with an inquiry, but it seems that the dates are approximate. The only way to determine them is to search methodically through the Polish-language newspapers published in Chicago between 1945 and 1960, which remains a daunting task.
8. In 1891, Kierst was a student at Warsaw University when he was arrested by the Russian police and sentenced to two and a half years in prison and a three-year-exile to Russia. Intriguingly, he must have compiled the dictionary during these years. For a former political prisoner to attract a publisher, however, would have been absolutely impossible, which prompted Kierst to turn to Oskar Callier, a secondary school teacher with some experience in practical lexicography.
9. It is worthy of mention that, between 1895 and 1926, Kierst made an attempt to compile another dictionary. Entitled *Dokładny słownik angielsko-polski i polsko-angielski w dwóch częściach z wymową wyrazów angielskich według najnowszych źródeł opracowany ...* (1915-16), it was put out by Warsaw's Księgarnia Mazowiecka, but the publication was discontinued.
10. The same dictionary was later published by J. Lorenz, a Moravian publisher, who, in 1929,

had commissioned Stanisławski to compile a large pocket dictionary embracing 45 printed sheets.

11. It is important to mention that Stanisławski was educated in England from an early age. He graduated from St. Michael's College in Hitchin, Herefordshire, in 1910 and was practically a native speaker of English.
12. This article is dedicated to the dictionary-making activity undertaken by, and aimed at, an immigrant community in the United States. Since Poles constituted one of many ethnic groups there, it would be worthwhile studying the topic from a comparative perspective, as one of the reviewers rightly suggests. This, however, would exceed the scope of this article, opening a broad research area with new challenges.
13. This also concerns thematic dictionaries. After World War II, for example, thousands of Polish soldiers and civilians remained in the United Kingdom, where they had to find ways to make a living. A number of bilingual thematic dictionaries appeared at that time, including *Angielsko-polski słownik spawalnicy ...* (1946) by Moszoro, *Słownik ślusarza angielsko-polski* (1946) and *Słownik betoniarza i zbrojarza polsko-angielski i angielsko-polski* (1947), to help them acquire specialist English vocabulary (see also Łukasik 2017).
14. Here, Micklethwait is referring to monolingual dictionaries, but bilingual ones were equally, if not more, indispensable.
15. In 1840, Sobolewski was himself engaged in compiling an English–Polish dictionary. He never saw it to fruition, however, having been discouraged by Count Adam Jerzy Czartoryski, whom he had approached for advice and help (Podhajecka 2016b: 330). Sobolewski's manuscript, extant in the alphabet range B–E, is found in the holdings of the Polish Library in Paris.
16. Despite many efforts undertaken by American Polonia to save its ancestors from oblivion, very little is known of most immigrants, particularly those whose names rarely appeared in the ethnic press.
17. Paryski was supposedly the author of a practical manual for his salesmen, *Katechizm dla agenta oświatowego Wydawnictwa Ameryki-Echa*, whom he called "educational agents" as their task consisted in the "dissemination of learning" [M.P.] (19--: 9). This "catechism" is organized on a question and answer basis, providing the agents with information on how, and why, to carry out their work. For example, in Section 6: The relationship between the agents and the Publishing Company, we read: "4. What is the next step in the agent's business career? Permitting accomplished agents, as shareholders, to benefit from the Company's profits. 5. Are these good promotion prospects? Extremely good, as the Company, developing thanks to the agents' fair and conscientious work, will soon have a turnover in millions [of dollars]" [M.P.] (19--: 17).
18. Only two copies of Paryski's pocket dictionary are available today: one in the Library of the Ossoliński National Foundation in Wrocław, and the other in the Sterling Memorial Library of Yale University in New Haven, Connecticut.
19. *Practical Phonics*, modelled on "the orthoepy of Webster's Dictionary", was published in America by Esmond de Graff, an established teacher and educator.
20. The system was not originally Webster's, as Micklethwait (2005: 134) shows. Indeed, the *Royal Standard English Dictionary* (1788) by William Perry may well be treated as a parent dictionary in this respect.

21. The use of "alphabet fatigue", a feature of large dictionary projects (Coleman 2008: 68), in relation to small bilingual dictionaries may seem bizarre, but there is no better term. It is clear that other dictionary-makers were also determined to finish their works as quickly as possible. Paryski must have been exhausted after the compilation of the English–Polish part, which led him, firstly, to reduce the number of headwords (9,587 as opposed to 14,527 in the English–Polish volume) and, secondly, to make a rigid selection of equivalents in the Polish–English part.
22. The guide-book was entered at the Office of the Register of Copyrights in Washington, D.C., in 1905, but was published in 1906. This is confirmed by Maryański himself; on the back cover of his book, he explains that the print was delayed for reasons beyond his control.
23. Franciszek Chłapowski, who would later become professor of medicine at Poznań University and chair of the Poznań Association of the Friends of Sciences, is a reliable source of information.
24. This may obviously be a pure coincidence, but, in 1895, the famous Polish writer and seaman Joseph Conrad "seems to have invested almost all of his money in South African goldmines" and lost it soon afterwards (Hampson 2012: 125). Dealings in gold shares would have been quite relevant, as Maryański managed gold mines in the United States, where he risked "his hard won fortune" (Paszkowski 1987: 304), and in Australia, where he once again decided to risk both his capital and good reputation (see Paszkowski 1987: 305-309).
25. The advertisement is available from the following website: [http://www.fultonhistory.com/Process%20small/Newspapers/Newspapers%20%20Out%20of%20NY/Utica%20NY%20Slowo%20Polskie%20\(The%20Polish%20Word\)/Utica%20NY%20Slowo%20Polskie%20\(The%20Polish%20Word\)%201913.pdf](http://www.fultonhistory.com/Process%20small/Newspapers/Newspapers%20%20Out%20of%20NY/Utica%20NY%20Slowo%20Polskie%20(The%20Polish%20Word)/Utica%20NY%20Slowo%20Polskie%20(The%20Polish%20Word)%201913.pdf) (accessed 29 September 2018).
26. According to Szumkowski's estimates, his dictionary included 12,000 headwords in the Polish–English part and 18,000 in the English–Polish part.
27. Rejall selected 4,000 words, but Jesień introduced some changes to the list, adding a handful of words (e.g. *die 'sztanca, kostka do gry, l.mn. dice (dajs)*), at the same time deleting others (e.g. *dime*).
28. Of course, this term should not be understood as equivalent with present-day "learner's dictionary" (see Hartmann and James 2001: 82-83), which includes features facilitating production in the TL.
29. A bilingual dictionary for inexperienced learners of English would call, whenever possible, for single-word equivalents rather than paraphrases. Taking this into account, *surprise* could be fairly effectively paired with three Polish equivalents: 'niespodzianka', 'zaskoczenie' and 'zdziwienie'.
30. The only reference to this edition, which is not available in any library, is found in Bilikiewicz-Blanc et al. (1991: 132).
31. Smulski was considered one of the "most brilliant and far-sighted" Polish Americans (Słownik Liliena ..., 1944: 3). Being aware of the importance of a modern Polish–English and English–Polish dictionary, he commissioned the task of compiling such a reference work to two of his talented collaborators, Henryk Setmajer and T.M. Wilde, paying them good salaries. The result of his bold initiative was, however, only the small volume under analysis here.

32. With 600,000 headwords, *Webster's Second* was the largest monolingual English dictionary until the beginning of the twenty-first century and was regarded by users, even after the third edition appeared in 1961, as "the dictionary *par excellence*" (Landau 2001: 86). This may have resulted, to some extent, from the fact that the editor of *Webster's Third* took a less prescriptive approach to the compilation process, including substandard forms such as *ain't*, which sparked a wave of criticism throughout the United States (for which see Morton 1994).
33. Mizwa (1961: v) calls Lilien "a lawyer by training and lexicographer by avocation", but his legal qualifications are difficult to confirm.
34. In fact, Lilien gained wide experience in the press because, between 1887 and 1907, he was a co-owner of *Kurier Lwowski*, a popular weekly serving the city of Lvov (Mazurek 2006: 196).
35. He was, among other things, a member of the Linguistic Society of America (Proceedings of the Linguistic Society ..., 1953: 7), a corresponding member of the Polish Institute of Arts and Sciences in America (Pawlikowski et al. 1945: 428), a board member of Polish People's University (*Pamiętnik dwudziestopięcioletniego jubileuszu ...*, 1933: 19), and one of the sponsors of *Poles in America* (Tomczak 1933: 6). He also lectured on both linguistic and literary subjects (e.g. Lednicki 1976: 363-364).
36. Julius Chylinski was diocesan dean and pastor of St. Peter's parish in Stevens Point (Bolek 1943: 73). As issues of *Stevens Point Journal* (<https://www.newspapers.com>) indicate, he actively engaged in a range of enterprises, but, somewhat surprisingly, no information of St. Peter's Foundation is available today.
37. Mizwa was probing whether or not the Kosciuszko Foundation should invest money in Lilien's project. The support, it seems, would have been more generous if the dictionary had received a positive recommendation from Prof. George R. Noyes of the University of California (letter to Mizwa of 13 November 1945).
38. The letter of 17 February 1949 concerning *Lilien's Dictionary* was written by Czeslaw Milosz, the 1980 Nobel Prize winner for literature, the then cultural attaché of the Embassy in Washington. This is no mere coincidence, as Franaszek (2017: 257-258) explains. He makes it clear that Milosz recognized the full worth of Lilien's undertaking and went to great lengths to find some support for it: "I worked out a detailed plan for providing support to Ernest Lilien, the author and editor of a monumental Polish-English dictionary ... I can only express my disappointment and regret that although the project was supported by the Embassy in Poland, it has not met with approval (Embassy Report, August-September 1946)".
39. Lilien's correspondence provides a useful clue as to which edition of *Webster's Second* he had at his disposal: it was the 1944 edition with "A Pronouncing Biographical Dictionary" appended to it. Lilien was flabbergasted to see Copernicus' name in it spelled erroneously as *Kopernicki* (letter to Mizwa of 12 September 1946).
40. In his letter to Mizwa of 29 April 1945, Lilien states that he resorted to the 1896 edition of Funk and Wagnalls *A Standard Dictionary ...*, even though an updated version titled *A New Standard Dictionary ...* (1916) would have been a much better source. Yet, this dictionary was used less sparsely than *Webster's Second*.
41. OED3 explains that the word *Hottentot* is considered both archaic and offensive, so it is usually replaced by the word *Khoekhoe*. *Bushman*, denoting a member of an aboriginal people in Southern Africa, apparently comes from Dutch *boschjesman* applied by the Dutch colonists in South Africa to the Khoisan peoples living in the "bush".

References

Dictionaries

- Arct, M.** 1916. *Słownik ilustrowany języka polskiego*. Vols. 1–3. Warsaw: Wydawnictwo M. Arcta.
- Bulas K. and F.J. Whitfield.** 1959. *The Kościuszko Foundation Dictionary: English–Polish, Polish–English*. Vol. 1: *English–Polish*. The Hague: Mouton.
- Bulas K., F.J. Whitfield and L.L. Thomas.** 1961. *The Kościuszko Foundation Dictionary: English–Polish, Polish–English*. Vol. 2: *Polish–English*. The Hague: Mouton.
- Choźko, A.** 1870. *Dokładny słownik polsko–angielski, czerpany z najlepszych źródeł krajowych i obcych ...* (Part 1). *A Complete Dictionary, English and Polish* (Part 2). Berlin: Neufeld & Henius.
- Choźko, A.** 1890. *Chodzki Alexandra Dokładny słownik polsko–angielski i angielsko–polski, czerpany z najlepszych źródeł krajowych i obcych ...* (Part 1). *Alex. Chodzko's A Complete Dictionary English and Polish and Polish and English. Compiled from the Dictionaries ...* (Part 2). Chicago, IL: Drukiem i nakładem W. Dyniewiczza.
- Choźko, A.** c. 1950 [?]. *Chodzki Alexandra Dokładny słownik polsko–angielski, czerpany z najlepszych źródeł krajowych i obcych: a mianowicie ze słowników polskich Lindego, Mrongoviusa i Ropelewskiego* (Part 1). *Alex Chodzko's A Complete Dictionary, English and Polish, Compiled from the Dictionaries of Johnson, Webster, Walker, Fleming and Tibbins, etc.* (Part 2). Chicago: Polish American Publishing Co.
- Flügel, J.G.** 1830. *A Complete Dictionary of the English and German and German and English Languages, Containing All the Words in General Use*. Vols. 1–2. Leipsic: Printed for A.G. Liebeskind.
- Funk and Wagnalls New Standard Dictionary of the English Language upon Original Plans ... Complete in One Volume.* 1916. New York/London: Funk and Wagnalls.
- Hartmann, R.R.K. and G. James.** 2001. *Dictionary of Lexicography*. London/New York: Routledge.
- Kierst, W.** 1915–1916. *Dokładny słownik angielsko–polski i polsko–angielski w dwóch częściach z wymową wyrazów angielskich według najnowszych źródeł opracowany ...* [Instalments 1–6]. Warsaw: Księgarnia Mazowiecka.
- Kierst, W.** 1984. *A Dictionary English–Polish and Polish–English / Trzaski, Everta i Michalskiego słownik angielsko–polski i polsko–angielski. Część pierwsza angielsko–polska* (Vol. 1). *Second part: Polish–English* (Vol. 2). New York: Saphrograph.
- Kierst, W. and O. Callier.** 1961. *English–Polish and Polish–English Dictionary / Słownik języka polskiego i angielskiego*. Chicago: Drukiem "Dziennika Związkowego".
- Lach-Szyrma, K.** 1828. *Słownik angielsko–polski ułożony przez K. L-S. dla użytku młodzieży Instytutu Politechnicznego*. Warsaw: Drukarnia Gałęzowskiego i Komp.
- Moszoro, K.W.** 1946. *Angielsko–polski słownik spawalniczy z omówieniem stosowanych lub przyjętych terminów oraz rysunkami*. London: Wojskowy Instytut Techniczny.
- Neilson, W.A., T.A. Knott and P.W. Carhart (Eds.).** 1934. *Webster's New International Dictionary of the English language*. 2nd edition. Springfield, MA: Merriam-Webster.
- OED3 = Simpson, J., M. Proffitt et al. (Eds.).** 2000–. *The Oxford English Dictionary*. 3rd edition. Oxford: Oxford University Press. <http://www.oed.com/>.
- Ogilvie, J.** 1859–1860. *The Imperial Dictionary of the English Language: A Complete Encyclopedic Lexicon, Literary, Scientific, and Technological*. Vols. 1–3. London: Blackie and Son.

- Perry, W.** 1788. *The Royal Standard English Dictionary, in Which the Words Are Not Only Rationally Divided into Syllables, Accurately Accented, Their Part of Speech Properly Distinguished, and Their Various Significations Arranged in One Line, but Likewise ...* London: Printed for, and sold by J. Murray ... and J. Bell and J. Dickson.
- Słownik warszawski* [Warsaw Dictionary] = **Karłowicz, J., A.A. Kryński and W. Niedźwiedzki (Eds.)**. 1900–1927. *Słownik języka polskiego*. Vols. 1–8. Warsaw: Nakładem prenumeratorów i Kasy im. Mianowskiego [and others].
- Słownik betoniarza i zbrojarza polsko–angielski i angielsko–polski*. 1947. London: Inspektorat Szkolenia P.K.P.R.
- Słownik ślusarza angielsko–polski*. Cz. 1: *Narzędzia i materiały*. Cz. 2: *Czynności ślusarskie*. 1946. [Great Britain]: Biuro Badań Technicznych Saperów.
- Stanisławski, J.** 1933. *An English–Polish and Polish–English Pocket–Dictionary / Słowniczek angielsko–polski i polsko–angielski*. Třebíč, Moravia: J. Lorenz.
- Stanisławski, J.** 1993. *McKay’s English–Polish and Polish–English Pocket–Dictionary*. [Montreal]: Tormont.
- Webster, N., W.G. Webster and W.A. Wheeler.** 1871. *A Primary School Dictionary of the English Language, Explanatory, Pronouncing and Synonymous. With an Appendix Containing Various Useful Tables*. New York/Chicago: Ivison, Blakeman, Taylor & Co.
- Webster’s Collegiate Dictionary* (1919). 3rd edition. Springfield, MA: G. & C. Merriam Co.
- Webster’s Second* = **Neilson, W.A., T.A. Knott and P.W. Carhart (Eds.)**. 1934. *Webster’s New International Dictionary of the English Language*. 2nd edition. Springfield, MA: Merriam-Webster.
- Whitney, W.D. and B.E. Smith (Eds.)**. 1895. *The Century Dictionary: An Encyclopedic Dictionary of the English Language*. Vols. 10. New York: The Century.

Other references

- Adamska-Sałaciak, A.** 2014. Bilingual Lexicography: Translation Dictionaries. Hanks, P. and G.-M. de Schryver (Eds.). 2014. *International Handbook of Modern Lexis and Lexicography*: 1-11. Berlin: Springer.
- Antonowicz, J.** 1788. *Grammatyka dla Polaków chcących się uczyć angielskiego języka krótko zebrana przez ... Bazyliana prowincyi litewskiéy za pozwoleniem zwierzchności pierwszy raz pod prasę oddana*. Warsaw: W Drukarni Nadworney J.K. Mci i P.K. Edu.
- Barry, A.** 2008. Reading the Past: Historical Antecedents to Contemporary Reading Methods and Materials. *Reading Horizons* 49(1): 31-52.
- Bilikiewicz-Blanc, D. et al.** 1991. *Polonika zagraniczne: bibliografia za okres od września 1939 do 1955 roku*. Vol. 3: R–Ż. Warsaw: Biblioteka Narodowa.
- Bolek, F. (Ed.)**. 1943. *Who’s Who in Polish America*. A Bibliographical Directory of Polish-American Leaders and Distinguished Poles Resident in the Americas. 3rd edition. New York: Harbinger House.
- Boyer, A.** 1729. *The Compleat French–Master for Ladies and Gentlemen. Containing I. A New Methodical French Grammar. II. A Well Digested and Copious Vocabulary. III. Familiar Phrases and Dialogues, on All Manner of Subjects ...* London: Printed for Samuel Ballard et al.
- Burroughs, E.R.** 1922. *Tarzan wśród małp*. Translated from English by Władysław Kierst. Warsaw: Trzaska, Evert i Michalski.

- Chłapowski, F.** 1884. Truskawiec: we wschodniej Galicyi. *Kłosa: Czasopismo Ilustrowane Tygodniowe* of 30 August–11 September 1884, 39(1002): 171-173. <http://www.wbc.poznan.pl/dlibra/publication?id=170061&tab=3> (accessed 29 September 2018).
- Coleman, J.** 2008. *A History of Cant and Slang Dictionaries*. Volume III: 1859–1936. Oxford: Oxford University Press.
- Collison, R.L.** 1955. *Dictionaries of Foreign Languages: A Bibliographical Guide to the General and Technical Dictionaries of the Chief Foreign Languages, with Historical and Explanatory Notes and References*. London: Hafner Publishing.
- Czartoryski-Sziler, P.** (n.d.) Wielcy zapomniani: Władysław Bielza — wielki piewca polskości. *Nasz Dziennik*. <http://www.lwow.home.pl/naszdziennik/belza2.html> (accessed 29 September 2018).
- Emans, R.** 1968. History of Phonics. *Elementary English* 45(5): 602-608.
- Erdmans, M.P.** 2013. Acculturation and Persistence of the Polish American Community in Connecticut, 1870–2010. Mazurkiewicz, A. (Ed.). 2013. *East Central Europe in Exile*. Volume 1: *Transatlantic Migrations*: 217-234. Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Franaszek, A.** 2017. *A Biography: Miłosz*. Edited and translated by Aleksandra and Michael Parker. Cambridge, MA/London: Belknap Press of Harvard University Press.
- Graff, E.V. de.** 1881. *Practical Phonics. A Comprehensive Study of Pronunciation, Forming a Complete Guide to the Study of the Elementary Sounds of the English Language, and Containing Three Thousand Words of Difficult Pronunciation, with Diacritical Marks According to Webster's Dictionary*. Syracuse, NY: C.W. Bardeen Publisher.
- Grzegorzczak, P.** 1967. *Index lexicorum Poloniae*. Warsaw: Państwowe Wydawnictwo Naukowe.
- Hargraves, O.** 2011. *Culture Shock! Chicago. A Survival Guide to Customs and Etiquette*. Tarrytown, NY: Marshall Cavendish Editions.
- Hartmann, Reinhard R.K.** 2007. *Interlingual Lexicography. Selected Essays on Translation Equivalence, Contrastive Linguistics and the Bilingual Dictionary*. Tübingen: Max Niemeyer.
- Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.)**. 1991. *Wörterbücher. Ein internationales Handbuch zur Lexikographie / Dictionaries. An International Encyclopedia of Lexicography / Dictionnaires. Encyclopédie internationale de lexicographie*. Vol. 3. Berlin/New York: Walter de Gruyter.
- Jaroszyńska-Kirchmann, A.** 2015. *The Polish Hearst: Ameryka-Echo and the Public Role of the Immigrant Press*. Urbana: University of Illinois Press.
- Kantowicz, E.R.** 1975. *Polish-American Politics in Chicago, 1880–1940*. Chicago: University of Chicago Press.
- Kąsinowski, B.** 1913. Władysław Belza *1847 +1913. *Literatura i Sztuka. Dodatek do Dziennika Poznańskiego* 5(6): 81-83.
- Landau, S.** 2001. *Dictionaries. The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.
- Lednicki, W.** 1976. *Adam Mickiewicz in World Literature. A Symposium*. Westport, CT: Greenwood Press.
- Lewandowski, I.** 1992. Rykaczewski Erazm Edward (1803–1873). *Polski Słownik Biograficzny* 33(3): 472-474.
- Lewański, R.C.** 1959. *A Bibliography of Slavic Dictionaries: With a Supplement of Lusatian and Polabian Dictionaries*. Vol. 1: Polish. New York: New York Public Library.

- Lilien, E.** 1944. Preface. *Lilien's Dictionary. Part I: English–Polish Lilien's Dictionary / Ernesta Liliena słownik. Cz. 1: Angielsko–polski*. Buffalo: Drukiem Dziennika dla Wszystkich.
- Lokański, H.** 1920. *Sześć lat wojny Polskiej a Polacy w Ameryce*. Chicago: [no publ.].
- Łukasik, M.** 2017. Polish Specialised Lexicography during WWII. Lipczuk, R., M. Lisiecka-Czop, K. H. Ramers (Eds.). 2017. *Sprache und Wörterbücher in Theorie und Praxis. Lexikografische und text-linguistische Fragestellungen*: 105-124. Hamburg: Dr Kovač.
- Maciejewski, J. and Z. Szweykowski (Eds.)**. 1982. *Bibliografia literatury polskiej Nowy Korbut. Literatura pozytywizmu i Młodej Polski*. Vol. 16. Warsaw: Państwowy Instytut Wydawniczy.
- Majewski, K.** 2003. *Traitors and True Poles. Narrating a Polish-American Identity 1880–1939*. Athens, OH: Ohio University Press.
- Maryański, M.** 1882. Kopalnie Truskawieckie. *Gazeta Lwowska* of 22 July 1882, 72(166): 3-4. <https://jbc.bj.uj.edu.pl/dlibra/publication/54184/edition/48401/content?ref=desc> (accessed 29 September 2018).
- Mazurek, J.** 2006. *Kraj a emigracja. Ruch ludowy wobec wychodźstwa chłopskiego do krajów Ameryki Łacińskiej*. Warsaw: Instytut Studiów Iberyjskich i Iberoamerykańskich.
- Micklethwait, D.** 2005. *Noah Webster and the American Dictionary*. Jefferson, NC/London: McFarland & Company.
- Mizwa, S.P.** 1961. Preface. Bulas, Kazimierz and Francis J. Whitfield (Eds.). 1961. *The Kościuszko Foundation Dictionary English–Polish Polish–English. Volume 1: English–Polish*. The Hague: Mouton & Co.
- Morton, H.C.** 1994. *The Story of Webster's Third: Phillip Gove's Controversial Dictionary and Its Critics*. New York/Cambridge: Cambridge University Press.
- [Obituary of] Ernest Lilien. 1952. *The Milwaukee Journal* of 16 June 1952: 23. <https://news.google.com/newspapers?nid=1499&dat=19520616&id=GQ8iAAAAIBAJ&sjid=XH4EAAAAIBAJ&pg=6843,128624&hl=en> (accessed 29 September 2018).
- Osselton, N.E.** 2007. Alphabet Fatigue and Compiling Consistency in Early English Dictionaries. Considine, J. and G. Iamartino (Eds.). 2007. *Words and Dictionaries from the British Isles in Historical Perspective*: 81-91. Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Pamiętnik dwudziestopięcioletniego jubileuszu Polskiego Uniwersytetu Ludowego w Stanach Zjednoczonych / 25th Jubilee Memorial Record of the Polish People's University in the U.S.* 1933. Chicago, ILL: [no publ.].
- Paryski, A.A.** 19--. *Katechizm dla agenta oświatowego wydawnictwa Ameryki-Echa*. Toledo, OH: Wydawnictwo Ameryki-Echa.
- Paszkowski, L.** 1987. *Poles in Australia and Oceania 1790–1940*. Sydney: Australian National University Press.
- Paszkowski, L.** 2008. *Polacy w Australii i Oceanii 1790–1940*. Toruń/Melbourne: Towarzystwo Przyjaciół Archiwum Emigracji.
- Pawlikowski, J. et al.** 1945. General Development of the Institute. *Bulletin of the Polish Institute of Arts and Sciences in America* 3(3/4): 422-453.
- Podhajecka, M.** 2016a. *A History of Polish–English / English–Polish Bilingual Lexicography (1788–1947)*. Opole: Wydawnictwo Uniwersytetu Opolskiego.
- Podhajecka, M.** 2016b. Szkic z dziejów leksykografii dwujęzycznej: Paweł Sobolewski i jego słownik angielsko–polski ... (1840). *Prace Filologiczne* 68: 323-344.

- Proceedings of the Linguistic Society of America at the Twenty-Seventh Annual Meeting Cambridge 28–29 December 1952. *Bulletin* 26, 1953. *Language: Journal of the Linguistic Society of America* 29(2): 6-20.
- Pula, J. et al. (Eds.).** 2011. *The Polish American Encyclopedia*. Jefferson, NC/London: McFarland & Company.
- Scherer, P.** 1946. Review of: *Lilien's Dictionary. Part I: English–Polish*, fascicles 1–6 by Ernest Lilien. *Language* 22(3): 265-266.
- Stinchfield-Hawk, S.** 1928. *The Psychology of Speech*. Boston: Expression.
- Tomczak, A.C. (Ed.).** 1933. *Poles in America. Their Contribution to a Century of Progress*. A Commemorative Souvenir Book Compiled and Published on the Occasion of the Polish Week of Hospitality, July 17 to 23. Chicago, IL: Polish Day Association.
- Wojan, K.** 2013. *Język angielski w polskiej leksykografii. Słowniki przekładowe lingwistyczne i encyklopedyczne wydane w latach 1782–2012*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Archival materials

- Chojnacki, W.** (n.d.) Antoni A. Paryski (1864–1935) — wydawca polski w Ameryce. The Archives of the Polish Museum of America in Chicago, manuscript.
- Lilien, E.** to S. Mizwa. Letter of 8 June 1944. The Archives of the Kosciuszko Foundation, XVII.10. Ernest Lilien.
- Lilien, E.** to S. Mizwa. Letter of 29 April 1945. The Archives of the Kosciuszko Foundation, XVII.10. Ernest Lilien.
- Lilien, E.** to S. Mizwa. Letter of 12 September 1946. The Archives of the Kosciuszko Foundation, XVII.10. Ernest Lilien.
- Miłosz, C.** to Towarzystwo Uniwersytetów Robotniczych [Society of People's Universities]. Letter of 17 February 1949. The Archives of Modern Records in Warsaw, No. 566/66/49.
- Noyes, G.R.** to S. Mizwa. Letter of 13 November 1945. The Archives of the Kosciuszko Foundation, XVII.10. Ernest Lilien.
- Słownik Liliena — cenne dzieło naukowe na wychodźstwie. 1944. *Dziennik dla Wszystkich* of 30 January 1944: 3. The Archives of the Kosciuszko Foundation, XVII.10. Ernest Lilien.
- Sobolewski, P.** [1840]. *English and Polish Dictionary, Containing All Words and Phrases of General Use (With the Pronunciation of Every English Word According to Walker) to Which is Also Added a Complete Table of Irregular Verbs of the English Language / Słownik Angielsko–Polski, zawierający w sobie wszystkie słowa i frazesy w powszechnym używaniu ... (B–E)*. The Polish Library in Paris, manuscript.

Semi-automating the Reading Programme for a Historical Dictionary Project

Tim van Niekerk, *Dictionary Unit for South African English, Rhodes University, Grahamstown, South Africa (dsae@ru.ac.za)*

Johannes Schäfer, *Department of Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany (johannes.schaefer@uni-hildesheim.de)*

and

Ulrich Heid, *Department of Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany (heidul@uni-hildesheim.de)*

Abstract: This paper describes the resources and software procedures used or developed in a major enabling step towards the revision of the scholarly reference work *A Dictionary of South African English on Historical Principles* (DSAE, Silva et al. 1996), namely the semi-automatic generation of a digitally-sourced lexical database on which new and updated dictionary entries will be based; as well as the addition, in parallel, of a new corpus of South African English (SAE) to the project. Drawing on online data sources and an extensive list of known SAE word forms, we have developed a software toolchain to gather, encode, annotate and collate textual sources, producing: (i) a 3.1-billion part-of-speech-annotated corpus of South African English; (ii) a lexical database of illustrative quotations for over 20,000 known SAE word forms, available for selection at the entry-revision stage; and (iii) a list of potential new variant spellings and headword inclusion candidates. These steps replace, where recent electronic sources are concerned, the mechanical aspects of quotation gathering, normally undertaken manually through a reading programme requiring years of teamwork to acquire sufficient coverage (cf. Hicks 2010).

Keywords: CORPORA, DICTIONARY WORKFLOWS, HISTORICAL LEXICOGRAPHY, LANGUAGE VARIETIES, LEXICAL DATABASES, READING PROGRAMMES, SOUTH AFRICAN ENGLISH

Opsomming: Die semi-outomatisering van die leesprogramme van 'n historiese woordeboekprojek. Hierdie artikel beskryf die hulpbronne en sagtewareprosedures wat gebruik word of ontwikkel is in 'n belangrike bemagtigingstap na die hersiening van die vakkundige naslaanwerk *A Dictionary of South African English on Historical Principles* (DSAE, Silva et al. 1996), naamlik die semi-outomatiese generering van 'n leksikale databasis van digitale bronne waarop nuwe en bygewerkte woordeboekinskrywings gebaseer sal wees; asook die gelyktydige toevoeging van 'n nuwe korpus van Suid-Afrikaanse Engels (SAE) tot die projek. Gebaseer op aanlyn data-

bronne en 'n uitgebreide lys bekende SAE woordvorme, het ons 'n sagteware nutsketting ontwerp vir die versameling, enkodering, annotering en vergelyking van teksbronne, wat gelei het tot die skep van (i) 'n 3.1-biljoen woordsoortgeannoteerde korpus van Suid-Afrikaanse Engels; (ii) 'n leksikale databasis van illustratiewe aanhalings vir ongeveer 20,000 bekende SAE-woordvorme, wat by die hersieningsfase van die inskrywings beskikbaar is vir seleksie; en (iii) 'n lys van potensieel nuwe variante spellings en moontlikhede vir trefwoordseleksie. Wat onlangse elektroniese bronne betref, vervang hierdie stappe die meganiese aspekte van die versameling van aanhalings, wat gewoonlik met die hand met behulp van 'n leesprogram wat jare se spanwerk vereis om voldoende dekking te verkry, gedoen word (cf. Hicks 2010).

Sleutelwoorde: KORPORA, WOORDEBOEKWERKSVLOEI, HISTORIESE LEKSIKOGRAFIE, TAALVARIËTEITE, LEKSIKALE DATABASISSE, LEESPROGRAMME, SUID-AFRIKAANSE ENGELS

1. Role of quotations in the dictionary

A Dictionary of South African English on Historical Principles (DSAE), Silva et al. 1996) is a diachronic variety dictionary, first published as a single-volume print dictionary spanning about 800 pages and available as a pilot online edition at <http://dsae.co.za> since 2014. It closely resembles the *Oxford English Dictionary (OED)* in the design of its entries as well as its research processes, but focuses solely on South African English (SAE) from its origins in the late 17th Century onwards. The first edition of the *DSAE* was a long-term project involving three Editors-in-chief and 24 editorial staff and research assistants (excluding volunteer readers) over a period of 25 years. The result was a historical dictionary containing 4 600 main entries documenting about 17 500 word forms including headwords, plural forms, orthographic variants, compounds, phrases and derivatives. Of paramount importance are its evidential quotations (variously named contexts, citations or, informally among project staff, 'quotes'). The quotations, drawn from monographs, periodicals, letters, manuscripts, ephemera and other sources are bibliographically referenced: while "in most other kinds of [monolingual] dictionary, attribution is rare ... historical dictionaries generally provide information about the source and date of the quotation" (Atkins and Rundell 2008: 455). Much of the *DSAE*'s compilation process was therefore directed towards an ongoing reading programme. With the help of numerous volunteer readers, approximately 300,000 index card citations were collected as illustrative evidence for dictionary entries, their sense-divisions as they evolve through time, and nested lemmas. Of these about 45,000 quotations were included in the printed version of the dictionary, resulting in an average of 10 quotations per entry and producing a full running text of about 1,5 million words. The object was akin to the *OED*'s, following the principle that "[t]he dictionary should set forth the life history of each single word" (Willinsky 1994: 225), prompting an empirical methodology "based on the analysis of quotations from many textual sources" which "interprets the meanings of words in relation

to historical evidence of their past usage" (Brewer, 2007: 239). The DSAE's inclusion policy was fundamentally quotation-driven: without 'quots' to present within the dictionary as attestations of usage, the compilers could not draft an entry or sense division. See Figure 1 for an example of the preponderance of citation evidence in a typical entry.

aardvark /ɑ:dfɑ:k/ *n.* Forms: *a.* **aardvaark**, **aardvark**, **aard-varké**, **aard-varken**; *β.* **erdvark**, **erdverk**. Also with initial capital. Pl. -s, -e, or unchanged; (*obs.*) -en. [S. Afr. Du., fr. Du. *aarde*, *erd* earth + *vark* pig. (The modern Afk. form is *erdvark*.)] The ant-eater *Orycteropus afer* of the Orycteropodidae, an insectivorous burrowing mammal of nocturnal habits with a long, tapering muzzle and sparsely-haired body; ANTBEAR, ANT-EATER sense 1; EARTH-HOG; EARTH-PIG. Also *attrib.*

a. 1786 G. FORSTER tr. *A. Sparrman's Voy. to Cape of G.H. I.* 270 The *aard-varken*, or earth-pig, which, probably, is a species of *manis*. 1795 [see ANT-EATER sense 1]. 1827 G. THOMPSON *Trav.* II. 86 The Aardvark is about four feet and a half in length, and occasionally is found to weigh upwards of 100 lbs. It lives entirely upon ants. 1847 J. BARROW *Reflect.* 146 The aard-varké, or earth-hog (the *Myrmecophaga Capensis*), is also very common, undermines the ground, and seldom appears but in the night. 1878 T. J. LUCAS *Camp Life & Sport* 86 In the category of strange creatures to be found in this district, I must not omit the ant-bear, or 'aard-vark' (earth pig), which not only inhabits the frontier, but is spread over all parts of the interior. 1896 R. WALLACE *Farming Indust. of Cape Col.* 68 They [*sc.* termites] are greedily sought after and *devoured* by a large ungainly looking quadruped with a long snout, called the ant-eater or 'aard-vark'. 1901 W. L. SCLATER *Mammals of S. Afr.* II. 220 The aard-vark. use their tails to thump the ground near the ants' nest and so cause a panic within. make an opening in the side of the ant-heap and then collect the ants by means of their sticky tongues. 1929 [see ANT-EATER sense 1]. c1936 S. & E. Afr. *Tr. Bk. & Guide* 1101 Nearly every ant-heap in the karoo has a widely gaping mouth on its Southern side, this point of attack being selected by the *aardvark* either because it is next to the habitation of the queen-ant or because the structure is not baked quite as hard as where it is exposed to the full rays of the sun. 1949 H. C. BOSMAN in L. Abrahams *Uuto Duct* (1963) 149 He was the kind of white man who, if he was your neighbour, would think it funny to lead the Government tax-collector to the aardvark-hole that you were hiding in. 1988 C. & T. STUART *Field Guide to Mammals* 162 The Aardvark resembles no other mammal occurring in southern Africa, with its long pig-like snout, elongated tubular ears, heavily muscled kangaroo-like tail and very powerful, stout legs which terminate in spade-like nails. 1990 J. KNAPPERT *Aquarian Guide to Afr. Mythology* 19 *Aardvark*. In African folklore the aardvark, or ant-bear, has a good name not only because it is unafraid of armies of soldier ants but also because it digs diligently searching for food all night, an example and model for lazy cultivators. 1991 M. NEL in *Personality* 11 Mar. 26 Like the Aardvark's sense of smell for ants, Nick's knowledge of theatre is intuitive.

β. 1796 E. HELME tr. *F. Le Vaillan's New Trav.* III. 392 This ant-bear is called in the colonies *erd-verkan* (earth hog). 1924 [see ANTBEAR]. 1959 L. G. GREEN *These Wonders* 207 That creature of obscure origin, that champion tunneller of the veld, the erdvark or ant-eater. This pig-shaped freak is not rare, but is seldom captured.

Figure 1: Example entry *aardvark* from the print edition of *A Dictionary of South African English on Historical Principles* (Silva et al. 1996) showing quotations from 1786 to 1991, separated by orthographic pattern (*aard-* vs *erd-*spelling)

2. The need for new quotations

Following the publication of the first edition of the print DSAE, after which the lexicography unit's focus shifted to synchronic dictionary projects not requiring citations, quotations continued being collected as part of an ongoing background reading programme, but on a much smaller scale. By 2004, index cards and various intermediate electronic wordprocessing formats had been replaced by an electronic lexical database allowing the capture, editing and annotation of quotation records in XML (eXtensible Markup Language) format for future use. Subsequently, as an early step towards revision of the historical dictionary, a data verification project was initiated to correct transcription errors in the 45,000 quotations used in the DSAE's first edition, also stored in the new data-

base. (For details of this painstaking and resource-intensive process, including additional information about source types and methodological considerations impacting on quotation collection, see Hicks 2010.) By 2017, the electronic database contained only about 9,000 new quotations, however. This small number is deceptive: it also contained a high proportion of new headword candidates (over 2,000). Nevertheless, quotation collection had suffered due to intervening dictionary projects, or the digitisation stages of the *DSAE*, having taken priority. A persistent limiting factor was that a large-scale reading programme requires staff to co-ordinate it, and capturing quotations manually, even with the help of assistants, is a highly labour-intensive task.

Nevertheless, just as the latest *OED* revision dedicated "a vast amount of well-directed energy" towards gathering new quotations (Brewer 2007: 241), so the *DSAE* revision requires increased data holdings of post-1995 citations. This applies not only to quotations for new SAE words not included in the first edition, but equally to new quotations for words already described in it, for several reasons: (1) recent citations for all entries should at least be reviewed, if not always necessarily included, to ensure that entries and their sense divisions are still up-to-date; (2) to support a focused review and potential redrafting of those entries labelled *rare*, *historical*, *obsolete*, *obsolescent* or *nonce* usage based on the limited evidence available at the time of compilation (bearing in mind that at that stage attestations could not be discovered via electronic retrieval systems); and (3) from the point of view of training a new team of lexicographers unfamiliar with the historical entry model and its complex styling policies, it may be preferable to begin by updating existing entries before drafting new ones.

3. Typical quotation-gathering stages

The selection of quotations to be reproduced in dictionary entries at the entry drafting stage will probably always require a human eye, and the current project does not attempt to replace editorial judgement. Most of the preceding stages of quotation gathering are, however, laborious and mechanical, namely:

- (1) accessing and reading (scanning) texts for SAE words
- (2) capturing quotations containing these words
- (3) capturing date and source information
- (4) verifying capture against sources to correct capture errors
- (5) recording the relationships of word forms to parent dictionary entries (e.g. adding IDs, canonical forms and noting orthographic variants)
- (6) anticipating as-yet unknown orthographic variants and repeating 1–5 above on discovery of new illustrative quotations.

In the toolchain described below all these stages are either wholly or partly automated, substantially increasing the dictionary project's quotation holdings,

now drawn from recent corpus sources, while dramatically reducing the labour involved. Additionally, we perform further computational steps to highlight potential new SAE terms within the corpus.

4. Input data sources

In 2009 it was reported that "there is no large corpus to represent South African English" (Pienaar and De Klerk 2009: 356) and, apart from proprietary, unfinished, or very small special-purpose corpora of under 1 million words, no others were available to suit the *DSAE*'s quotation-gathering requirements prior to the current collaboration. Additionally, in order to apply a toolchain to process the corpus, track relationships between word forms and extract quotations, a full dataset is required (rather than, for example, a web interface to a corpus allowing individual searches).

In building the SAE corpus, we draw on two sources of data, a newspaper corpus and a generic web corpus.

4.1 Newspaper Corpus

The newspaper corpus was created for quotation-gathering purposes from a suite of Perl programs¹ customised to crawl seven South African online newspapers between 2015 and 2017. After the resulting articles were fed through a parser to strip HTML markup, a further pre-processing step removed corpus noise such as boilerplate headers and footers unrelated to the article at hand, producing a corpus of about 6,5 million sentences or 100 million tokens. See Table 1 for specific counts across sources.

Although not as large as the web corpus described below, the newspaper corpus on its own provides a source of SAE quotations far exceeding the research data formerly available to the dictionary project. It also preserves contextual information in the original HTML versions of the articles such as embedded author details, when indicated, and typographic features such as italic font. (Italicisation of word forms is sometimes useful as an indicator that the author possibly regards a SAE word as a borrowed form, helping the lexicographer judge assimilation.) These features could not, however, be retained in the corpus-encoded and lexical database versions of the data since corpus encoding and other automatic processing steps required plaintext as input. The toolchain did, however, automatically add links to the source HTML as meta-data, allowing the lexicographer to consult the original source if desired.

4.2 Web Corpus

The second dataset is a generic web corpus generated from .za domain sources by the NLP Group of the Computer Science Department at Leipzig University,

as part of its CURL (Crawling Under-Resourced Languages) project (see Goldhahn et al. 2012). The dataset was supplied already split into individual sentences and it does not distinguish between source types (e.g. newspapers vs blogs). Preprocessing steps such as the removal of HTML markup had already been performed on these data along with sentence segmentation. The order of the sentences is scrambled but each has an accession date and source URL, meeting the minimum requirements for an electronic citation. While a corpus of sentences may not satisfy the needs of linguists requiring more context for written utterances, this format is suitable for the historical dictionary project which requires only brief attestations. The Leipzig/CURL strategy of splitting articles into sentences was likewise adopted with the newspaper corpus described in 4.1 above. This was done for the sake of uniformity across the consolidated SAE corpus, and to simplify other automated processing steps including corpus encoding. The resulting corpus amounts to approximately 150 million sentences or 3 billion tokens, averaging 20 words per sentence. Table 1 below provides specific counts for 2011 through 2014.

Type	Subcorpus source	Number of sentences	Number of tokens
Newspapers:	<i>BusinessLIVE</i>	2,762,984	40,643,811
	<i>Daily Maverick</i>	320,273	7,331,568
	<i>DispatchLive</i>	117,752	2,481,062
	<i>Independent Online</i>	258,598	5,438,759
	<i>SowetanLIVE</i>	1,835,773	20,206,205
	<i>The Citizen</i>	368,182	7,881,382
	<i>TimesLIVE</i>	834,656	15,934,246
<i>Subtotal (Newspapers)</i>		<i>6,498,218</i>	<i>99,917,033</i>
Web (generic):	.za Domains 2011	3,870,783	74,114,784
	.za Domains 2012	2,784,879	53,248,634
	.za Domains 2013	50,191,936	1,031,432,748
	.za Domains 2014	91,728,781	1,823,257,689
<i>Subtotal (Web)</i>		<i>148,576,379</i>	<i>2,982,053,855</i>
SAE Corpus Totals		161,572,815	3,081,970,888

Table 1: Sentence and token counts for the Newspaper and Web subcorpora of the SAE corpus

5. Toolchain and its output

The overall toolchain is illustrated in Figure 2. Having described the input data sources, processing stages and the resulting tools and datasets are elaborated below.

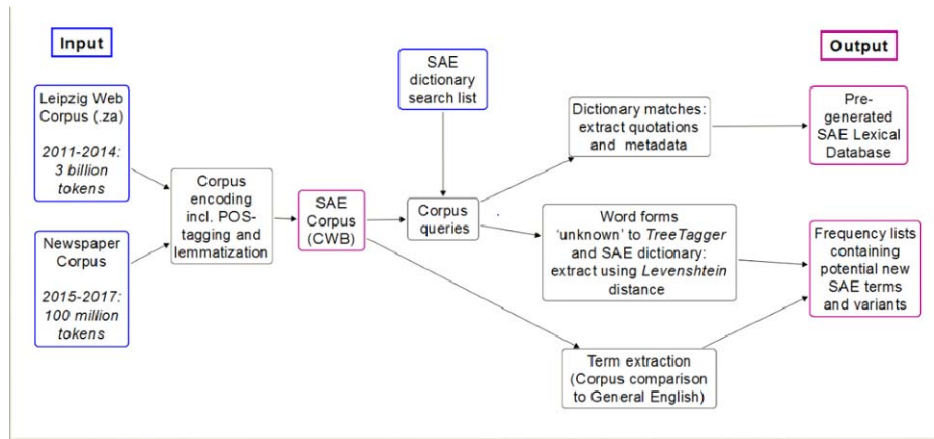


Figure 2: The full software toolchain showing the main inputs, outputs and processing stages

5.1 Annotated corpus and corpus query system

The toolchain introduces a corpus querying system to the *DSAE*'s set of research tools. Previously the project used a lexically and bibliographically annotated but comparatively miniscule lexical database of individually-captured quotations, or else Internet search engines. The lexical database, since it only contains quotations already captured, does not allow the discovery of new words or new senses of general English not yet recorded, and neither does this database nor general-purpose Internet searches allow queries according to linguistic attributes of word forms. Additionally, Internet sources are ephemeral, and continued access to quotations could previously only be assured by capturing them manually in the lexical database.

The SAE corpus solves these basic problems by providing a snapshot of the Internet across .za domains in a linguistically-annotated dataset that remains immutable even if the source web pages become inaccessible. In the preparatory stages, the dataset was part-of-speech-tagged (POS-tagged) and lemmatised using the *TreeTagger*², and loaded into the IMS Open Corpus Workbench (CWB)³. Past experiments at the lexicography unit showed that other concordancers like *Wordsmith* (Scott 2017) and *AntConc* (Anthony 2018) were not well-suited to the project's long-term needs. While these systems are user-friendly and helpful to many linguists, they "are designed to work with plain-text corpora ... generally of rather small extent [and] they lack built-in support for complex annotation" (Evert and Hardie 2011: 2). Such annotation, along with search indexing, is required for sophisticated and efficient querying of very large datasets like the current one. The CWB and the query language implemented by its Corpus Query Processor (CQP) provide advanced search facilities such that *DSAE* editors can now disambiguate new usages from well-

known ones. For example, the colloquial SAE verb *vrek* meaning 'die' is already described in the dictionary, with 11 quotations. The last one, dated 1990, reads: "English goes from bad to worse as ... Prince Charles rewrites Shakespeare — and Hamlet *vreks*" (*DSAE*, *vrek*, v.). The other use of *vrek* as an intensifier is not, however, recorded in the *DSAE*. Examples from the corpus show it paired with adjectives, e.g. *vrek dangerous* (extremely dangerous) or *vrek happy* (extremely happy), suggesting a dictionary update for this word. To find such cases previously, the editors would have had to resort to Internet phrase searches to find attestations, requiring that they imagine for themselves the possible alternative adjectives for *dangerous*, *happy* and so forth — a hit-and-miss methodology. Instead, the CQP tool now allows editors quickly to locate examples of *vrek* followed by any adjective.

At the same time, because the TreeTagger depends on an English lexicon that is relatively (and in some respects inevitably) unaware of SAE word forms, corpus annotations are sometimes incorrectly applied or simply lacking. For example, *vrek* is sometimes incorrectly tagged as a proper noun according to its predictive model, and because the tagger's lexicon does not contain an entry for this word or its inflections, the past participial form *vrekked* is never recognised as being associated with the lemma form. Searches for this word would sometimes therefore fail to return examples when queries are constrained strictly to part of speech or the canonical form. In these cases more relaxed query constraints over multiword contexts would, however, still typically produce significant numbers of usefully disambiguated results, and on such a large dataset, the advantages of this new research tool remain numerous.

5.2 Semi-automatically generated lexical database

5.2.1 General overview

The toolchain produced a second major result, namely the creation of a pre-generated lexical database compatible with its existing one. This resulted in a dramatic expansion of the project's electronically-encoded data holdings from about 9,000 quotations (captured manually between 2004 and 2017) to a gross count of about 147 million. This figure does include inordinately large sets of examples for common SAE words (e.g. 1315 quotations for *aardvark*) and SAE terms which overlap with general English (e.g. over 12,000 examples for *robot*, the SAE term for 'traffic light'). The quotations were extracted by matching a list of 21,718 SAE word forms against the entire corpus. This list, which was also used by the toolchain for other purposes, is described in more detail below (see 5.2.2 Input: SAE dictionary search list).

Table 2 shows frequencies of matches against the SAE search list by number of examples found. Despite occasional high-frequency matches attributable to terms which are also general English, most of the word forms searched for

are specific to the SAE lexicon and do not entail such overlap. Even if, for argument's sake, only 0.2% of the newly gathered quotations were considered, this would still approximate to the roughly 300,000 quotations gathered manually in the history of the project since it was established almost 50 years ago. When the corpus-derived quotations are reduced to 100 examples per word form, the total is about 540,000.

Number of quotations found per word	Number of words in search list
0	10,734
1-500	8,457
501-1000	526
> 1000	2,001
Total number of search terms	21,718

Table 2: Frequencies of matches against the SAE dictionary search list by number of quotations found

The lexical database with its new inclusions has a specific purpose for the dictionary: it differs from the CWB corpus in that it stores quotations already preformatted in the XML markup used to present quotations in the dictionary entries, with additional metadata for workflow and editing processes. The lexical database itself is designed to be interoperable with XML dictionary editing software and automatically renders quotations as HTML (Hypertext Markup Language) for published output. The XML output of the toolchain was therefore modelled in such a way that it was not only compatible with the existing database schema, but mirrored the basic hierarchical structure of the dictionary, facilitating further interoperability.

Figure 3 shows the lexical database's editing interface, with an automatically-generated record for *erdvark*, variant spelling of *aardvark*. The *e-* spelling was last recorded in the dictionary in 1959 (see Figure 1), making this 2014 instance a rare but valuable quotation. The record was generated by mapping the toolchain's XML output to the lexical database format via an XSL (Extensible Stylesheet Language) transformation. This data conversion process was fully automated, including the insertion of bibliographical information and annotations associating the *erdvark* quotation with various lexical attributes of its parent entry, as well as the more common spelling, ensuring easy retrieval during subsequent workflow stages. Roles such as *inputter*, *reader* and *annotation group author* have been performed by the toolchain and are therefore marked SYSTEM. The original source URL and the toolchain's corpus file are also included for reference.

The screenshot shows a web-based form for creating a lexical entry. At the top, 'Inputter' and 'Reader' are both set to 'SYSTEM'. Below this, 'QuotationYear' is '2014', and 'PublicationDate' is 'day: 17', 'month: 09', 'year: 2014'. 'AuthorInfo' has a red 'X' icon. The 'URL of web source' is 'http://www.entrepreneur.co.za/donald-trump-how-to-turn-fear-int-o-faith/'. The 'Corpus file' is 'SAE2-e00015.xml (eng-za_web_2014)'. The 'Excerpt 1 of 1' has a 'Content Type' of '-None-'. The 'Quotation Text' is 'My guess, it hit a **erdvark** hole at high speed, or a hard landing.' Below this is a 'Catchwords' table for 'erdvark' (1 TOTAL). The table has columns for FORM, FORM TYPE, P.O.S., In DSAEHIST1, DSAEHIST1 ID, In SACOD2, and In DSAE(JB)4. The entry shows 'aardvark' as the FORM, 'variant' as the FORM TYPE, 'UNKNOV' as P.O.S., 'YES' as In DSAEHIST1, 'e00015' as DSAEHIST1 ID, 'UNKNOV' as In SACOD2, and 'UNKNOV' as In DSAE(JB)4. Below the table is an 'Annotation Group 1 by SYSTEM' section with a 'Catchword annotation for: erdvark' and a note '▶ is a variant form of: aardvark'. At the bottom, 'Proofreader' and 'Annotator' are 'NOBODY', and 'Final Reviewer' is 'NOBODY' with an 'Inclusion Status' of 'PENDING'.

FORM	FORM TYPE	P.O.S.	In DSAEHIST1
▶ erdvark	aardvark	variant	UNKNOV
	DSAEHIST1 ID	In SACOD2	In DSAE(JB)4
	e00015	UNKNOV	UNKNOV

Figure 3: An automatically-generated lexical database entry for *erdvark*, variant spelling of *aardvark*, only requiring approval

The main remaining task of the human editor is to review the quotation and, if he or she considers it to be useful and reproducible in the dictionary, to update its *inclusion status* attribute from its default value PENDING to ACCEPTED. Acceptance requires that the quotation be checked for errors on the part of the author (but this proofreading stage no longer requires verification against the original source since capture was automated during corpus creation, removing the possibility of transcription errors). Further annotations may also be optionally added.

Of course the editor should view all quotations with a critical eye, and in the current example the source URL happens to have been removed from the Internet since 2014. As with print-era ephemeral sources (leaflets, temporary signage and so forth) this does not mean the quotation cannot be cited. In this case the SAE corpus itself can ultimately be referenced: one of the advantages of the corpus is that it preserves ephemeral sources. It would also be desirable to note in the bibliographical metadata that this source comes from what used to be an industry news site, in this case for business entrepreneurs. This can be indicated manually in the lexical database interface via a SOURCE TYPE attribute not shown in the example. Such metadata could be added automatically in future by adding a subcomponent to the toolchain which queries a categorised list of the most frequently-cited domain names.

5.2.2 Input: SAE dictionary search list

The toolchain generates quotation records by matching word forms against the corpus using a simple string-matching process, drawing on a dictionary search list of 21,718 previously-documented SAE words. About 19,000 of these were drawn from the DSAE's existing XML dataset which distinguishes between lemma types, namely *headwords* versus forms derived from these headwords (*variant spellings, plurals, compounds, derivatives* and other forms such as phrases). Included in this search list were hypothetical software-generated spellings for multi-word lexical items which could occur as orthographic variants due to hyphenation or spacing changes. For example, from the headword *mealie-meal* (901 corpus matches), *mealie meal* and *mealiemeal* were generated, producing 136 and 57 matches respectively. A further 2,393 new word forms were added to the DSAE list from existing post-2004 electronic holdings and categorised simply as *catchwords*. Although the catchword sub-list did not encode relationships between canonical and other forms, it brought valuable new or potentially-new words to the quotation-mining process. Table 3 shows the composition of the search list.

Type of word form	Number of items
Headword	6,057
Plural of headword	843
Variant spelling	7,444
Compound, derivative or other nested lemma form	4,981
Catchword (new words)	2,393
Total word forms	21,718

Table 3: Composition of SAE dictionary search list

Subsequent components of the toolchain centred on word forms based on their similarity to items in the SAE dictionary search list, or on the TreeTagger's failure to recognise them, in order to isolate lists of potential new SAE words or variants without dependence on pre-existing lexical knowledge.

5.3 Semi-automatic discovery of spelling variants and headword candidates

5.3.1 Analysis of new headword candidates unrecognised by the TreeTagger

Since the TreeTagger (see 5.1 above) relies on a general English lexicon for POS-tagging and lemmatisation, it assigned many words a default 'proper noun' POS value based on its probabilistic model, and an 'unknown' lemma value. As a precursor to further processing, a list of tokens with unknown lemmas tagged as proper nouns with a minimum frequency of 100 was analysed manually to assess the correctness of these POS-tags. This list amounted to 416 unique tokens. Review by an editor found 268 (64%) to have been correctly tagged as proper nouns. Further normalisation steps were also performed to improve overall results.

Normalisation and exclusion steps

Because list filtering was automated and frequency-based, two kinds of corpus preprocessing were undertaken before using the TreeTagger, to limit the number of irrelevant results. Firstly, numerous Unknowns took the form of a cardinal number followed by an alphabetical character as used in measures, e.g. *5 m*. Since the TreeTagger does not split these tokens and therefore cannot lemmatise them correctly, these were normalised in the corpus to add intervening whitespace, producing e.g.: *5 m*. Performing this step on an experimental sub-corpus of about 2 million records, one per line, resulted in 2% of records being changed. Given that the Unknowns are a subset of otherwise correctly-lemmatized general English, this represented a significant reduction of corpus noise. A second preprocessing step used SAE proper noun exclusion lists to reduce the number of irrelevant Unknowns (typically proper nouns do not form part of the DSAE's inclusion policy). A comprehensive list of South African Geographical Names⁴ and a selective list of personal names, together totalling 47,987 single-word items and 1,027 multi-word proper names, were excluded. This list resulted in 625,787 unwanted corpus matches being filtered out.

5.3.2 Detection of new variants based on word similarity

The SAE dictionary search list included a list of documented variant spellings which were matched against quotations in the corpus. This left potential new

or previously undocumented variant forms which could be detected based on orthographic similarity between words in the corpus and words in the dictionary search list. Similarity was calculated using the Levenshtein distance algorithm, "a measure of the similarity between two strings ... the source string (*s*) and the target string (*t*). The distance is the number of deletions, insertions, or substitutions required to transform *s* into *t*" (Gilleland n.d.).

This computationally-intensive process produced lists of word forms from the dictionary search list, each with a sub-list of similar words found in the corpus. These sub-lists were annotated and sorted first by Levenshtein distance measure, then by frequency in the corpus. See Table 4 for example data generated this way, showing potential variants of the SAE word *imphepho* (the name of a medicinal plant).

Words similar to <i>imphepho</i> (a medicinal plant), corpus frequency: 48		
Word form	Frequency	Levenshtein distance
<i>impepho</i>	77	1
<i>imphepho</i>	13	1
<i>mphepho</i>	4	1
imphephu	3	1
iphepho	3	1
<i>mpepho</i>	16	2
iphepha	15	2
iphupho	5	2
mphephu	5	2

Table 4: New variant candidates extracted from the corpus based on word similarity (likely candidates italicised)

Ordinarily, researching potential variant spellings is a painstaking process fraught with uncertainty. The lexicographer cannot easily anticipate all possible spelling permutations of borrowings from the several languages acting on English in the exceptionally multilingual context of South Africa. The pre-generated orthographic profile shown in Table 4 reduces labour, guesswork and subjectivity substantially, including the exclusion of imagined permutations which are not attested in the corpus, allowing quick evaluation of data in a single view. Being presented with multiple unfamiliar word forms may, at the same time, prompt unproductive corpus searches into what are found to be unrelated word forms not counting as valid SAE usage (e.g. code-switching). These may come as undesirable distractions increasing the burden of research. The lexicographer may, however, counter this with editorial judgment to compensate the Levenshtein algorithm's blindness to certain patterns, for example by

noting in this case that the most likely valid variant spellings are those which do not produce a vowel change.

The variant-tables also provide a quick indication of the relative currency of the headword's spelling form. For instance, the current example shows that one of the variants identified (*impepho*, 77 matches) occurs more frequently than the initial word form supplied as the search term (*imphepho*, 48 matches), suggesting that the former should be considered not as a variant but as the more likely headword candidate.

For shorter words, a Levenshtein maximum distance of 2 or 3 was found to be most productive in identifying new variant spellings. For longer words, typically compounds, a maximum distance of 4 or 5 was found to be useful. In the latter cases, variations due to spacing, hyphenation or lack thereof accounted for initial orthographic permutations, followed by further permutations possibly prompted by borrowing from a different language for part of the multiword item. For example, a search for *karretjie people* (SAE for a nomadic people who travel in animal-drawn carts or (Afrikaans) *karretjies*), with a maximum distance of 5, illustrated such alternations in language borrowing (*mense* is Afrikaans for 'people'):

- (1) *karretjiepeople* (without space, Afrikaans + English)
- (2) *karretjie-people* (hyphenated, Afrikaans + English)
- (3) *karretjiemense* (direct borrowing of Afrikaans compound)
- (4) *karretjie-mense* (ditto, hyphenated)
- (5) *karretjiesmense* (with Afrikaans-style plural marker)

The hyphenation in (4) above (frequency: 4) likely represents Anglicisation in SAE, since hyphenation of compounds is not typical in Afrikaans. Likewise the plural form in (5) (frequency: 4) would probably not have been anticipated by a native English-speaking lexicographer. The direct borrowing from Afrikaans in (3) was found to be most frequent (39 corpus matches).

5.3.3 Detection of new headword candidates based on word similarity

Because the word forms matched against had already been filtered and therefore tended to produce words specific to SAE, a side-effect of the variant-detection process was that unrelated but new headword forms were sometimes uncovered. For example, a search for variant spellings of *bogadi* (a traditional African wedding gift) returned a 'similar' word *moladi* (a system for rapid and inexpensive wall construction, designed and used in South Africa). Levenshtein distance was 2 with 120 corpus matches, suggesting headword candidate status. While such results were difficult to predict given that they were incidental to the actual purpose of this toolchain component, they presented an additional means of lexical acquisition and they are flagged for the editor's attention by high frequency values.

5.3.4 Detection of headword candidates using term extraction

In the final stage of the toolchain, standard term extraction techniques (cf. Ahmad et al. 1994) were used to detect potential new headword candidates. The term extractor *TrEx_v5.9_sae*⁵ was used. It compared the SAE corpus with the British National Corpus (BNC), since the latter would likely show lower frequencies for SAE tokens, and confined analysis to terms with a minimum corpus domain frequency of 10. This process compared the relative frequency of tokens in each corpus and extracted those more prominent in the SAE corpus. As output, the tool produced 15 lists, each representing a part-of-speech pattern, with the term, its frequency, an example quotation and other statistical data. Of these statistical rankings the most relevant was its *termhood* value. Termhood measures "try to identify candidate terms which are used [or] specialized in the domain as technical terms" (Schäfer 2015: 49), and the tool was used to test the hypothesis that this measure would also highlight SAE-specific words. Given the very large scale of the SAE corpus, and because general English terms had not yet been filtered out, the lists produced were unwieldy, generating a startling total of 2,615,854 candidate terms. Ranking these candidates using a combination of corpus frequency and termhood value, however, made new headword candidates accessible, and identified this toolchain component as a useful new mechanism in semi-automatic lexical acquisition. Some example new noun headword candidates discovered this way are:

- (1) *braairoom* (an entertainment room used for indoor barbecues)
- (2) *mokoro* (a type of canoe used in Botswana)
- (3) *miombo (woodland)* (a Southern African vegetation type).

6. Re-orientation of reading programme prompted by semi-automation

The preceding discussion of the data resources newly available to the historical dictionary project, and the algorithms and output of the toolchain, together suggest a long-term review of the project's workflow and policies during its revision stage. Topics either not mentioned or only lightly touched on in this paper — being detailed and beyond its scope — are inclusion policy, criteria for SAE status, entry revision prioritisation, and the role of the lexicographer in assessing evidence. The new data and tools impact on all of these questions. For example, the print edition was compiled in a period when electronic sources were only starting to become accessible. Its inclusion policy for headwords and its high proportion of variant spellings may have been based on the reasonable assumption that more evidence existed than was available in the project's index card database. Now, faced with a massive influx of new data — albeit only for 2011–2017 sources — should it perpetuate the same inclusive approach when drafting new entries? Already over 2,300 new headword candidates had been identified prior to the semi-automation process, or roughly

half the number of headwords in the *DSAE*'s first edition (a 25-year project). Given the scale of new data, the number of new headword candidates may also increase dramatically, requiring prioritisation, probably also best done with further computational methods. In order to cope with large scale data the question becomes: which types of lexicographical tasks can be delegated to machines?

Responding to the analogous question "Will there be lexicographers in the year 3000?", posed in 1998 by Gregory Grefenstette, Michael Rundell observes that:

From the standpoint of the editor and publisher, the shift to automation offers the prospect of producing a more diverse range of lexical resources without the enormous costs associated with conventional dictionary-making. It seems likely that, for the time being, there will be a central role for skilled lexicographers and editors. But their role is changing, from selecting and synthesising information, to 'editing' and validating choices already made by software. (Rundell 2012: 17)

The semi-automation of the *DSAE*'s reading programme prompts such a change in roles. The separation of automatic data collection processes from the closing steps requiring human judgement, as detailed above (5.2 Semi-automatically generated lexical database), are ample illustration of a shift from 'selecting and synthesising' quotations to 'editing' (if necessary) and 'validating' them.

The transition extends beyond validation, however, in that the toolchain's dependence on predetermined, categorised word forms signals a more general change in data collection strategy: project staff should orient their collection efforts towards lists. Whereas previously reading for neologisms or updated quotations typically involved (1) opportunistic reading across a wide variety of sources, (2) checking back against the database to avoid duplication, and (3) searching for further attestations if necessary to establish the currency of a potential new word candidate, the task is now simply to find a single example. On capturing a single instance of a potential new word, preferably with its canonical and inflected forms distinguished, the toolchain will be far faster and more comprehensive in sourcing new quotations. Likewise, if a word form is already known to the toolchain, it will already have extracted all possible quotations from the corpus, and to source new examples manually would duplicate effort. Likewise, orthographically-similar spelling forms could be detected by the toolchain as described above, again using the new list item as a starting point.

7. Conclusion

This paper has described the acquisition of new electronic data sources, their encoding as a very large, queryable part-of-speech-annotated SAE corpus, the subsequent dramatic expansion of a historical lexical database, and the provision of new headword and variant spelling candidates using a computational linguistic toolchain. The next steps for the dictionary project prior to revision

involve incorporating these results most usefully into future workflows. For example, the sometimes overwhelming numbers of quotations now provided for certain words could potentially be filtered for easier evaluation or disambiguated to reveal new, undocumented patterns of usage. Similarly, additional data sources could be added to the toolchain. These would, however, be improvements on a major development for a previously under-resourced dictionary project where data holdings were concerned; several highly-enabling steps towards semi-automatic 'reading' for lexicographic evidence have already been taken.

8. Acknowledgements

Thank you to Ms Heike Stadler, University of Hildesheim, Germany for generously adapting and maintaining her suite of Perl programs between 2015 and 2017 for the purposes of the collaborative research described in this article. Ultimately this produced the input data for the 100-million-token newspaper corpus described in 4.1 above. We also gratefully acknowledge the supply of data from the NLP Group, Department of Computer Science, University of Leipzig, Germany (see 4.2).

9. Endnotes

1. Developed by Ms Heike Stadler, University of Hildesheim, Germany for the collaboration described in this article. Please see 8. Acknowledgements.
2. The TreeTagger is made "freely available for research" at <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.
3. Available open-source under the Gnu General Public Licence at <http://cwb.sourceforge.net/>.
4. Released in 2011 by Statistics South Africa (see <http://www.statssa.gov.za/?p=1341>) and supplied to the project for research use.
5. Originally developed as part of the project described in Schäfer 2015 (see p. 87).

10. References

- A Dictionary of South African English*. [Online]. Available: <http://dsae.co.za>.
- Ahmad, K., A. Davies, H. Fulford and M. Rogers.** 1994. What is a Term? The Semi-automatic Extraction of Terms from Text. Snell-Hornby, M., F. Pöschhacker and K. Kaindl (Eds.). 1994. *Translation Studies: An Interdiscipline*: 267-278. Amsterdam: John Benjamins.
- Anthony, L.** 2018. *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University.
- Atkins, B.T.S. and M. Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Brewer, C.** 2007. *Treasure-house of the Language: The Living OED*. New Haven/London: Yale University Press.

- 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. Accessed at <http://eprints.lancs.ac.uk/62721/> [07/27/18].
- Gilleland, M.** n.d. *Levenshtein Distance, in Three Flavors*. Accessed at <https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm> [25/7/2018].
- Goldhahn, D., T. Eckart and U. Quasthoff.** 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. 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): 759-765. Istanbul, Turkey: European Language Resources Association. Accessed at <https://pdfs.semanticscholar.org/1b56/0f892432fb853d233c92f9294640bc91de3c.pdf>.
- Hicks, S.** 2010. Firming up the Foundations: Reflections on Verifying the Quotations in a Historical Dictionary, with Reference to *A Dictionary of South African English on Historical Principles*. *Lexikos* 20: 248-271. Accessed at <http://lexikos.journals.ac.za/pub/article/view/142> [03/12/2017].
- Pienaar, L. and V. de Klerk.** 2009. Towards a Corpus of South African English: Corraling the Sub-varieties. *Lexikos* 19: 353-371. Accessed at <http://lexikos.journals.ac.za/pub/article/view/444> [03/12/2017].
- Rundell, M.** 2012. The Road to Automated Lexicography: An Editor's Viewpoint. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 15-30. Oxford: Oxford University Press.
- Schäfer, J.** 2015. *Statistical and Parsing-based Approaches to the Extraction of Multi-word Terms from Texts: Implementation and Comparative Evaluation*. BSc Thesis. Stuttgart: Institute for Natural Language Processing (IMS), University of Stuttgart.
- Scott, M.** 2017. *WordSmith Tools*. Stroud: Lexical Analysis Software.
- Silva, P., W. Dore, D. Mantzel, C. Muller and M. Wright (Eds).** 1996. *A Dictionary of South African English on Historical Principles*. Cape Town: Oxford University Press.
- Statistics South Africa.** 2011. *South African Geographical Names Database*. See <http://www.statssa.gov.za/?p=1341>.
- Willinsky, J.** 1994. *Empire of Words: The Reign of the OED*. Princeton, N.J.: Princeton University Press. Accessed at <https://books.google.co.za/books?id=UvCbv3ckRDkC&dq> [07/27/18].

Objectivity, Prescription, Harmlessness, and Drudgery: Reflections of Lexicographers in Slovenia*

Alenka Vrbinc, *Faculty of Economics, University of Ljubljana,
Ljubljana, Slovenia (alenka.vrbinc@ef.uni-lj.si)*

Donna M.T.Cr. Farina, *Department of Multicultural Education, New Jersey
City University, New Jersey, USA (dfarina@njcu.edu)*

and

Marjeta Vrbinc, *Department of English, Faculty of Arts, University of
Ljubljana, Ljubljana, Slovenia (marjeta.vrbinc@ff.uni-lj.si)*

Abstract: This contribution reports on a study that set out to paint as complete a picture as possible of the context and content of modern Slovenian lexicography. We aimed to discern the philosophical underpinnings, the most noteworthy accomplishments, and the main projects of Slovenian dictionary work as presented by our seven subjects, who are all prominent members of the lexicographic community. We sought specialists who work on synchronic topics and concentrate more on the standard language and terminology rather than on dialectal variation and other lexicographic topics that are of more interest to scholars than to educated lay persons. The interview script consisted of thirteen narrative questions, designed to allow the interviewees to reflect in as much depth as possible on their daily practice as well as on their underlying vision of what lexicography or terminography is. This article discusses the development and influences of Slovenian lexicographic theory and presents part 1 of the results of this study: the views of the practicing lexicographers on whether they perceive their lexicographic work as drudgery and what they see as the essential nature of their role in society — how the dictionary maker can be a force for good and avoid any potential for harm.

Keywords: HARMLESS DRUDGE, DRUDGERY, HARM, HARMLESSNESS, INTERVIEW, LEXICOGRAPHER, LEXICOGRAPHIC PHILOSOPHY, LEXICOGRAPHIC PRINCIPLES, MONOTONY, REPETITIVENESS, TEDIOUS

Opsomming: Objektiviteit, voorskriftelikheid, onskadelikheid en sleurwerk: Beskouings van leksikograwe in Slowenië. In hierdie bydrae word verslag

* A highly abbreviated summary of this article, "Slovenian Lexicographers at Work," will appear in the *Proceedings of the XVIII EURALEX International Congress, 17–21 July 2018: Lexicography in Global Contexts*.

gedoen oor 'n studie waarin gepoog is om so 'n volledig moontlike beskrywing te gee van die konteks en inhoud van die moderne Sloweense leksikografie. Ons het probeer om die filosofiese boustene, die noemenswaardigste prestasies, en die belangrikse Sloweense woordeboekprojekte soos voorgehou deur ons sewe respondente, wat almal prominente lede van die leksikografiese gemeenskap is, weer te gee. Ons het vakkundiges gekies wat aan sinchroniese onderwerpe werk en meer op die standaardtaal en -terminologie konsentreer as op dialektiese variasie en ander leksikografiese onderwerpe, wat van meer belang is vir die vakkundige as vir die opgevoede leek. Die onderhoud het bestaan uit dertien narratiewe vrae, wat ontwerp is om die respondente toe te laat om so volledig moontlik weer te gee wat hul daaglikse praktyke is sowel as wat hul onderliggende visie van die leksikografie en terminografie is. Hierdie artikel bespreek die ontwikkeling en invloed van Sloweense leksikografiese teorie en gee deel 1 van die resultate van hierdie studie weer: die beskouings van die praktiserende leksikograwe oor of hulle hul leksikografiese werk as sleurwerk ervaar en wat hulle as die wesensaard van hul rol in die gemeenskap beskou — hoe die woordeboekmaker 'n goeie mag kan wees en enige potensiele skade kan vermy.

Sleutelwoorde: ONSKADELIKE WERKESEL, SLEURWERK, SKADE, ONSKADELIKHEID, ONDERHOUD, LEKSIKOGRAAF, LEKSIKOGRAFIESE FILOSOFIE, LEKSIKOGRAFIESE BEGINSELS, EENTONIGHEID, HERHALING, VERVELING

1. Introduction

To a certain degree, dictionaries are created and delivered in similar ways worldwide. Some lexicographers are aware of others' work and become familiar with new ideas via conferences and publications. Bilateral and multilateral lexicographic work takes place between organizations (such as AFRILEX, ASIALEX, DSNA, and EURALEX) or else between academies of science (such as the Austrian or Slovenian academies). Despite this seemingly favorable state of affairs, many lexicographers still labor alone without a deep awareness of what others in the field are doing, even when similar dictionaries are being created in other countries. Working on a dictionary is by its nature solitary, so to some extent not so much has changed since 1755, when Samuel Johnson, the great English lexicographer, humorously defined the word *lexicographer* as a "harmless drudge." While some lexicographers can network frequently through conference attendance and have time to keep abreast of the state of the art through publications, others are hard pressed to keep up with the demands on their time imposed by the tyrannic words of their focus language. In such circumstances, the average dictionary maker may be barely aware of the existence of international lexicographic thought.

The purpose of the present study is to break this solitude and provide a glimpse into the world of lexicographers whose practices may not be well-known. To our knowledge, there have been no in-depth studies based on intensive, extensive interviews with the lexicographers of any country or culture. In the present work, we are examining Slovenian lexicography through the eyes of the seven Slovenian lexicographers whom we interviewed; our hope is that

other researchers will replicate this work to allow insight into practices prevailing in other countries. This type of reflection within the discipline of lexicography will aid, we suggest, in the advancement of theory globally.

2. The setting of Slovenian lexicography

The Republic of Slovenia is a country of over two million people, located in Central Europe. One of the six republics of the former Yugoslavia, Slovenia declared independence in 1991. Slovenian, the most widely spoken language in the country, is classified genetically as a South Slavic language along with other languages spoken both within the former Yugoslavia and beyond it. Although Slovenian has a relatively small number of speakers, it nevertheless has a significant lexicographic tradition; this history, like that of many other traditions (cf. Béjoint 2016; Farina and Durman 2009; Fontenelle 2016) began with needs arising from contact between languages and cultures. In the case of Slovenian, the main contact was with the German language within the Central European cultural context.

Contact with the cultures of Central Europe influenced the eventual organization of Slovenian lexicographic work. The Slovenian Academy of Sciences was founded in 1938; within it, the Institute of the Slovenian Language — where lexicographic projects are ongoing today — was established in 1945. The modern Slovenian Academy focuses on monolingual lexicography but not all monolingual work takes place exclusively within it. In 2004, the independent Trojina, Institute for Applied Slovenian Studies, was founded in Ljubljana. Through grant funding, Trojina collaborates on projects with other institutions engaging in lexicographic work, at the University of Ljubljana and beyond. Since Slovenia achieved its independence in 1991, public interest in the national language has increased. The number of monolingual projects has grown within the Academy of Sciences; there are existing dictionaries or ongoing projects on phraseology, orthography, synonymy, and terminology (to name some). In order to field an increasing number of questions from the public about language, the Academy maintains an active online consulting service. The Trojina Institute has its own online tools that are utilized to engage Slovenian speakers to the fullest extent possible in deeper reflection on their language.

Slovenian bilingual lexicographic work is conducted outside the walls of the Academy of Sciences. Presently there are pairings of Slovenian with a greater number of languages than was the case historically. For example, there now exist recent dictionaries of Slovenian with Czech, Dutch, English, French, German, Italian, Polish, Russian, Serbo-Croatian, and Spanish. Unfortunately, just as the public's interest in bilingual lexicographic tools has increased, Slovenian publishing houses have ceased to publish such dictionaries. For this reason, as one of our interviewees indicates, the future of Slovenian bilingual lexicography is unclear.

3. Development and influences of Slovenian lexicographic theory

In the history of lexicography, prefaces and other front matter have usually provided some insight into a given dictionary's compilation principles (Shapiro 2017), but they have seldom been forthcoming enough to fully guide specialists or the general user. For example, Landau (2001: 64) and Béjoint (2010: 68-76) discuss Samuel Johnson's theory with references to his preface, while Jackson (2002: 42-46) points out the additional theoretical benefit of Johnson's 1747 *Plan of a Dictionary of the English Language*. Farina and Durman (2012: 9) contrast the original preface by Baudouin de Courtenay in his revision of an early twentieth-century Russian dictionary, with the more detailed explanations he provided in later writings — when he was trying to defend his highly-criticized compilation decisions. Slovenian lexicography has followed the same typical historical movement toward providing ever-increasing theoretical information. While the front matter to the first volume of *The Dictionary of Standard Slovenian* (Bajec et al. 1970) gives a detailed explanation of how to use the dictionary, this is almost impossible for a lay person to decipher. Since the 1970s but particularly in the new century, there has been a constant stream of scholarly work putting forward an underlying philosophy of what general Slovenian lexicography should be (for example: Gantar 2015; Gliha Komac et al. 2015; Gorjanc et al. 2015; Gorjanc et al. 2017; Ledinek et al. 2015; Snoj 2004; Srebnik 2015; and Žagar Karer 2011).

Both contemporary monolingual and bilingual lexicography within Slovenia have been deeply influenced by British lexicographic theory; the lexicographers interviewed for this study mentioned Sue Atkins, Patrick Hanks, R.R.K. Hartmann, Adam Kilgarriff, Michael Rundell, and John Sinclair. The interviewees also demonstrate a wide reading across many linguistic and lexicographic cultures. They mentioned Sylviane Granger (Belgium); Gilles-Maurice de Schryver (working in Belgium and South Africa); Rufus Gouws and Danie Prinsloo (South Africa); František Čermák (former Czechoslovakia and Czech Republic); Herbert Ernst Wiegand (Germany); Ute Römer (working in Germany and the United States); Dwight Bolinger, Don McCreary, Erin McKean, and Ben Zimmer (United States); Ladislav Zgusta (working in former Czechoslovakia and then the United States); Anna Wierzbicka (Poland and Australia); Juri Apresjan (former Soviet Union and Russian Federation); and Bo Svensén (Sweden). In the realm of modern terminography, the Slovenian tradition has been most influenced by the classical Vienna school of terminology.

4. Ensuring the future of Slovenian lexicographic work

For the authors of the present article, there is a striking contrast between the governmental and societal nurturing of lexicographic endeavors that take place in the small country of Slovenia, versus the almost entirely independent and commercial practice of the United States (as well as many other countries, such

as Germany, the Netherlands, and the UK), where there is little to no government funding of dictionary work. In Slovenia there are university courses designed to introduce graduate students to lexicographic theory; such courses are rare across the United States. At the University of Ljubljana alone, there are two graduate courses on monolingual lexicography; there is also a short graduate course on bilingual lexicography. At the University of Maribor, a much smaller institution than the University of Ljubljana, there is a graduate course on lexicography and another course that treats dictionaries as a cultural practice. What is more, the official curriculum for all public and private Slovenian high schools has several components intended to familiarize students with dictionaries and their purposes; there is a question about dictionaries on the official high school final exam. Certainly, the visibility of both high school and university programs of dictionary study is an important factor both in maintaining the interest of the general public in dictionaries and in the Slovenian language, and in ensuring that lexicography will remain a viable discipline as well as a career field for some.

Slovenia has taken other steps to ensure the future development of lexicographic practice and theory. Since 1985, the Young Researchers Program has selected talented master's and doctoral students to work in industry, university departments, and institutes both within the Academy of Sciences and beyond; lexicography is one of many fields of study to benefit from this program. By training the future cadre of practicing lexicographers, the program has helped move forward the professionalism of the discipline. Four out of the seven interviewees for this project — as well as two authors of this article — began their lexicographic careers within the Young Researchers Program.

5. Aims of the study

This study set out to paint as complete a picture as possible of the context and content of modern Slovenian lexicography. We aimed to discern the philosophical underpinnings, the most noteworthy accomplishments, and the main projects of Slovenian dictionary work as presented by our seven subjects, who are all prominent members of the lexicographic community. For this study only seven persons were interviewed, so we do not claim to present a comprehensive picture; our findings would most likely require revision if additional subjects were consulted. Nevertheless, because we interviewed lexicographers working on different projects and within several institutions, who have different duties and approaches that vary significantly, we do claim that this study captures some of the most important issues in Slovenian lexicography today. This study should be of interest to lexicographers worldwide who want to reflect upon their own practice, their country's or culture's practice of making dictionaries. Through a look at the work lives of Slovenian lexicographers, dictionary makers internationally stand to gain a better understanding of what they most want to do at home to improve our field. Lexicographic practice

around the globe would benefit if other researchers replicated this study or used components of it as a departure point for the examination of other lexicographic cultures. Finally, apart from the more immediate aims of this work, we hope that the Slovenian lexicographers who were kind enough to participate will gain from the reflection they engaged in during the interviews, as they continue to pursue excellence in their future work.

The extensive interviews of this study yielded copious data, which the present article does not cover in its entirety. Here, in part 1 of our findings, we address drudgery in lexicographic work and the potential of the lexicographer to do harm. Future reports will treat other important topics revealed in the interviews.

Four overarching research questions drove our thinking in the full study and informed the creation of the interview script:

1. What is the philosophical and intellectual framework governing the work of Slovenian lexicographers? What ideas do they all share — across different institutions and projects — as they engage in making dictionaries?
2. What are the main areas of concern and common significant problems that inform the work of Slovenian lexicographers?
3. What do the lexicographers consider both the main strengths and the weaknesses of their current efforts in dictionary creation? What would they most like to change about their practice?
4. What are the differences among our interviewees in their conception of what lexicography is all about?

Approximately sixteen hours of interviews provided us with information related to the above questions. The present article focuses mostly on Research Question 1, with some elements of 4: What do the lexicographers think about before they even sit down to work; what are their reflections on the most important underlying ideas that drive how they perform their duties. A future article will focus more on Research Questions 2 and 3: the specific projects, challenges, and practices of the lexicographers.

6. The interview script

Since this project was designed to be replicable in other cultures and countries, the full interview script appears in the Appendix for the use of other researchers. The script consisted of thirteen narrative questions, designed to allow the interviewees to reflect in as much depth as possible on their daily practice as well as their underlying vision of what lexicography or terminography is. (In other words, the script was designed to assist us in answering the overarching questions above.) It took two hours or more to cover all of the questions in the script with each person. The first two interview questions as well as Script

Questions 7–9 provided us with personal background information as well as information about the lexicographers' daily work, projects, and accomplishments: How did they end up "doing" lexicography and what does a "normal" day look like for them; what project takes up most of their time presently and what product(s) has/have given them the most satisfaction? Script Questions 4–6 treated the philosophical and theoretical underpinnings to their work. We chose to approach this topic with several detailed questions phrased in different ways, in order to appeal to individual styles and thought processes. In addition, because the objective of the grant that funded this work (see Acknowledgements) is to foster collaboration between scientists in the United States and Slovenia and to encourage future cooperative projects, we asked directly in Script Question 6 about any U.S. sources, theories, or practices that may have influenced the Slovenian lexicographers' work. While one interviewee may have said more about (for example) Script Question 5 and another may have elaborated most on Script Question 4, overall we sought and received a comprehensive picture of each person's lexicographic or terminographic world view. Script Questions 10–12 dealt with the problems and constraints the lexicographers face commonly as they strive to deliver high-quality products to dictionary users. Finally, Script Question 13 asked the subjects to recommend different ways in which international cooperation could take place and how it might improve lexicographic practice everywhere.

While all of the interview questions (see Appendix) inform the present article directly or indirectly, two of them, Script Questions 3a and 3b, are our main focus here:

3. The famous English lexicographer, Samuel Johnson, defined the word *lexicographer* thus, in 1755: "a writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words."
 - a. We would like to know, first: What elements of your own work do you consider "drudgery," hard, menial, or monotonous work?
 - b. Second, do you think the lexicographer is "harmless?" Does he or she play an invisible, unnoticed social role, or the opposite? How are lexicographers significant to the society of which they are a part?

Interview Question 3a turned out to be less significant than we expected. As will be shown in 9. Lexicography as drudgery? (below), while the lexicographers had opinions on the tedious or monotonous aspects of their work, this is not an issue that preoccupies their thinking, most likely because technology has truly diminished drudgery in modern lexicography. On the other hand, Interview Question 3b (discussed in 10. Harmless or harmful?) gets to the heart of the Slovenian lexicographers' most pressing concerns. They think about the role they play in society and about what they must do to fulfill this role, in

order to satisfy their users. The analysis presented here is most dependent on the answers our volunteers supplied to Interview Question 3b.

7. The selection of interview subjects

In order to select whom to invite for interviews, we first considered how lexicographic work is organized in Slovenia and what the different contexts are where such work is taking place. First, within the Research Center of the Slovenian Academy of Sciences and Arts there is the Fran Ramovš Institute of the Slovenian Language. The goal of this institute is to compile linguistic materials for the creation of high-quality resources on the Slovenian language. This institute specializes in the following areas: lexicology, etymology, onomastics, dialectology, terminology, and historical dictionaries. In addition to work within the Academy of Sciences, there are ongoing lexicographic projects in a variety of units at the University of Ljubljana (for example, in the Faculty of Arts, the Faculty of Social Sciences, and the Faculty of Computer and Information Science). There is also, for example, an ongoing collaborative project within the Faculties of Arts at the University of Ljubljana and the University of Maribor, in cooperation with the independent Institute of Ethnic Studies in Ljubljana. There are projects led by Trojina, Institute for Applied Slovenian Studies, usually in cooperation with other units.

The focus of this research was on those aspects of lexicographic work that have the greatest significance for the general public rather than areas that might attract primarily language specialists. As a result, there are etymologists, dialectologists, and other lexicographic specialists in Slovenia who were not interviewed because their work is beyond the purview of this study. We wished to discern how the lexicographers interviewed envisage and relate to the users of the contemporary Slovenian language who are the consumers of their products. We sought specialists who work on synchronic topics, and who concentrate on the standard language and terminology rather than on dialectal variation and other topics that are of more interest to scholars than to educated lay persons. We were interested in finding out how "traditional" or not the views of the Slovenian lexicographers are toward their language; to what extent are they accepting of language change and documenting that change in their dictionaries? How do they relate to borrowings into Slovenian from a variety of languages? We also wanted to know what the lexicographers thought about their dictionary users: What is the vision of "the user" that they have in mind when seated at their computers engaging in lexicographic work?

Our request for assistance was well received and we had an adequate number of volunteers; all are prominent lexicographers representing a broad spectrum of work. Only seven persons were interviewed; therefore, this should not be considered a representative sample of the views and thoughts of all of Slovenian lexicography. Due to time constraints and availability of lexicographers, not all specialists could be asked and not all were able to volunteer. This

study should be considered a sampling of thought-provoking views prevailing within the evolving and viable modern Slovenian lexicographic tradition.

8. Our interview subjects

Operating from our script of questions, we interviewed seven Slovenian lexicographers who, collectively, address through their work most of the significant issues facing synchronic theoretical lexicography today. Our interviewees were not anonymous participants. Due to their positions and influence in the field, their reflections are quoted and cited here so that these ideas might advance lexicography worldwide. The interviewees had the option at all times to provide information "off the record," information that is not directly associated with them in what follows. Over the course of an interview lasting two hours or more, the lexicographers were free to make specific comments that would not be directly attributed to them in any subsequent oral or written discussion. In reality, we received very few "off the record" comments; the seven interviewees were candid and forthcoming with their views. What follows is an introduction to the interviewees and their areas of expertise.

Apolonija Gantar is a researcher in the Department of Translation of the Faculty of Arts, University of Ljubljana. She currently collaborates with several different academic and research institutions on projects dealing with: collocations, a new grammar of Slovenian, and non-standard Internet Slovenian.

Nataša Jakop works in the Lexicological Section of the Fran Ramovš Institute within the Academy of Sciences. She is in charge of phraseology for the third edition of *The Dictionary of Standard Slovenian* [Slovar slovenskega knjižnega jezika], a project begun in 2016.

Iztok Kosem is affiliated with Trojina, the Institute for Applied Slovenian Studies; he also is a researcher in the Faculty of Arts at the University of Ljubljana. He works with several institutions on projects concerning: a Hungarian–Slovenian dictionary, collocations, and a new grammar of Slovenian.

Nina Ledinek is the Head of the Lexicological Section of the Fran Ramovš Institute; she coordinates the work on *The Dictionary of Standard Slovenian* and also worked on the improvement of the FRAN online dictionary portal.

Jerica Snoj began her lexicographic career during the final stages of preparation of the first edition of *The Dictionary of Standard Slovenian* (1970–1991). Today she works on the new (third) edition. From 1991, she participated in the planning and production of *Slovenian Orthography* (Toporišič et al. 2001), which established the norms for the written Slovenian language. After fifteen years, her *Dictionary of Slovenian Synonyms* came to fruition (Snoj et al. 2016). Among our interviewees, Dr. Snoj is the lexicographer with the longest experience in the field of general as well as special-purpose lexicography.

Anita Srebnik is an instructor of Dutch in the Department of German, Dutch and Swedish in the Faculty of Arts at the University of Ljubljana. She is an independent lexicographer who authored the *Slovenian–Dutch European Dic-*

tionary (2006) and the *Dutch–Slovenian Dictionary* (2007), intended for Slovenian learners of Dutch.

Mojca Žagar Karer is the Head of the Terminological Section of the Fran Ramovš Institute. She has worked on numerous terminological dictionaries, including the *Dictionary of Theatre Terms* (Sušec Michieli et al. 2007), the *Dictionary of Automated Control Systems and Robotics* (Karba et al. 2014), and an ongoing dictionary of legal terminology.

9. Lexicography as drudgery?

Our interview question (3a) on drudgery was intended to encourage interviewees to speak about what people sometimes would rather not talk about with a stranger: the more unpleasant or undesirable aspects of their work. We guessed that the interviewees would prefer not to complain to us. We assumed they would certainly consider some aspects of lexicographic work to be drudgery (even considering modern technology) and through discussion of such a general topic might begin to speak about both the positive and negative aspects of their work.

The description of the dictionary maker as a drudge, thanks to Samuel Johnson, is familiar to almost every lexicographer. The topic of drudgery has been discussed often in the lexicographic literature, whether or not the words *drudge* or *drudgery* are actually used. Recently, Kory Stamper discussed the difference between art and craft in lexicography, and argued that "craft" — because it implies repetition — is a more accurate depiction of dictionary making than "art," which often connotes instantaneous inspiration and creation:

... "[C]raft" implies care, repetitive work, apprenticeship, and practice. ... Defining is the mental equivalent of free throws in basketball: anyone can stand at the free-throw line and sink one occasionally; everyone gets lucky. But the pro is the person who stands at the free-throw line for hours, months, years, perfecting that one motion until it is as fail-safe as humanly possible. ... Craft takes time, both internal and external. You need patience to hone your skill; you need a society willing to wait (and pay) for that skill. (2017: 256).

The repetitive and never-ending nature of lexicographic work is also mentioned by Landau (2001: 396): "Making a dictionary is like painting a bridge: by the time one coat of paint has been applied, the bridge is in need of another." Algeo, while acknowledging the inherent drudgery of the work during the print-dictionary era, emphasizes the dictionary maker's social value: "Although they are relatively anonymous, lexicographers as a class enjoy some of the same popular trust and respect as physicians... . Lexicographers do a real good in recording the language" (1985: 357). This is a recognition shared by Roberts, who, in his foreword to Sharp (2012), notes that "... producing dictionaries is no mere harmless drudgery. ... [D]ictionaries have a crucial role in helping to advance a common language, and to bring at least a degree of order to a

cacophony of voices" (viii). On the other hand, Schäfer (1984: 196) expressed optimism that "A computerized dictionary should take the 'harmless drudgery,' if not the drudge, out of lexicography." Finally, Sokolowski (2014: 287-288) considers whether the real drudgery might be the lack of knowledge about how the user actually benefits from lexicographic endeavor:

But when they look up particular words, which words are they looking up? The privacy of the act has meant that, for nearly all of the history of published dictionaries, only the users have known. Lexicographers and publishers could never have known whether their labors on any given word were read often — or never. This might make for a grim perspective on one's life's work ("harmless drudge," indeed), but it is obviously understood by all dictionary makers that in order for a dictionary to be generally useful, it must contain all the specific information about words that is likely to be needed. This is the true pact between the user and the dictionary: whenever you have questions, here are answers.

The repetitiveness, the anonymity, and the social significance of dictionary work occupy the thoughts of Slovenian lexicographers just as they occupy their colleagues globally. Among our seven interviewees, the interview question on drudgery resulted in one "no" and six "yes" responses. Four of these were a resounding "yes," while two interviewees gave a "yes, but ..." answer that focused less on the drudgery itself and more on suggestions for mitigating the amount of drudgery in lexicographic work.

The sole terminographer among our interviewees was the only person to answer an unequivocal "no" to the drudgery question. This is not so surprising given that the work approach of terminography is radically distinct from that of other realms of lexicography. Monolingual as well as bilingual lexicographers, phraseologists as well as compilers of synonym and other types of dictionaries, compare contexts of word use or study sense discrimination and composing apt dictionary entries. In contradistinction, the terminographer's work, in the words of Mojca Žagar Karer, Head of the Terminological Section of the Fran Ramovš Institute, is much more "dynamic" and is highly interactive. She does not find any of her tasks to be monotonous because she is engaged constantly with experts from different fields. It is the experts who labor over the definitions (because these definitions have to be precise from the perspective of their field) and Dr. Žagar Karer and other terminographers then edit them. Terminographers do not work alone, in "peace and quiet;" they are constantly on the phone or on email coordinating terminological work or checking fine points in the definitions completed by others. If the terminological work at hand is bilingual or multilingual (which is the norm), Dr. Žagar Karer would most likely need to consult with several different experts to hit upon a general consensus about the most felicitous way for the Slovenian language to convey accurately a concept from the terminology of another language. In short, the terminographer is more like an editor than a lexicographer.

Among those four who provided an emphatic "yes" to our drudgery question were Nataša Jakop and Jerica Snoj, both of the Fran Ramovš Institute

in the Academy of Sciences. They said that *all* lexicographic work, *all* phases of dictionary making are drudgery! Nina Ledinek of the Fran Ramovš Institute and Anita Srebnik, a bilingual lexicographer, used the word "monotonous" to describe many aspects of lexicographic work. Dr. Snoj mentioned the repetitive nature of the work; each task must be performed thousands of times, for as many words as are being investigated; Dr. Jakop pointed out that monotony can lead to waning concentration, a single moment of which can lead to an error: For example, a feminine noun can be labeled mistakenly as neuter. Dr. Ledinek emphasized how difficult it is to analyze a word with numerous concordance lines in a corpus and multiple meanings; there is lots to describe! She noted how extremely difficult it is to be consistent, systematic, and coherent when treating grammatical patterns and collocates. It is also challenging to describe what the standard language is and what the norm is, or to try to describe similar things (i.e. taxonomic sets such as mammals, days of the week) in a unified way. Finally, Dr. Srebnik, who, of these four interviewees is the only one who compiled her dictionary independently, contributed one not-strictly-lexicographic aspect of her work as additional drudgery: fundraising. She was forced to raise money on her own in order to convince the publisher to put her Dutch–Slovenian dictionary into print. Dr. Srebnik stressed that Slovenia needs much better support for bilingual lexicographic work.

Our two "yes, but ..." answers came from lexicographers who acknowledge that many aspects of lexicographic work are drudgery, but whose remarks focused more on how to lessen its amount in lexicographic work. Apolonija Gantar, a researcher at the University of Ljubljana, works on semantic description and discrimination of senses; she acknowledges that this is challenging but not menial work — what is monotonous is the transfer of such work into a database. Dr. Gantar quoted the subtitle of Michael Rundell's conference address (2009: 9): "First banish the drudgery ... then the drudges." She noted that the dictionary is no longer a book; users now expect much more than they did from the print dictionaries of the past. Web-based dictionaries can include lengthy semantic descriptions, grammar, examples, exercises, etymology, phraseology, and other types of information. This is logical: the space limitations of print dictionaries did not allow for all of these possibilities. Dr. Gantar is interested in the roles that automatization and crowd-sourcing play now and can play in the future in reducing the amount of drudgery in lexicography.

Iztok Kosem, a researcher at Trojina, the Institute for Applied Slovenian Studies, and at the University of Ljubljana, has had as his focus over the past five years how to get drudgery out of lexicographic work. He works on identifying the menial and routine tasks of lexicography in order to reduce them. He mentioned GDEX, "Good Dictionary Examples" (Sketch Engine | GDEX n.d.), an electronic tool that takes all available corpus examples and ranks their suitability for a specific meaning or sense according to predetermined criteria.¹ With the assistance of GDEX, for example, 300 concordance lines from a corpus could be reduced to only the twenty best contexts for the lexicographer to peruse, thus significantly reducing drudgery and saving time. Dr. Kosem con-

siders that the advent of GDEX is a big step forward in lexicographic work; as corpora have grown to a billion or more words, the problem of too many examples has become ever greater. The answers of Drs. Gantar and Kosem appear to contradict the prediction of Ladislav Zgusta: "The lexicographer has been called a harmless drudge by Dr. Johnson, and he will not advance to a harmless electrician" (1971: 357).

While our subjects had diverse views on exactly how much drudgery is involved in lexicographic work, there was consensus that they find their work extremely rewarding. Jerica Snoj commented that, in the course of the work the lexicographer reaches insights into the language that no one else has — because no one, not even well-educated native speakers, can see linguistic phenomena in quite the same way. And, these insights are what help one to endure. Dr. Snoj stated: "It is a gift for all your suffering but you must be serious in your work to get this satisfaction; otherwise, you can't reach this stage of insight and there will be only suffering! You must invest a lot to reach this satisfaction."

10. Harmless or harmful?

The *Merriam-Webster Unabridged* defines *harmless* as: "free of or lacking capacity or intent to injure : innocuous." Samuel Johnson, in his formulation "a harmless drudge," was making a statement about the lack of capacity of the dictionary writer to do harm. However, our Slovenian interviewees had clearly given extensive thought to whether the lexicographer has the potential to be *harmful*; or, in the words of the *Merriam-Webster Unabridged*: "damaging, troublesome, injurious." The interviewees were very concerned with what for them was the essential nature of their role in society — how the dictionary maker can be a force for good and avoid any potential for harm. For the three authors who undertook this interviewing research, this focus by the seven lexicographers on their ethical position was one of the most interesting findings. The sections below explore this topic in detail.

We discovered a variety of opinions among our interviewees concerning objectivity in lexicography and the relationship of objectivity to harm. Should the lexicographer be objective, describe the language and present it to the user as it is (so that users can evaluate the material and draw their own conclusions), or should the dictionary maker prescribe to users and guide them in what the lexicographer considers to be the best forms of expression in the language? While speaking about Malay dictionary work, Jacobson (1991: 214-215) frames the issue thus:

[There is] some doubt as to what actual role a dictionary should play. Should it be an instrument to prescribe a set of forms that is ruled as standard, correct, good or else should it be one that merely describes the forms frequently used and leaves it then up to the dictionary user to determine which choice is the appropriate one in light of the situation at hand?

On the other hand, Landau (1989: 32) questions whether there is room for doubt:

All dictionaries based on usage — and all competently done dictionaries must be based on usage — are descriptive. Prescription is impossible to distinguish from bias. Any preferred usage or condemnation of existing usage necessarily reflects the educational or cultural background of the editor Such judgments ... have no place in coloring definitions in a general dictionary any more than editorial opinions belong in straight news articles in the morning newspaper.

Jacobson (1991: 214-215) does not see dictionaries as being limited only to descriptivism — but even if they were, they would nevertheless exercise influence on the norms:

The words that appear in a dictionary represent the correct notations according to the standard norm at a given time and a given place. Therefore, the dictionary in question ... [becomes] the guide for the use of the language that is 'good' or correct. ... Usually, this norm will be accepted for its use if the dictionary is accepted as an authority. ... [Or] the dictionary is considered a recorder of the use of the language without making any judgment according to good or bad So, words, good or bad, need to be recorded. However, the dictionary will (still) become the standardizer of language.

Landau agrees that dictionaries have a standardizing role, whether their editors want them to or not. Despite the goal of objective description, dictionaries reflect "the views and prejudices of the established, well-educated, upper classes" (1989: 303). "One can no more pretend that dictionaries are culturally neutral than one can pretend that any other utilitarian object such as a door-knob or clothes hanger is culturally neutral and without any particular design" (1994: 39). In fact, dictionaries are "powerful forces for the preservation and dissemination of a distinctly cultivated form of expression" (1989: 303).

When our interviewee Apolonija Gantar was previously employed at the Fran Ramovš Institute, she was confronted regularly with the issue of objectivity, because one of her duties was working in the consulting service for the public. Dr. Gantar remembers that, even in those instances where she was not fully satisfied with an answer she provided, the users believed her due to their perception of her status. While Dr. Gantar considers that "people have to take responsibility for their own language and take part in the [lexicographic] decisions," she is aware that most "people don't want gray areas — they want a straightforward answer" as to whether something is "correct" or "incorrect."

Interviewee Nina Ledinek considers that people often consult the dictionary to see what is "right" (even though linguists do not want to encourage this attitude). Another interviewee sees users as going to the lexicographer for a "definite," "black and white" answer. This is the tension inherent in lexicographic work, a tension apparent both to the interviewees and to their colleagues outside of Slovenia. While the users want a dictionary that guides them, lexicographers cannot move away from objective description. Moving

toward prescription risks failing to depict how most people actually talk and write, which would result in dictionaries of no use and with no credibility or authority.

Iztok Kosem advocates for an objective approach to lexicographic work. He does not see lexicographers as harmless but as individuals with power whose responsibility to the user can be abused. Dr. Kosem views the lexicographer as a mediator between all the complexity of language and the final explanation that appears in the dictionary. This mediating role can be quite influential: If a word does not appear in the dictionary, users might believe that it does not exist at all, or they might be suspicious of it. They might also be suspicious of the dictionary because it omits a word they like — and then they would just go to Google. From Dr. Kosem's perspective, lexicographers have a duty *not* to be prescriptive. It is the description that really matters, finding the relevant information (evidence) for the users and delivering it quickly to them.

Nataša Jakop is also an advocate for a more descriptive approach. She considers that, as a single individual, the lexicographer is invisible and harmless, but in order to avoid becoming harmful, lexicographers must be objective; they must forget about beliefs and feelings and consider the linguistic material as objectively as the biologist looks at insects. If lexicographers cannot do this and insert their own [prescriptive] views, especially without looking at the linguistic material, then they would become harmful.

Apolonija Gantar noted that while there is no single objective interpretation of what a language is, nevertheless the lexicographer must still strive toward objectivity. A well-developed initial plan and conceptualization of the dictionary to be compiled can contribute to the overall objectivity of the final work. On the other hand, a too-rigid adherence to an initial plan could be harmful, if some specific set of objective data indicates that you need to do things differently. An example of this, according to Dr. Gantar, is the treatment of gerundial forms in the first edition of the *Dictionary of Standard Slovenian*. There, gerunds were only described syntactically, with no accompanying lexical explanation. The editors at the Academy of Sciences realized it was a mistake but did not change it, despite the fact that some gerunds in Slovenian are not semantically linked to the verb of origin in a transparent manner, so that a strictly syntactic definition will be obscure. [For example: *skakanje*: *glagolnik od skakati* 'a gerund from [the verb] *to jump*' instead of: 'the process of jumping; a gerund from [the verb] *to jump*'. Dr. Gantar's comment shows that the goals of objectivity and descriptive accuracy, despite the lexicographers' best intentions, can be quite elusive.

While Nina Ledinek, like Nataša Jakop, considers that lexicographers are not visible, she emphasizes that they must be socially responsible and sensitive to the different groups in society: Just this, the fact that they must demonstrate sensitivity, shows that lexicographers do play a significant role. Dr. Ledinek maintains that the *Dictionary of Standard Slovenian* does and should have a normative value; their language has connected Slovenians throughout their history — a

history which until recently has always been that of a minority people surviving in larger regimes. Dr. Ledinek's comments bring home the descriptive challenge posed by a language like Slovenian with only two million speakers; while objectivity is still very much in the focus of Slovenian lexicographers, they also must consider the role of their language very differently than would any lexicographer of English. Anita Srebnik notes that other languages bring the outside world to Slovenia and allow Slovenians to communicate when they cross any border. Slovenia might be small but it cannot live without exchange, and an asset of its people is the ability to learn other languages well. Her comments bring to light the important relationship of Slovenian to other languages, as depicted in its bilingual dictionaries. Bilingual lexicography takes on a special significance in the case of such a (relatively) small language.

Dr. Srebnik finds it deplorable that the public regards only some dictionaries as conveyers of the norm, as authorities on the language. For the Slovenian media, she maintains, this authority only accrues to the work of the Academy of Sciences, when in reality there are many other worthy and authoritative projects. In her eyes, it is the media (rather than the lexicographer) that causes harm because it limits the focus — and attributes the power and authority — to a small number of lexicographers and projects. In particular, Dr. Srebnik faults the lack of status and authority for bilingual lexicography; in reality, bilingual lexicographers treat not just equivalence in two languages but also connotation and cultural differences. Dr. Srebnik's point about the societal status of bilingual dictionaries highlights something that is often overlooked: It is not only monolingual, but also bilingual dictionaries that have a role in the maintenance of the norm, and the power to do (or not do) harm.

Mojca Žagar Karer, the sole terminographer of our study, sharply distinguishes her practice from that of lexicographers and has a very different take on the whole notion of objectivity. For Dr. Žagar Karer, it is clear: Lexicography is more subjective and therefore might not be harmless. Because lexicographers write definitions and analyze meaning themselves, they are subjective; in other words, definition writing and meaning analysis, as non-descriptive activities, have a potential for harm. Terminographers, in her perspective, must be objective because they must be credible for the subject field and for the society. They are trying to create quality language resources which are useful for translators, language editors, and others. As was mentioned, Dr. Žagar Karer's work role is closer to that of an editor than a lexicographer, in that she gathers the terminological definitions written by specialists in a given field and edits toward reaching consensus among those she consults. While Dr. Žagar Karer's perception of objectivity is reasonable, in the case of terminography, the "burden" of objectivity does not disappear but is simply transferred from the terminographer/editor to those field specialists who actually write the definitions. It is reasonable to suppose that, given their lack of lexicographic experience, some field specialists do inadvertently bring their personal beliefs, perceptions, and prescriptive ideas to definition writing, what for them is a relatively new

endeavor. If two field specialists were to disagree about which of two terms is the best to designate a given concept, then certainly we would have two persons striving toward objectivity of description who come up with different results.

Jerica Snoj stressed that, regardless of how they are regarded (or ignored), lexicographers are very important for the society; their dictionaries bring the description of language to users, thereby helping users to express their thoughts in an appropriate way. When a new dictionary appears, a new insight into the language is opened up. Dr. Snoj considers that the dictionary has a very important role in exploring the possibilities of a language; Nataša Jakop cites the significant role it plays in the preservation of cultural heritage. Dr. Jakop's point is of special significance for the lexicography of any language with a relatively small number of speakers: Preservation for such languages is crucial.

Whether visible or invisible, whether harmless, whether a drudge, the lexicographer is *the* source of insight into a given language. The responsibility to provide these insights to users in the most ethical way possible is something that all of our interviewees agree on.

11. Conclusions

It has been more than 260 years since Samuel Johnson defined *lexicographer* as a "harmless drudge." Our interviews with seven working Slovenian lexicographers reveal many opinions on the viability of his definition today, and the insights of these interviewees are significant for the development of lexicographic theory broadly construed. The Slovenian lexicographers, all distinguished and experienced modern practitioners, accept some implications of Johnson's metaphor while they categorically reject others. First, they certainly acknowledge that some aspects of their work can be tedious, despite the more pervasive use of technology today. While their strong commitment and their focus on the end result of lexicographic endeavor allow them to accept drudgery as part of the picture, the interviewees are acutely aware that repetitive work has pitfalls, such as the possibility for attention to wane and mistakes to be introduced. Because of the potential deleterious effects of monotony on the quality of final lexicographic products, some of the interviewees actively work toward the development of new technologies to replace the hard, repetitive and routine lexicographic work that is still done by people.

The Johnsonian notion of "harmless drudge" contains not just tedium but also anonymity. Slovenian lexicographers know that the dictionary maker usually labors in isolation, unknown to the public. What is of more concern to our interviewees than anonymity is the lack of understanding in the public of what the lexicographer actually does. The lack of public awareness can contribute to an overestimation of the lexicographer's authority, which in turn may lead to the disengagement of the public from interest in the Slovenian language. After all, if it is only the lexicographers who know the language, then there is noth-

ing for the educated language user to think about or do except follow the "advice" that (they think) the dictionary is trying to give. Conversely, as the bilingual lexicographer in the group of interviewees pointed out, a lack of public awareness can undermine the valuing of dictionary work by the media or by society at large — to the detriment of production of sorely needed bilingual and monolingual dictionaries.

While they concede the reality of problems engendered by drudgery and anonymity, the Slovenian lexicographers interviewed would reject outright the idea that the dictionary writer is *a priori* "harmless." Because the interviewees have reflected extensively on the social implications of their profession, they perceive many possibilities for harm and are motivated to avoid it. It is the ethical responsibility of the lexicographer to the dictionary user that is the most important preventative of harm. If a lexicographer were to ignore or misrepresent language facts as represented in a corpus or other lexicographic source and veer away from linguistic description, this imposition of personal bias would most certainly be socially harmful.

The serious discussion engaged in during this study by the seven Slovenian specialists should not leave the reader with the impression that for them, lexicography is a grim and onerous business; quite the contrary. Certainly, as one interviewee put it, lexicography requires a tremendous persistence because, despite constantly improving facilities and research tools, there is still a lot of menial work. Surely, media portrayals and the society's general misapprehensions about what lexicography is complicate the already-challenging work of linguistic description. Nevertheless, the six Slovenian lexicographers and one terminographer spoke frequently about "satisfaction": the satisfaction of gaining real insight into the language, the satisfaction of meeting the language needs of the users, and the satisfaction of helping users to engage more fully with a language that is such an important part of Slovenian identity.

Endnote

1. For more on how Sketch Engine | GDEX works and what makes for a good corpus example for lexicography, see Kilgarriff et al. (2008: 426): Examples of criteria mentioned are typicality — an example should exhibit "frequent and well dispersed patterns of usage;" informativeness — the example should "elucidate the definition;" intelligibility — the example should avoid "difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context." See also Atkins and Rundell (2008: 458-461).

Acknowledgements

This study could not have taken place without the full cooperation and consummate generosity of our seven Slovenian lexicographer-interviewees. The

participants were candid, forthcoming, reflective, insightful. Interviewing them was not like work but was a pleasure. Some of the busiest people in Slovenia — its lexicographers — took the time to move our research forward. As this study reveals, they bring to their practical lexicographic tasks a highly developed sense of ethics and responsibility. Now, by participating in this study, they bring the same sense of duty to furthering the theoretical development of lexicography internationally. Many thanks to: Apolonija Gantar, Nataša Jakop, Iztok Kosem, Nina Ledinek, Jerica Snoj, Anita Srebnik, and Mojca Žagar Karer! We would also like to thank Marko Snoj, the director of the Fran Ramovš Institute of the Slovenian Academy of Sciences, for welcoming us there. We hope to see you again in the future.

The authors acknowledge the project, Lexicographic exchange as a way of building bridges between Slovenian and American lexicographic philosophy, governing principles, goals, and work tools, No. BI-US/16-17-053, which was financially supported by the Slovenian Research Agency. They also acknowledge the approval (20 February 2017) of the New Jersey City University (NJCU) Institutional Review Board for the Protection of Human Participants in Research. Donna Farina thanks NJCU for travel support to Ljubljana, Slovenia, as well as released time support from its Separately Budgeted Research program. The authors thank NJCU, in particular Tamara Cunningham, Assistant Vice President for Global Initiatives, for providing housing and hospitality to Alenka Vrbinc and Marjeta Vrbinc during their research visit to the United States.

References

- Algeo, John.** 1985. Harmless Drudgery. *American Speech* 60(4): 357-361.
- Ammon, Ulrich and Marlis Hellinger (Eds.).** 1991. *Status Change of Languages*. Berlin/New York: Walter de Gruyter.
- Atkins, Beryl T. Sue and Michael Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Bajec, Anton et al.** 1970. *Slovar slovenskega knjižnega jezika* [Dictionary of Standard Slovenian]. Vol. 1, A-H. Ljubljana: Državna založba Slovenije.
- Béjoint, Henri.** 2010. *The Lexicography of English*. Oxford: Oxford University Press.
- Béjoint, Henri.** 2016. Dictionaries for General Users: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 7-24.
- Bernal, Elisenda and Janet DeCesaris (Eds.).** 2008. *Proceedings of the XIII EURALEX International Congress, Barcelona, 15–19 July 2008*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.
- Cowie, Anthony P. (Ed.).** 2009. *The Oxford History of English Lexicography. Volume I: General-purpose Dictionaries*. Oxford: Oxford University Press.
- Durkin, Philip (Ed.).** 2016. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press. [Online version.]

- Farina, Donna M.T.Cr. and George Durman.** 2009. Bilingual Dictionaries of English and Russian in the Eighteenth to the Twentieth Centuries. Cowie, Anthony P. (Ed.). 2009: 105-126.
- Farina, Donna M.T.Cr. and George Durman.** 2012. "Academic Hooliganism" or "False Gold"? The Reception of Baudouin de Courtenay's Russian Dictionary. *Dictionaries: Journal of the Dictionary Society of North America* 33: 1-41.
- Fontenelle, Thierry.** 2016. Bilingual Dictionaries: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 44-61.
- Gantar, Polona.** 2015. *Leksikografski opis slovenščine v digitalnem okolju* [Lexicographic Description of Slovenian in a Digital Environment]. (Zbirka Sporazumevanje). First edition, e-publication. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gliha Komac, Nataša, Nataša Jakop, Janoš Ježovnik, Simona Klemenčič, Domen Krvina, Nina Ledinek, Tanja Mirtič, Andrej Perdih, Špela Petric, Marko Snoj and Andreja Žele (Eds.).** 2015. *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika* [Preliminary Conceptualization of a New Explanatory Dictionary of Standard Slovenian]. Različica 1.1. Ljubljana: Inštitut za slovenski jezik Frana Ramovša; Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (Eds.).** 2015. *Slovar sodobne slovenščine: problemi in rešitve* [Dictionary of Modern Slovene: Problems and Solutions]. (Zbirka Prevodoslovje in uporabno jezikoslovje). First edition. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (Eds.).** 2017. *Dictionary of Modern Slovene: Problems and solutions.* (Book series Prevodoslovje in uporabno jezikoslovje). First edition, e-edition. Ljubljana: Ljubljana University Press, Faculty of Arts. [This is the translation of Gorjanc et al. 2015.]
- Granger, Sylviane and Magali Paquot (Eds.).** 2009. *eLex2009: eLexicography in the 21st Century: New Challenges, New Applications.* Conference abstracts of eLex2009, 22–24 October 2009. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université de catholique de Louvain.
- Harmful.** n.d. *Merriam-Webster Unabridged* [online]. Retrieved from <http://unabridged.merriam-webster.com/unabridged/harmful>.
- Harmless.** n.d. *Merriam-Webster Unabridged* [online]. Retrieved from <http://unabridged.merriam-webster.com/unabridged/harmless>.
- Jackson, Howard.** 2002. *Lexicography: An Introduction.* London/New York: Routledge.
- Jacobson, Rodolfo.** 1991. In Search of Status: Bahasa Malaysia for National Unification. Ammon, Ulrich and Marlis Hellinger (Eds.). 1991: 200-226.
- Karba, Rihard, Gorazd Karer, Juš Kocijan, Tadej Bajd, Mojca Žagar Karer and Tanja Fajfar (Eds.).** 2014. *Terminološki slovar avtomatike* [Dictionary of Automated Control Systems and Robotics]. Zbirka Slovarji. Ljubljana: Založba ZRC.
- Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý.** 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Bernal, Elisenda and Janet DeCesaris (Eds.). 2008: 425-432.
- Landau, Sidney I.** 1989. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.
- Landau, Sidney I.** 1994. The Expression of Changing Social Values in Dictionaries: Focus on Family Relationships. Little, Greta D. and Michael Montgomery (Eds.). 1994: 32-39.

- Landau, Sidney I.** 2001. *Dictionaries: The Art and Craft of Lexicography*. Second Edition. New York/ Cambridge: Cambridge University Press.
- Ledinek, Nina, Kozma Ahačič and Andrej Perdih.** 2015. *Fran: slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, Vodnik* [A Guide to Fran: The Dictionaries from the Fran Ramovš Institute of the Slovenian Language in the Research Center of the Slovenian Academy of Sciences and Arts]. (Zbirka Fran). Različica 1.0. Ljubljana: Založba ZRC.
- Little, Greta D. and Michael Montgomery (Eds.).** 1994. *Centennial Usage Studies*. (Publication of the American Dialect Society 78.) Tuscaloosa: University of Alabama Press.
- Rundell, Michael.** 2009. The Road to Automated Lexicography: First Banish the Drudgery ... Then the Drudges? Granger, Sylviane and Magali Paquot (Eds.). 2009: 9-10.
- Schäfer, Jürgen.** 1984. The History of Ideas and Cross-Referencing in the Future EMED. *Dictionaries: Journal of the Dictionary Society of North America* 6: 182-198.
- Shapiro, Rebecca (Ed.).** 2017. *Fixing Babel: An Historical Anthology of Applied Lexicography*. Lewisburg: Bucknell University Press.
- Sharp, Gene.** 2012. *Sharp's Dictionary of Power and Struggle: Language of Civil Resistance in Conflicts*. Oxford: Oxford University Press.
- Sketch Engine | GDEX.** n.d. <https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/gdex>.
- Snoj, Jerica.** 2004. *Tipologija slovarske večpomenskosti slovenskih samostalnikov* [The Lexicographic Treatment of Polysemous Nouns in Slovenian]. (Zbirka Linguistica et philologica). Ljubljana: Založba ZRC, ZRC SAZU.
- Snoj, Jerica, Martin Ahlin, Branka Lazar and Zvonka Praznik.** 2016. *Sinonimni slovar slovenskega jezika* [Dictionary of Slovenian Synonyms]. First edition. Ljubljana: Založba ZRC.
- Sokolowski, Peter.** 2014. The Dictionary as Data. *Dictionaries: Journal of the Dictionary Society of North America* 35: 287-298.
- Srebnik, Anita.** 2006. *Slovensko-nizozemski evropski slovar* [Slovenian–Dutch European Dictionary]. (Zbirka Evropski slovarji). Ljubljana: Cankarjeva založba.
- Srebnik, Anita.** 2007. *Nizozemsko slovenski slovar = Nederlands Sloveens woordenboek* [Dutch–Slovenian Dictionary]. (Slovarji DZS). First edition. Ljubljana: DZS.
- Srebnik, Anita.** 2015. *Jezikovnotehnološki postopek obračanja dvojezičnih slovarjev* [The Technology and Linguistics behind the Process of Reversing Bilingual Dictionaries]. Praha: Verbum.
- Stamper, Kory.** 2017. *Word by Word: The Secret Life of Dictionaries*. New York: Pantheon Books.
- Sušec Michieli, Barbara, Marjeta Humar, Katarina Podbevšek, Slavka Lokar, Edi Majaron, Viktor Molka, Janko Moder, Miran Herzog, Ana Kocjančič and Mojca Žagar Karer (Eds.).** 2007. *Gledališki terminološki slovar* [Dictionary of Theatre Terms]. Zbirka Slovarji. Ljubljana: Založba ZRC.
- Toporišič, Jože, Franc Jakopin, Janko Moder, Janez Dular, Stane Suhadolnik, Janez Menart, Breda Pogorelec, Kajetan Gantar, Martin Ahlin and Milena Hajnšek-Holz (Eds.).** 2001. *Slovenski pravopis* [Slovenian Orthography]. Ljubljana: Založba ZRC.
- Žagar Karer, Mojca.** 2011. *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije* [Terminology from the Text to the Dictionary: Analysis of Electro-technical Terminology]. (Zbirka Linguistica et philologica 26). Ljubljana: Založba ZRC, ZRC SAZU.
- Zgusta, Ladislav.** 1971. *Manual of Lexicography*. Prague: Academia / The Hague/Paris: Mouton.

Appendix

Interview script

Beginning of interview

We want to thank you very kindly for agreeing to work with us on this project. Our working title is: "Slovenian Lexicographers at Work." Our goal is to add to the worldwide understanding of what lexicographic work is by focusing on work in this country. We consider that the practices in Slovenia should be known and will prove relevant to lexicographers everywhere.

As indicated by the statement you signed, your remarks are not anonymous; we would like to mention you by name and highlight your ideas in any resulting publications. But, on the other hand, if any specific remark you make is not one that you want attributed to you by name, just tell us that it is "off the record." In that case, we would quote you or cite you generally, using language such as: "Some of our interviewees considered that"

Questions

1. First of all, can you tell us a little bit about yourself? Why were you attracted to the field of lexicography? How did you end up doing what you do today?
2. Can you describe your daily work as a lexicographer? What are the main activities that you do on a daily, weekly, or monthly basis? What aspects of your work do you like best?
3. The famous English lexicographer, Samuel Johnson, defined the word *lexicographer* thus, in 1755: "a writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words."
 - a. We would like to know, first: What elements of your own work do you consider "drudgery:" hard, menial, or monotonous work?
 - b. Second, do you think the lexicographer is "harmless?" Does he or she play an invisible, unnoticed social role, or the opposite? How are lexicographers significant to the society of which they are a part?
4. What is the philosophical and theoretical framework that governs your work? In other words, what is the "umbrella" of ideas under which you do everything that you do?

(Follow-up to Question 4, if needed: What are the "big" ideas that influence how you go about your habitual work as a lexicographer?)

5. Can you explain what are the two or three driving principles that govern your work as a lexicographer? How do you think about these principles as you engage in the minute tasks which lexicographers of necessity must perform?
6. The two previous questions tried to understand more the theoretical and philosophical basis for your lexicographic work. Now we wish to ask: Can you name any theories or practices used in other countries, including the U.S., that inform your own lexicographic work? Or, perhaps when you formulated the principles of your work you incorporated some ideas from abroad?
7. Related to the previous question, have you joined any lexicographic organizations such as the Dictionary Society of North America or EURALEX? Do your memberships of this type affect your work? How?
8. Can you describe two or three of the current projects that you are involved with? We are looking to describe, as completely as possible, what is going on today in Slovenian lexicography. We are also very interested in any future projects that are in the planning stages.
9. In recent years, what are the most noteworthy accomplishments in the work of you and your immediate colleagues?
10. It goes without saying that lexicographic work takes place in the real world and is subject to the usual constraints and challenges of any practical work. In particular, there are always budgetary constraints, but not only budgetary. We would like to know: How is your work challenged by a variety of circumstances; what are the challenges and constraints?
11. Can you name the major strengths of your work situation? What is a best practice for you and your colleagues (e.g., access to different information/sources, user-friendly dictionary-making software, cooperation with IT specialists and/or corpus linguists and/or experts from other fields, etc.)? What affects most positively the compilation of your dictionaries?
12. If you could change one thing about the circumstances of your lexicographic work, what would it be? If you could change one feature of the lexicographic philosophy/theory that underpins your work, what would it be?
13. Could you offer us some suggestions? How do you think the cooperation and exchange of ideas between Slovenian and American lexicographers can be encouraged? Do you consider that more cooperation would improve lexicographic work in Slovenia, the U.S., and beyond?

Towards Chinese Learner's Dictionaries for Foreigners Living in China: Some Problems Related to Lemma Selection

Mei Xue, *Department of Foreign Language,
China University of Mining and Technology (Beijing), China;
and Centre for Lexicography, University of Aarhus, Denmark
(meix99@yahoo.com)*

and

Sven Tarp, *Sino-Danish Sindberg Centre of Lexicography,
Translation and Business Communication, Guangdong University
of Finance, China; International Centre for Lexicography,
Universidad de Valladolid, Spain; Department of Afrikaans and Dutch,
University of Stellenbosch, South Africa; and Centre for Lexicography,
University of Aarhus, Denmark (st@cc.au.dk)*

Abstract: During the past decades, various dictionaries for foreign learners of Chinese have seen the light. Except for one picture dictionary which is almost completely ignored in the academic literature, none of these dictionaries has taken into account the special needs which foreigners living in China and learning Chinese may have. This contribution will discuss these needs with special focus on lemma selection. We argue that foreigners living in China, in order to meet their lexicographical needs, require additional words typically occurring in social contexts in which they often find themselves, whether or not these words have a high corpus-frequency. As a solution we therefore recommend a set of selection criteria that combines corpus frequency and context relevance. Finally, we discuss how logfiles reflecting user behaviour can be used as a new and very reliable empirical source for lemma selection for an online Chinese learner's dictionary.

Keywords: CHINESE LEARNER'S DICTIONARIES, LEMMA SELECTION, SOCIAL CONTEXTS, CORPUS FREQUENCY, CONTEXT RELEVANCE

Opsomming: Op weg na Chinese aanleerderswoordeboeke vir buitelanders wat in China woon: Enkele probleme verwant aan lemmaseleksie. Gedurende die afgelope dekades het verskeie woordeboeke vir vreemdetallearders van Chinees verskyn. Buiten een prentewoordeboek wat byna heeltemal in die akademiese literatuur geïgnoreer is, het geeneen van hierdie woordeboeke die spesiale behoeftes wat buitelanders wat in China woon en Chinees aanleer, mag hê, in ag geneem nie. In hierdie artikel word hierdie behoeftes, met spesiale fokus op lemmaseleksie, bespreek. Ons argumenteer dat buitelanders wat in China woon, addisionele woorde benodig wat tipies voorkom in sosiale kontekste waarin hulle hulself dikwels bevind, ongeag of hierdie woorde 'n hoë korpusfrekwensie het of nie. As oplossing hiervoor beveel

ons 'n stel seleksiekriteria aan wat korpusfrekwensie en konteksrelevansie kombineer. Laastens bespreek ons hoe loglêers wat gebruikersgedrag weerspieël, as 'n nuwe en baie betroubare empiriese bron vir lemmaseleksie vir 'n aanlyn Chinese aanleerderswoordeboek gebruik kan word.

Sleutelwoorde: CHINESE AANLEERDERSWOORDEBOEKE, LEMMASELEKSIE, SOSIALE KONTEKSTE, KORPUSFREKWENSIE, KONTEKSRELEVANSIE

1. Introduction

The last three decades have witnessed a rapidly increasing worldwide interest in learning Chinese. According to the recent statistics issued by Confucius Institute Headquarters (Hanban), as many as 100 million people around the world were learning Chinese as a foreign or second language by 2015 (Yang and Zhang 2017). This growing population of non-native learners of Chinese cannot be seen isolated from China's increased role and projection in the world. But apart from that, there may be many specific reasons why people decide to learn Chinese. They may have Chinese ancestors and aspire to re-establish the relations with their roots. They may live next to a Chinese speaking community in their own country and need Chinese as a means of communication. They may want to study Chinese because they plan to visit China and are interested in its rich history and culture. In 2014, about 26 million foreigners visited China (Liu 2016). Finally, some foreigners may already, for various reasons such as business, study or family, live in China for a shorter or longer period. In fact, Song (2013) estimates that in 2013 there were several million foreigners who were either registered as foreign residents or were staying in China, a number which will probably grow in the nearby future. Most of these people may wish to learn Chinese in order to manage in their daily and professional life.

All non-native learners of Chinese may demand specially designed dictionaries to assist the learning process, but their needs and expectations may not be exactly the same when they are living inside versus outside China. In this article, we will argue that this is a distinction to which too little attention has been paid, especially in practical dictionary making. We will therefore discuss Chinese learner's dictionaries with special focus on the needs which foreigners living in China may experience in terms of the required lemmata. As a conclusion, we will present some principles that can guide the selection of lemmata in an *online Chinese learner's dictionary* that takes advantage of the available technology. However, in order to put the discussion into perspective we will start with a brief excursion into the Western tradition of learner's dictionaries.

2. The western tradition

The term *learner's lexicography* was coined in Britain as a direct result of the

pioneering work of H. Palmer, M. West and A.S. Hornby and the publication of the first dictionaries specifically designed to assist foreign learners of English, cf. Cowie (1999). With the gradual development of English as a lingua franca in a large part of the world in the years following the Second World War, the monolingual English learner's dictionaries almost obtained a cult status and strongly influenced the making of similar dictionaries elsewhere in the world. In this period, learner's dictionaries saw the light in countries like Germany, France and Spain; cf. Hernández (1989), Zöfgen (1994), Wiegand (1998), Welker (2008), among others. The languages spoken in these three countries are all big languages in terms of the number of native speakers as well as the hundreds of thousands, if not millions, of learners interested in studying them not only inside but also outside the geographical areas where they are traditionally spoken. Most foreign learners of English, for instance, are studying this second language in their native countries. This situation differs dramatically from the situation in other European countries, like the Scandinavian countries, with a relatively small number of native speakers. In these countries, only a limited number of foreign learners have shown interest in learning the respective languages beyond their national borders, probably due to their limited communicative value at an international level. This also influenced the lexicographical terminology used in these countries. In Denmark, for instance, during many years there was no Danish equivalent to the English *learner's dictionary* until it was coined by Tarp (1999) and even today this new term (*lørnerordbog*) has not been used in any dictionary title. However, with the massive immigration of foreign workers starting in the 1960s, and later the arrival of many refugees escaping endless wars and natural disasters, all of them in need of learning the language used in their new country, a new type of dictionary began to see the light. This development resulted in a new and much more successful term being spontaneously coined, namely *immigrant's dictionary*, cf. Pálfi and Tarp (2009).

The immigrant's dictionary can be defined as a variant or subtype of learner's dictionaries specially adapted to the needs of immigrants and refugees living inside the geographical area where their new second language is spoken. It differs in various ways from the British *Big Five*, i.e. the prestigious learner's dictionaries published by Oxford, Collins Cobuild, Macmillan, Longman and Cambridge. Most immigrant's dictionaries are bilingual, either monoscopal or biscopal. The most emblematic of these dictionary projects is undoubtedly the Swedish *Lexins Svenska Lexicon* which is available both on paper (in a series of bilingual dictionaries) and online where it currently can be accessed from 20 different languages representing the biggest foreign language communities in Sweden, cf. Gellerstam (1999) and Hult (2016). The number of L2 lemmata in immigrant's dictionaries varies considerably but is generally much smaller than the ones treated in the monolingual English dictionaries mentioned above. These lemmata have frequently been selected according to criteria taking into account the very specific needs of the immigrants in their new life. In this regard, an immigrant's dictionary published in Spain in 2011

with the title *Bienvenidos* (Welcome) describes itself as the immigrants' and refugees' "first Spanish dictionary" and writes the following in its Introduction:

It contains about 3 000 frequently used words that have been selected from a set of communicative situations that intend to cover the needs of daily life and to assist the development of the new speakers' linguistic competence (understanding and expressing themselves) and, in this way, to facilitate their full integration into social, work and family life (Martín 2011: ix).

The Spanish immigrant's dictionary is monolingual which, of course, limits its value, but it is worth noting that half of the dictionary consists of thematic tables illustrating the communicative situations mentioned, whereas the other half is a traditional alphabetically structured wordlist with definitions of each word. In this way it can be accessed both through the wordlist and the thematic tables. There is little doubt that this design makes it highly useful to most learners of Spanish at the very beginner's level, especially if it is used in combination with a Spanish language course. However, the limited vocabulary (as well as the title) suggests that its usefulness will be reduced proportionally with a growing proficiency level, and that it, after a few months, will have to be replaced by another type of learner's dictionary, preferable a bilingual one as argued by Lew and Adamska-Salaciak (2015).

3. The discussion on learner's dictionaries in China

With a very few exceptions, the Chinese tradition of making dictionaries for foreign learners of Chinese started in the late 1970s. These dictionaries were, as a rule, based upon independent reflections by Chinese lexicographers and scholars, and they were only to a limited degree influenced by the traditions in other countries. If the increasing Chinese learning population all around the world is taken into consideration, it could be argued that the number of Chinese learner's dictionaries designed to serve foreign learners is rather limited. Wei, Geng and Wang (2014) have examined the dictionaries published in China from 1978 to 2008 and conclude that there are 21 Chinese learner's dictionaries for foreign learners. However, a study conducted by Wei and An (2014: 71-72) shows that about 45 Chinese dictionaries for foreign learners have been produced since 1980. Among these dictionaries can be mentioned the one helping foreigners with the Chinese Proficiency Test (HSK) (Liu 2000), the one illustrating the use of Chinese function words (Lü 1980), and the ones helping intermediate foreign learners learn Chinese (Lu and Lü 2006a; Shi and Wang 2011; Zheng 2009). These dictionaries are either monolingual or bilingual/bilingualized, most of the latter with English as the auxiliary language (Zhang 2010: 33). Despite their respective focus and characteristics, these dictionaries are all aimed at assisting foreigners in learning Chinese.

In this light, it is rather awkward that a number of empirical studies reveal that foreign learners inside and outside China are generally not aware of the

existence of many of the Chinese learner's dictionaries published in China and, hence, seldom use them (Liu 2014; Hao and Wang 2013; Xie and Li 2012; Yang 2015). Many of these dictionaries gather dust in libraries and are mainly used for research purposes (Jin 2015; Liu 2014). A study carried out by Yang (2015) shows that only about 9 percent of the foreign learners of Chinese, even the ones studying in China, use Chinese learner's dictionaries published in China. On the other hand, nearly 95 percent of the Chinese learners of English are using one of the *Big Five* British learner's dictionaries (Liu 2014). Facing such an unfortunate status with regard to learner's dictionaries, Lu and Lü (2006b) criticize that many of the so-called Chinese learner's dictionaries are nothing but the reduced versions of distinguished dictionaries designed for native Chinese, like the *Xinhua Dictionary* and *Modern Chinese Dictionary*. Such criticism has been echoed by many Chinese lexicographers (Cai 2011; Li 2013; Liu 2014; Wang 2009; Yang 2016).

The sharp contrast between the status of Chinese and English learner's dictionaries has spurred wide discussion among Chinese lexicographers on the concept, design and principles to be used in the compilation of learner's dictionaries that are specifically aimed at foreign learners of Chinese (Jin 2015; Wang 2009; Yang 2016). In an overview of the studies conducted into Chinese learner's dictionaries between 1984 and 2013, Jin (2015: 35) concludes that during the past 30 years the research has mainly focused on describing and evaluating dictionary articles from the point of view of linguistics, although there is a certain tendency to shift the focus to their usefulness in terms of the target users' needs. There is, however, a manifest lack of research on the actual needs of foreign learners as dictionary users and on the integration of modern information technology into the conception and compilation of Chinese learner's dictionaries (Jin 2015: 36). In a situation where there is an increased focus on dictionary users and where the English learner's dictionaries are rapidly moving from the printed to the digital media (Rundell 2015), Chinese learner's dictionaries to a great extent still stay in a comfort zone.

A few studies have touched upon the issue of foreign learners' actual needs in the process of planning and compiling learner's dictionaries, but without providing further specifications. Zheng (2004: 92-93) points out that it is necessary to make a distinction between the foreign learners of Chinese living in China and those who study this language in other geographic areas of the world. However, he does not explicitly elaborate on the different lexicographical needs which these two groups of foreign learners may experience, or how dictionaries should respond to such needs. Yang (2016) proposes four principles to guide the compilation of a Chinese learner's dictionary for foreign learners, namely intelligibility, utility, comprehensiveness, and explicitness. These principles, which are formulated at a very high level of abstraction, involve all the data selected for the planned dictionary and require reasoning and step-wise specifications in theory and practice. It is always easy to state that foreign learners' needs should be attended to in academic research, but

fragmented suggestions and ideas that are too abstract are not sufficient to plan a modern high-quality learner's dictionary in the real sense of the word. Efforts should rather be made to develop a *coherent framework* which can guarantee the production of a learner's dictionary that responds to the actual needs and expectations of the foreseen dictionary users. It is, hence, imperative to have a clear understanding of the concept of a learner's dictionary in terms of foreign language learning.

4. An unnoticed dictionary: *My Chinese Picture Dictionary*

We will now have a brief look at a dictionary that was published in 2008, namely *My Chinese Picture Dictionary* (Wu 2008). In the scholarly discussion of the principles that should guide the design of Chinese dictionaries for foreign learners, this dictionary goes almost unnoticed. It is, for instance, not mentioned by Wei et al. (2014) who claim to offer the most comprehensive overview of dictionaries published in China between 1978 and 2008, and neither is it included in a recent overview study conducted by Wei and An (2014). This rather unnoticed existence is surprising, inasmuch as the dictionary reflects a new and different approach to Chinese learners' lexicography.

My Chinese Picture Dictionary consists of a thematic section which makes up the bulk of the dictionary as well as two indexes in English and Chinese Pinyin, respectively. The vocabulary treated is structured in 15 main themes, each further subdivided into a number of topical units. There are a total of 142 such units which are all represented in graphic tables covering various aspects of daily life such as personal information, family, school, work, shopping, dining, hospital, transportation and travel, etc. (Figure 1 provides an example of how the thematic units are represented in the dictionary). Each of the figures consists of an illustration where the words representing the different phenomena are written with Chinese characters as well as in Chinese Pinyin and English. The figures can be accessed either through the list of thematic content in the front matter of the dictionary or through one of the two indexes in the back matter where the English and Chinese Pinyin words treated in the figures are organized alphabetically.

As can be seen, in its overall structure *My Chinese Picture Dictionary* has many similarities with the Spanish immigrant's dictionary *Bienvenidos* mentioned above, as both make an endeavour to serve foreign learners' actual needs in various social situations in which they may find themselves in China or Spain. However, compared to the wordlist in its Spanish counterpart, the two wordlists, or indexes, in *My Chinese Picture Dictionary* are much more primitive in the sense that they do not offer definitions or any kind of grammatical data, not even part of speech. In addition, the overwhelming majority of selected words are nouns whereas there are few verbs and adjectives and no function words. It goes without saying that these problems, and others which will be identified in the following discussion, reduce its usefulness for foreigners

living in China despite its innovative approach to Chinese learners' lexicography.

5. The concept of a learner's dictionary

Foreign or second language (L2) learning is a complex process and learner's dictionaries are conceived to assist learners in different situations or contexts related to this learning process. The situations in which foreign learners turn to dictionaries are generally of a communicative or cognitive character as defined by the lexicographical *Function Theory* which will constitute the theoretical framework for the following reflections, cf. Tarp (2008). Communicative situations include *text reception* in L2, *text production* in L2, as well as *translation* from L1 into L2 or vice versa, whereas the study and assimilation of L2 *vocabulary* or *grammar* are the most relevant cognitive situations. These learning situations may vary according to the chosen didactic methods. As stated by Tarp (2004), the great challenge to learner's lexicography is to conceive and compile dictionaries that can assist learners in as many aspects of the language learning process as possible. Hence, users' needs, which may occur in the specific types of user situation, should be the starting point for learner's lexicography.

The focus on foreign learners' needs is time-honoured in learner's lexicography. Learner's lexicographic needs occur in concrete situations and are basically determined by these situations and simultaneously shaped by the learners' personal characteristics as user types. A number of variables have been identified to define the profile of foreign learners and investigate how they influence learners' lexicographic needs in concrete situations. Among these variables, the most important and relevant variables are foreign learners' proficiency in L2, native language and cultural background, age and learning circumstances (Tarp 2008). An advanced learner can in most cases resort to L2 definitions to solve his or her comprehension problems, whereas a beginner may need L1 equivalents or explanations to solve the same type of problem. A Thai learner of Chinese may have no difficulty in identifying water spinach, but a Danish speaker may wonder what it is. In short, a learner can have different lexicographic needs in different user situations and different learners in the same user situation could differ in their needs.

The purpose of a learner's dictionary is to satisfy its target users' needs. In this respect, the advent of the new information and communication technologies can help lexicography move closer than ever before to providing personalized and individualized service as claimed by Rundell (2010) and Tarp (2011), among others, a goal that can only be fully achieved in context-aware integrated information tools like e-readers and writing assistants, cf. Tarp et al. (2017). Hence, it is imperative that lexicographers planning a new dictionary project should have a coherent understanding of a homogenous group of learners' needs in specific user situations as well as their relevant characteristics as users, including their age (adult or child), first language, cultural background and L2 proficiency level. Without reference to the target users' specific

needs, the discussion on defining styles, examples and other type of data contained in a dictionary will be fruitless and futile in the end.

Conceiving a learner's dictionary is a complex process and involves a holistic understanding of the functions of the concerned dictionary in terms of its users and their needs in particular situations or social contexts. This article will focus on expounding the issue of lemma selection for Chinese learner's dictionaries aimed at assisting L2 (Chinese) text production. A distinction is made between foreign learners living in and outside China, as the general circumstances in which a foreign language is learned constitute an important variable that influences the learners' lexicographic needs in specific situations.

6. Lemma selection in the conception of Chinese learner's dictionaries

The issue of lemma selection has always been central in learner's lexicography and Chinese learner's dictionaries are no exception in this regard. The main questions concerning lemma selection for learner's dictionaries are summarized by Tarp (2008: 174) as follows:

1. How big should the lemma stock in learner's dictionaries be?
2. Which criteria and principles should guide lemma selection?
3. Which empirical basis should lemma selection be based on?

With these three questions in mind, in the following section we will briefly examine the practice of lemma selection in some major Chinese learner's dictionaries for foreigners. Frequent references will be made to *My Chinese Picture Dictionary*, given its unique organization of lemmata according to social contexts that foreign learners would encounter when they live in China. Based on the analysis, proposals will be made to respond to the three questions.

6.1 The present lexicographic practice with regard to lemma selection

There are two official lists of characters which are most frequently used as empirical basis for lemma selection in dictionaries for foreign learners of Chinese. The first one is *The Outline of Chinese Vocabulary and Chinese Character Level* published by the National Office for Teaching Chinese as a Foreign Language (2001) and includes 8,822 Chinese words falling into four language levels. This word list is used as vocabulary curriculum for the Chinese Proficiency Test (HSK), an international standardized test of Chinese language proficiency which assesses non-native Chinese speakers' ability to use Chinese in their daily, academic and professional lives. The other official list of characters is the *List of Frequently Used Characters in Modern Chinese* elaborated by the State Language Commission (1988). This list contains a total of 3,500 Chinese characters structured in two sections, one with the 2,500 most frequent characters and another containing the 1,000 characters that come next in frequency.

With reference to the above-mentioned official lists of characters, *The Commercial Press Learner's Dictionary of Contemporary Chinese* (Lu and Lü 2006a), considered by Jiang (2006) to be one of the best Chinese learner's dictionaries, has a lemma stock of 2,400 Chinese characters to which should be added about 10,000 multi-character words selected as sublemmata. *A Learner's Chinese Dictionary* (Zheng 2009) includes 3,000 characters as lemmata and an additional 32,000 multi-character words and expressions presented as sublemmata. *A Dictionary of Chinese Usage* (Liu 2000) offers 8,822 single-character and multi-character words as lemmata, i.e. exactly the same amount as *The Outline of Chinese Vocabulary and Chinese Character Level* referred to above.

Wang and Liu (2014) have examined the lemma stock in eight major Chinese learner's dictionaries for foreign learners. The two authors show how these dictionaries generally claim to select lemmata with reference to the above official lists of characters but are quite divergent regarding the number of lemmata actually included. The philosophy underpinning the principles of sticking to the official teaching curriculum seems to be the belief that internalizing the knowledge of the basic vocabulary is the stepping stone for learning Chinese. This philosophy seems, to a great extent, to ignore the fact that foreign learners' needs for Chinese vocabulary may arise in authentic social situations rather than in educational contexts, especially when they are living in China.

The extensive exposure to various aspects of life in China prompts foreigners to demand a wide range of vocabulary specific to their personal situations. They need to go to local markets to buy food and vegetables, and they may also need to deal with residential issues in the local police station. Quite a number of Chinese characters relevant to realistic social situations may be ranked low-frequency in the language-teaching curriculum.

Considering the distance between Chinese and other languages, there are several words and expressions describing common social phenomena typical for Chinese society, for instance the complex system of addressing forms. These phenomena may be absent in other cultures and language communities, but does this mean that foreign learners should ignore the corresponding vocabulary, since it is not part of their language and culture? Or should they just learn it for receptive purpose, since they may never have to use these words? In order to answer these questions, it is necessary to consider the geographic and linguistic communities where the targeted foreign learners live, when it comes to selecting lemmas for Chinese dictionaries targeting this user segment. Although part of the academic literature emphasizes the importance of considering foreign learners' daily life and study in China when selecting lemmata, no further and detailed elaborations on this challenge have yet been made (Li 2013: 36; Wang 2009: 569; Wang and Liu 2014: 73; Yang 2016: 47). Wang and Liu (2014: 73), for instance, explicitly state that the core vocabulary in the official curricula cannot be incorporated directly as lemmata in learner's dictionaries and that lemma selection requires practical experience and expertise from the lexicographers. However, the abstract selection principles of frequency, common errors, and levels of core

vocabulary suggested by these two authors do not solve the problem. It is therefore imperative to develop practical methods that are easier to handle.

The publication of *My Chinese Picture Dictionary* seems to be a practical response to the abstract discussion on lemma selection for Chinese learner's dictionaries. The strength of this dictionary is the presentation of 4,200 lemmata organized in 15 thematic social contexts defined in the dictionary (See figures 1 and 2 in this regard). The thematic organization of lemmata inevitably challenges the rigid levels of the wordlists designated in the official curriculum, although the preface states that the '15 thematic units are categorized according to the *International Curriculum for Chinese Language Education* published by the Office of Chinese Language Council International' (Wu 2008).

The fact that all the words included in *My Chinese Picture Dictionary* are illustrated with pictures makes it easy for non-native speakers to identify the referents and associate the vocabulary with things and phenomena in their social life. This is especially helpful to newly arriving foreigners who want to learn Chinese and become familiar with Chinese culture. Disregarding the improper translations in some cases, the bilingual dimension with its inclusion of English equivalents attached to the presented Chinese words also lowers the threshold to use this dictionary, at least for the users who are native speakers of English or have a certain proficiency level in this language. The important question of access to the words treated in the illustrations is solved by the appended English and Chinese indexes.

However, the lack of a clear definition of users and functions of *My Chinese Picture Dictionary* results in many problems, which to a certain extent reduces its value in practical use. Consequently, we will briefly discuss some of these problems because of their relevance for our vision of a Chinese learner's dictionary for foreigners living in China.

First of all, the social contexts treated in *My Chinese Picture Dictionary* do not always seem to be relevant to the envisaged user group. Some contexts like *Construction Sites* and *Martial Arts* are quite distant from most foreign learners' daily life and the words grouped under these topics are therefore not the most relevant to their actual needs. For instance, twenty-four words, both single-character and multi-character, are listed in connection with *Martial Arts*, but one may wonder in what social contexts foreign learners will have contact with the specific vocabulary describing the movements in martial arts like 二指禅 (*two-finger meditation*) 形意拳 (*intent-shaped fist*), etc. Even average Chinese people seldom have contact with these moves of martial arts in their daily life.

Secondly, the depiction of the social contexts tends to be skewed in the dictionary. For instance, people generally go to the local police station to deal with civil issues, like applying for residential permission or registration, but the words provided under *Police Station* mainly focus on crime and violence, such as 谋杀 (*to murder*), 绑架 (*to kidnap*), 手铐 (*handcuffs*), and 警棍 (*police baton*), and therefore deviate from the daily routines in China; cf. Figure 2. Moreover, the number of items listed in some thematic contexts seems to be too modest in

comparison with the expected user needs. This is, among others, the case with the fruit and vegetables shown under *Nutrition* as well as the food referred to in connection with *Western Restaurant*.

Thirdly, the words and terms presented in some thematic units like *Hospital* and *Skv* appear to be too specialized even for native Chinese speakers. Anatomical terms like 上腔静脉 (*superior vena cava*), 腓骨 (*fibula*) or 趾骨 (*phalanges*) are medical terms and unfamiliar to people who are laymen within this field. The same holds true for technical terms like 对流层 (*troposphere*), 同温层 (*stratosphere*), 积雨云 (*cumulonimbus cloud*) which are rather challenging for laymen and rarely appear in daily communication.

In the fourth instance, pictorial illustrations in general greatly limit the classes of the words included in the dictionary as not all words, particularly verbs, adjectives and adverbs, can be illustrated in a simple and easily understandable way. As a result, most vocabulary presented in *My Chinese Picture Dictionary* are nouns, of which multi-character words make up the majority. This may block foreign learners' mastery of the single Chinese characters, as Chinese characters are independent meaning units and very productive in combination usage.

Lastly, but not less important, the marking of word classes is missing in the dictionary. This leaves foreign learners almost helpless in terms of identifying part of speech of the presented lemmata. The lack of grammatical data and collocations also greatly reduces the value of *My Chinese Picture Dictionary* in communicative situations. As the dictionary claims to demonstrate different Chinese social contexts by means of pictures, it is not easy to understand why the 'new vocabulary' should be 'useful not only in Chinese culture but also in Western societies' (cf. Preface in Wu 2008). Given the limited vocabulary illustrated by pictures in typical Chinese settings, the dictionary seems to be too ambitious when it comes to 'help students to learn and use Chinese words to talk about different cultures' (cf. Preface in Wu 2008).

In summary, *My Chinese Picture Dictionary* opens new perspectives for lemma selection in Chinese dictionaries for foreign learners living in China. It seems, nonetheless, that the practical method of selecting lemmata according to specific social situations requires further reflections and refinements in order to overcome the problems identified above. The selected themes do not, to a large extent, represent the most typical social contexts in real life. The dictionary describes its target users as 'students', a generic term that tends to blur the profile of the users. The critical remarks put forward in this and the previous sections suggest that the principles applied to select lemmata are not sufficiently well-considered to achieve the desired high-quality learner's dictionary for non-native speakers living in China.

6.2 Some proposals

The analysis in the previous sections indicates that two main criteria have been

used to select lemmata for existing Chinese learner's dictionaries, i.e. frequency based on corpora, and something that could be called context-relevance. As none of these criteria applied separately seem to fully meet the needs of adult foreigners living in China, this article proposes that the selection of lemmata for an *online Chinese learner's dictionary* targeted at this audience should follow a combination of the two criteria mentioned, namely *frequency* and *relevance*. The application of these two criteria will be discussed in the following sections.

First, it goes without saying that the criterion of frequency provides solid empirical evidence for the occurrence of a word in actual language use. Using corpora to assist lemma selection is widely practiced in dictionary-making; cf. Rundell and Kilgarriff (2011), Hanks (2012), among many others. As shown in the previous sections, most Chinese learner's dictionaries claim to make reference to the national teaching curriculum, various corpora or frequency dictionaries containing the most frequent Chinese characters and words. Against the rapid development of corpora and the ever-increasing size of these, the two frequency dictionaries of modern Chinese words published in 1986 and 1990 tend to be outdated. The current representative corpora of modern Chinese are generally organized and constructed by national institutions or universities, like the corpus of modern Chinese (50 million characters), the CCL corpus (307 million characters), the corpus of Chinese texts by international students in Beijing (1 million characters), the interlanguage corpus of Chinese from global learners (50 million characters) and the HSK dynamic composition corpus (4 million characters) (Zhang 2015).

The availability of these corpora undoubtedly provides a huge amount of data, which offers a solid empirical basis for the frequency information about the candidate lemmata for Chinese learner's dictionaries. Statistic measures can be used to identify and select the words which remain stable in terms of their corpus coverage, their time sensitivity and diachronic classification. The fact that such words have a stable occurrence in the corpora indicates that they express vigour and versatility. Finally, the national teaching curriculum can also be considered a reliable source for candidate lemmata.

However, corpus frequency cannot be taken as the only criterion to select or exclude a lemma, despite the essential role played by the statistical significance in lemma selection. There is possible divergence between the high-frequency words in corpora and the words required in specific social situations as it has been argued by Guo et al. (2014) as well as Zhang (2015). Few foreigners will read a dictionary from one end to another in order to learn Chinese. Often, they are driven to Chinese dictionaries by practical problems in specific communicative situations and acquire the knowledge of Chinese from dictionaries incidentally. The discussion in the previous chapter indicated that the vocabulary provided under topical units such as *police station* and *hospital* may not be the most relevant to foreigners living in China whereas other words may be much more relevant to them in these contexts. The criterion of relevance should therefore be applied as an additional criterion when it comes to determining

whether a Chinese word should be included in a Chinese learner's dictionary.

The criterion of relevance is referring to the likelihood of a word occurring in one of the social situations which foreigners living in China most typically encounter in their day to day life. Although not among the most frequent words in a corpus, such a word may nonetheless be frequent and typical in the mentioned situations and therefore relevant as a lemma candidate in a learner's dictionary for this specific segment of users. As illustrated above, the likelihood of foreigners needing the vocabulary related to civilian services is much higher than their needs for the words about violent crimes in Chinese police stations. Hence, the vocabulary related to civilian services should be prioritized in the lemma selection for foreign beginner learners in China without ignoring that related to various sorts of crime. The same applies to the high-culture vocabulary like the words related to martial arts or other specialized subject fields. This does not mean that Chinese learner's dictionaries should limit the lemmata to low-culture survival words. It is to be understood that foreign language learning is a continuum and foreign learners' needs for vocabulary vary in numbers and scopes in the continuum of learning Chinese, starting from the survival needs and advancing to the needs for specialized and high-culture vocabulary.

In short, the more typical a word is in social situations in which foreigners learning Chinese in China frequently find themselves, the more often they will have contact with it although it may not display the same degree of frequency in a corpus. In order to identify words often appearing in relevant social contexts, it is first of all requisite to determine the respective contexts. The Council of Europe (2001) defines four social domains in the Common European Framework of Reference for Languages (CEFR): personal, public, educational and occupational. These domains can also be used to determine the various social situations typical of Chinese society. With reference to the CEFR framework, Tseng (2014: 27) has further specified 12 situations: personal data, work, education, housing, family and environment, daily routines, relaxation, interpersonal relationship, travelling, body and health, shopping, and food. Each of these situations or main themes can be further subdivided into a number of topical units as was the case with the thematic tables in *My Chinese Picture Dictionary*. It may be assumed that the words typically occurring in these contexts are relevant to foreigners living in China, even if they rank low in the general corpora. These words should therefore be selected as lemmata, for instance based on a collection of texts covering each of the situations in question and using the criterion of relevance.

Finally, the size of the lemma stock is subjective to the proficiency stages through which foreign-language learning develops. A learner's dictionary can be designed to assist its users in the first phase of foreign-language learning or to follow them until a more advanced proficiency level. There is therefore no absolute number of lemmata that can be recommended for a learner's dictionary. It all depends on its purpose and specific user segment. If the dictionary is

primarily conceived to assist learners at the beginner's level with production of Chinese text, then a reduced vocabulary may satisfy the learners' needs in this respect. However, if the dictionary is also supposed to cover the learners' needs in relation to text reception, and if the learners are living in China and exposed to Chinese every day, then a much bigger vocabulary is required even for beginners.

Among Chinese scholars there are various proposals as to the size of the lemma stock relevant to foreign learners of Chinese. Li (1999: 58), for instance, suggests that the national teaching curriculum should cover 10,000 to 12,000 words in order to meet foreigners' communicative needs in Chinese. Guo et al. (2014: 12) propose that 13,000 words could decently satisfy foreigners' needs. Zhang (2015) proposes 4,000 words for Chinese learner's dictionaries targeted at foreign beginners, an additional 6,000 for intermediate learners and a further 10,000 for advanced learners of Chinese. In total, the Chinese learner's dictionary should include about 20,000 words according to Zhang (2015: 42). As a starting point, the proposed size of 20,000 lemmata seems feasible and reasonable in a printed dictionary for foreign learners of Chinese. However, when it comes to future online dictionaries the problem may have to be approached in a different way as we will see below.

7. Perspectives

Dictionaries are human-made tools designed to assist possible users looking for information in order to solve different types of problem, as they have been defined in the *Routledge Handbook of Lexicography* by Tarp (2018). This suggests that people consult dictionaries when they have specific information needs and that the dictionaries should contain the corresponding lexicographical data, including the relevant lemmata, whereas the inclusion of superfluous data and lemmata can be regarded as a waste of time and money. In this respect, the best way to satisfy user needs in terms of lemma stock is to include the words which users actually look up. Until recently, lexicographers have generally only been able to guess the words that are relevant to their specific users. They have therefore resorted to indirect selection criteria like corpus frequency and context relevance as discussed above. At present, these selection criteria may still be recommended for dictionaries designed to assist foreign learners of Chinese living in China. However, these criteria are about to change radically in the nearby future as a much more reliable empirical basis is being developed, namely *logfiles* which trace user behaviour in dictionary consultation. Logfiles can be used either as a supplementary (see below) or as a primary source for lemma selection. When we speak about "radical change" we refer to the latter, i.e. the use of logfiles as the primary source for lemma selection which *totally replaces the corpus* as the main empirical basis for this purpose.

Once a high-quality online dictionary has been produced and used for some time, such logfiles will provide reliable evidence of the items which dic-

tionary user actually look up. Studies of logfiles show that there is not a complete correspondence between the most frequent words in a corpus and the words most frequently looked up in dictionaries. Bergenholtz and Norddahl (2012), for instance, have shown that some Danish words, which are very frequent in the corpus, are seldom or never looked up in a big online dictionary with more than hundred thousand lemmata whereas other words with a low corpus-occurrence are frequently consulted by the users after a total of more than 20 million lookups.

There is little doubt that logfiles will increasingly be used as an empirical basis for the selection of specific lexicographical data categories such as lemmata. In this respect, the frequency of the words appearing in the logfiles, or just the appearance itself, will become the basic criterion for lemma selection as it is currently the case in the Spanish–English–Spanish *Diccionarios Valladolid-Uva* (under production) which do not use corpora at all but only logfiles as the primary empirical basis (personal information). However, before logfiles can be used as a reliable empirical basis for future Chinese learner's dictionaries, a number of requirements have to be fulfilled. First of all, at least one high-quality learner's dictionary designed from scratch for the digital media should be produced and made available online. Then a statistically significant number of lookups should have been made, for instance 20 million. In addition, if the new dictionary is planned to serve foreigners learning Chinese in China, the logfiles used as empirical basis should make allowance for a distinction between users (learners) living inside and outside China. Finally, and in order to make an even better product, it should be possible to distinguish between lookups related to text production and text reception, respectively. In this respect, some lemmata could be given extra treatment with the inclusion of additional data categories in order to assist text production whereas others could focus on explanations with a view to supporting text reception. This would save time for lexicographers and result in a more focused lexicographical product.

Until this nearby future becomes reality, the combination of the criteria of frequency and relevance discussed above can be recommended when lemmata are selected to compile new digital learner's dictionaries for foreigners learning Chinese in China. But even so, the existing technology already allows for a gradual transition to new selection criteria as well as new publication methods; cf. Bergenholtz and Johnsen (2005), De Schryver (2013), Trap-Jensen et al. (2014), among others. The possibility of constant updating in the online media allows for a flexible publication process where the first version (or "edition") of a web-based dictionary can be made available to its users when a certain percentage of articles covering the most frequent and relevant lemmata have been finished, for instance, 20–30 percent. This could for example be the 4,000 words which Zhang (2015) recommends for a Chinese dictionary for foreign learners at the beginner's level. Once this number of articles has been completed, the lexicographers can continue working in two directions: (1) follow the established work plan and compile articles based on the selected lemma stock, and

(2) simultaneously study the logfiles (on a daily or weekly basis), detect words looked up by the users but still not treated in the dictionary and straightaway compile the corresponding dictionary articles, whether or not the words in question are included in the originally selected lemma stock. Such a methodological procedure will undoubtedly put real user needs at the centre of the lexicographical compilation process.

Finally, it can be said that the new disruptive computer and information technologies open new horizons to lexicography as a millennial cultural practice. Modern lexicographers — and publishers — should take full advantage of these technologies and adapt their methods accordingly. Lemma selection, from being a once-and-for-all decision in printed dictionaries, has been transformed into a dynamic endeavour which, in principle, can continue for years even after the first version of an online dictionary has been published. Continuous refinement and adaptation to the users' real needs should be the guiding principle also for online Chinese learner's dictionaries aimed at assisting non-native speakers living and learning Chinese in China.

Acknowledgements

Thanks are due to the China Scholarship Council for funding the project (Grant No. 201606435015) as well as to the Spanish Ministry of Economy and Competitiveness for funding the project "La Teoría Funcional de la Lexicografía: Diseño y Construcción de Diccionarios de Internet" (Ref. FFI2014-52462-P) in which this article is theoretically embedded.

References

- Bergenholtz, H. and B. Norddahl.** 2012. Ordbogsartikler som ingen læser. *LexicoNordica* 19: 207-223.
- Bergenholtz, H. and M. Johnsen.** 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34: 117-141.
- Cai, Y.Q.** 2011. The User-friendly Principle in the Compilation of Dictionary for Chinese Language Learning. *Lexicographical Studies* 2: 67-77.
- Council of Europe.** 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Available at: www.coe.int/lang. Accessed 9 June 2017.
- Cowie, A.P.** 1999. Learners' Dictionaries in a Historical and a Theoretical Perspective. Herbst, T. and K. Popp. (Eds.). 1999. *The Perfect Learners' Dictionary (?)*: 3-13. Tübingen: Max Niemeyer.
- De Schryver, G.-M.** 2013. The Concept of Simultaneous Feedback. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 548-556. Berlin: Walter de Gruyter.
- Gellerstam, M.** 1999. LEXIN — lexikon för invandrare. *LexicoNordica* 6: 3-18.
- Guo, X.L., X.S. Ma and K.T. Li.** 2014. Comparative Study of Chinese Characters and Words Based on Language Situation in China. *Journal of Beihua University (Social Sciences)* 15(3): 10-13.

- Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hao, Y.X. and Z.J. Wang. 2013. A Study on the Requirements of Learners in L1 Environment for Chinese Language Learners' Dictionary. *TCSOL Studies* 3: 50-57.
- Hernández, H.H. 1989. *Los diccionarios de orientación escolar. Contribución al estudio de la lexicografía monolingüe española*. Tübingen: Max Niemeyer.
- Hult, A-K. 2016. Ordboksanvändning på nätet. En undersökning av användningen av Lexins svenska lexicon. Gothenburg: Institutionen för svenska språket.
- Jiang, L.Sh. 2006. Preface. Lu, J.J. and W.H. Lü (Eds.). 2006. *The Commercial Press Learner's Dictionary of Contemporary Chinese*. Beijing: The Commercial Press.
- Jin, P.P. 2015. Thirty-year Researches on Chinese Learner's Dictionaries. *The Journal of Yunnan Normal University: Foreign Language Teaching and Research* 13(3): 27-37.
- Lew, R. and A. Adamska-Salaciak. 2015. A Case for Bilingual Learners' Dictionaries. *ELT Journal* 69(1): 47-57.
- Li, Q.H. 1999. The Issue of Quantity of Vocabulary on the Outline of Chinese. *Language Teaching and Research* 1: 50-59.
- Li, Y. 2013. On the Compilation of General-purpose Chinese Dictionaries for Foreign Learners of Chinese. *Lexicographical Studies* 5: 34-39.
- Liu, L.L. 2000. *A Dictionary of Chinese Usage: 8000 Words Chinese Proficiency Test Vocabulary Guideline*. Beijing: Beijing Language and Culture University Press.
- Liu, S.T. 2014. Systematic Research on the Structural Features of Chinese Learner's Dictionaries for Foreigners. *Social Sciences in China*, 17 February: A8.
- Liu, Y. 2016. It is Not Easy to Find Jobs in China. *People's Daily Overseas Edition*, October 17, 2016. Retrieved from http://paper.people.com.cn/rmrhwb/html/2016-10/17/content_1719150.htm.
- Lu, J.J. and W.H. Lü. 2006a. *The Commercial Press Learner's Dictionary of Contemporary Chinese*. Beijing: The Commercial Press.
- Lu, J.J. and W.H. Lü. 2006b. The Compilation of a Monolingual Learner's Dictionary of Chinese as a Foreign Language: A Venture and Some Considerations. *Chinese Teaching in the World* 1: 59-69.
- Lü, S.H. 1980. *800 Words of Modern Chinese*. Beijing: The Commercial Press.
- Martín, I.L. (Ed.). 2011. *Bienvenidos. El primer diccionario de español*. Madrid: Octaedro.
- National Office for Teaching Chinese as a Foreign Language. 2001. *The Outline of Chinese Vocabulary and Chinese Character Level*. Beijing: Jingji Kexue Press.
- Pálfi, L.-L. and S. Tarp. 2009. Lernerlexikographie in Skandinavien — Entwicklung, Kritik und Vorschläge. *Lexicographica* 25: 135-154.
- Rundell, M. 2010. What Future for the Learner's Dictionary? Kernerman, I.J. and P. Bogaards. (Eds.). 2010. *English Learners' Dictionaries at the DSNA 2009*: 169-175. Jerusalem: Kdictionaries.
- Rundell, M. 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.
- Rundell, M. and A. Kilgarriff. 2011. Automating the Creation of Dictionaries: Where Will It All End? Meunier, F., S. de Cock, G. Gilquin and M. Paquot. (Eds.). 2011. *A Taste for Corpora. In Honour of Sylviane Granger*: 257-281. Amsterdam/Philadelphia: John Benjamins.
- Shi, G.H. and S.X. Wang. 2011. *A Chinese Dictionary for Learners and Teachers*. Beijing: The Commercial Press.

- Song, Q.C.** 2013. A Demographic Sociological Analysis of Foreigners and Hong Kong, Macao, and Taiwan Residents in Mainland China. *The Journal of Shandong University: Philosophy and Social Sciences* 2: 89-99.
- State Language Commission of China.** 1988. *List of Frequently Used Characters in Modern Chinese*. Beijing: Language & Culture Press.
- Tarp, S.** 1999. Lærnerordbøger for indvandrere og andet godtfolk. *LexicoNordica* 6: 107-132.
- Tarp, S.** 2004. Basic Problems of Learner's Lexicography. *Lexikos* 14: 222-252.
- 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.** 2011. Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, P.A. and H Bergenholtz (Eds.). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 54-70. London/New York: Continuum.
- Tarp, S.** 2018. The Concept of Dictionary. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 237-249. London: Routledge.
- Tarp, S., K. Fisker and P. Sepstrup.** 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.
- Trap-Jensen, L., H. Lorentzen and N.H. Sørensen.** 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 2(2): 94-113. <https://doi.org/10.4312/slo2.0.2014.2.94-113>. Accessed 4 July 2018.
- Tseng, W.H.** 2014. Classification on Chinese 8,000 Vocabulary. *Teaching Chinese as a Second Language* 16: 23-35.
- Wang, H.Y.** 2009. What Chinese Dictionaries are Expected by Foreign Learners. *Chinese Teaching in the World* 4: 567-575.
- Wang, X. and S.D. Liu.** 2014. The Study on Lemma Selection for Export-oriented Learner's Dictionaries. *Ludong University Journal: Philosophy and Social Sciences* 31(3): 69-74.
- Wei, J.X. and H.L. An.** 2014. The Review of Compilation and Research on Export-oriented Chinese Learning Dictionary. *Journal of Guangdong Ocean University* 34(5): 70-75.
- Wei, X.Q., Y.D. Geng and D.B. Wang.** 2014. *Lexicography in China (1978–2008)*. Beijing: The Commercial Press.
- Welker, H.A.** 2008. *Panorama geral da lexicografia pedagógica*. Brasilia: Thesaurus Editora.
- Wiegand, H.E. (Ed.).** 1998. *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von Langenscheidts Großwörterbuch Deutsch als Fremdsprache*. Tübingen: Niemeyer.
- Wu, Y.M. (Ed.).** 2008. *My Chinese Picture Dictionary*. Beijing: The Commercial Press.
- Xie, H.J. and L. Li.** 2012. An Investigation into the Publication and Using Condition of CFL Chinese Learner's Dictionaries. *Ludong University Journal: Philosophy and Social Sciences* 29(1): 62-68.
- Yang, H.** 2015. *The Investigation of Use of the Chinese Dictionary Status Quo of International Students in China*. Unpublished Master's thesis, Chongqing Normal University, Chongqing, China.
- Yang, J.H.** 2016. On the Four Principles of Compiling Chinese Dictionaries for Foreign Learners. *Lexicographical Studies* 1: 45-51.
- Yang, N. and X.D. Zhang.** 2017. The Mania of Learning Chinese: A Bonus to Overseas Chinese. *People's Daily Overseas Edition*, April 17, 2017. Retrieved from <http://world.people.com.cn/n1/2017/0417/c1002-29214843.html>.

- Zhang, B.L.** 2015. Compiling Design of Chinese as A Second Language Learning Dictionary Series. *Bilingual Education Studies* 2(1): 37-44.
- Zhang, X.M.** 2010. An Exploratory Discussion on Chinese Learners' Dictionaries in China. *Lexicographical Studies* 3: 27-37.
- Zheng, D.O.** 2004. On Chinese Learner's Dictionaries for Foreigners. *Chinese Teaching in the World* 4: 85-94.
- Zheng, S.P.** 2009. *A Learner's Chinese Dictionary*. Beijing: Foreign Language Teaching and Research Press.
- Zöfgen, E.** 1994. *Lernerwörterbücher in Theorie und Praxis. Ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*. Tübingen: Niemeyer.

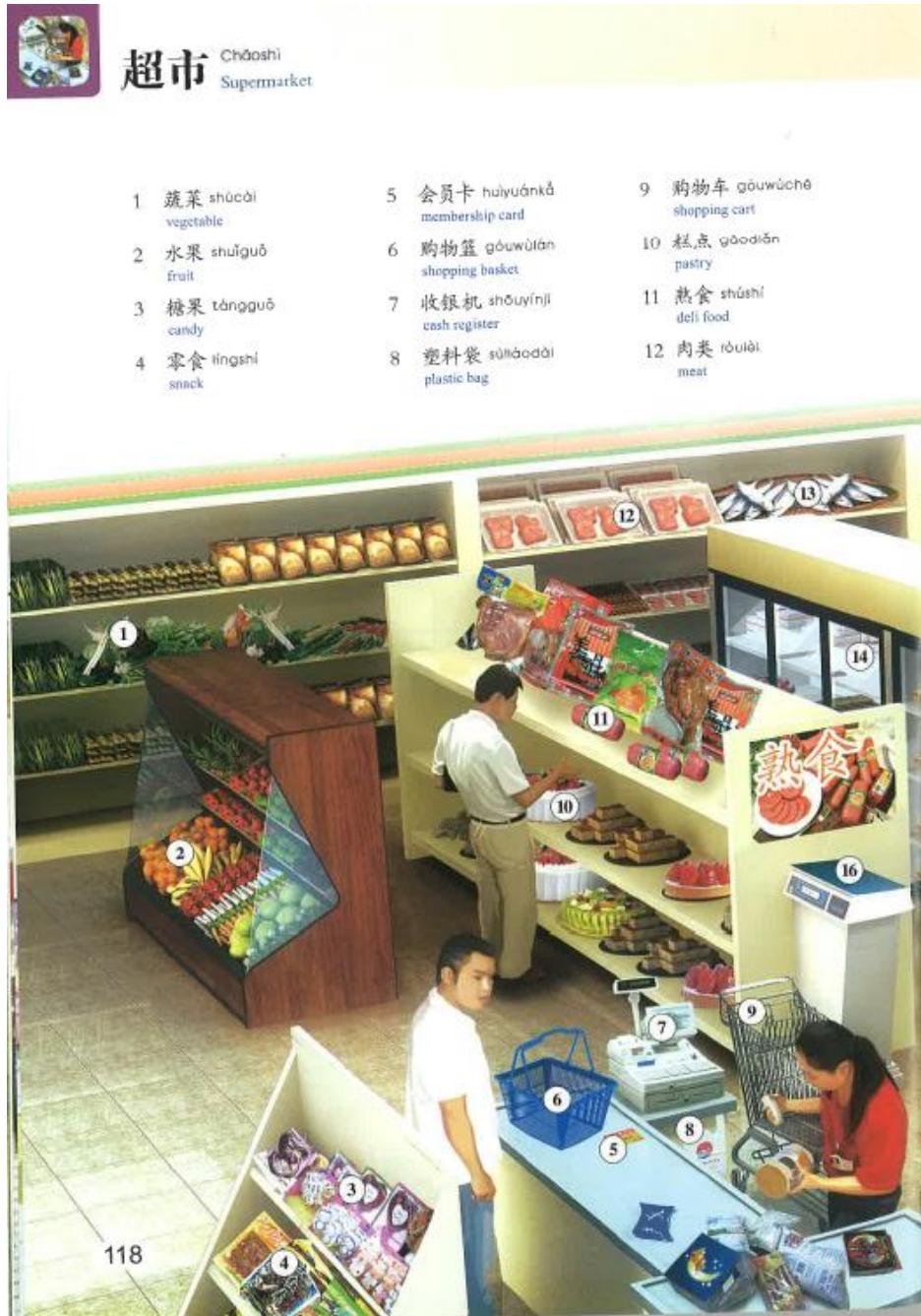


Figure 1: Illustration from *My Chinese Picture Dictionary: Supermarket*

- | | | |
|--|---|--|
| 1 公安局 gōngānjú
police station | 13 拘留 jiūliú
to detain | 23 警车 jǐngchē
police car |
| 2 监狱 jiānyù
prison | 14 手铐 shǒukào
handcuffs | 24 警察 jǐngchá
police officer |
| 3 小偷 xiǎotōu
thief | 15 手枪 shǒuqiāng
pistol | 25 目击者 mùjīzhě
witness |
| 4 抢劫 qiǎngjié
to rob | 16 警帽 jǐngmào
police hat | 26 警徽 jǐnghuī
police emblem |
| 5 绑架 bāngjià
to kidnap | 17 钢盔 gāngkuī
helmet | 27 警笛 jǐngdí
siren |
| 6 报警 bàojǐng
to call the police | 18 警棍 jǐnggùn
police baton | 28 受害者 shòuhàizhě
victim |
| 7 谋杀 móushā
to murder | 19 对讲机 duìjiǎngjī
walkie-talkie | 29 逮捕 dǎibù
to arrest |
| 8 审问 shěnwèn
to interrogate | 20 警服 jǐngfú
police uniform | 30 便衣警察 biànyī jǐngchá
plain-clothes police |
| 9 犯罪嫌疑人 fánzù xiányíren
suspect | 21 警用摩托车 jǐngyòng mótuōchē
police motorcycle | 31 证据 zhèngjù
evidence |
| 10 法庭 fǎtīng
court | 22 交通警察 jiāotōng jǐngchá
traffic police | 32 警犬 jǐngquǎn
police dog |
| 11 审判 shěnpan
to try | | |
| 12 做笔录 zuò bǐlù
to take a statement | | |

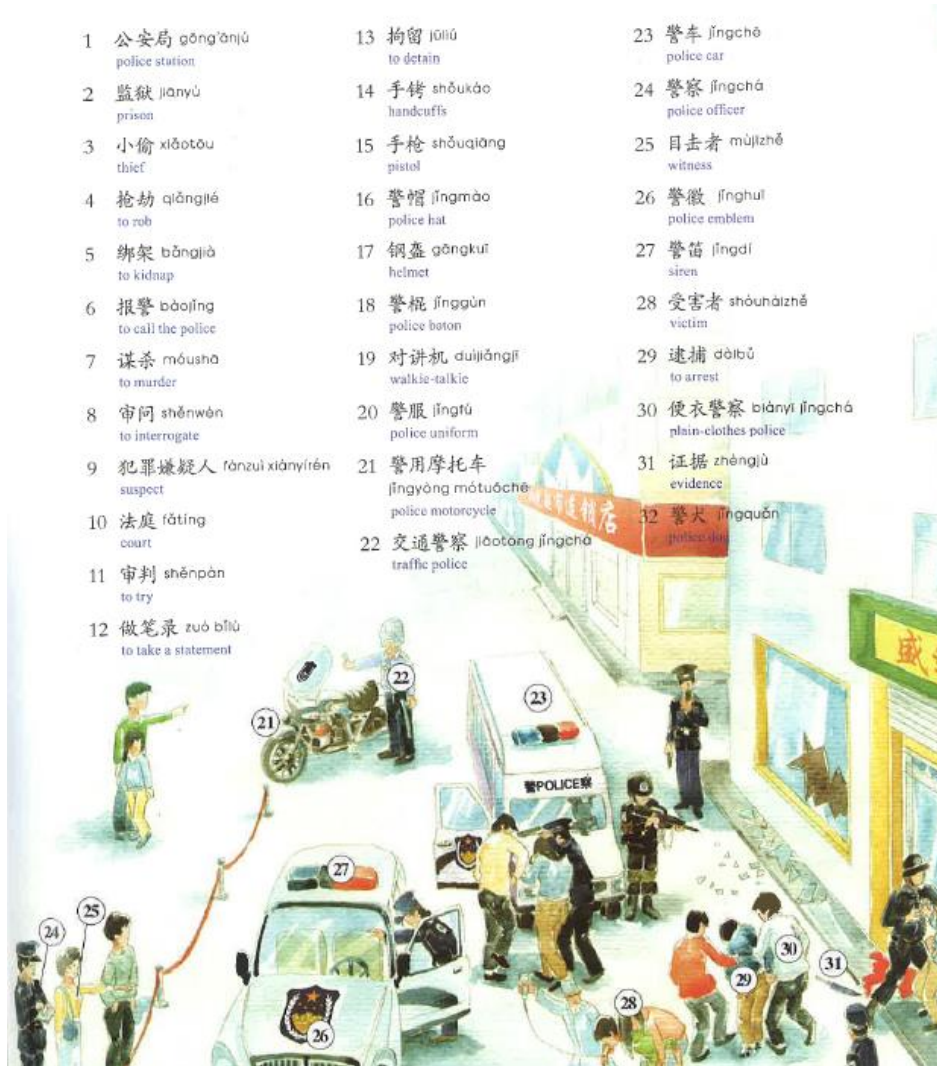


Figure 2: Illustration from *My Chinese Picture Dictionary: Police Station*

Enhancing the Learnability of Chinese–English Dictionaries for Chinese as a Foreign Language Learners: The Neglected Legacy of Robert Morrison in His Compilation of *Wuche Yunfu* (1819)

Ying Ye, *Bilingual Dictionary Research Centre, School of Foreign Studies, Nanjing University; School of Foreign Languages, Nanjing University of Chinese Medicine, Jiangsu, China (clare_ye@163.com)*

Xiangqing Wei, *Bilingual Dictionary Research Centre, School of Foreign Studies, Nanjing University, Jiangsu, China (dicweixiangqing@163.com)*
and

Wenlong Sun, *Bilingual Dictionary Research Centre, School of Foreign Studies, Nanjing University, Jiangsu, China (287971655@qq.com)*

Abstract: In previous studies on learner lexicography, design features of both the content and presentation of learner's dictionaries are the two major research concerns. The quality assessment of learner's dictionaries also covers the two dimensions. Terms used for evaluating them are respectively "usability" or "availability" for the former and "findability" or "accessibility" for the latter. However, the lexicographical construction of "learnability", which takes into account the users' reference and learning needs, remains virtually unexplored either theoretically or practically. Compared to the features of dictionary design mentioned above, "learnability" as the design philosophy of learner lexicography is worth more serious consideration. The present paper aims at exploring the lexicographical notion of "learnability" by way of introducing the neglected legacy of Robert Morrison in his compilation of *Wuche Yunfu* (五车韵府) (1819)¹, which is characterized by a high degree of learnability illustrated in the dictionary entries. Morrison's pioneering efforts may help with the conceptual clarification of "learnability" in compiling learner's dictionaries, bilingual ones in particular. Moreover, it is hoped that the recognition of Morrison's lexicographical practice will be beneficial to the future production of better Chinese–English dictionaries for non-native Chinese learners.

Keywords: LEARNERABILITY, LEARNER LEXICOGRAPHY, CHINESE AS A FOREIGN LANGUAGE LEARNERS, WUCHE YUNFU, CHINESE–ENGLISH DICTIONARIES, BILINGUAL

DICTIONARIES, LEXICOGRAPHICAL PRACTICE, LEXICOGRAPHICAL NOTION, LEARNING LOAD, LEARNERS' NEEDS

Opsomming: Verhoging van die leerbaarheid van Chinees–Engelse woordeboeke vir aanleerders van Chinees as vreemde taal: Die vergete nalatenskap van Robert Morrison in sy samestelling van *Wuche Yunfu* (1819). In vorige studies van aanleerderleksikografie is ontwerpkenmerke van beide die inhoud en aanbieding van aanleerderswoordeboeke die twee belangrikste navorsingsaspekte. Die kwaliteitsbepaling van aanleerderswoordeboeke dek ook hierdie twee dimensies. Terme wat gebruik word vir hul evaluering is onderskeidelik "bruikbaarheid" of "beskikbaarheid" vir eersgenoemde en "vindbaarheid" of "toeganklikheid" vir laasgenoemde. Die leksikografiese begrip "leerbaarheid", wat die gebruikers se verwysings- en aanleerdersbehoefte in ag neem, bly egter eintlik teoreties en prakties onontgin. Vergeleke met die kenmerke van woordeboekontwerp waarna hierbo verwys is, behoort "leerbaarheid" as die ontwerpfilosofie van aanleerderleksikografie ernstiger oorweeg te word. In hierdie artikel word gepoog om die leksikografiese konsep van "leerbaarheid" te ondersoek met behulp van die vergete nalatenskap van Robert Morrison in sy samestelling van *Wuche Yunfu* (五车韵府) (1819)¹, wat gekenmerk word deur 'n hoë mate van leerbaarheid soos geïllustreer in die woordeboekinskrywings. Morrison se baanbrekerswerk kan van hulp wees met die konseptuele verheldering van "leerbaarheid" in die samestelling van aanleerderswoordeboeke, veral tweetalige aanleerderswoordeboeke. Bowenal word daar gehoop dat die erkenning van Morrison se leksikografiese praktyk tot voordeel van die toekomstige produksie van beter Chinees–Engelse woordeboeke vir nie-moedertaal Chinese aanleerders sal wees.

Sleutelwoorde: LEERBAARHEID, AANLEERDERSLESIKOGRAFIE, LEERDERS VAN CHINEES AS VREEMDE TAAL, *WUCHE YUNFU*, CHINEES–ENGELSE WOORDEBOEKE, TWEETALIGE WOORDEBOEKE, LEKSIKOGRAFIESE PRAKTYK, LEKSIKOGRAFIESE KONSEP, WERKSLADING, AANLEERDERSBEHOEFTE

1. Introduction

Compared to the large number of English–Chinese (E–C) dictionaries meant for English as a Foreign Language (EFL) learners, the number of Chinese–English (C–E) dictionaries available in China for Chinese as a Foreign Language (CFL) learners is rather small, which contrasts remarkably with the increasing popularity of CFL learning worldwide. For the already published C–E dictionaries for CFL learners, few of them are found satisfactory (Wang 2008). Yang (2015) conducted a questionnaire survey of CFL learners in two Chinese universities, and the results showed a predominant preference for dictionaries published outside China with regard to the dictionaries used by CFL learners. Some participants of the survey complained about the quality of current CFL learner's dictionaries, especially those published in China. The main complaint was the lack of helpful information for their Chinese language learning. Actually, lexicographers always concern themselves with the usefulness of the dictionaries

they produce. For example, Li (2013) discussed practicality in compiling Chinese dictionary for CFL learners. Yang (2016) stated that compiling Chinese dictionaries for CFL learners should follow four basic principles: simplicity, practicality, comprehensiveness and explicitness. However, a fundamental problem lies in the fact that lexicographers in many cases fail to capture the specific learning needs of users, which are often deeply rooted in their learning process. Or in other words, the learnability of learner's dictionaries has not been fully explored, either as lexicographical conceptualization or for practical purposes.

This article is intended to draw more attention to the term "learnability", a lexicographical construct that has in fact long been overshadowed by the high frequency of some similar terms used in the lexicographical literature, such as "usability", "practicality" or "availability". The latter ones are used for designing or judging dictionaries in general while the former is specifically meant for learner's dictionaries. Nevertheless, the exact conceptual content of "learnability" still remains unclear or unspecified. In this article, "learnability" is defined, from the perspective of learner lexicography, as how much useful the information is for learning and how easily the users can learn the needed information. A learner's dictionary should be not only user-centred as all dictionaries do, but also learning-centred in the way that the user's learning process is the centre of lexicographer's attention. Instead of "usability", "practicality" or "availability", the authors of this article consider "learnability" as a more appropriate term that labels learner lexicography. More importantly, we are going to further illustrate the concept of "learnability" through Morrison's actual practice in writing *Wuche Yunfu* (五车韵府) (1819), which is a lexicographical product made nearly two centuries ago and renowned for its unfailing popularity with generations of western CFL learners.

As "the first Chinese–English dictionary widely used by people both in the East and in the West" (Wu and Zheng 2009: 3), it was found "highly detailed and was well received, being acclaimed as the best Chinese dictionary in a European language" (Ryu 2009: 8). It was also used as "the base for publications of multilingual dictionaries in Japan and Korea" (Ryu 2009: 1). Even today, *Wuche Yunfu* is still used as a reference book. It is regarded by many scholars as an encyclopedia for its comprehensive coverage of Chinese culture (Wu and Zheng 2009).

As mentioned above, Morrison's *Wuche Yunfu* has often been given much credit for its success in helping CFL learners to learn Chinese. In other words, the degree of learnability of this very dictionary is quite distinct from that of other CFL dictionaries.

The lexicographical success of Morrison's *Wuche Yunfu* can be illustrated typically by one specific entry article, compared with its counterparts in some contemporary CFL learner's dictionaries, either monolingual or bilingual. Take the entry of the Chinese character 精 (pronounced as "jing" in Chinese, and literally means "refining" in English) as an example. Compared to the article selected

from *The Commercial Press Learner's Dictionary of Contemporary Chinese* (2007), the most representative monolingual CFL learner's dictionary in contemporary China, Morrison's lexicographical treatment is obviously more helpful for CFL learners. Though *The Commercial Press Learner's Dictionary of Contemporary Chinese* in its preface claims to target CFL learners, it actually fails to achieve its lexicographical goal.

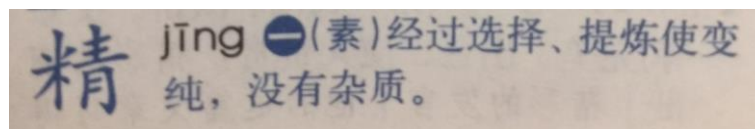


Figure 1: Entry of 精 in *The Commercial Press Learner's Dictionary of Contemporary Chinese* (2007: 373)

As shown in Figure 1 above, the daily Chinese character 精 is defined as "经过选择、提炼使变纯，没有杂质" (literally means "to purify sth. through selection or refinement" in English) (2007: 373). Obviously, the wording of this definition is abstract and rather difficult for intermediate or even advanced CFL learners to comprehend. There are also two terms used in this definition, namely 提炼 (literally means "to refine" in English) and 杂质 (literally means "impurity" in English), which are even more difficult for CFL learners. In fact, on the vocabulary list of HSK², 精 is a Chinese character of Level Four for intermediate CFL learners whereas 提炼 is a word of Level Six³. The Chinese word 杂质 is not actually found on the vocabulary list of HSK.

Similarly, as can be seen from the Figure 2 below, the selected article from the bilingualized version of *Xinhua Dictionary* (2013), which is also aimed at CFL learners, the entry of the headword 精 does not meet the CFL learners' reference needs. The lexicographical information presented in this article is oversimplified with only a few English equivalents and some short verbal illustrations. By this way of explaining the headword 精, the related cultural connotation of this Chinese character is lost, which does not help with CFL learners' understanding of Chinese farming culture in general. To be more specific, the left part of this Chinese character 精 is 米 (pronounced as "mi" in Chinese, and literally means "rice") and the right part of this Chinese character 精 is 青 (pronounced as "qing" in Chinese, and literally means "golden age"). The original meaning of the character 精 is the combination of the meaning of 米 and 青, referring to "selecting first-class rice" in agriculture.

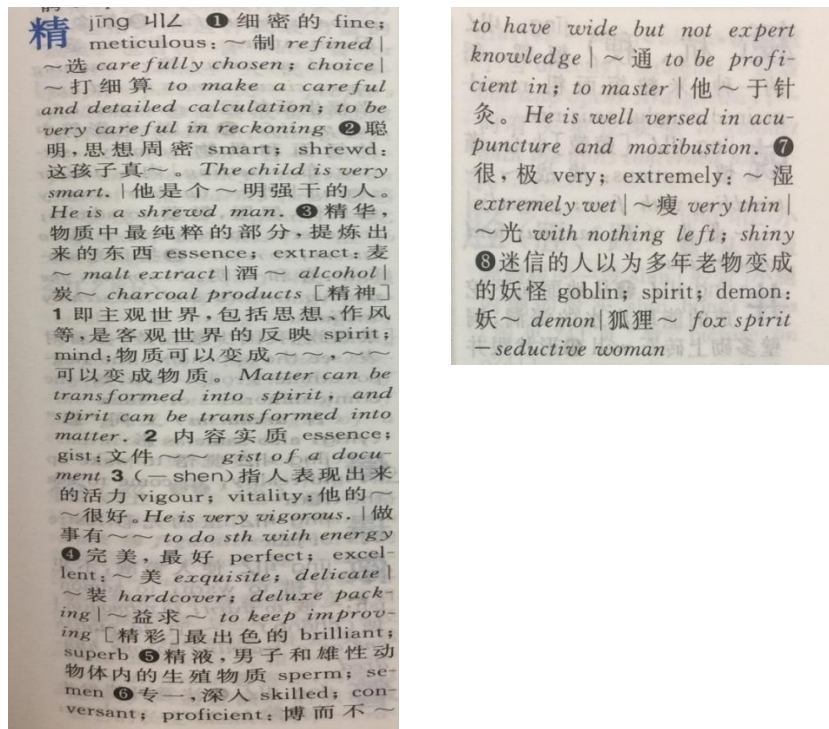


Figure 2: Entry of 精 in *Xinhua Dictionary* (2013: 359)

However, contrastingly, an entry article of the same Chinese character 精 in Morrison's *Wuche Yunfu*, seems to be more informative and interesting for its target users (see Figure 3). The Chinese character 精 in this C–E dictionary is defined first as "From rice and pure. To cleanse grain; the pure part of anything" (Morrison 1819: 915). Morrison explained first of all the original meaning of 精 in simple English words, which shows clearly the close relation between this Chinese character and the farming culture in China. Some English equivalents are also given to facilitate CFL learners' comprehension of this character. Morrison further illustrated the abstract senses "true ether; spiritual; subtile fluid; essence; essential; the semen of animals" based on the users' understanding of how rice is processed in China. In other words, he presented the related cultural information in the definitions, which helps CFL learners understand the inseparable connection between the Chinese culture and the Chinese language. With such dictionary definitions of greater comprehensibility and less cognitive burden, Morrison in his compiling *Wuche Yunfu* made great efforts to assist CFL learning.

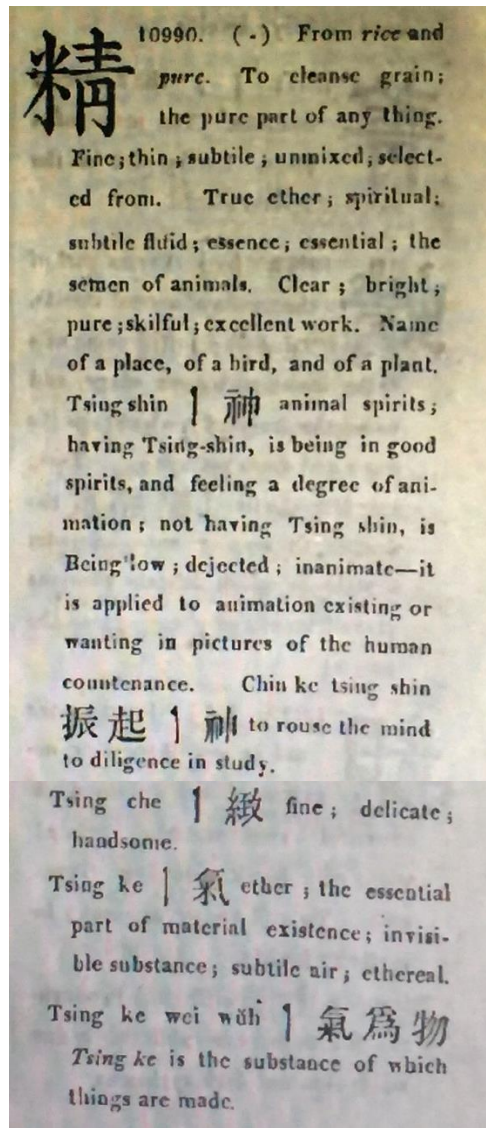


Figure 3: Part of the entry of 精 in *Wuche Yunfu* (Morrison 1819: 915)

It is clear from the comparison above that the lexicographical content as well as presentation of Morrison's *Wuche Yunfu* is more user-friendly. In his actual compilation of the dictionary, Morrison did take into consideration "the quality or fact of being learnable" (OED 1989: 768). Thus, the present study is intended to explore the essence of learnability in learner lexicography by examining how Morrison actually realized it during the process of compiling *Wuche Yunfu*. It is

hoped that this study can shed some light on the compilation of future Chinese–English learner's dictionaries, those for CFL learners in particular. More importantly, with the development of learner lexicography, the concept or notion of learnability has to be emphasized and further clarified.

2. Learnability as the design philosophy of learner lexicography

Ever since the early 20th century, lexicographers have been endeavoring to improve the quality of learner's dictionaries, focusing mainly on lexicographical design. They conducted researches into the "findability" of learner's dictionaries, which is concerned with finding the target lexical items and the related information in the dictionary; they carried out studies on the "accessibility" of learner's dictionaries, which is "the relative ease with which information can be located in a reference work" (Hartmann and James 2000: 2). They also investigated the "usability" of learner's dictionaries, which deals with the use of lexical items correctly in terms of grammar, pragmatics, etc. (Chon 2008) or users' preference in accomplishing their writing tasks (Laufer and Levitzky-Aviad 2006). The lexicographical terms such as "findability", "accessibility", "usability" and "availability" are frequently used to evaluate the content and presentation of learner's dictionaries. For instance, Bogaards and Van der Kloot (2001) studied verb complementation in the dictionaries. They compared the usefulness of the information presented in LDOCE3, CIDE, and COBUILD2 by examining two major dimensions: findability and usability. Lew and Dziemianko (2006) discussed the usefulness of a new defining model for foreign learners, mentioning the importance of accessibility. Faaß et al. (2014) discussed their means to achieve data accessibility. Alzi'abi (2017) presented changes of the general layout of DSAEHist online platform, which were made to improve the usability of data. Nevertheless, these terms used as the criteria for evaluating the design of learner's dictionaries, are too general in that these terms can be used to evaluate many types of dictionaries, which cannot make the peculiarities of learner lexicography clear. That is to say, learner lexicography can seek its specific criteria to make itself distinctive from other dictionary types, which is closely related to the origin of learner's dictionary.

The very nature of a learner's dictionary is "aimed primarily at non-native learners of a language" (Hartmann and James 2000: 82); therefore, the fundamental design philosophy of learner lexicography is essentially learning-centered.

Scholars in the field of language learning often take into account the role of learnability in the process of language acquisition and foreign/second language learning. For instance, Pinker (1989) mentioned the relationship between language learnability and second language acquisition. In his view, "learnability" mainly concerned learners' ability to learn, especially in the process of learning a language. Bertolo (2001) did an overview of the literature in the field of language learning and referred to "learnability" as the ability to learn, which

can be influenced by learning environment and learners. Some scholars also mentioned learnability in their discussion about the nature of L2 lexical learning. For example, Siepmann (2006) stated the ways of determining the core vocabulary for non-native language learners with reference to "learnability". According to him, "learnability" is one of the criteria for selecting learning content.

Comparatively speaking, the term of "learnability" has, up to now, not been much explicated in the field of learner lexicography. Isamu (2001) is perhaps the one who discusses the term learnability most in the lexicographical field. He mentions the term "learnability" when he talks about the *Idiomatic and Syntactic English Dictionary* by Hornby (ISED), which was introduced into Japan when the Japanese people adopted a hostile attitude towards the English language. In order to help Japanese EFL learners, the compilers put a lot of effort in to ensure a high learnability of the dictionary. They provided users with simplified definitions, explicit grammatical labels, detailed information about usages and collocations, frequency-based sense ordering, rich illustrative phrases and sentences, useful phonetic information and pictorial illustrations, etc. To reduce the cognitive load, terms with low frequency were also excluded. Isamu (2001) used the term "learnability" to describe an important feature of a learner's dictionary though he did not actually offer the definition of it. In his opinion, a learner's dictionary should provide its user with all the essential and necessary information needed in L2 learning; what is presented in the dictionary text has to be easily accessible and comprehensible. However, Isamu's (2001) understanding of learnability is one part of what learnability is in this article.

In fact, the philosophy of learnability has recently been reiterated by Wei et al. (2014) when they addressed the issue of designing English–Chinese bilingual learner's dictionaries. Wei et al. (2014) summarized three major design features that lexicographers should take into consideration. Firstly, the information contained should suit the learning needs of the learners. Secondly, the content should help with specific learning activities of the learners, such as reading, writing or translating. Thirdly, the arrangement of the lexicographical information should be able to help with the learning process of the learners. These features reflect one crucial requirement for a learner's dictionary, namely, "learnability". Actually, the emergence of early English monolingual learner's dictionaries "sprang from experience of linguistic analysis and from a particular approach to language pedagogy", and "the linguistic information of a certain specificity and depth had been brought to light and only special dictionaries could capture its fullness and complexity" (Cowie 1999: 1). In other words, learner's dictionaries are designed for making easier the learning process of particular groups of learners. The motivation for learner lexicography is learning and the learnability of the dictionaries should be the top priority.

Interestingly, although learnability has not been explicitly put forward to refer to the attention compilers pay to the language learning process, it has

always been the design philosophy behind the compilation practice ever since the birth of learner's dictionaries.

From 1935 to 1942, three influential dictionaries, *The New Method English Dictionary* (NMED) (1935), *A Grammar of English Words* (GEW) (1938) and ISED (1942), were published. The compilers of these three bodies of work applied their teaching experience and research results in the creation of their dictionaries with the aim of benefiting learner's language study. According to Cowie (1999), the Vocabulary Control Movement, pedagogical grammar, and phraseology had exerted the greatest influence on the dictionary compilation at that time. To be more specific, the Vocabulary Control Movement leads to dictionary compilers' conception of a core vocabulary for English language learners. The research into pedagogical grammar reminds dictionary compilers of the importance of syntactic information, English verb patterns in particular. The study on English phraseology benefits dictionary compilers in dealing with the information that enables learners to produce idiomatic English. These three language-learning-oriented linguistic studies pushed forward the emergence of learner lexicography with a distinctive feature, namely learnability.

With the development of researching second language learning⁴ and linguistic theories, the second generation of learner's dictionaries put more emphasis on content and design for the sake of learners' reception and production (Cowie 1999). In this period, the *Longman Dictionary of Contemporary English* (1978) in its compilation introduced a controlled defining vocabulary to deal with learners' difficulty in comprehending the lexical items. "During the late 1980s, EFL lexicographers kept in balance the two long-established functions of the learner's dictionary — its role as a storehouse of meanings and its role as an activator of language use and vocabulary development" (Cowie 1999: 173). Tarp (2004), in his functional theory of lexicography, has already stressed user-orientated compilation. Compilers should consider the language learner's proficiency level, level of general culture, feature of his culture, learner type and age. Along with the development of computer science and corpus-based lexical studies, analysis of users' needs has become specified and can fulfill users' more specific requirements. Nevertheless, only when computer science and corpus-based lexical studies are developed on the basis of learning need analysis can learner's dictionary truly satisfy the learning needs.

3. Essential components of "learnability" in learner lexicography

As mentioned above, learnability refers to how much useful the information is in learning and how easily the users can learn the information. In general, both the content and the presentation of a learner's dictionary should be learning-driven. Specifically speaking, learnability in bilingual learner lexicography should be reflected at least in the following three respects.

3.1 Learning load controlled

From the perspective of learnability, dictionary compilers should primarily consider three points: the target learners' cognitive capacity, their reference habits as well as skills. Briefly speaking, what has been offered in the dictionary should not increase the user's cognitive load. Ever since the birth of learner lexicography, dictionary compilers have long been attempting to do so by means like controlling defining words and providing usage labels. However, for bilingual learner's dictionaries, not all of these traditional practices are useful. When starting to learn a foreign language, non-native learners more often than not have already formed their own system of language and culture. How to make use of the existing knowledge or language system to reduce the cognitive load will have a great impact on the learning effect in that comprehension is the key to knowledge acquisition (Cao 1991).

3.2 Learning needs specified

In general, dictionaries are designed to meet their users' reference needs. Learner's dictionaries are meant for learners whose reference needs may vary at different stages of their learning process. To make clear the specific group of users' learning needs is crucial in compiling learner's dictionaries. Learnability, compared with user-oriented need analysis in the functional theory of lexicography, stresses learning-centered need analysis. That is to say, the aim of need analysis is to specify learning needs: how much useful information the dictionary should contain and how easily learners can learn the specific information. The functional theory of lexicography includes many factors to analyze users' needs; however, many users, especially beginners, are not clear about their needs and it is difficult to collect related information through questionnaires. Learner's dictionaries are designed to assist learning and they can follow the theories about the learning process and learning rules which have been tested through empirical studies in the field of language learning; therefore, learnability can reflect the regular demands in the learning process.

3.3 Integrated-learning oriented

Language learning involves understanding the linguistic system of the target language, including sense, forms, pronunciation, etc. Different aspects of language learning have to be well integrated, which benefits learners in the long term. In this case, learner's dictionaries should also be integrated-learning oriented, making systematic language learning possible. Hence the learnability of learner's dictionaries is achieved.

The above three aspects explain the major concerns of learnability in learner's dictionaries. However, in the existing literature and actual lexico-

graphical practice, lexicographers' attention is usually given to the first aspect, namely the acceptability of the learning content. The other two aspects have actually been ignored. For instance, Isamu's (2001) summary of learnability in ISED only focused on the necessary and comprehensible information provided for language learners. Therefore, it is essential that learnability is further analyzed and summarized based on the observation of some successful dictionaries. Morrison's *Wuche Yunfu*, introduced in the first section, was a popular dictionary for CFL learners in the Qing dynasty (1644–1912), proving its effectiveness in helping non-native learners with their Chinese language study. Presumably, it should reflect the above elements of learnability.

4. Robert Morrison's efforts in enhancing the learnability of *Wuche Yunfu*

As discussed above, learnability should be the most distinguishing feature of a learner's dictionary, either monolingual or bilingual. Though Morrison did not in fact know how to compile a learner's dictionary, he successfully made his dictionary very popular among generations of CFL learners, with a high degree of learnability. What has been achieved by Morrison is worth a detailed textual analysis.

4.1 Controlling the learning load

The compilation of *Wuche Yunfu* was based on Morrison's experience of using Chinese dictionaries and his practice of compiling the first volume of *A Dictionary of the Chinese Language* (1815) which was later found to be difficult for CFL learners. Thus, Morrison intentionally made some modifications in this dictionary to reduce the target learners' load of learning Chinese through two compiling means.

4.1.1 Making use of learners' existing knowledge

To lighten the CFL learner's cognitive burden, Morrison in his compilation of *Wuche Yunfu* made full use of their mother tongue as well as the knowledge they had already acquired. Being different from most modern dictionaries for non-native learners, which usually immerse users in the target language, *Wuche Yunfu* is a bilingual dictionary in which lemmata and sub-lemmata are written in Chinese, and the other parts of the entries are written in English. By defining the headword in the user's native language, *Wuche Yunfu* facilitates learners' comprehension of the lemmata.

As held by Adamska-Salaciak and Kernerman (2016: 271), "the acquisition of new knowledge proceeds via relating it to the knowledge one already possesses". Morrison in his compilation of *Wuche Yunfu* did help CFL learners with

the understanding of new linguistic knowledge based on their general knowledge. The definition of the Chinese headword 滯 (pronounced as "zhi" in Chinese and literally means "stagnate") is a case in point. In Figure 4, the Chinese headword 滯 is defined in English as "water impeded; some hindrance to the circulation of fluids" (Morrison 1819: 41) first. This original meaning of 滯 is easy for users to understand since it simply describes a kind of natural phenomenon. Then, based on this original meaning, Morrison further introduces its metaphorical meaning, i.e. "a stoppage in the human system" (Morrison 1819: 41), which is closely connected with the knowledge of Traditional Chinese Medicine (TCM). By using general knowledge as the basis and English as a presentation tool, Morrison manages to assist the users with their understanding of the difficult senses of the character. Moreover, he also avoided interrupting the users' learning with an overwhelming amount of new cultural knowledge. Instead, the shared metaphorical mechanism helps CFL learners grasp the idea of TCM.

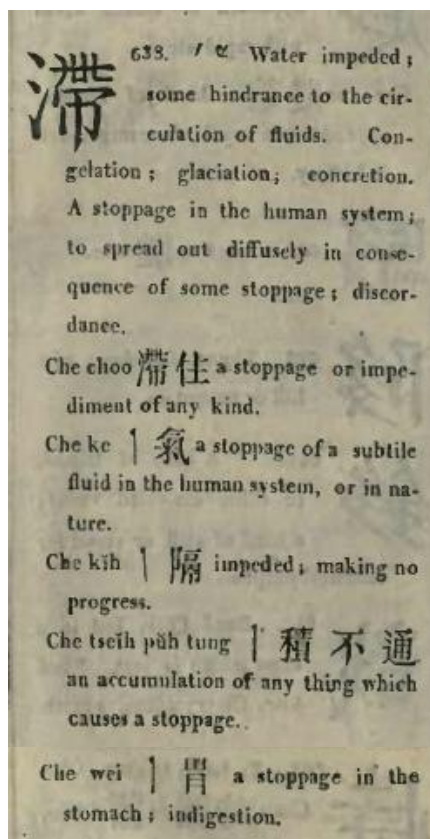


Figure 4: Part of the entry of 滯 in *Wuche Yunfu* (Morrison 1819: 41)

4.1.2 Conforming to learners' reference skills and habits

Morrison was clear about the target users of this dictionary because he had been a learner himself when he first came to China and he knew what a beginner needed when consulting a dictionary. He criticized existing dictionaries for the way of arranging Chinese characters according to the number of strokes used to write the character, which was rather difficult for CFL learners to locate the character (The Chinese language adopts ideographic writing while English uses phonetic writing; the two languages normally use two different arrangements in the dictionaries). In his opinion, when CFL learners heard a new character, they could only look for the character according to its pronunciation instead of its orthography; hence, they could not find it in the dictionary easily. To help the target users, Morrison, in the first place, adopted an alphabetical wordlist in *Wuche Yunfu*, i.e. arranged the entry words according to the alphabetical order of their *Pinyin*, the Romanization form of Chinese pronunciation, which solved this problem easily.

In the second place, Morrison provided some auxiliary lists in *Wuche Yunfu* to help learners of different reference habits locate lexical items. In the first part of *Wuche Yunfu*, Morrison presented information about the orthography of the Cantonese dialect (different from the orthography used in other parts of China), and in the table of this orthography, he provided information about the Cantonese pronunciations opposite which were the spellings that he used in the dictionary. In this way, non-native Chinese learners in Canton could locate the characters according to the pronunciation they heard in real communication. And they could get used to the system adopted in the dictionary easily. In the second part, there was a table of radicals (the compositional part of a Chinese character, which usually expresses the meaning of the whole character), an index of characters under the radicals, a list of various forms of these characters, and an index of English words which were linked by numbers to the corresponding Chinese characters in the part of the syllabic arrangement.

Through the dictionary arrangement and lists, Morrison reduced the learners' difficulty in using this dictionary. Some Chinese dictionaries for non-native learners nowadays actually intend to cope with the target users' needs in this respect as well. For instance, *The Commercial Press Learner's Dictionary of Contemporary Chinese* (2007) provides lemmata with the alphabetic sequence, and it also contains indexes of syllables, strokes, radicals as well as single-component characters (characters that develop from ancient painting and cannot be separated into parts). However, even though it intends to provide accesses for various kinds of learners, it does not fully take learners' reference skills into consideration. The index of single-component characters is quite confusing because it is a difficult grammatical phenomenon in the Chinese language. Few non-native Chinese learners can understand the single-component character easily, though the compilers have realized the difficulty and have explained what single-component characters are in the preface.

4.2 Meeting specific learning needs

Morrison set a good example in helping users find the specific information they need in their language learning. As mentioned earlier, the target users of *Wuche Yunfu* were originally missionaries who needed to communicate with the local people in China. In the preface of the dictionary, Morrison stated that "the author's object has been, and the intention of the Dictionary ought to be, to communicate the language to Europeans" (1819: viii). In other words, the dictionary was intended to enable users to learn the Chinese language and its culture. Morrison achieved his goal by three means: the provision of core vocabulary, design of phonetic symbols and adoption of a mixed model of explanation.

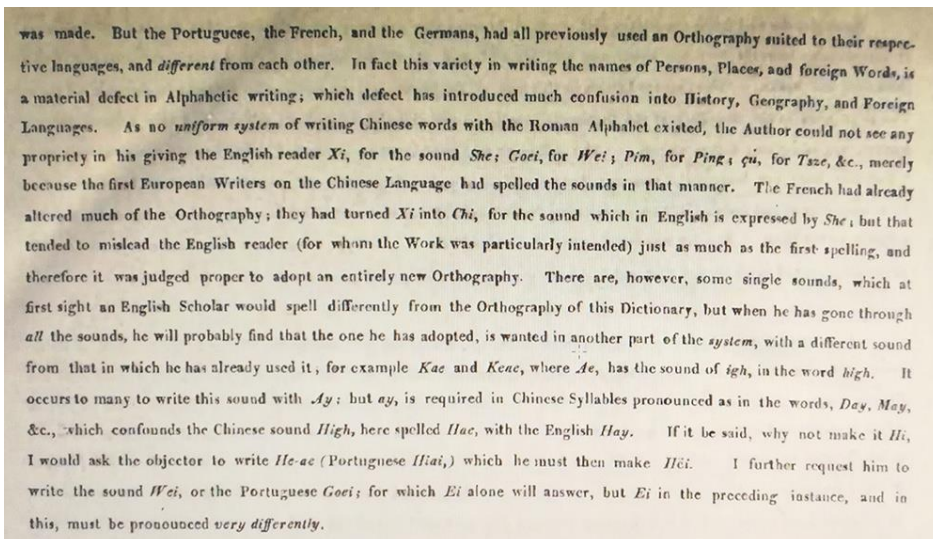
Firstly, Morrison only provided the core lexical items in the dictionary, which were enough to satisfy learners' needs for daily communication. The first part of *A Dictionary of the Chinese Language* (1815) contains about 40,000 characters and Morrison reduced the number in *Wuche Yunfu*. According to the research conducted by Yang (2012), the number of headwords in *Wuche Yunfu* is 12,674. This shows that Morrison intended to make non-native learners focus on the core vocabulary, which can well meet their basic needs in daily communication.

Secondly, Morrison considered information on pronunciation a priority to help CFL learners communicate orally. He tried to satisfy this need by devising a system of pronunciation and tones that was somewhat easier for the users to accept and learn. Morrison modified the phonetic transcription for the convenience of westerners. Otherwise, it would be very hard for them to produce certain vocal sounds, which were very different from those in their mother tongues. Actually, Morrison found that people who had arrived in China earlier from other countries had altered the pronunciation of some Chinese characters to suit their own languages; however, he was determined to keep Chinese in its original form and "adopt an entirely new orthography" (Morrison 1819: ix) by taking the whole Chinese language system into consideration. There was no Putonghua (standard spoken Chinese) then; nevertheless, Morrison himself developed a system of phonetic symbols based on the English pronunciation system.

Adopting a totally new orthography would only cause trouble for those beginners. To avoid this, Morrison made some changes to some single sounds to facilitate western beginners of Chinese. In the preface of *Wuche Yunfu*, he gave the example of "Kae" and "Kene" to explain the reason for such modifications (Morrison 1819: ix) (see Figure 5). As for the example 精 in Figure 3, "j" was modified in the way of "ts" to make it easier for westerners to pronounce. Morrison also stressed the importance of pronunciation and reinforced users' learning of it by providing pronunciation for related lexical chunks in the sub-entries. The pronunciations of the whole lexical chunks were indicated before the Chinese written forms. In each sub-entry, the pronunciation of the lexical

chunk was placed at the very beginning, for example, "Tsingke", followed by the Chinese character 精气 in traditional Chinese characters. The actual pronunciation of 气 is "qi" in Pinyin⁵, but it was not easy for westerners to pronounce. Morrison also changed it into "ke", which was pronounced by westerners in a way quite similar to its Chinese pronunciation. It is certain that Morrison made transcription changes whenever target users might find it problematic to pronounce.

Meanwhile, Morrison devised his own way of marking the different tones of Chinese characters (Morrison 1819: xiii) (see Figure 6) for CFL beginners. In Chinese, there are four tone symbols, high-level tone (first tone), rising tone (second tone), falling-rising tone (third tone) and falling tone (fourth tone). In the early 19th century, Morrison was already aware that he should use a set of symbols to help learners distinguish different tones of the characters that are pronounced similarly in Chinese. For example, the tone marker "—" of 精 is placed in [] at the very beginning of the entry. This is especially crucial for CFL learners because English is not a tonal language and westerners often find it difficult to distinguish the four tones of Chinese characters. It cannot be denied that Morrison's endeavor was a novel try at that time since no acknowledged standard pronunciation system was created to represent Chinese pronunciation features then. This applicability was very successful and the phonetic system invented by Morrison was the basis of the Wade-Giles Romanization system, which has been popular until very recently (Yang 2014), proving the effect of this method on CFL learners.



was made. But the Portuguese, the French, and the Germans, had all previously used an Orthography suited to their respective languages, and *different* from each other. In fact this variety in writing the names of Persons, Places, and foreign Words, is a material defect in Alphabetic writing; which defect has introduced much confusion into History, Geography, and Foreign Languages. As no *uniform system* of writing Chinese words with the Roman Alphabet existed, the Author could not see any propriety in his giving the English reader *Xi*, for the sound *She*; *Goei*, for *Wei*; *Pim*, for *Ping*; *cu*, for *Tze*, &c., merely because the first European Writers on the Chinese Language had spelled the sounds in that manner. The French had already altered much of the Orthography; they had turned *Xi* into *Chi*, for the sound which in English is expressed by *She*; but that tended to mislead the English reader (for whom the Work was particularly intended) just as much as the first spelling, and therefore it was judged proper to adopt an entirely new Orthography. There are, however, some single sounds, which at first sight an English Scholar would spell differently from the Orthography of this Dictionary, but when he has gone through *all* the sounds, he will probably find that the one he has adopted, is wanted in another part of the *system*, with a different sound from that in which he has already used it, for example *Kae* and *Keae*, where *ae*, has the sound of *igh*, in the word *high*. It occurs to many to write this sound with *Ay*; but *ay*, is required in Chinese Syllables pronounced as in the words, *Day*, *May*, &c., which confounds the Chinese sound *High*, here spelled *Hac*, with the English *Hay*. If it be said, why not make it *Hi*, I would ask the objector to write *He-ae* (Portuguese *Hiai*;) which he must then make *Hi*. I further request him to write the sound *Wei*, or the Portuguese *Goei*; for which *Ei* alone will answer, but *Ei* in the preceding instance, and in this, must be pronounced *very differently*.

Figure 5: Picture of the Examples Mentioned in the Preface of *Wuche Yunfu*

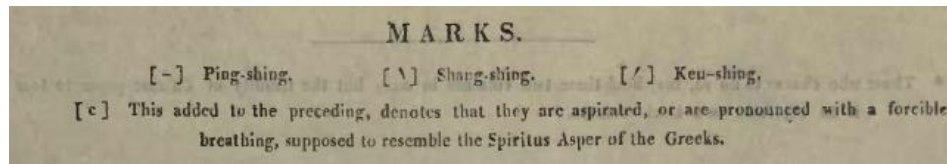


Figure 6: Picture of Marks in the Preface of *Wuche Yunfu*

Thirdly, Morrison found that in the process of learning Chinese, CFL learners need to understand the Chinese culture to avoid making any mistake in this respect in communication. Hence, Morrison provided clues to remind the learners of the cultural elements associated with Chinese characters whenever possible.

For instance, Traditional Chinese Medicine (TCM), which has been practiced in China for nearly two thousand years, is an indispensable part of traditional Chinese culture, which is deemed as a key part in the transmission of the Chinese culture to the west. However, TCM is different from western medicine in many aspects. Westerners who have been brought up by concepts of western medicine may find it difficult to comprehend concepts in TCM (Wiseman and Zmiewski 1989). In turn, it is difficult to find equivalents in the western languages, leaving westerners in a kind of predicament in studying TCM-related lexical items and in cultural transmission. Therefore, studying TCM-related expressions in Morrison's *Wuche Yunfu* can help us have a clearer picture of how Morrison helped the non-native learners avoid mistakes concerning these Chinese characters.

Take some characters referring to typical Chinese medicines as an example (see Table 1).

Table 1: Concepts Concerned with "Medicines"

Chinese Medicine Terms	Definition
药	From plant and to harmonize. Medicinal plants; medicines; to heal; an ingredient, applied to various compositions made up as medicines are.
阿胶	mule or asses glue, a famous Chinese medicine
硼砂	borax sub borate of soda, used in medicine
枇杷叶	the leaves of the loquat tree, used as a medicine to treat coughs

Chinese medicines are vastly different from western medicines. For the general term of medicine in Chinese, 药, Morrison described the unusual contents in Chinese medicines. For some culture-specific medicines, he simply pointed out what the substance actually is in English, such as "mule or asses glue" for 阿胶. Then he further explained that it was some kind of Chinese medicine, which makes the specialty of its nature apparent to the users. 枇杷叶 is defined as "the leaves of the loquat tree, used as a medicine to treat coughs", which not only explains what the three-character chunk refers to, but also provides CFL learners with its main function in the Chinese culture. In a word, he put "Chinese" and "medicine" in the explanation to prevent CFL learners from making mistakes in their communication.

These three aspects show that Morrison considered it a priority to meet CFL learners' needs in language learning. He provided a core vocabulary, devised a system of pronunciations and tones, and helped learners avoid mistakes in cultural communication. In this way, Morrison achieved his goal through content design, focusing on learners' needs of cross-cultural communication.

4.3 Activating an integrated learning

As a language of ideographical type, the Chinese language is rather difficult for learners whose native tongue is of phonographical nature. To help CFL learners, learner's dictionaries must activate an integrated learning of the linguistic system of the Chinese language. The most challenging part of learning this system for CFL learners is the Chinese character. The major difficulties confronting CFL learner from the west concern recognizing, memorizing and using Chinese characters. The interrelationship among the form, pronunciation and meaning of the Chinese character is crucial for CFL learning.

In *Wuche Yunfu*, Morrison attempted to emphasize an integrated learning through revealing the connection between these three aspects, especially the relationship between form and meaning. This can be seen from his stress over forms, description of the connection between form and meaning, and the presentation of these three aspects in specific entries.

Firstly, Morrison emphasized the unique status of form in the Chinese language. Form, an indispensable part of Chinese characters, is closely related to their meanings. Just as he stated in the preface, "a knowledge of abbreviated forms must be acquired in the same way as a knowledge of the running hand in any Alphabetic Language, where the scope and connexion assist the Reader in determining for what the imperfectly formed letter is intended" (Morrison 1819: vi). Morrison often provided different written forms for the headwords, building a tangible connection among the variants and saved the users from further trouble with distinguishing the relationship between these variants in real communication.

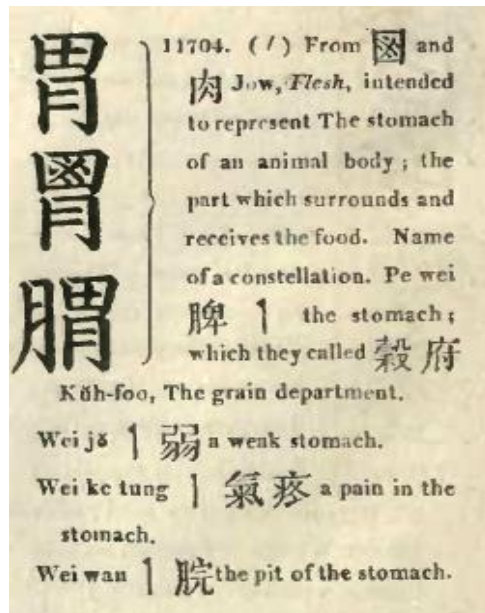


Figure 7: Entry of 胃 in *Wuche Yunfu*

For the character 胃 (Morrison 1819: 978) (see Figure 7), Morrison gave three different forms, which were commonly used by the Chinese people at that time. He did not provide all the forms those characters had, because a full list of different forms of writing might frustrate the target users at the beginning of their study, or even confuse them in their production of the language. He also avoided providing forms that were questionable and just inserted characters with correct forms that were used by a majority of the local Chinese people, and used braces to connect these forms.

Secondly, Morrison intentionally displayed the relation between the form and the meaning. In terms of 胃, he provided the motivation for the formation of this character as "intended to represent the stomach of an animal body; the part which surrounds and receives the food", which vividly explained the close connection between the meaning and the form or structure of the character. In Figure 8, the character 跌 (meaning "fall" in English) is explained as "from foot and to miss or lose" (Morrison 1819: 828), which is closely associated with the formal composition of this character. That is, the left part of the character means "foot" in Chinese, and the right part denotes "miss". The relationship between the form and the meaning of the character not only helps learners memorize the character, but also enhances their concentration on this feature of Chinese characters, which, in the long run, facilitating their learning of the Chinese language as a whole.

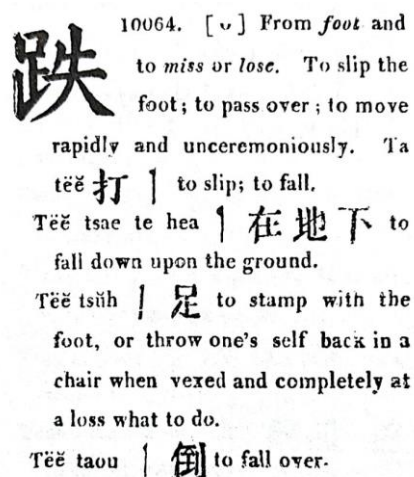


Figure 8: Entry of 跌 in *Wuche Yunfu* (Morrison 1819: 828)

Thirdly, Morrison provided pronunciation, form, and meaning together, intending to make learners realize that these three constitute Chinese characters. This can be seen clearly in the entry and sub-entries of 胃. Below the three variant forms of the character on the left, Morrison labeled the pronunciation. On the right, he provided equivalents and explanations of the senses. Then, he described several symptoms and diseases in the sub-entry. He used phonetic symbols first and traditional Chinese characters next to the pronunciations. After that, he gave brief explanations instead of western medicine equivalents in English. For example, 脾胃 (the stomach); 胃弱 (a weak stomach); 胃气疼 (a pain in the stomach), and 胃脘 (the pit of the stomach). With this way of arrangement, the users learn the pronunciation, form, and meanings together. As presented in the entry or sub-entries, Morrison attempted to fix in users' minds a notion that the acquisition of the Chinese language requires an integrated learning of all these three aspects, which is vastly different from the learning of English.

5. Implications on the compilation of future C–E learner's dictionaries

In previous parts, the major elements of learnability have been analyzed and further explored with substantiation in the historical text of Morrison's *Wuche Yunfu*. It is "learnability", the design philosophy behind Morrison's lexicographical practice that makes *Wuche Yunfu* a successful bilingual dictionary, which has benefited non-native learners for many generations. In recent decades, the craze for learning the Chinese language has provided new momentum to the compilation and publication of Chinese–English learner's dictionary.

ies. Based on the previous discussion of learnability and how it is reflected in Morrison's dictionary, some insights can be summarized concerning the compilation of such dictionaries.

First of all, to guarantee the learnability of a learner's dictionary, compilers need to reduce the target learners' learning load and pay special attention to bridging the gap between user's native language and the target language. When the target users are non-native beginners, compilers can make full use of their mother tongues, the knowledge they have or other means to control the learning load.

Secondly, to achieve greater learnability, compilers should provide guidance concerning learning contents that can satisfy learners' specific needs. More often than not, CFL learners are not sure about what they should learn in order to achieve their learning goals. Morrison provided a core vocabulary, devised pronunciation and tone marking systems and endeavored to prevent learners from making mistakes in terms of culture-dependent characters. All these efforts conform to western learners' purposes to learn Chinese in their daily communicative settings.

Last but not least, a bilingual learner's dictionary is a pedagogical tool and its compilers should bear in mind that "while learning meaning is undoubtedly an essential initial step, more precisely this involves developing a link between form and meaning" (Schmitt 2014: 27). The design philosophy of learnability in learner lexicography means that compilers need to activate an integrated learning of the target language. Only in this way can dictionaries to a large extent, help learners comprehend various aspects of the target language and learn the language efficiently. Morrison accomplished it by stressing the basic features of the Chinese language, that is, the close relationship between form, pronunciation, and meanings. The knowledge of the target language's features can motivate learners' deeper understanding of the language and benefit their learning process.

In a word, learnability as the design philosophy of learner lexicography should be the fundamental conception and practical basis of compilers' subjectivity. Based on thorough analyzes of the reference needs, skills and habits of the target users, lexicographers of learner's dictionaries will achieve a high degree of learnability in their compilation practice.

6. Concluding remarks

As held by Adamska-Salaciak and Kernerman (2016), some old dictionaries may seem obsolete; however, the principles these dictionaries follow are never out of date. This is especially true for Robert Morrison's *Wuche Yunfu* (1819). Many of his lexicographical efforts are pioneering and effective in promoting CFL learning. However, unfortunately, previous studies on Morrison's *Wuche Yunfu* are mostly confined to the discussion about its historic influence in cultural transmission. Morrison's great contribution to CFL learner lexicography

had been to a large extent ignored, if not totally forgotten. The literature review shows that a detailed analysis of the text of Morrison's *Wuche Yunfu* is still lacking, even though it does serve as a good example of a learner's dictionary. The rich experience of compiling C–E dictionaries for CFL learners in this leading lexicographical work is definitely worth exploring.

Learnability as a design philosophy is crucial to the success of learner lexicography. It can be seen from the existing literature that learnability has been greatly ignored in the lexicographical field. Nevertheless, related research should be encouraged in that learnability is the foundation and premise of improvement on the design features of learner's dictionaries. Based on the description of learnability in the field of second language acquisition and bilingual learner lexicography, the authors illustrated the connotation and three dimensions of learnability in the lexicographical field, which are reducing the learning load, meeting the learning needs and skills and activating the integrated learning.

These dimensions have been well reflected in *Wuche Yunfu*, which win huge popularity among CFL learners for generations. In the dictionary, Morrison adopted several approaches to ensure learnability out of his learning experience. His compiling philosophy exceeded his time and left a legacy for contemporary compilers, especially those who write learner's dictionaries for CFL learners worldwide. What lexicographers need to do is to improve learnability by any possible means so as to produce a better dictionary to facilitate learner's second or foreign language learning.

Acknowledgements

We are grateful to Elsabé Taljard, her colleague and two anonymous reviewers for their insightful comments on the earlier drafts of this paper.

Endnotes

1. *Wuche Yunfu* (五车韵府), *A Dictionary of the Chinese Language Part II*, was compiled and published in 1819, which was, according to the preface written by Morrison, founded on the original Chinese version of *Wuche Yunfu* compiled by a Chinese lexicographer.
2. HSK is the abbreviation of *Hanyu Shuiping Kaoshi*, which refers to the only official standard Chinese proficiency test for non-native Chinese speakers. It offers a ranking system of vocabulary lists for learners of different proficiency level.
3. Level four of HSK means the learner has a vocabulary of about 1,200 Chinese characters and he can communicate with native Chinese fluently. HSK (level 4) tests candidates' Chinese capacity, which corresponds to level 4 of the international Chinese language ability standard and level B2 of *A Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEF). Level six means the learner has a vocabulary of more than 5,000 Chinese characters and he can read Chinese newspapers, watch Chinese movies and deliver speeches

in Chinese. HSK (level 6) tests the candidates' Chinese ability, which corresponds to level C2 of CEF.

4. Ellis defines Second Language Acquisition (SLA) as the study on "the way in which people learn a language other than their mother tongue, inside or outside of a class room" (1997: 3). In this paper, L2 and FL are used with this definition.
5. Pinyin refers to the standardized phonetic symbols for Chinese characters in contemporary China.

References

A. Dictionaries

- Lu, Jianji and Lü, Wenhua (Eds.).** 2007. *The Commercial Press Learner's Dictionary of Contemporary Chinese* (商务馆学汉语词典). Beijing: The Commercial Press.
- Morrison, Robert (Ed.).** 1819. *A Dictionary of the Chinese Language. Part II (Wuche Yunfu 五车韵府)*. Macao: Printed at the Honorable East India Company's Press, by P.P. Thoms.
- Simpson, J.A. and E.S.C. Weiner.** 1989. *The Oxford English Dictionary*. (OED) Second edition. Oxford: Oxford University Press.
- Xinhua Dictionary (Chinese-English Edition)*. 2013. Beijing: The Commercial Press.

B. Other literature

- Adamska-Salaciak, Arleta and Ilan Kernerman.** 2016. Introduction: Towards Better Dictionaries for Learners. *International Journal of Lexicography* 29(3): 271-278.
- Alzi'abi, Safi Eldeen.** 2017. Guessing Verb-Adverb Collocations: Arab EFL Learners' Use of Electronic Dictionaries. *Lexikos* 27: 50-77.
- Bertolo, Stefano.** 2001. A Brief Overview of Learnability. Bertolo, Stefano (Ed.). 2001. *Language Acquisition and Learnability*: 1-14. Cambridge: Cambridge University Press.
- Bogaards, P. and W. van der Kloot.** 2001. The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography* 14(2): 97-121.
- Cao, Nanyan.** 1991. *Cognitive Learning Theory*. Kaifeng: Henan Education Press.
- Chon, Y.V.** 2008. The Electronic Dictionary for Writing: A Solution or a Problem? *International Journal of Lexicography* 22(1): 23-54.
- Cowie, A.P.** 1999. *English Dictionaries for Foreign Learners. A History*. Oxford: Clarendon Press.
- Ellis, Rod.** 1997. *Second Language Acquisition*. Oxford: Oxford University Press.
- Faaß, G., Sonja E. Bosch and Rufus H. Gouws.** 2014. A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages. *Lexikos* 24: 94-115.
- Hartmann, R.R.K. and Gregory James.** 2000. *Dictionary of Lexicography*. Beijing: Foreign Language Teaching and Research Press.
- Isamu, Hayakawa.** 2001. *Methods of Plagiarism: A History of English-Japanese Lexicography*. Tokyo: Jiyusha.

- Laufer, Batia and Tamar Levitzky-Aviad.** 2006. Examining the Effectiveness of 'Bilingual Dictionary Plus' — A Dictionary for Production in a Foreign Language. *International Journal of Lexicography* 19(2): 135-155.
- Lew, Robert and Anna Dziemianko.** 2006. A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and Its Usefulness for Conveying Syntactic Information. *International Journal of Lexicography* 19(3): 225-242.
- Li, Yan.** 2013. On the Compilation of General-purpose Chinese Dictionaries for Foreign Learners of Chinese. *Lexicographical Studies* 5: 34-39.
- Pinker, Steven.** 1989. *Learnability and Cognition*. M.A. thesis. Cambridge: MIT Press.
- Ryu, Hyun-Guk.** 2009. Robert Morrison's Influence on Translation, Printing, and Publishing in Asia. *Design Discourse* 4(2): 1-13.
- Schmitt, N.** 2014. *Researching Vocabulary: A Vocabulary Research Manual*. Beijing: Foreign Language Teaching and Research Press.
- Siepmann, Dirk.** 2006. Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects. *International Journal of Lexicography* 19(1): 1-39.
- Tarp, Sven.** 2004. Basic Problems of Learner's Lexicography. *Lexikos* 14: 222-252.
- Wang, Yajun.** 2008. Translation of Culture-specific Words in the Chinese–English Dictionary for Non-native Chinese Learners. *Journal of Xiamen University of Technology* 3: 104-108.
- Wei, Xiangqing et al.** 2014. *A Research on the Designing Features of Bilingual Learner's Dictionaries*. Beijing: Foreign Language Teaching and Research Press.
- Wiseman, Nigel and Paul Zmiewski.** 1989. Rectifying the Names: Suggestions for Standardizing Chinese Medical Terminology. Unschuld, Paul U. (Ed.). 1989. *Approaches to Traditional Chinese Medical Literature*: 55-66. Dordrecht: Kluwer Academic Publishers.
- Wu, Xian and Liren Zheng.** 2009. Robert Morrison and the First Chinese–English Dictionary. *Journal of East Asian Libraries* 147: 1-12.
- Yang, Han.** 2015. *The Investigation of Use the Chinese Dictionary Status quo of International Students in China*. M.A. thesis, Chongqing Normal University.
- Yang, Huiling.** 2012. *Tradition of Chinese–English Dictionaries in the 19th Century: Genealogy Research of Morrison's, Williams's and Giles' Chinese–English Dictionaries*. Beijing: The Commercial Press.
- Yang, Huiling.** 2014. The Making of the First Chinese–English Dictionary: Robert Morrison's Dictionary of the Chinese Language in Three Parts (1815–1823). *Historiographia Linguistica* 41(2–3): 299-322.
- Yang, Jinhua.** 2016. On the Four Principles of Compiling Chinese Dictionaries for Foreign Learners. *Lexicographical Studies* 1: 45-51.

Corpus-Based Research on Terminology of Turkish Lexicography (CBRT-TURKLEX)*

Erdoğan Boz, *Center for Lexicography, Turkish Language and Literature Department, Eskişehir Osmangazi University, Eskişehir, Turkey (erdoganboz@ogu.edu.tr)*

Ferdi Bozkurt, *Turkish Language Department, Anadolu University, Eskişehir, Turkey (ferdib@anadolu.edu.tr)*

and

Fatih Doğru, *Turkish Language and Literature Department, Eskişehir Osmangazi University, Eskişehir, Turkey (fdogru@ogu.edu.tr)*

Abstract: In this paper, we introduce an ongoing lexicographic corpus project. The Center for Lexicography, abbreviated as SÖZMER, was established under the aegis of Eskişehir Osmangazi University to support lexicographical projects. SÖZMER decided to initiate a corpus-based Turkish lexicography project. This project will be the first stage of the endeavour aimed at preparing a specialized dictionary for Turkish lexicography. The primary aim of the project is to prepare an electronic corpus for researchers of Turkish lexicography. The secondary aim of the project is to obtain a word list of Turkish lexicographic terms. This paper presents a description of the process of data collection and the methodology employed for building a specialized corpus. The study contains an outline of the project background, needs, problems, and the phases of corpus building.

Keywords: TURKISH LEXICOGRAPHY, TERMINOLOGY, CORPUS LINGUISTICS, DICTIONARY, DATA COLLECTION, DATABASE, TERM EXTRACTION

Opsomming: Korpus-gebaseerde navorsing op terminologie van die Turkse leksikografie (CBRT-TURKLEX). In hierdie artikel word 'n lopende leksikografiese projek bekend gestel. Die Sentrum vir Leksikografie, afgekort tot SÖZMER, is onder die vaandel van die Eskişehir Osmangazi Universiteit tot stand gebring om leksikografiese projekte te ondersteun. SÖZMER het besluit om 'n korpus-gebaseerde Turkse leksikografieprojek te inisieer. Hierdie projek

* An earlier version of this article was presented at the 3rd International Conference of Lexicography, which was hosted by SÖZMER (Eskişehir Osmangazi University Center for Lexicography) in Eskişehir Osmangazi University, Turkey, 1-3 November 2016. This work was supported by Scientific Research Projects Coordination Unit of Eskişehir Osmangazi University. Project number 2016-019056.

sal die eerste fase vorm van die strewe wat die skep van 'n gespesialiseerde woordeboek vir Turkse leksikografie ten doel het. Die primêre oogmerk van die projek is om 'n elektroniese korpus vir navorsers van die Turkse leksikografie voor te berei. Die sekondêre oogmerk van die projek is om 'n woordelys van Turkse leksikografiese terme te verkry. In hierdie artikel word 'n beskrywing gegee van die proses van dataversameling en die metodologie wat gebruik word vir die bou van 'n gespesialiseerde korpus. 'n Oorsig word gegee van die projekagtergrond, behoeftes, probleme, en die fases van korpusbou.

Slutelwoorde: TURKSE LEKSIKOGRAFIE, TERMINOLOGIE, KORPUSLINGUISTIEK, WOORDEBOEK, DATAVERSAMELING, DATABASIS, TERMONTTREKKING

1. Introduction

The first dictionary work in Turkish began with Mahmut Kashgar. He started writing his *Divânu Lüğati't-Türk* (Dictionary of Turkish Languages) in January 1072 and completed it in February 1074. Turkish lexicography has a long tradition spanning over centuries; however, it is found to be deficient in many aspects, including the realm of theoretical studies which are still not adequate. To date, there is no handbook of lexicography for Turkish lexicographers. Especially considering that for English, there are many handbook studies including Zgusta (1971), Jackson (2002), Van Sterkenburg (2003), Atkins and Rundell (2008), and Svensén (2009). The main reason for the delay in Turkish lexicographical research is the fact that academic institutions that would support field research and researchers have still not reached the desired numbers or the scientific levels. The Turkish Language Institute (*Türk Dil Kurumu*), which was established in 1932, is regarded as a milestone for linguistic research in Turkey. Furthermore, studies in the field of Turkish lexicography began to acquire a scientific character with the establishment of the Turkish Language Institute. Various studies related to the field of Turkish lexicography have been carried out by Turkish researchers (Levend 1957; Parlatur 1995; Aksan 1998 et al.). These studies have made considerable contributions to the development of the Turkish lexicographic literature. Various problems have been discussed in this process, but there are crucial unsolved problems in the field of Turkish lexicography. One of these problems is that a standardized terminology accepted by field experts has not yet been established. The first study about problems in Turkish lexicography was carried out by Tietze in 1976.

Other researchers such as Aksan (1990), Boz (2006), Boz (2011), Bozkurt (2017) have published various studies on Turkish lexicography, however, standardization of the specialized terminology of Turkish lexicography — both practical and theoretical — have not been provided by these studies.

Language for specific purposes (LSP) dictionaries such as those by Hartmann and James (1998), Burkhanov (1998) and the glossaries appended at the end of research studies such as those by Robinson (1983), Van Sterkenburg (2003) and Jackson (2013) have been very useful in standardization of lexicographical

terminology.

To date, no significant research has been published covering all terms related to Turkish lexicography. The absence of a comprehensive list of terminology or a dictionary of Turkish lexicography has given rise to standardization problems among researchers.

Despite the increase in the number of research and educational centers such as universities and research institutes, especially in the period of the Turkish Republic, terms in the field of Turkish lexicography could not be gathered together, and the usage of lexicographical terms was not presented scientifically and systematically.

Instead of studies utilizing intuitive approach; studies that will allow the use of corpus linguistics, statistics, and computer-aided linguistics operation modes will generate more objective and more scientific results. Hence, in recent years, it has given rise to the so-called "corpus revolution" (Rundell and Stock 1992; Bergenholtz and Tarp 1995; Krishnamurthy 2002, 2008; Hanks 2012). A systematic, principled, scientific terminology study needs to be carried out by researchers for the development of the quality of the texts in the field of Turkish lexicography.

Term preference in cases of multiple terms for a single concept in Turkish lexicography is based on subjective approaches, or small discussions in academic communities of several people. Hence, extensive studies in the field of lexicography will increase the quality of terminology usage. Furthermore, there is no Turkish lexicography platform where researchers can agree on the usage of lexicographical terms by analyzing the tendencies in the corpus. Bowker and Pearson (2002: 12) state that "A special purpose corpus is one that focuses on a particular aspect of a language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers)." Therefore, an LSP corpus for Turkish lexicography is important with regard to providing term unity in the field of Turkish lexicography.

2. Aim of CBRT-TURKLEX

The main aim of the CBRT-TURKLEX is to build a lexicographical corpus for researchers that consists of master dissertations, doctoral theses, published presentations, news, books, articles, and reviews about the field of Turkish lexicography.

The secondary aim of the project is to obtain a word list of Turkish lexicographic terms, and to determine polysemy, synonymy, and term preferences among authors.

3. Method of CBRT-TURKLEX

The CBRT-TURKLEX project consists of five main phases.

3.1 Determination of corpus content and scope

There is no academic journal which relates only to Turkish lexicography in Turkey. However, there are many academic journals addressing grammar and linguistics research studies. Topics related to Turkish lexicography are generally published in the linguistics and grammar journals.

The articles, published presentations, books, master dissertations, doctoral theses, news and reviews were considered for CBRT-TURKLEX by the project researchers. Texts produced between 1932 (the year of the establishment of the Turkish Language Institution) and 2016 (the year of the project initiation) were collected for the corpus.

The texts containing the keywords "sözlük" (dictionary), "lügat" (dictionary, an old usage), "sözlükbilim" (lexicography), "sözlük bilim" (lexicography), "sözlükbilimi" (lexicography), "sözlük bilimi" (lexicography), "sözlükçülük" (synonym with lexicography), "leksikografi" (lexicography) were included in the corpus. A total of 1003 texts were identified as a result of this search. The types and the number of the texts included in the corpus are presented in Table 1.

<i>Text Type</i>	<i>Number of Texts</i>
<i>Master dissertations</i>	39
<i>Doctoral theses</i>	12
<i>Published presentations</i>	310
<i>News</i>	21
<i>Books</i>	3
<i>Articles</i>	468
<i>Reviews</i>	150
Total	1003

Table 1: Text types included in the corpus database

3.2 Digitization of printed texts

Some of the specified texts were in print format and others were in portable document format (PDF). Printed texts were transferred to the digital medium by means of optical character recognition (OCR) scanning. Texts in PDF were converted to OCR format by Abbyy Finereader 11© software.

In the process of conversion to OCR format, information such as bibliography, name of the journal, and page number in each text were deleted. An article page which was imported into Abbyy Finereader 11© software is presented in Figure 1.

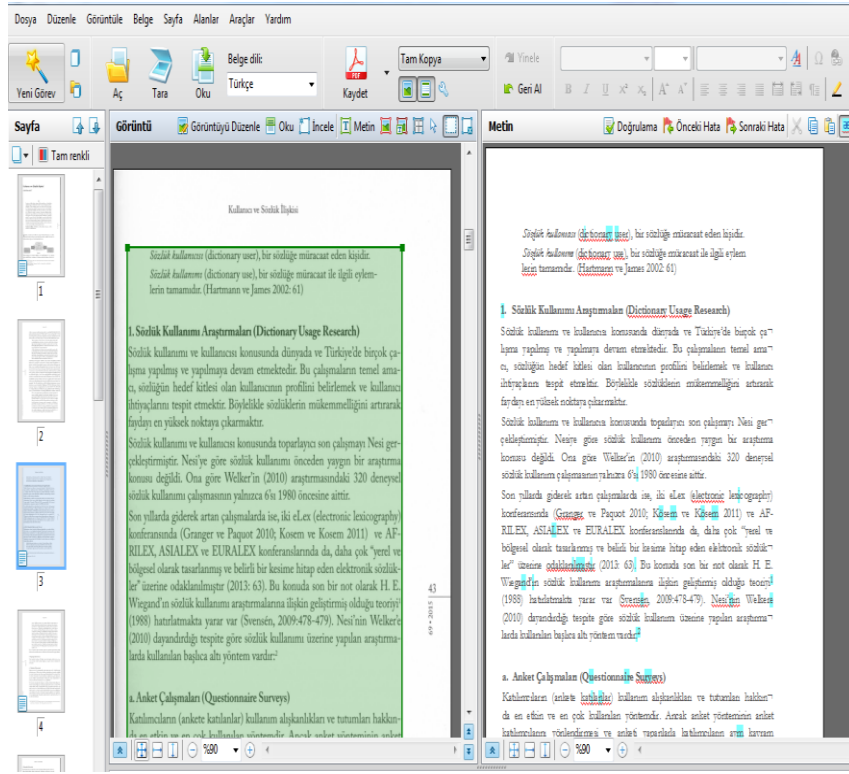


Figure 1: Converting PDF files to OCR format

The text page contains details such as the name of the article "Kullanıcı Sözlük İlişkisi", the number of the page, the year of the publication and the volume of the journal in which the article was published. These details were not included in the text corpus due to the software considering this information as junk.

3.3 Uploading of texts into the corpus

Once the conversion phase was complete, the machine-readable texts were uploaded into the corpus in the following stages.

3.3.1 Determination of metadata for the corpus

Information about the texts in the corpus means metadata, in other words metadata is data about data. This information may include the title, author, publisher and date of a written text, or details of the speakers in a spoken text (Baker et al. 2006: 115). Authors' names/last names and the publication year of

the text were identified as the metadata in the corpus for Turkish lexicography. The metadata screen is shown in Figure 2.

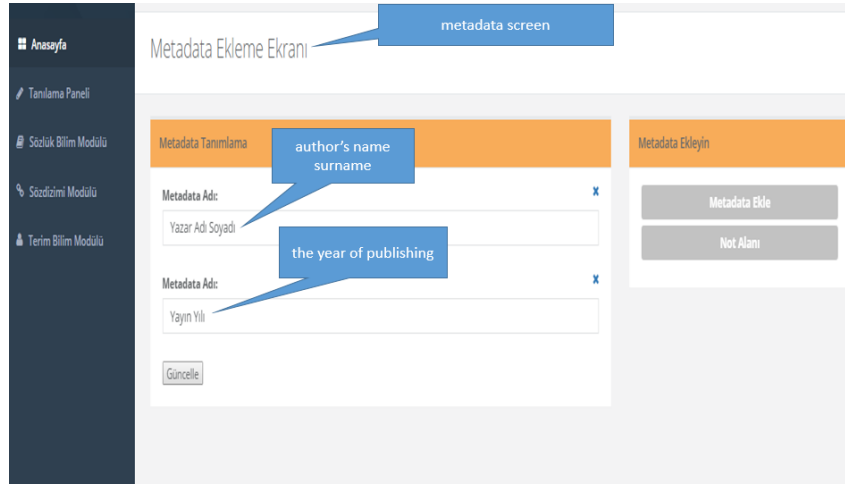


Figure 2: Metadata of the texts

3.3.2 Determination of layers for the corpus

In corpora, it is necessary to decide at the beginning on correct clustering of the texts for reporting corpus findings (Kupietz 2016: 68-70). The texts related to the field of lexicography are classified into seven different types in the database as shown in Figure 3.

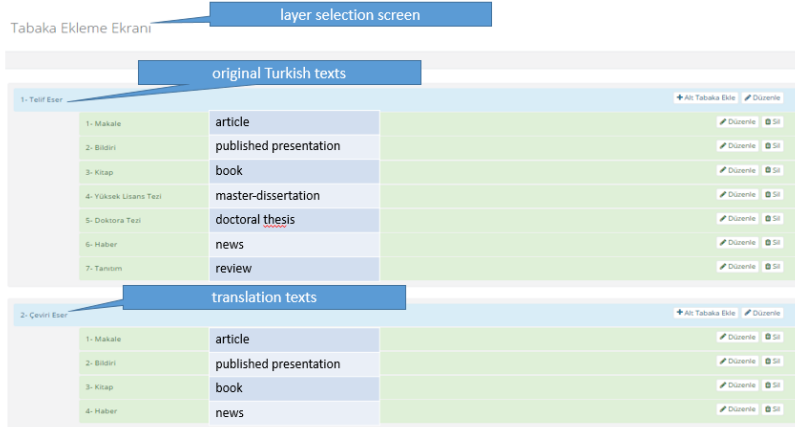


Figure 3: Layer selection screen

Layers of the corpus are articles, published presentations, books, masters-dissertations, doctoral theses, news and reviews. It is possible to report the frequency and dispersion of the terms according to the text types through these layers.

3.4 Lemmatizing of words

Francis and Kučera (1982: 1) define a lemma as a 'set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling. Inflected forms of WALK as a lemma are given by Francis and Kučera. These are *walk*, *walked*, *walking* and *walks*.

A lemmatization process was necessary for CBRT-TURKLEX due to the fact that Turkish is an agglutinative language. There are two kinds of suffixes in this language. Some of the suffixes are inflectional suffixes and the others are derivational suffixes. Derived words are accepted as separate lemmas, but inflected ones are not considered as separate lemmas.

Various inflected forms for the lemma SÖZLÜK (dictionary) lemma are given in Figure 4.

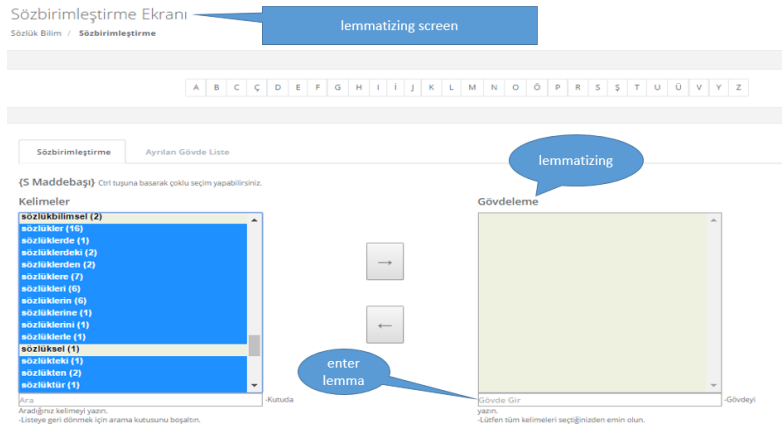


Figure 4: Lemma selection

As shown in Figure 5, "sözlükler" (dictionaries), "sözlüklerde" (in dictionaries), "sözlüklerdeki" (that in dictionaries), "sözlüklerden" (from dictionaries), "sözlüklere" (to dictionaries), "sözlükleri" (dictionaries, accusative form), "sözlüklerin" (of dictionaries), "sözlüklerine" (to their dictionaries), "sözlüklerini" (their dictionaries, accusative form), "sözlüklerle" (with dictionaries), "sözlükteki" (that in dictionary), "sözlükten" (from dictionary), "sözlüktür" (is dictionary). As shown in Figure 5, "SÖZLÜK" (dictionary) is the lemma of these inflected forms.

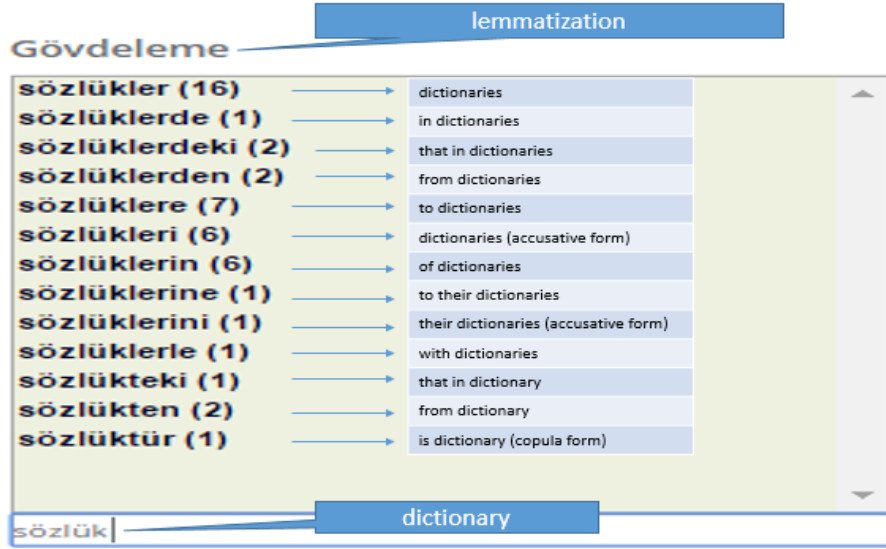


Figure 5: Screen for lemmatization

3.5 Tagging terms (identification and extraction of terms)

Some of the words in the corpus such as "bu" (this) and "güzel" (beautiful) cannot be lexicographic term candidates. At this stage, term candidates related to the field of lexicography will be selected from the sample sentences by means of "term extraction tab". For instance, the word "genel" (general) can be a lexicographic term or not, according to context.

A sample sentence which includes the word "genel" is illustrated in Figure 6. The sentence is not marked since the "genel" word is regarded as a non-lexicographic term.

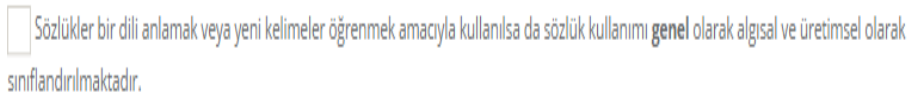


Figure 6: Term extraction tab (□ is not a term)

A sample sentence which includes the word "genel" is illustrated in Figure 7. The sentence is marked since the word "genel" is regarded as a lexicographic term.

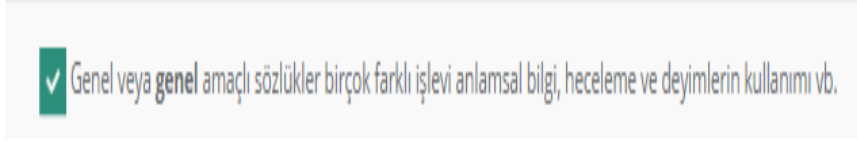


Figure 7: Term extraction tab (is a term)

This decision procedure was followed for all of the term candidates in the corpus.

Not only single-word terms but also multi-word terms appear in the field of Turkish lexicography. Collocations, in which two or more words constitute or enter into a syntactic unit, also had to be marked in the corpus (Bergenholtz and Tarp 1995: 118).

Collocations were determined with collocation screen as can be seen in Figure 8. Word collocations could be listed on the screen. Collocational relations could be provided for the left and right of the center word.



Figure 8: Collocation screen

As can be seen in Figure 9, the query for the word "genel" was input as n-4. The four words to the left of "genel", "yola çıkılarak tek dilli" turned as results and were shown in bold in the query screen. As a result of this query the word "sözlük" to the right of "genel" was deduced to be related by the researcher. The lexicographical term in this context was determined as "tek dilli genel sözlük" meaning "monolingual general dictionary".

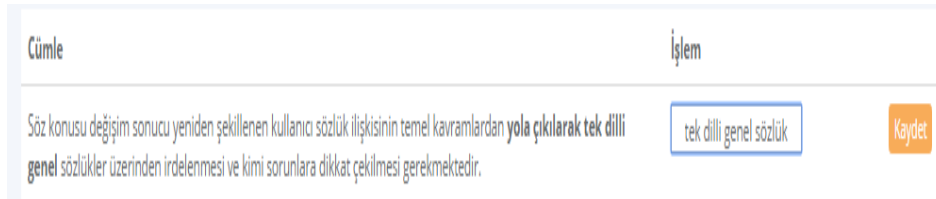


Figure 9: Collocational words query screen

Tagging terms in the corpus is conducted by multiple project researchers to eliminate individual mistakes and decisions based on intuition. Figure 10 shows the screen for notes. The project researcher's decision, whether a word is a term or not, can be followed in the notes screen.

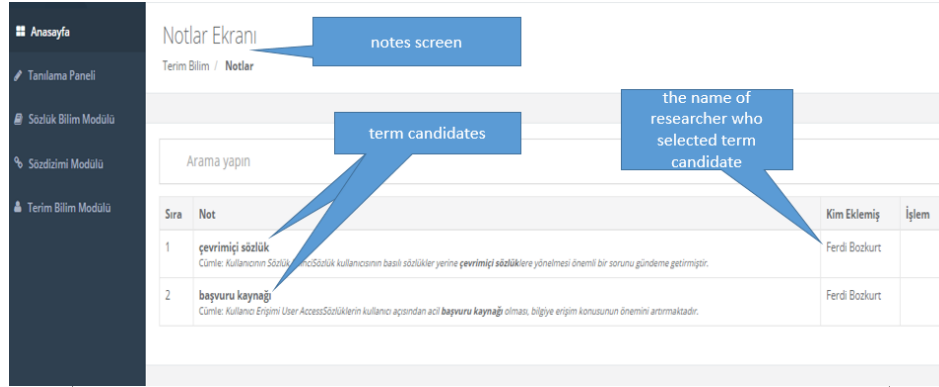


Figure 10: Notes screen

4. Conclusion

In this article a research project, namely Corpus-Based Research on Terminology of Turkish Lexicography, has been presented. The project is conducted by the Center for Lexicography at Eskişehir Osmangazi University.

The processes for the determination of the terms within the scope of the study are presented in this article. Totally 1003 texts were determined on the field of Turkish lexicography. 329 texts were in printed form. These were scanned to PDF. 674 texts were already in PDF. All of the PDF texts were converted to OCR format.

The corpus was built on October 10th, 2017. The website of the corpus is available at www.tsd.ogu.edu.tr for lexicographers. The corpus contains 1003 texts. It comprises 42.831 sentences, 703.986 orthographic words, and 86.368 types.

The frequency, dispersion, and the author's word preferences of term candidates were examined in the corpus. 1.616 lexicographic terms were determined in the corpus by the project researchers.

Future Work

A Dictionary of Turkish Lexicography will be compiled through the corpus.

Acknowledgement

1. The software of the corpus was utilized by Assoc. Prof. Dr. Bülent ÖZKAN's "TÜBİTAK-SOBAG - 114E791 Do It Yourself Corpora" project.
2. The project was completed on October 10th, 2017. The website of the corpus is available at www.tsd.ogu.edu.tr.

References

- Aksan, D. 1990. *Her Yöniyle Dil*. Ankara: Türk Dil Kurumu Yayınları.
- Aksan, D. 1998. Türklere Sözlükçülük, Bugün Türkiye`de Sözlük. *Kebikeç Dergisi* 6: 115-118.
- Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Baker, P., A. Hardie and T. McEnery. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Bergenholtz, H. and S. Tarp. 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries*. Vol. 12. Amsterdam/Philadelphia: John Benjamins.
- Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.
- Boz, E. 2006. Sözlük ve Sözlükçülük Sorunu. *Türkçenin Çağdaş Sorunları*: 9-46. İstanbul: Divan Yayınları.
- Boz, E. 2011. Leksikografi Teriminin Tanımı ve Türkçe Karşılığı Üzerine. *Dil ve Edebiyat Araştırmaları Dergisi* 4: 9-14.
- Bozkurt, F. 2017. *Sözlükselleşme: Genel Sözlükler için Sözlük Birim Seçimi*. İstanbul: Kesit Yayınları.
- Burkhanov, I. 1998. *Lexicography: A Dictionary of Basic Terminology*. Wydawn: Wyższej Szkoły Pedagogicznej w Rzeszowie.
- Francis, W.N. and H. Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.
- Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hartmann, R.R.K. and G. James. 1998. *Dictionary of Lexicography*. London/New York: Routledge.
- Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*. London: Bloomsbury.
- Jackson, H. 2002. *Lexicography. An Introduction*. London/New York: Routledge.
- Krishnamurthy, R. 2002. The Corpus Revolution in EFL Dictionaries. *Kernerman Dictionary News* 10: 23-27.
- Krishnamurthy, R. 2008. Corpus-driven Lexicography. *International Journal of Lexicography* 21(3): 231-242.
- Kupietz, M. 2016. Constructing a Corpus. Durkin, Philip (Ed.). 2016. *The Oxford Handbook of Lexicography*: 62-75. Oxford: Oxford University Press.
- Levend, A.S. 1957. Türkçe Sözlük Üzerine. *Türk Dili* VI(67): 365-367.
- Parlatır, İ. 1995. Türkçe Sözlük Çalışmaları ve Sorunlarımız. *Türk Dili. Dil ve Edebiyat Dergisi* I(517): 3-19.
- Robinson, J. 1983. A Glossary of Contemporary English Lexicographic Terminology. *Dictionaries* 5: 76-114.
- Rundell, M. and P. Stock. 1992. The Corpus Revolution 3. A Consideration of the Prospects and Potential of Corpus-and-concordance Lexicography (third article of three). *English Today, The International Review of the English Language* 8(4): 45-51.
- Svensén, B. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-making*. Cambridge/New York: Cambridge University Press.
- Tietze, A. 1976. Problems of Turkish Lexicography. Householder, Fred W. and Sol Saporta (Eds.). 1976. *Problems in Lexicography*: 263-272. Third edition. Bloomington: Indiana University.

Van Sterkenburg, P. 2003. *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins.

Zgusta, L. 1971. *Manual of Lexicography*. Berlin/New York: Walter de Gruyter.

Web-based Exploration of Results From a Large European Survey on Dictionary Use and Culture: ESDexplorer

Sascha Wolfer, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (wolfer@ids-mannheim.de)*

Iztok Kosem, *Department of Translation, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia (iztok.kosem@ff.uni-lj.si)*

Robert Lew, *Faculty of English, Department of Lexicography and Lexicology, Adam Mickiewicz University, Poznań, Poland (rlew@amu.edu.pl)*

Carolin Müller-Spitzer, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (mueller-spitzer@ids-mannheim.de)*

Maria Ribeiro Silveira, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (ribeiro@swhk.ids-mannheim.de)*

Abstract: We present ESDexplorer (<https://owid.shinyapps.io/ESDexplorer>), a browser application which allows the user to explore the data from a large European survey on dictionary use and culture. We built ESDexplorer with several target groups in mind: our cooperation partners, other researchers, and a more general public interested in the results. Also, we present in detail the architecture and technological realisation of the application and discuss some legal aspects of data protection that motivated some architectural choices.

Keywords: SURVEY, DATA COLLECTION, DATA PROCESSING, DATA PRESENTATION, DATA ANALYSIS, TECHNOLOGY AND ARCHITECTURE, TARGET GROUP, PLOT, BROWSER APPLICATION, ESDexplorer

Opsomming: Webgebaseerde verkenning van die resultate van 'n omvattende Europese opname van woordeboekgebruik en -kultuur: ESDexplorer. Ons stel ESDexplorer (<https://owid.shinyapps.io/ESDexplorer>), 'n webblaaier-toepassing wat die gebruiker toelaat om die data van 'n omvattende Europese opname van woordeboekgebruik en -kultuur te verken, bekend. Met die bou van ESDexplorer het ons verskeie teikengroepe in gedagte gehad: ons samewerkingsvennote, ander navorsers, en 'n meer algemene publiek wat in die resultate sou belangstel. Ons bespreek ook die argitektuur en tegnologiese totstandkoming van die toepassing in

besonderhede en brei uit oor enkele regsaspekte rakende databeskerming wat sommige argitektuurkeuses gemotiveer het.

Sleutelwoorde: OPNAME, DATAVERSAMELING, DATAVERWERKING, DATA-AANBIEDING, DATA-ANALISE, TEGNOLOGIE EN ARGITEKTUUR, TEIKENGROEP, GRAFIEK, WEBBLAAIERTOEPASSING, ESDEXPLORER

1. Introduction

On 8 May 2017, a large-scale survey on dictionary use¹ was launched in 26 European countries and Brazil². The main goal of the survey was to provide an up-to-date picture on dictionary use and culture (particularly) across Europe. This has been by far the most extensive dictionary-related survey to date, both in terms of the sheer number of participants and in terms of the breadth of coverage along national and linguistic dimensions. Due to its large scale, the survey presented particular challenges with regard to data collection, processing, and presentation. A core group of four researchers (Iztok Kosem, Robert Lew, Carolin Müller-Spitzer and Sascha Wolfer) drafted the general part of the survey. The general part consisted of 13 questions that were accompanied by 11 questions eliciting personal data from the participants (henceforth referred to as "meta-variables"). Around 60 researchers all over Europe (so-called "local partners") translated this original English version of the questionnaire into their local languages and helped to disseminate the survey in their countries. After all the translations were completed, different language versions were implemented in the online survey system Unipark Questback at the Institute for the German Language in Mannheim.

The local partners were given the opportunity to create local parts of the survey in their native language consisting of up to five short questions. These local parts were only presented to the participants from the respective countries and are not covered in this contribution or available in ESDexplorer.

Between 8 May and 9 July 2017, 9,373 participants completed the survey. Figure 1 shows the distribution of participants over countries and the professional status of the participants. All data is accessible in raw format to the core group. The local partners were given access to the raw data³ from participants from the respective country and — if present — the raw data from their local part. An analysis of the survey data (Kosem, Lew, Müller-Spitzer, Ribeiro Silveira, Wolfer et al. 2018) and one article in German, mainly covering the German local part, has already been published (Müller-Spitzer, Ribeiro Silveira, Wolfer, Kosem and Lew 2018). A Slovene paper by Arhar Holdt (2018) focuses on the Slovene perspective.

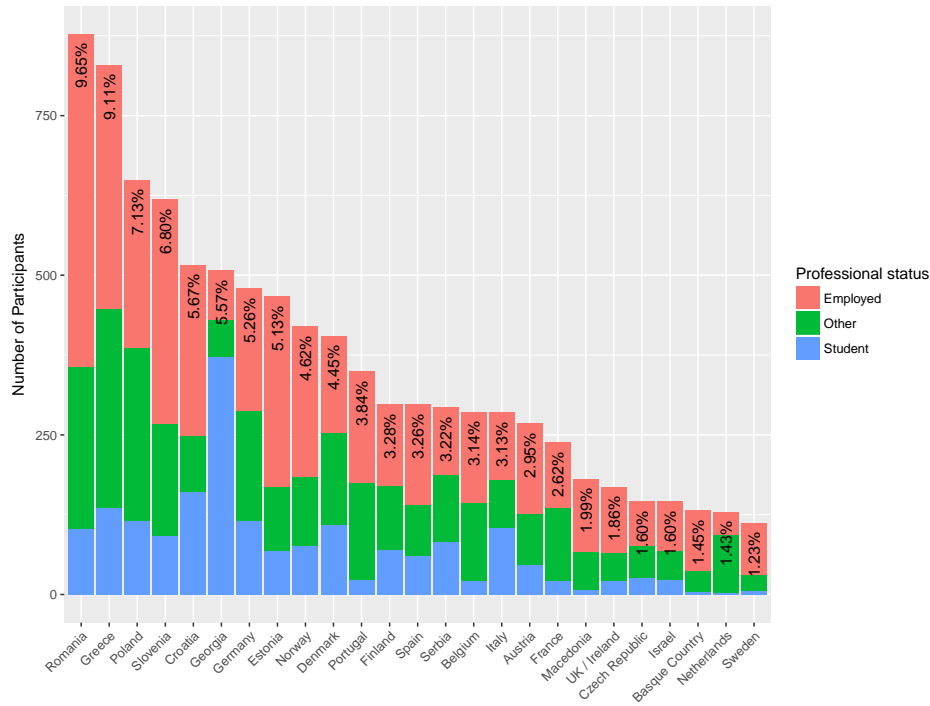


Figure 1: Number (*y*-axis) and percentages of participants per country. The bars are divided by professional status of the participants ("Student" does not contain "Ph.D. student", Ph.D. students are counted as "Other").

In the following section, we introduce ESDexplorer from the perspective of the user. In section 3, we go into more detail regarding the groups that we had in mind when designing ESDexplorer. Section 4 describes the technology that was used in building ESDexplorer, and briefly explains the technical mechanism of the server-side calculations which are not visible to the user. In section 5, we conclude this contribution with a summary.

2. ESDexplorer

The application is available under <https://owid.shinyapps.io/ESDexplorer>. In ESDexplorer, data from 11 questions from the general part of the survey is accessible in aggregated form. Due to the declaration of consent given by the participants, we are not permitted to make available the raw data which, in principle, could be used to trace back answers to individuals⁴.

Please refer to Figure 2 for an overview of ESDexplorer's user interface, with labels identifying its main elements.

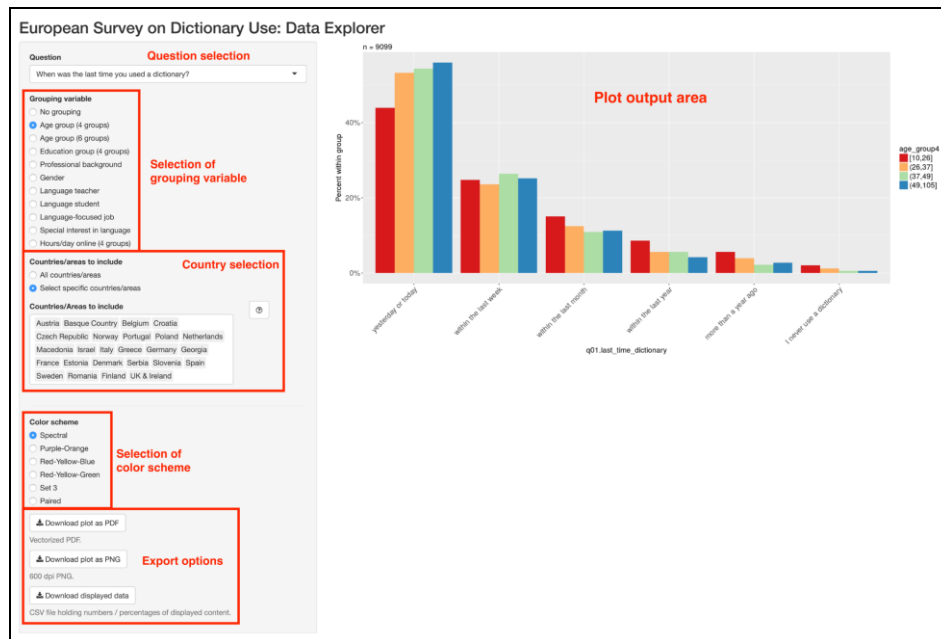


Figure 2: An annotated screenshot of ESDexplorer's user interface. The annotated labels appear in *italics* in the text.

The left-hand side with the greyish background is reserved for user input, whereas the right-hand side of the screen (*Plot output area*) presents the output given by the system. The input area on the left-hand side includes a number of elements. First, the user needs to select a specific question (*Question selection*). Eleven questions from the general part are available. Two questions from the general part are not available for selection. The first of these was an open-ended question asking participants which monolingual dictionaries they used. Such an open question has to be coded manually before the results can be meaningfully presented in a visual format; for example, responses such as "oed.com", "www.oed.com", "OED online", "oed on the web" and so on need to be mapped to a single entry⁵. The other question not currently available for analysis in ESDexplorer is "How much are you willing to spend on a good monolingual dictionary of [your language] (in [your currency])?". Here, currency conversion would have to be included to obtain reasonable results. Also, other factors like variation in purchasing power between the participating

countries may need to be considered. Since this has not yet been done, we decided to exclude this question from ESDexplorer.

Optionally, data analysis can be grouped by a meta-variable describing the status of participants (*Selection of grouping variable*). For the age of participants, two granularities (four and six groups) are available. Two meta-variables that were included in the survey are not available for analysis in ESDexplorer: native language and device usage (participants were asked to indicate all devices they used on a daily basis; the options given were: desktop computers, laptops, tablets, and smartphones. We did not include native language because the list of available native languages was very long (44 items). A visualization with that many categories would not be legible at all. Device usage was not included because there is no straightforward way to represent all the different combinations that are possible for the four options. Since one of the main goals of ESDexplorer is to represent the information in a clear and compact way, it seemed like a good decision not to include this meta-variable. If "No grouping" (first option) is selected, the bars in the output plot are collapsed to one grey bar per answer category (x -axis) and the y -axis switches to counts instead of percentages within the group. The overall percentages are then also annotated above the bars.

If the user wants to exclude certain countries from their analysis they can do so with the *Country selection*. In the selection list, any subset of countries can be selected. The "n = [number]" above the plot tells the user how many participants contributed their data to the current plot. Consequently, this number decreases when fewer countries are selected.

If a grouping variable is selected, the user might choose between six different color schemes (*Selection of color scheme*) to accommodate the context in which the plot might be used. Users can export the data currently shown in the *Plot output area* with three *Export options*. PDF export provides a vector-based graphic that can be scaled to any size. The second download option is a high-resolution (600 DPI) PNG file. With the last option, the user can download a comma-separated data file containing all the counts or percentages that underlie the plot currently shown.

3. Motivation and target groups

We had three groups of users in mind when designing ESDexplorer. The initial idea to create the application was to help those local partners that do not have training in data representation, manipulation and analysis to access the data in an easy and straightforward way. The application could thus serve as a starting point for our local partners to conduct preliminary analyses that might lead to publications in their local language. With ESDexplorer, the partners can access the data from the general part of the survey and use the plots generated by the application to document and compare answers from any combination of countries. The plots generated by the system can be used for their own publications,

either directly (as PNG files) or with little extra manipulation, e.g. using the CSV files with Excel or similar software.

The second group that we had in mind is a broader scientific community. Researchers from outside the consortium that co-operated in the survey might check the plausibility of more fine-grained analyses presented in publications that are based on this data. It has to be said, though, that this does not satisfy the broadest requirements for reproducible research to the full extent. To do so, we would have to provide all the data on an individual level, i.e. what we referred to as "raw data" above. However, due to data protection issues and the declaration of consent we asked our participants to acknowledge, we are not allowed to make available the raw data to anyone who was not part of the consortium. Nevertheless, we believe that ESDexplorer at least allows for detailed plausibility checks of analyses that are presented elsewhere (e.g. in journal papers).

The third target group is the broader, non-scientific audience that might be interested in lexicographic research. With ESDexplorer, these users have an easy-to-use point-and-click interface where they can learn something about the culture and use of monolingual dictionaries in their own country or all over Europe. It may also be the case that the participants of the study are interested in the final results. With the application, they can explore the data by themselves without the researchers functioning as "gatekeepers".

Our online visualization system can also have a didactic application beyond lexicography: ESDexplorer might serve as a model example for university teachers to illustrate the visualization possibilities for questionnaire studies. Since the application shows the results of a questionnaire study, it might nicely complement theoretical discussions about questionnaire design.

4. Technology and architecture

ESDexplorer is built using the R (R Core Team 2018) package "shiny" (Chang et al. 2017). With this package, one can build web applications without extensive web development skills. The basic architecture of a Shiny application consists of two scripts written in R code. One script controls the behaviour of the user interface with the input elements (so-called "widgets") and outputs (mostly plots and downloadable data). All the input widgets that are used in ESDexplorer come with a standard Shiny installation and can be readily used. The other script determines what is going on "behind the scenes". Here, the developer controls data management and statistical computation on the server. The computation underlying the graphs is not done in the user's browser but almost exclusively on the server itself. The only computations that are running in the user's browser are for showing the output and getting user input from the widgets and passing them along to the server.

Interestingly, the server script uses a data set that is very close to the raw data, i.e. the data on an individual level. This data set is used to aggregate the

data so that it can be displayed in the graph that the user requested. Due to the encapsulated nature of the computations and the raw data itself, the user cannot access this raw data file (nor can they access the server script, but this is less critical). This is necessary because, as indicated in section 1, we are not allowed to disclose the individual data.

Whenever a parameter is changed on the left-hand side (i.e. the user input section), the plot is updated. This is thanks to the reactive nature of the Shiny environment: whenever the user changes something in the user interface, the server script detects this change and a new calculation is triggered. For very large data sets (e.g., with several million cases or a large number of variables), this process might be slow. But with our dataset of 9,373 rows (= participants) and roughly 100 columns (= variables), this is no problem for real-time server-side calculations.

Luckily, the Shiny environment comes with its own session management system, so the developer of the application does not have to deal with the challenge of several users accessing the application at once.

While Shiny and R itself are free software, a Shiny application still has to be hosted on a Shiny server, so that users can access it online. At the moment, ESDexplorer is hosted with shinyapps.io, a service provided by RStudio, the company that also created the Shiny package. This is a proprietary service with a free tier. This free tier, however, only includes very limited usage. Hence, we chose the cheapest paid option to host ESDexplorer at the moment. An alternative is to host your own Shiny server using the open-source Shiny server, which is also available from RStudio. ESDexplorer might be transferred to such a solution in the long term.

5. Summary

ESDexplorer is a browser application where users can explore the results of the 2017 European Survey on Dictionary Use. The users can use grouping variables in their analysis and subset the data by country. With the application, we hope to reach three target groups: our local partners, the broader scientific community in lexicography and related disciplines, and the general public. ESDexplorer is implemented in Shiny, a framework for the dynamic and user-adaptive presentation of data.

Endnotes

1. A full list of participating researchers and countries can be accessed at <http://www.elexicography.eu/events/european-survey-on-dictionary-use/> [last access on August 8th, 2018].
2. Brazil has been included primarily to be able to compare the Portuguese and Brazilian answers.

3. With "raw data", we refer to the individual questionnaires. Technically, the raw data is one large table with all completed questionnaires stored in rows and all the variables in columns.
4. Tracing back a specific questionnaire to an individual is still highly unlikely using the raw data. However, through a combination of country, native language, age, years of formal education and profession, it is theoretically possible. With aggregated data, it is definitely impossible to "track" single individuals.
5. Users gave 6,697 different answers (types) to this question (each user was allowed to enter five dictionaries). Altogether, 15,663 answers (tokens) were given. The OED example in the text would contain 4 tokens and 4 types that would have to be mapped to a single type by manual coding.

References

- Arhar Holdt, Š.** 2018. The Attitude of Language Users Towards General Monolingual Dictionaries: The Slovene Perspective. *Slovenščina 2.0*, 6(1): 1-36. Retrieved from http://slovenscina2.0.trojina.si/arhiv/2018/1/Slo2.0_2018_1_01.pdf.
- Chang, W., J. Cheng, J. Allaire, Y. Xie and J. McPherson.** 2017. shiny: Web Application Framework for R (Version 1.0.5). Retrieved from <https://CRAN.R-project.org/package=shiny>.
- Kosem, I., R. Lew, C. Müller-Spitzer, M. Ribeiro Silveira and S. Wolfer et al.** 2018. The Image of the Monolingual Dictionary Across Europe: Results of the European Survey of Dictionary Use and Culture. <https://doi.org/10.1093/ijl/icy022>.
- Müller-Spitzer, C., M. Ribeiro Silveira, S. Wolfer, I. Kosem and R. Lew.** 2018. Eine europaweite Umfrage zu Wörterbuchbenutzung und -kultur: Ergebnisse der deutschen Teilnehmenden. *Sprachreport 2* (2018): 26-35. Retrieved from <http://pub.ids-mannheim.de/laufend/sprachreport/pdf/sr18-2.pdf#page=28>.
- R Core Team.** 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>.

Which Learning Tools Accompanying the Paid Online Version of *LDOCE* Do Advanced Learners of English Find Useful?

Bartosz Ptasznik, *University of Warmia and Mazury, Olsztyn, Poland*
(bartosz.ptasznik@uwm.edu.pl)

and

Robert Lew, *Faculty of English, Department of Lexicography and Lexicology,
Adam Mickiewicz University, Poznań, Poland* (rlew@amu.edu.pl)

Abstract: The aim of the report is twofold: to (1) briefly describe the learning tools of the *Longman Dictionary of Contemporary English* (LDOCE6) which are available to English learners in the paid online version of the dictionary (sixth edition); and (2) present the results of the questionnaire that was conducted on 114 students of English at the University of Warmia and Mazury in Olsztyn. The participants completed a questionnaire in which they were asked to assess the usefulness of the learning tools of the paid online version of LDOCE6. The first section of the paper introduces the reader to the five major British monolingual learners' dictionaries on the market and the most prominent features of LDOCE. The second section is a description of the learning tools available to learners of English in the paid online version of LDOCE6. The following section elaborates on the earliest questionnaire studies conducted in the field of dictionary use, and some of the problematic aspects of this research method are discussed. The report ends with a presentation of the results of the questionnaire and a brief discussion.

Keywords: LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH, ONLINE DICTIONARIES, QUESTIONNAIRES, LEARNERS' DICTIONARIES

Opsomming: Watter leerhulpmiddels in die betaalde aanlyn weergawe van die *Longman Dictionary of Contemporary English* vind gevorderde leerders van Engels nuttig? Die doel van hierdie artikel is tweërlei: om (1) kortliks die leerhulpmiddels van die *Longman Dictionary of Contemporary English* (LDOCE6) wat tot die beskikking van Engelse leerders in die betaalde aanlyn weergawe van die woordeboek (sesde uitgawe) is, te beskryf; en om (2) die resultate wat verkry is uit die vraelys wat aan 114 studente van Engels aan die Universiteit van Warmia en Mazury in Olsztyn voorgelê is, weer te gee. Die deelnemers het 'n vraelys waarin hulle gevra is om die bruikbaarheid van die leerhulpmiddels van die betaalde aanlyn weergawe van LDOCE6 te beoordeel, voltooi. Die eerste afdeling van hierdie artikel stel die leser bekend aan die vyf belangrikste Britse eentalige aanleerderswoordeboeke wat tans beskikbaar is en ook aan die mees prominente kenmerke van LDOCE. Die tweede afdeling is 'n beskrywing van die leerhulp-

middels wat tot die beskikking van die Engelse leerders in die betaalde aanlyn weergawe van LDOCE6 is. In die volgende afdeling word verder uitgebrei oor die vroegste vraelysstudies wat uitgevoer is in die woordeboekgebruiksveld en 'n paar van die problematiese aspekte rondom hierdie navorsingsmetode word bespreek. Ten slotte word die resultate van die vraelys weergegee met 'n kort bespreking daarvan.

Sleutelwoorde: *LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH, AANLYN WOORDEBOEKE, VRAELYSTE, AANLEERDERSWOORDEBOEKE*

1. Evolution of the Longman Dictionary of Contemporary English

LDOCE is one of the five major British monolingual learners' dictionaries (the others being: *Cambridge Advanced Learner's Dictionary*, *Collins COBUILD Advanced Learner's Dictionary*, *Macmillan English Dictionary for Advanced Learners*, *Oxford Advanced Learner's Dictionary of Current English*) on the market. Having become an instantly recognizable learners' dictionary due to the introduction of a controlled defining vocabulary and its own grammatical system in its first edition (1978), LDOCE has evolved into a remarkably user-friendly dictionary within a span of approximately 36 years. Since its original 1978 publication, the dictionary has undergone several important changes, such as the introduction of frequency information, incorporation of signposts, explicit presentation of grammatical information (Bogaards and van der Kloot 2002), increase in the number of entries and meanings, use of colour, and it has become more corpus-oriented. Most importantly, though, with the advent of the computer era and a profound decline in the importance of print dictionaries in EFL¹ lexicography, LDOCE has not lagged behind the competition and Longman publishers have put considerable effort into meeting the 21st century English learners' needs by making the dictionary available online (free or paid version). Nowadays, online dictionaries have simply more to offer to learners of English than their book-form counterparts — more information (for example, more meanings and examples), faster access to meaning, instant cross-references, native speaker voice recordings, or the inclusion of multimedia content being just a few of the advantages that online dictionaries hold over the paper medium.

2. *Longman Dictionary of Contemporary English* (6th edition) paid online learning tools

LDOCE6 offers free and paid online access, however, there is no doubt that the paid version has a lot more to offer to learners and Longman gives learners the opportunity to register for a 30-day free trial. Longman lexicographers have made an effort to adjust to the needs of advanced students of English by adding more information to the fee-based online dictionary. As a result, by accessing the full package, dictionary users can look up over 300,000 words, meanings and phrases, an additional 82,000 collocations (147,000 collocations

altogether) and 30,000 synonyms, antonyms and related words (48,000 altogether), and have access to an additional one million corpus examples. Beyond these core offerings, the following learning tools are available: the Longman Vocabulary Checker, Grammar Centre, Video Library, Study Centre, Culture Dictionary, Thesaurus Dictionary, Exam Practice and the Pronunciation page.

The Longman Vocabulary Checker offers the possibility to find out which words should be learned from a random text and what the difficulty level of vocabulary is. By pasting a selected text in the box provided, one can have the Longman Vocabulary Checker highlight the words from the text based on different vocabulary lists: words from the Academic Word List, or words from the Longman Communication 9000, selecting high frequency, mid frequency, or lower frequency words. To be more precise, one can see which words are and which are not part of the wordlist that was selected, what the total word count is and what the percentage of words from the selected wordlist is. Numbers, symbols and proper names are ignored by the Vocabulary Checker.

The Grammar Centre includes the Grammar Guide and Communication Guide features familiar from the book-form dictionary. Also, the Grammar Centre contains an "Intermediate Practice", "Advanced Practice" and "Scores" page. By and large, the first two pages have video presentations, interactive exercises and practice, diagnostic, progress and exit tests of selected grammar points, while the "Scores" page keeps track of the learner's results from different tests and exercises and measures students' progress. After taking the tests, learners can always get feedback on their performance by checking what the correct answers are. The "Intermediate Practice" and "Advanced Practice" pages deal with the following main grammar topics: "Intermediate Practice" — (1) Adjectives and adverbs, (2) Future forms, (3) Verbs with *-ing* forms and infinitives, (4) Passive forms, (5) Word combinations; "Advanced Practice" — (1) Modal verbs, (2) Conditionals, subjunctives and the "unreal" past, (3) Reported speech, (4) The grammar of formal English, (5) The grammar of spoken English. These grammar topics have been further divided into more specific grammar areas. For example, the part that focuses on adjectives and adverbs on the "Intermediate Practice" page covers the following grammar points: Adjectives and *-ed/-ing* forms; Order of adjectives before nouns; Comparison of adjectives; Big, small, and equal comparisons; Adverbs and adverb phrases; Adverbs and word order; Comparison of adverbs. What seems to be one of the most unique features of the Grammar Centre are the video presentations which contain thorough explanations of various grammar points, grammar patterns, examples, mistakes that should be avoided by learners, etc. Students can find this method of learning appealing, as the more traditional approach based on learning from books has been given priority in many schools, and could be perceived as tedious, or even outdated. Another advantage is exposure to the language of native speakers, which has always been highly valued by learners of English. Learners can listen to native English pronunciation and, at the same time, learn grammar rules. Moreover, if for some reason learners do not manage to keep up with the

pace of the video presentations, they can always replay the videos and listen to them again.

The Video Library is a collection of 44 video presentations containing monologs of native speakers of English and their conversations on various topics. For example, there are video presentations on how to ask for information, describe a sporting event, express ambitions, give directions, make a complaint, order food in a restaurant, report an event, talk about work and computers, summarize events in a film, book a train ticket, etc. It is possible to search the videos by title, topic or keyword. After listening to the conversations, learners can complete the transcript below the video presentations by typing the missing words into the gaps. In this way, students learn various expressions in English that should be used in a particular context. Once the students have done the task, they can next check the correct answers.

The Study Centre is another learning tool of LDOCE6 online. The Study Centre gives learners the opportunity to brush up on their grammar and vocabulary skills, as well as improve their knowledge of synonyms, collocations, register and culture. For example, on the "Register" page, learners can practice phrasal verbs, or they can learn idioms on the "Vocabulary" page. Learners will encounter here different test question types, for example, multiple-choice, or matching tasks. In addition, just like in the Grammar Centre, the system records the learners' scores.

The Culture Dictionary, which can be accessed by clicking on the Culture tab, has more than 9,000 encyclopedic entries. The entries in this dictionary provide users with cultural information which refers to people, places and events. As an example, one can find words like *Adidas*, *Daffy Duck*, *Orange Bowl*, *Tagalog*, or *Facebook*. These entries can also be looked up in the main LDOCE6 online section — the Dictionary tab (Dictionary page).

The Thesaurus Dictionary might come in handy when trying to enhance one's production skills which are needed for writing assignments, in-class essays, oral presentations, debates, etc. For example, when trying to find a synonym for *angry*, one can type in the word (concept/heading) *angry* in the search box and discover that the word *angry* has been divided into fifteen more specific meaning categories: (1) feeling angry; (2) feeling extremely angry; (3) angry for a short time; (4) angry because something is unfair or wrong; (5) words for describing an angry meeting, argument etc; (6) to get angry; (7) to make someone angry; (8) to deliberately make someone angry; (9) making you angry; (10) to behave in a very angry way; (11) often behaving in an angry, unfriendly way; (12) unfriendly and quiet because you are angry; (13) easily annoyed; (14) angry feelings; (15) to try to make someone less angry. By clicking, for example, on the fifth meaning category, one not only learns the meanings of words like *furious*, *stormy*, *uproar*, *heated*, etc., but also such entries may contain information about the pronunciation of these words, context in which they should be used (example sentences), part of speech, etc. The Thesaurus Dictionary also provides a list of the concepts related to the concept *angry*, such

as *disappointed*, *violent*, *insult*, *revenge*, etc., or its opposite (*calm*), and by clicking on those words the user will be directed to the appropriate concept.

The Exam Practice page helps potential candidates prepare for the following exams: FCE (First Certificate in English), CAE (Certificate in Advanced English), CPE (Certificate of Proficiency in English), IELTS (International English Language Testing System), TOEIC (Test of English for International Communication), PTE Academic (Pearson Test of English Academic). It contains practice materials in the style of particular exams. For example, the exercises for the FCE and CAE exams allow to practice one's reading, listening and use of English skills. Also, the Exam Practice page informs users from which specific books the learning materials have been adapted and it lists other books that could help in preparing for the exams. The "Scores" tab records the learners' scores.

The Pronunciation page (tab) contains exercises in the following areas: stress, syllables, sound recognition, British and American English pronunciation. Also, there are dictation exercises. Learners can listen to words that are played in either British or American English and then type those words in the box to check for correct spelling. Similarly, learners can listen to sentences instead of individual words, type the sentences in the box and check for correct answers.

However, it is the Dictionary² page (tab) that is the most basic element of LDOCE6 online and one that users will probably spend most of their time with, searching for meanings of words and phrases. This tab is the online equivalent of the information that can be found in the middle matter of the print dictionary and users can find here the same entries (entries containing the same information, or entries having the same entry structure) which appear in the book-form dictionary. What makes the Dictionary page even more special than the paper dictionary is the fact that learners can listen to the American and British English pronunciation of every word in the dictionary (or the 88,000 example sentences that have been recorded), check the meaning of every word in an entry's definition or example sentence by double clicking on that word, click on "Entry menu" links in polysemous entries that allow to guide users to the part of the entry (specific meaning) learners are interested in, check the etymology of a word ("Word origin" link) or inflections for all irregular and regular verbs in the dictionary ("Verb Table" link). In addition, the "Examples" link at the top of an entry allows users to browse through additional examples from the corpus (1 million example sentences) or other Longman dictionaries (80,000 example sentences), the "Collocations" link provides a list of collocations for the entry one is reading, collocations from other dictionary entries or from the Longman Corpus Network (the "Collocations" link also shows how these collocations are used), the "Thesaurus" link makes it possible to view a Thesaurus box from the entry, see information from the *Longman Language Activator*, or access the word sets option when a word forms part of a word set (definitions and examples of how words are used are also shown), and clicking on the "Phrases" link allows to see the phrases from the entry or other diction-

ary entries (the "Phrases" link also shows how these phrases are used). Last but not least, the Dictionary page has an "Advanced Search" function, which allows to find words that one is looking for, for example, by part of speech, frequency level, or register. Interestingly, it is even possible to search for entries that have pictures, or Collocations boxes, Grammar notes, etc.

Finally, LDOCE6 online gives learners the opportunity to download the Longman dictionary applications to one's iPhone, iPad or iPod Touch.

Given the fact that this paper presents the results of a questionnaire in the final section of the paper, the following section briefly elaborates on the earliest questionnaire studies conducted in the field of dictionary use and some of the problematic aspects of this research method are discussed.

3. Questionnaires in dictionary use research

The first questionnaires which were concerned with dictionary use research were surveys conducted by Barnhart (1962), Quirk (1974), Tomaszczyk (1979) and Béjoint (1981). In his questionnaire, Barnhart asked teachers of American college students to express their opinions about how they thought their students used dictionaries. In the following surveys, however, starting with Quirk's study in 1974, dictionary users were asked to report on their dictionary use patterns and not their teachers. As far as Tomaszczyk's questionnaire is concerned, it was conducted with a view to learning whether American and Polish college students, who were foreign students, preferred to use monolingual or bilingual dictionaries during dictionary consultation. Tomaszczyk discovered that bilingual dictionaries were superior to monolingual dictionaries, as well as the fact that dictionary users tended to consult dictionaries primarily for the meaning of words and their spelling. Tomaszczyk's study is considered by dictionary use researchers to be the first and one of the most important questionnaire studies. The aim of Béjoint's questionnaire was to learn how monolingual dictionaries are used by French students of English. Béjoint's study gave the following results: (1) dictionaries are consulted mainly for the meaning of unknown words; (2) students very rarely study the information that can be found in the front matter of dictionaries; (3) more than half of Béjoint's participants admitted to not using dictionary codes; (4) approximately 75% of the participants were content with their dictionaries.

Significantly, a few researchers have expressed their concern about questionnaire research in dictionary use studies. One problem lies in the reliability of questionnaire reports. Hatherall (1984: 184) is one researcher who raised doubts about questionnaire studies: "Are subjects saying here what they do, or what they think they do, or what they think they ought to do, or indeed a mixture of all three?". Crystal (1986) is of the opinion that dictionary users who complete questionnaires are not able to remember in detail what exactly happens during dictionary consultation. The second problem deals with the language used in questionnaires. According to Lew (2004: 40-41), questionnaires

which are prepared in the participants' target language, for example the English language for native speakers of Polish, can often lead to misunderstandings of questions and instructions. Hence, they ought to be prepared in the participants' mother tongue because only in this way can questionnaire respondents fully understand the questions they are being asked in questionnaires and avoid vagueness in the foreign language.

Notwithstanding these problematic issues, questionnaires are a source of valuable knowledge for dictionary use researchers. Questionnaires can indicate the direction for further research, they can often be treated as a starting point for researchers and may lead to conclusions on what additional studies need to be conducted in a specific research field. Lew (2002) reaches the conclusion that there is certainly a place for questionnaires in dictionary use research. Whether a questionnaire is successful or not depends on how well it is prepared by researchers. Finally, Lew suggests that it is highly desirable for dictionary use researchers to acquire knowledge about how to design questionnaires even from other fields of study, like sociometry or psychometry (Berdie and Anderson 1974; Bradburn, Sudman and Blair 1979; Oppenheim 1992; Sudman and Bradburn 1982), in which researchers' work is also concerned with the design of questionnaire manuals.

The following section demonstrates the results of the questionnaire which was completed by students of English at the University of Warmia and Mazury in Olsztyn.

4. Questionnaire — results and discussion

The questionnaire was carried out on 114 students of English at the University of Warmia and Mazury in Olsztyn. The participants were third and fourth year students. Their English language proficiency level had been assessed by their academic teachers as B2 to C1 by the Common European Framework of Reference for Languages standards and the students had considerable experience of dictionary use, as it was necessary for them to use dictionaries throughout their studies, especially in their practical English courses (for example, writing classes) and BA and MA seminars that they were attending. The participants were asked to complete a short questionnaire in paper format in which they were asked to assess the usefulness of the learning tools (Dictionary page, Culture Dictionary, Thesaurus Dictionary, Study Centre, Pronunciation page, Exam Practice, Grammar Centre, Video Library, Longman Vocabulary Checker) of the paid online version of LDOCE6. The research question which the study attempted to answer was the following:

- Which paid online learning tool of the *Longman Dictionary of Contemporary English* do advanced learners of English find useful?

A Likert-type rating scale was adopted for the questionnaire (USEFUL, RATHER USEFUL, DIFFICULT TO SAY, RATHER NOT USEFUL, NOT USEFUL). Also, the partici-

pants were asked to elaborate on the choices they had made in the comments section. Importantly, the whole questionnaire was delivered in the participants' native language, except for the specific names of learning tools which were listed in the table, as providing students with Polish translations of the learning tools in this specific case could have been misleading for the participants. Before the participants were asked to fill out the questionnaire, they received one-hour training on how English learners can use the learning tools for learning English. The results of the questionnaire are given below.

Table 1: Results of the questionnaire

TOOL	USEFUL	RATHER USEFUL	DIFFICULT TO SAY	RATHER NOT USEFUL	NOT USEFUL
Dictionary page	100	14	0	0	0
Culture Dictionary	22	26	42	20	4
Thesaurus Dictionary	100	14	0	0	0
Study Centre	50	44	16	4	0
Pronunciation page	68	38	6	2	0
Exam Practice	44	56	10	4	0
Grammar Centre	62	42	8	2	0
Video Library	4	16	36	44	14
Vocabulary Checker	14	48	36	12	4

The results of the questionnaire could be divided into three groups³:

- (1) MOST USEFUL LEARNING TOOLS: Dictionary page, Thesaurus Dictionary;
- (2) USEFUL LEARNING TOOLS: Study Centre, Pronunciation page, Exam Practice, Grammar Centre;
- (3) LEAST USEFUL LEARNING TOOLS: Culture Dictionary, Video Library, Longman Vocabulary Checker.

There is no doubt that the Dictionary page and Thesaurus Dictionary are the most useful learning tools for dictionary users (87.7% of the participants said that the Dictionary page and Thesaurus Dictionary were useful). This finding is not at all surprising. First of all, the Dictionary page is the most basic learning tool of the paid online version of LDOCE6. Whenever dictionary users do not understand a word, it is obvious that the first step lexicographers would normally expect them to take is to consult the meaning of this word in the Dictionary tab. The Dictionary page contains information about meaning, grammar, pronunciation, collocations, example sentences, etc. In other words, the chances are that all the information learners need about a given word can be found in the Dictionary page. In addition, the Dictionary page has its own thesaurus tab from which learners can access useful information about synonyms of words. Second, the Thesaurus Dictionary seems to be invaluable for advanced-level dictionary users who strive to improve their English language production

skills. Learners can definitely benefit from this tool when writing essays, formal and informal letters, or when preparing oral presentations, speeches, etc. One of the main problems that more proficient language learners encounter is that they tend to be repetitive in their word choices, with insufficient recourse to appropriate synonyms. The Thesaurus Dictionary attempts to alleviate this problem by providing dictionary users with a list of concepts, which allow them to discover new words and expressions by presenting them in the form of a word list along with their meanings and even example sentences under those broader concepts mentioned above. For example, one learns that instead of saying *obey the rules* the expressions *comply with the rules* or *abide by the rules* can be used. In this way, the Thesaurus Dictionary encourages more varied word choices when communicating or writing in the target language. This seems to be the Thesaurus Dictionary's biggest advantage.

The Study Centre, Pronunciation page, Exam Practice and Grammar Centre have all been rated rather positively by the participants; however, these tools appear to be less useful for dictionary users than the Dictionary page and Thesaurus Dictionary (43.9% of the respondents rated the Study Centre as useful and 38.6% as rather useful, 59.6% rated the Pronunciation page as useful and 33.3% as rather useful, 38.6% rated the Exam Practice page as useful and 49.1% as rather useful, and 54.4% rated the Grammar Centre as useful and 36.8% as rather useful). Some of the participants made comments in the questionnaire that they would use the Exam Practice page only before taking the CAE or CPE exams, and complained about the small number of tests included under each specific exam page (FCE page, CAE page, CPE page, etc.). Many participants criticized the Grammar Centre for being too theoretical and for containing only basic information about grammar rules. This, indeed, may be a problem, as users of English monolingual learners' dictionaries tend to be more advanced with regard to their linguistic skills and, hence, they endeavor to find new types of information rather than just read about things they already know. Despite dictionaries not being grammar books, it does seem like there is room for more detailed explanations of grammar rules or exceptions to these rules in online dictionaries, as space constraints are not a problem in the case of this specific dictionary medium. As for the Pronunciation page, the participants thought that the exercises devoted to sound recognition and American and British English word pronunciation distinction are extremely useful. However, they also said that they would not really decide to do the exercises devoted to word syllables and stress, as these types of information are not normally the types of information learners want to acquire from dictionaries.

Importantly, the participants rated as least useful: the Culture Dictionary (19.3% of the participants said it was useful, 22.8% said it was rather useful, 36.8% said it was difficult to say and 17.5% said it was rather not useful), Video Library (3.5% of the participants said it was useful, 14% said it was rather useful, 31.6% said it was difficult to say and 38.6% said it was rather not useful) and Longman Vocabulary Checker (12.3% of the participants said it was useful,

42.1% said it was rather useful, 31.6% said it was difficult to say and 10.5% said it was rather not useful). These are some of the more important comments that the participants made:

- all the information (dictionary entries) in the Culture Dictionary can also be found in the Dictionary page
- the Video Library is a compilation of English dialogs which could be useful, but only for much less advanced students of English
- the Culture Dictionary could be useful for translating texts devoted to the topic of culture
- the Longman Vocabulary Checker is not useful as more advanced students of English are aware of which words belong to academic language and which do not
- the Culture Dictionary is similar to an encyclopedia, however, more useful information about the terms that have been defined in the Culture Dictionary could be found in other types of reference books
- the Culture Dictionary contains only additional information about the words that have been defined in the Culture Dictionary, more information about these words can be found on the Internet rather than in the Culture Dictionary
- the Longman Vocabulary Checker could be helpful when writing an academic essay

By and large, the lexicographic data in the Culture Dictionary has also been incorporated into the Dictionary page. This means that LDOCE6 users will most likely prefer to open the Dictionary tab and search for pertinent information there, rather than access it from the Culture Dictionary. In addition, some participants complained that the Culture Dictionary might not contain enough information about cultural terms and concepts, and that they would rather consult other resources for such information. However, some students acknowledged that the Culture Dictionary could be used during translation tasks or exercises. As far as the Longman Vocabulary Checker is concerned, it seems that its purpose remains unclear to dictionary users. Most participants did not really understand why they would want to use this tool, as the vast majority of users who decide to use LDOCE6 are more proficient in the target language, which at the same time means that a more proficient user is able to distinguish academic words from general ones in a text. Only a handful of participants perceived the Longman Vocabulary Checker as an advantage of LDOCE6 by saying that they would use it when writing an essay. As for the Video Library, most participants commented that some dictionary users could perhaps benefit from this learning tool, however, only those representing a much less advanced level of English.

Endnotes

1. English as a Foreign Language.
2. The Dictionary page has been treated as a learning tool as it provides information about word meanings, the context in which words are used (example sentences), collocations, grammar patterns, idioms, phrasal verbs, etc.
3. The learning tools which appear in specific groups (MOST USEFUL LEARNING TOOLS, USEFUL LEARNING TOOLS, LEAST USEFUL LEARNING TOOLS) have been listed in random order in their respective groups.

References

Dictionaries

- Delacroix, L. (Ed.)**. 2014. *Longman Dictionary of Contemporary English*. Sixth edition. Harlow: Longman.
- Delacroix, L. (Ed.)**. 2014. *Longman Dictionary of Contemporary English*. Sixth edition (free online version). Harlow: Longman. Available: <https://www.ldoceonline.com/>.
- Delacroix, L. (Ed.)**. 2014. *Longman Dictionary of Contemporary English*. Sixth edition (paid online version). Harlow: Longman. Available: global.longmandictionaries.com/.
- Deuter, M. (Ed.)**. 2015. *Oxford Advanced Learner's Dictionary of Current English*. Ninth edition. Oxford: Oxford University Press.
- McIntosh, C. (Ed.)**. 2013. *Cambridge Advanced Learner's Dictionary*. Fourth edition. Cambridge: Cambridge University Press.
- Procter, P. (Ed.)**. 1978. *Longman Dictionary of Contemporary English*. First edition. Harlow: Longman.
- Rundell, M. (Ed.)**. 2007. *Macmillan English Dictionary for Advanced Learners*. Second edition. Oxford: Macmillan Education.
- Sinclair, J. (Ed.)**. 2014. *Collins COBUILD Advanced Learner's Dictionary*. Eighth edition. London: HarperCollins.
- Summers, D. (Ed.)**. 2002. *Longman Language Activator*. Second edition. Harlow: Longman.

Other literature

- Barnhart, C.** 1962. Problems in Editing Commercial Monolingual Dictionaries. Householder, F.W. and S. Saporta (Eds.). 1962. *Problems in Lexicography*: 161-181. Bloomington: Indiana University.
- Berdie, D.R. and J.F. Anderson.** 1974. *Questionnaires: Design and Use*. Metuchen, NJ: Scarecrow Press.
- Béjoint, H.** 1981. The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* 2(3): 207-222.
- Bogaards, P. and W.A. van der Kloot.** 2002. Verb Constructions in Learners' Dictionaries. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17, 2002*: 747-757. Copenhagen: Center for Sprogteknologi, Copenhagen University.

- Bradburn, N.M., S. Sudman and E. Blair.** 1979. *Improving Interview Method and Questionnaire Design*. The Jossey-Bass Social and Behavioral Science Series. San Francisco: Jossey-Bass.
- Crystal, D.** 1986. The Ideal Dictionary, Lexicographer and User. Ilson R.F. (Ed.). 1986. *Lexicography: An Emerging International Profession*: 72-81. Manchester: Manchester University Press.
- Hatherall, G.** 1984. Studying Dictionary Use: Some Findings and Proposals. Hartmann, R.R.K. (Ed). 1984. *LEXeter '83 Proceedings: Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983*: 183-189. Tübingen: Max Niemeyer.
- Lew, R.** 2002. Questionnaires in Dictionary Use Research: A Reexamination. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17, 2002*: 267-271. Copenhagen: Center for Sprogteknologi, Copenhagen University.
- Lew, R.** 2004. *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex.
- Oppenheim, A.N.** 1992. *Questionnaire Design, Interviewing, and Attitude Measurement*. London/New York: Pinter Publishers.
- Quirk, R.** 1974. The Image of the Dictionary. Quirk, R. (Ed.). 1974. *The Linguist and the English Language*: 148-163. London: Edward Arnold.
- Sudman, S. and N.M. Bradburn.** 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. Jossey-Bass Series in Social and Behavioral Sciences. San Francisco: Jossey-Bass.
- Tomaszczyk, J.** 1979. Dictionaries: Users and Uses. *Glottodidactica* 12: 103-119.

APPENDIX

Oceń pożyteczność podanych narzędzi do nauki języka angielskiego słownika *Longman Dictionary of Contemporary English*. Wstaw "X" w odpowiednim miejscu tylko jeden raz dla każdego z podanych narzędzi do nauki języka angielskiego.

NARZĘDZIE DO NAUKI JĘZYKA ANGIELSKIEGO	POŻYTECZNE	RACZEJ POŻYTECZNE	TRUDNO POWIEDZIEĆ	RACZEJ NIEPOŻYTECZNE	NIEPOŻYTECZNE
<i>Dictionary page</i>					
<i>Culture Dictionary</i>					
<i>Thesaurus Dictionary</i>					
<i>Study Centre</i>					
<i>Pronunciation page</i>					
<i>Exam Practice</i>					
<i>Grammar Centre</i>					
<i>Video Library</i>					
<i>Vocabulary Checker</i>					

KOMENTARZE

Herbert Ernst Wiegand

08 Januarie 1936 – 03 Januarie 2018

Van tyd tot tyd smaak enige vakgebied die voorreg om 'n buitengewone deelnemer te hê wat op 'n buitengewone manier tot daardie vakgebied bydra. As vakgebied was dit leksikografie se voorreg en eer om so 'n deelnemer te hê. Op 3 Januarie 2018 het die leksikografiese gemeenskap hierdie toonaangewende lid verloor met die afsterwe van prof. Herbert Ernst Wiegand.

In die veld van die metaleksikografie was prof. Wiegand 'n leidende en hoogs produktiewe lid, 'n aktiewe navorser, 'n akademiese en wetenskapsorganiseerder, 'n lojale kollega en 'n vriend.

Wiegand se talle publikasies sluit artikels in wat op 'n groot verskeidenheid onderwerpe uit die leksikografie gerig was. Sy werk as enkelouteur vind 'n hoogtepunt in die publikasie van sy magnum opus *Wörterbuchforschung* (1998). Die hoofokus in sy uitgebreide publikasielys was woordeboekstrukture. Hy het 'n groot verskeidenheid strukture geïdentifiseer, geanaliseer en tot in die fynste besonderhede bespreek. Met hierdie bydrae het hy die vlak van die leksikografiese gesprek verhoog en dit as 'n erkende wetenskapsterrein gevestig.

Behalwe sy werk as navorser het Wiegand ook 'n reusebydrae gelewer as redakteur en mederedakteur van 'n aantal vaktydskrifte en boekreekse. Hy was mederedakteur van die tydskrifte *Zeitschrift für germanistische Linguistik* en *Lexicographica. International Annual for Lexicography. Revue Internationale de Lexicographie. Internationales Jahrbuch der Lexikographie*, van die boekreekse *Reihe Germanistische Linguistik*, die *Wörterbücher zur Sprach- und Kommunikationswissenschaft* en *Lexicographica. Series Maior*. In 1982 het hy besef dat daar 'n behoefte bestaan aan 'n omvattende reeks handboeke wat die verskillende subdissiplines van die taalkunde dek. Gevolglik het hy die reeks *Handbücher zur Sprach- und Kommunikationswissenschaft* gevestig. Hy was ook mederedakteur en redakteur van hierdie reeks boeke waarin die jongste stand van taalkunde en kommunikasiekunde weerspieël is. In deel 5.4 *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (2013) het Wiegand verskeie belangrike bydraes gelewer met 'n vernuwend fokus op van die woordeboekstrukture wat hy reeds in vroeëre publikasies bekendgestel het. Alhoewel hy dit uitdruklik stel dat sy bespreking gerig is op strukture in gedrukte woordeboeke het die oorgang van gedrukte na aanlyn leksikografie reeds grootliks gebaat by hierdie strukture, want met geringe aanpassings kan baie daarvan ook benut word in die beplanning en samestelling van aanlyn woordeboeke.

Nog 'n beduidende bydrae van Wiegand was die vierdelige *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung*. Hierdie publikasie kan beskou word as dié belangrikste leksikografiebibliografie. Alhoewel dit hoofsaaklik Duitse bronne verstrek, bevat dit ook talle verwysings na

bronne uit Engels, die Nordiese en die Romaanse tale.

Een van Wiegand se laaste groot projekte was sy deelname as hoofredakteur van die gesaghebbende *Wörterbuch zur Lexikographie und Wörterbuchforschung/Dictionary of Lexicography and Dictionary Research*. Die eerste deel van hierdie vierdelige projek is in 2010 gepubliseer en die tweede deel in 2017. 'n Belangwekkende doel van hierdie bron, waarin Duitse leksikografeterme gekoördineer word met hulle ekwivalente in Afrikaans, Bulgaars, Engels, Frans, Hongaars, Italiaans, Portugees, Russies en Spaans, is die standaardisering van leksikografiese terminologie.

Wiegand se akademiese loopbaan het in 1972 begin met sy aanstelling as professor vir teoretiese taalkunde aan die Philipps Universiteit van Marburg. Na 'n termyn aan die Universiteit van Düsseldorf het hy 'n aanstelling aanvaar aan die Ruprecht-Karls Universiteit van Heidelberg waar hy tot en met sy aftrede in 2004 'n professor ordinarius was.

Wiegand se rol in Afrilex en *Lexikos* mag nooit onderskat word nie. In 1996 was hy die eerste hoofspreeker by die "First International Conference of the African Association for Lexicography" aan die toenmalige Randse Afrikaanse Universiteit. Hy het gereeld in *Lexikos* gepubliseer en was ook 'n gewaardeerde lid van dié tydskrif se adviesraad.

HEW, sy voorletters kan ook 'n afkorting wees van "High Energy Wanderer", het welverdiende erkenning vir sy werk gekry, onder meer eredoktorsgrade van die Aarhus School of Business, die Universiteit van Sofia en die Universiteit Stellenbosch. Die belangrikste erkenning van sy werk sal egter die voortgesette invloed van sy indrukwekkende nalatenskap wees.

Prof. Wiegand was vir meer as twee jaar terminaal siek. Hy wou egter nie dat sy gesondheidstoestand in die akademiese gemeenskap bekend moes raak nie, want hy wou met sy navorsing voortgaan sonder om die teiken van kollegas se simpatie te wees. Slegs enkele kollegas was bewus van sy siekte en hy het hulle deurlopend ingelig gehou oor die implikasies van sy siekte vir die voortgaande werk. Van tyd tot tyd het hy my gevra om seker te maak dat persoon X of persoon Y na sy dood met 'n spesifieke deel van sy werk sal voortgaan. Hy het altyd presiese aanduidings gegee van watter werk hy self nog voor sy dood wou voltooi. 'n Maand voor sy sterwe het ek hom besoek en ons is dadelik na sy studeerkamer waar ons sy onvoltooide werk bespreek het. Hy het genoem dat hy graag nog 'n sekere stuk werk sou wou voltooi. Toe ek vier dae voor sy dood telefonies met hom gepraat het, het hy verskonend genoem dat hy ongelukkig nie daardie stuk werk kon voltooi nie. Sy werk het altyd 'n dominante rol in sy lewe gespeel en hy het tot sy dood passievol en entoesiasties daarmee voortgegaan.

Herbert Ernst Wiegand sal gemis word. Maar hy het 'n nalatenskap wat verseker dat sy werk steeds 'n wesenlike rol in die metaleksikografie sal speel.

Rufus H. Gouws
Universiteit Stellenbosch
Suid-Afrika
(rhg@sun.ac.za)

Herbert Ernst Wiegand 08 January 1936 – 03 January 2018

From time to time every academic subject field experiences the privilege of having an exceptional scholar contributing to that field in an exceptional way. The field of lexicography was privileged and honoured to have such a scholar. On the 3 January 2018 the lexicographic community lost this prominent member with the passing away of Prof Herbert Ernst Wiegand.

In the field of metalexicography Prof Wiegand has been a leading and highly productive scholar, a prolific researcher, an academic and scientific organiser, a loyal colleague and friend.

Wiegand's numerous publications include articles focusing on a wide variety of topics from the field of lexicography with his work as single author culminating in his magnum opus *Wörterbuchforschung* (1998). The dominant focus in his extensive publication list has been on dictionary structures. He identified, analysed and discussed a comprehensive selection of dictionary structures in minute details. With this contribution he elevated the lexicographic discussion and established it as an acknowledged scientific domain.

Besides his work as researcher Wiegand made a huge contribution as editor and co-editor of a number of scientific journals and book series. He was co-editor of the *Zeitschrift für germanistische Linguistik*, *Lexicographica*, *International Annual for Lexicography*, *Revue Internationale de Lexicographie*, *Internationales Jahrbuch der Lexikographie*, the book series *Reihe Germanistische Linguistik*, the *Wörterbücher zur Sprach- und Kommunikationswissenschaft* and *Lexicographica. Series Maior*. In 1982 he realised the need for a comprehensive series of text books covering all subfields of the broad discipline of linguistics. Consequently he co-founded the series *Handbücher zur Sprach- und Kommunikationswissenschaft*, and was co-editor and editor of this series of text books that reflect the state-of-the-art of linguistics and communication science. In Volume 5.4 *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (2013) Wiegand wrote a number of seminal contributions with an innovative focus on dictionary structures introduced in earlier publications. Although he explicitly stated that his discussion focused on structures of printed dictionaries the transfer from printed to online dictionaries has already benefited substantially from these structures because with minor adaptations many of them can be employed in the planning and compilation of online dictionaries.

Another significant contribution was Wiegand's four volume *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung*. This publication can be regarded as the most important bibliography of lexicography. Although it primarily provides German references it also includes refer-

ences from English and the Nordic and Romance languages.

One of Wiegand's last major endeavours was his participation as leading editor of the *Wörterbuch zur Lexikographie und Wörterbuchforschung/Dictionary of Lexicography and Dictionary Research*. The first volume of this four volume project was published in 2010 and the second volume in 2017. An important aim of this publication, in which German lexicographic terms are coordinated with equivalents in Afrikaans, Bulgarian, English, French, Hungarian, Italian, Portuguese, Spanish and Russian, is the standardisation of lexicographic terminology.

Wiegand's academic career started with his appointment in 1972 as professor for theoretical linguistics at the Philipps University of Marburg. After a period at the University of Düsseldorf he took up a position at the Ruprecht-Karls University of Heidelberg where he was Professor Ordinarius from 1977 until his retirement in 2004.

Wiegand's role in *Afrilex* and *Lexikos* may never be underestimated. In 1996 he was the first keynote speaker at the First International Conference of the African Association for Lexicography, held at the then Rand Afrikaans University. He regularly published in *Lexikos* and has also been a respected member of the advisory board of this journal.

HEW, the initials can also be read as High Energy Wanderer, received well deserved recognition for his work, including honorary doctorates from the Aarhus School of Business, the University of Sofia and Stellenbosch University. The main recognition, however, will be the continued influence of his comprehensive legacy.

Prof Wiegand was terminally ill for more than two years. He did not want the condition of his health to be made known in the scholarly community because he wanted to continue with his research and work without having to be the target of his colleagues' sympathy. Only a few colleagues were aware of his illness, and he kept them informed of his condition and the implications it had for the ongoing work. From time to time he would ask me to make sure that person X or person Y would continue with a specific part of his work once he had passed away. He always gave a clear indication of what he still hoped to complete before his death. A month before he passed away I visited him and he immediately took me to his study where we discussed his uncompleted work. He told me that he would still like to complete one specific part of that work. When I called him four days before his death he apologised for not having completed that section. His work played a dominant role in his life and he continued working with passion and enthusiasm until his death.

Herbert Ernst Wiegand will be missed. However, he left a legacy that will continue to play a major role in the field of metalexigraphy.

Rufus H. Gouws
Stellenbosch University
South Africa
(rhg@sun.ac.za)

Das Rumäniendeutsche in der Neuauflage (2016) des *Variantenwörterbuchs des Deutschen* Ioan Lăzărescu zum 65. Geburtstag gewidmet

Doris Sava, *Department für anglo-amerikanische und
germanistische Studien, Lucian-Bлага-Universität Sibiu, Sibiu,
Rumänien (doris.sava@ulbsibiu.ro)*

Zusammenfassung: Noch vor Jahresende 2016 ist eine völlig neu bearbeitete, erweiterte und aktualisierte zweite Auflage des *Variantenwörterbuchs des Deutschen* (VWB) erschienen, das bisher lexikografisch nicht kodifizierte standardsprachliche Besonderheiten in Rumänien, Namibia und Mexiko erfasst. Im Hinblick auf das veränderte Normdenken zur standardsprachlichen Regionalität ist es erfreulich, dass sich das VWB vornimmt, das gesamte gegenwärtige Varietätenspektrum am Rande und weit außerhalb des geschlossenen deutschen Sprachgebiets lexikografisch zu dokumentieren. Mit der Fokussierung auf die schriftsprachliche Seite der Standardvarietäten, der sich die Bestandsaufnahme im VWB widmet, belegt das Wörterbuch Unterschiede und damit auch Eigenheiten der Viertel(s)zentren, um sie von der zweifelhaften Einschätzung als Non-Standard zu befreien. In der Erstauflage des VWB (2004) wurden nur die lexikalischen Varianten der deutschen Standardsprache in den Ländern und Regionen, wo Deutsch offizielle und/oder Amtssprache ist, kodifiziert. Aufgrund dieses lobenswerten Bestrebens gilt es zu fragen, inwiefern eine überzeugende lexikografische Bearbeitung der in der Erstauflage zu Unrecht vernachlässigten Viertel(s)zentren gewährleistet wurde. Im Beitrag soll dies exemplarisch am Beispiel des *Rumäniendeutschen* aufgezeigt werden. Die Bewusstmachung typischer Erscheinungsformen außerhalb des deutschen Amtssprachengebietes im täglichen Gebrauch wird den Vergleich der Viertel(s)zentren miteinander fördern und auch dazu beitragen, dass auch *dieses Deutsch* intensiver wahrgenommen wird.

Schlüsselwörter: VARIANTENWÖRTERBUCH, VARIETÄTEN DES DEUTSCHEN, VIERTELZENTREN, LEXIKOGRAPHISCHE BESCHREIBUNG, RUMÄNIENDEUTSCH, RUMÄNISMUS

Abstract: *Romanian German in the 2016 edition of the German Variant Dictionary. Dedicated to Ioan Lăzărescu on his 65th Birthday.* Towards the end of 2016, a fully revised, extended and updated second edition of the *Variantenwörterbuch des Deutschen* (German Variant Dictionary, GVD) was published, covering hitherto not lexicographically coded peculiarities of the German language in Romania, Namibia, and Mexico. In view of the changed normative thinking about standard language regionality, it is gratifying that the GVD undertakes to lexicographically document the entire variety spectrum beyond the boundaries of the closed

German language area. Focussing on the written-language side of the standard varieties to which the inventory in the GVD is dedicated, the dictionary points out differences and thus peculiarities of the different varieties of the German language spoken in the quarter centres, in order to free them from their suspect assessment as non-standard. In the first edition of the GVD (2004), only the lexical variants of Standard German in countries and regions where German is the official and/or administrative language were coded. In view of this praiseworthy endeavor, it is important to question to what extent a convincing lexicographical treatment of the quarter centres, which had been unjustly neglected in the first edition, has been ensured. In the article, this will be exemplified on the basis of *Romanian German*. Raising awareness of typical manifestations in everyday use outside of the German official language area will promote the comparison of the quarter centres, and also contribute to the fact that *this German* is also perceived more intensively.

Keywords: GERMAN VARIANT DICTIONARY, VARIETIES OF GERMAN, QUARTER CENTRES, LEXICOGRAPHICAL DESCRIPTION, ROMANIAN GERMAN, ROMANIANISM

Nicht der Name des Werks soll dem Autor Achtung,
sondern das Werk dem Autor Gerechtigkeit verschaffen.
Denis Diderot

1. Vorbemerkungen

Ende 2004 ist nach einer Bearbeitungszeit von mehr als sechs Jahren das fast 1.000 Seiten und ca. 12.000 Lemmata umfassende *Variantenwörterbuch des Deutschen* (hinfort VWB) in seiner Erstauflage erschienen.¹ Im VWB wurde Rumänien als Viertel(s)zentrum² nicht erwähnt, hingegen Südtirol und Liechtenstein. Das Nachschlagewerk stellt Varianten der Standardsprache in den sogenannten „Vollzentren“ (Deutschland, Österreich und die deutschsprachige Schweiz), die sich dadurch auszeichnen, dass ihre standardsprachlichen Besonderheiten kodifiziert und damit autorisiert sind, denen der „Halbzentren“ (Liechtenstein, Luxemburg, Ostbelgien und Südtirol) gegenüber, wo Deutsch offizielle und/oder Amtssprache ist.³

Das Vorkommen des Deutschen in verschiedenen Ländern mit teilweise divergierenden standardsprachlichen Normen wird in der Fachliteratur unter dem Terminus *Plurizentrik/Plurizentrität* erfasst. In der Plurizentritätsforschung ist jedoch umstritten, inwieweit das Deutsche als plurinationale oder als pluriareale Sprache einzuordnen ist. Während der Begriff der *Plurinationalität* die standardsprachlichen Besonderheiten auf nationaler Ebene hervorhebt, bezieht sich der Begriff der *Pluriarealität* auf die areale Gliederung des Deutschen, die nicht mit nationalen Grenzen in Zusammenhang steht.⁴

Laut den Hinweisen zur Benutzung (S. XI-XVI) vermerkt das Nachschlagewerk Wörter und Wendungen des Standarddeutschen mit nationalen oder regionalen Besonderheiten.⁵ Für die Aufnahme eines Stichwortes waren Unterschiede im Vorkommen, d.h. in der Verbreitung, Bedeutung, Gebrauchsweise und Verwendungshäufigkeit in den gesichteten Quellen, in der Forschungs-

literatur und anderen Wörterbüchern ausschlaggebend. Das VWB erfasst damit erstmals die Varitäten der deutschen Standardsprache, wobei Dialekt und Umgangssprache nur dann berücksichtigt wurden, wenn die entsprechenden Wörter und Ausdrücke häufig in den ausgewerteten standardsprachlichen Quellen vorkamen. Diese wurden mit dem Vermerk „Grenzfälle des Standards“ aufgenommen. Fachsprachliches oder aus dem aktuellen Sprachgebrauch ausgeschiedenes Wortmaterial, Wörter und Wendungen aus der ehemaligen DDR, okkasionelle und sprecherindividuelle Bildungen blieben unberücksichtigt.⁶ Der Wörterbuchartikel bietet Informationen zur Grammatik, Aussprache, Lautschrift, Bedeutung, Etymologie, Angaben zur nationalen und regionalen Zuordnung (Varianten), Querverweise auf die gemeindeutschen, im ganzen deutschen Sprachgebiet geltenden Entsprechungen, sodass der Variantenreichtum des Deutschen deutlich wird. Auf diese Weise werden Unterschiede und Gemeinsamkeiten bezüglich der Standardsprache übersichtlich und benutzerfreundlich herausgehoben. Die Arealangaben bei den Lemmata sind nach Länderkürzeln⁷ angeordnet. Ein Verweis- und Ergänzungsapparat („Dazu“-Teil), der zum Stichwort gehörende Ableitungen und Komposita und weitere Zusatzangaben (Frequenz, Alter, Stilschicht, zusätzliche Bezeichnungen oder Synonyme) vermerkt, rundet den Wörterbuchartikel ab.

Seit Ende 2016 liegt eine komplett neu bearbeitete, stark erweiterte und aktualisierte Neuauflage des VWB vor.⁸ Diese Neuauflage erfasst nicht nur den Sprachgebrauch in Ländern und Regionen mit Deutsch als Amtssprache, sondern auch wichtige, bisher lexikografisch nicht kodifizierte, voneinander abweichende standardsprachliche Besonderheiten des Deutschen in Rumänien, Namibia, Nordamerika und Mexiko.⁹ In diesen vom geschlossenen deutschen Sprachraum entfernten Gebieten haben sich eigenständige Varietäten des Deutschen herausgebildet, die in der Varietätenlinguistik als Viertel(s)zentren gelten.¹⁰

Von einem plurizentrischen Ansatz ausgehend, stellt das Nachschlagewerk standardsprachliche Varianten der Länder mit Deutsch als Amtssprache (Bundesrepublik Deutschland, Österreich, Schweiz, Liechtenstein, Luxemburg, Ostbelgien und Südtirol) anderen spezifischen Ausdrücken des Deutschen als Regionalsprache gegenüber. Die theoretisch-methodologischen Voraussetzungen gehen auf das Konzept des Deutschen als plurizentrische Sprache und auf die Erscheinungsformen der Standardvarietäten zurück. Daher galt es zu fragen, welcher Grundwortschatz dominant oder teilweise als staatspezifisch zu werten ist und welcher über mehrere Staatsgebiete oder dessen Teile hinaus verbreitet ist.

Im Hinblick auf das veränderte Normdenken zur standardsprachlichen Regionalität ist es erfreulich, dass sich das VWB vornimmt, die Varietätenunterschiede des Deutschen und damit das Varietätenspektrum am Rande und weit außerhalb des geschlossenen deutschen Sprachgebiets lexikografisch zu dokumentieren und die Besonderheiten der Viertel(s)zentren, die historisch unter verschiedenen Bedingungen aufgekommen sind, gleichberechtigt zu

behandeln. Mit der Fokussierung auf die schriftsprachliche Seite der Standardvarietäten, der die Bestandsaufnahme im VWB gewidmet ist, belegt das VWB Unterschiede und damit auch Eigenheiten der Viertel(s)zentren, um sie von der zweifelhaften Einschätzung als Non-Standard zu befreien. Leider sind *diese* Varietäten auch in der Fachliteratur weniger berücksichtigt und ihre Besonderheiten nur vereinzelt beschrieben, jedoch bis dato nicht lexikografisch erfasst, erklärt und denjenigen der Voll- und Halbzentren gegenübergestellt worden.

Aufgrund des lobenswerten Bestrebens, das gesamte gegenwärtige Varietätenspektrum lexikografisch zu dokumentieren und der damit einhergehenden Informationsdichte des VWB in seiner Neubearbeitung gilt es zu fragen, ob es nicht ergiebiger wäre, statt einer Globaldarstellung einen Teilaspekt regionaler Vielfalt unter Berücksichtigung der Ausbildung von Normen und Formen des Sprachgebrauchs herauszugreifen und gesondert zu beleuchten. Generell soll daher geprüft werden, ob der Wörterbuchbenutzer, der mit punktuellen Anliegen zur intrasprachlichen Variation zum Wörterbuch greift, auch fündig wird. Die Aufgabe dieses Beitrags ist es daher zu prüfen, inwiefern eine überzeugende lexikografische Bearbeitung der in der Erstauflage (2004) zu Unrecht vernachlässigten Varietäten der Viertel(s)zentren gewährleistet wurde, speziell inwieweit die Kodifizierung einer interessanten Sprachlandschaft in einem historischen bedeutsamen deutschsprachigen Siedlungsgebiet gelungen ist. Im Folgenden soll dies exemplarisch am Beispiel eines ausgewählten Viertel(s)zentrums und des hier gebräuchlichen Standards *Rumäniendeutsch*¹¹ verdeutlicht werden. Die rumänische Variante der deutschen Standardsprache, das rumänische Deutsch oder die rumäniendeutsche Varietät, ist bruchstückhaft und eher aus kontaktlinguistischer Sicht beschrieben worden. Auch wenn die rumäniendeutsche Bevölkerung in den letzten drei Jahrzehnten sehr stark zurückgegangen ist, müssen die standardsprachlichen Besonderheiten des Deutschen in Rumänien erforscht werden. Die Berücksichtigung der *rumäniendeutschen Standardvarietät* und der Besonderheiten des Deutschen in Sprachinsellagen sind — über die variationslinguistische Relevanz und Herzensangelegenheit der Autorin dieses Beitrags hinaus — auch damit zu begründen, dass das Rumäniendeutsche ein ausgebautes Diasystem aufweist.¹²

Mit der Erfassung verschiedener Varianten der Standardsprache, die sich bei einer plurizentrischen Sprache entwickelt haben, ist dieses Standardwerk daher auch für die rumänische, amerikanische oder kanadische Sprachforschung relevant. Es deckt nicht nur spezielle Informationsbedürfnisse ab, sondern wird den Vergleich der Viertel(s)zentren miteinander sicherlich fördern und auch dazu beitragen, dass auch *dieses Deutsch* intensiver wahrgenommen wird.

Für die deutsche Standardvarietät in Rumänien war Ioan Lăzărescu von der Universität Bukarest der verantwortliche Experte. Es ist Ioan Lăzărescus Verdienst, die Bewusstseinsbildung vorangetrieben zu haben, sodass Rumänien zum Viertel(s)zentrum geklärt werden konnte und in einer breiteren Öffentlichkeit dem *Rumäniendeutschen* mehr Interesse entgegengebracht wird.

2. Variantenreichtum des Deutschen

2.1 Konzeptionelle Ausrichtung und Neuaufnahmen in der Neuauflage des VWB

Die Neuauflage 2016 geht auf eine trilaterale Forschungs Kooperation der Arbeitsstellen in Deutschland (Universität Duisburg-Essen), in Österreich (Universität Wien) und in der Schweiz (Universität Basel) zurück. Das Dreiländerprojekt zur Erweiterung und Verbesserung des VWB ist 2012 gestartet.¹³ Bei dem großformatigen Band handelt es sich jedoch *nicht* um einen ergänzten Nachdruck eines älteren Nachschlagewerks, sondern um ein völlig neues Wörterbuch. In der Erstauflage des VWB wurden eben *nur* die lexikalischen Varianten der deutschen Standardsprache in den nationalen Voll- und Halbzentren kodifiziert, wo Deutsch offizielle und/oder Amtssprache ist. Mit der Ausarbeitung eines *neuen* VWB, das regionale und nationale Besonderheiten der deutschen Sprache kodifiziert, können neue Einsichten in Differenzierungsprozesse der deutschen Standardsprache geboten werden. Das Wörterbuchteam verfolgt mit dieser Neuauflage vielseitige Ergebnisse, darunter ein besseres Verständnis zwischen verschiedenen deutschsprachigen Nationen; eine ausgewogene Berücksichtigung von Sprachbesonderheiten; die Erstellung einer Datenbasis für eine verlässlichere lexikografische Beschreibung der nationalen und regionalen Variation; die Lieferung einer empirischen Basis für wichtige theoretische Fragestellungen im Hinblick auf Variation und standardsprachorientierten Normen. Das Projektteam war nicht nur um eine neue Auflage des VWB bemüht, sondern auch an deren Bekanntmachung durch Publikationen, Vorträge und Lehrveranstaltungen.¹⁴

Für die Neuauflage wurde der gesamte Lemmabestand der Erstauflage (2004) wissenschaftlich und empirisch überprüft und um 2.500 Stichwörter und Wortvarianten als Zusatzangaben bereichert, wobei das Internet für die Ermittlung der Verbreitung und Häufigkeit der regionalen und nationalen Eigenheiten des Deutschen und von alternativen Wortvarianten wie auch für die Aktualisierung und das Zusammentragen neuer Belege ausgiebig genutzt wurde.¹⁵ Für die Erhebung schriftsprachlicher Varianten wurden unterschiedlich geartete, regional und inhaltlich vielfältige und elektronisch verfügbare Quellen ausgewertet. Als solide Grundlage für die Variantensuche fungierten hauptsächlich umfassende und aktuelle Korpora mit gedruckten standardsprachlichen Texten, welche die mehr oder weniger befriedende Beleglage und die Belegverdichtung für das 20.–21. Jh. bezeugen.

Die Bestandsaufnahme der jeweiligen Varietäten des Deutschen und die Prüfung der Stichwörter im Hinblick auf ihre Gebräuchlichkeit und Geltung im jeweiligen Gebiet ist der Zusammenarbeit mit achtzehn (Regional-)ExpertInnen, darunter auch erfahrene LexikografInnen, zu verdanken, wo die jeweiligen Varietäten des Deutschen gesprochen werden.

Ausschlaggebend für die Aufnahme alter und neuer Lemmakandidaten

war ein in den jeweiligen drei Arbeitsgruppen intern angewandtes Kriterienraster, das die Eruiierung nationaler oder regionaler Besonderheiten erlaubte: die korpusgestützte Gebrauchsfrequenz der als Lemma angesetzten Wörter in den qualitativ-quantitativ ausgewerteten umfangreichen elektronischen Zeitungskorpora und eine evidente stilistisch-varietätenlinguistische Markierung. Wenn die Abgrenzung nicht immer eindeutig vorgenommen werden konnte, wurde der Hinweis „Grenzfall des Standards“ angebracht. Hierzu gehören Wörter, die eigentlich dem Dialekt oder der Umgangssprache zuzuordnen wären, aber häufig in Standardtexten vorkommen. Die für die jeweiligen Viertel(s)zentren gültigen und typischen Ausdrücke wurden nur dann aufgenommen, wenn ihre Verbreitung in keinem anderen Zentrum belegt war. Die Ermittlung der neu hinzugekommenen Standardvarianten beruht auf der Auswertung von umfangreichen und aktuellen Sprachkorpora, die für die Erfassung der nationalen und arealen Lemma-Distribution geeignet sind.

Konkret schlagen sich die Neuerungen in einer verbesserten arealen korpusbasierten Lemmata-Verortung, in einer angemesseneren Bestandsaufnahme auch binnendeutscher Variation nieder wie auch in einer überarbeiteten Lauttabelle und einer sorgfältigeren Kennzeichnung der Grenzfälle des Standards und der Markierungspraxis.

Neben einer gründlichen Überarbeitung der Erstauflage bezweckt die Neuauflage des VWB „eine theoretische Neufassung und Neudefinierung zentraler Begrifflichkeiten“ wie z.B. „Standardsprachlichkeit“, „Standardsprache(n)“, „Grenzfälle des Standards“, „gemeindeutsch“ (S. XIII-XIV) auf empirischer Basis. Entgegen der hohen Dichte des Begriffs „national“ in der Erstauflage begegnet in der Wörterbucheinleitung der Neuauflage ein reflektierter Umgang mit fachlich-terminologischen Feindifferenzierungen, was sich auch im angemesseneren Pendant *regional-areal* äußert. Bei der Durchsicht des Nachschlagewerks nimmt es daher nicht Wunder, dass die Lexikografen und die beratenden Experten auf theoretisch-methodische Fragestellungen, die grundlegend für die Konzeption, Zielsetzung und Struktur des Wörterbuchs und die mit der variationslinguistischen Erhebung und Erforschung von Sprachphänomenen verbunden sind, großes Gewicht gelegt haben. Es gehört zur Gründlichkeit der Darstellung bei der konzeptionellen Herleitung und empirischen Fundierung auf die Legitimation der *Viertelzentren des Deutschen* anregend und fachkundig, dem aktuellen Forschungsstand entsprechend, eingegangen zu sein, sodass auch das interessierte Nichtfachpublikum hier einen guten Einstieg findet. Kapitel 4 informiert über Charakteristika der Voll-, Halb- und den „echten“ Viertel(s)zentren des Deutschen (S. XXXIX-LXIII) in Rumänien (S. LX), Namibia (S. LXI) und in den Mennonitensiedlungen in Mexiko (S. LXII). Hier haben sich spezifische und eigenständige Varianten des Deutschen, eben als *Standardvarietäten* anzuerkennende Formen herausgebildet, die eine für in diesem Areal lebende deutsche Minderheit normative Geltung aufweisen, demnach auch Modelltexte hervorgebracht haben und im öffentlichen Sprachgebrauch anerkannt sind, obwohl sie in Regelwerken nicht kodifiziert sind.¹⁶

Für diese Auflage wurde kein neues lexikografisches Erfassungskonzept entworfen; das gelungene Ausarbeitungsmuster der Erstaufgabe wurde beibehalten. Um auch den Aufbau des Wörterbuchartikels nachvollziehbar zu machen, hat das Wörterbuchteam die Anordnung der Angaben und deren Funktion innerhalb der zehn Artikelpositionen farblich gestaltet (s. den inneren Einbanddeckel). Die Neuauflage verzichtet jedoch auf die Namenartikel, unter denen in der 2004 erschienenen Erstaufgabe länderspezifische und regionaltypische (traditionelle) Personennamen oder inoffizielle geografische Namen für Städte und Landschaften gebucht wurden. Dafür bietet das VWB in seiner Neuauflage eine erhöhte empirische Fundierung des gesamten kodifizierten Sprachmaterials, die durch quantitative und qualitative Analysen umfangreicher Quellenkorpora gewährleistet wurde. Das VWB verdeutlicht, wie Wörter und Wendungen mit national oder regional eingeschränkter Verbreitung oder Differenzen im Gebrauch mit ihren gemeindeutschen Entsprechungen lexikografisch adäquat dargestellt werden können. Diese Leistung verdient — auch angesichts der kurzen Bearbeitungszeit von vier Jahren — alle Hochachtung.

Das VWB schließt mit seiner korpusbasierten Darstellung des national- und regionalspezifischen Wortschatzes der deutschen Standardsprache nicht nur eine lexikografische Lücke, sondern bietet auch neue Einsichten in die Varietätenvielfalt des Deutschen und in die Beurteilung von Variation. Das VWB leistet damit einen wertvollen Beitrag für das bessere Verständnis des Deutschen als plurizentrische Sprache.

2.2 Bearbeitungspraxis der Standardsprache zuzurechnenden lexikalischen Regionalspezifika. *Rumänismen* als Fallbeispiel

Das VWB basiert auf der Auswertung eines umfangreichen Quellenkorpus aus allen Ländern, in denen Deutsch nationale/regionale Amtssprache oder anerkannte Minderheitensprache ist, sowie des Internets als Belegquelle. Es präsentiert empirische Evidenz anhand interessanter Beispiele, die Besonderheiten des arealen Auftretens des Lemmakörpers verdeutlichen. Indem das VWB lexikalische Eigenprägungen als der Standardsprache zuzurechnenden lexikalischen Regionalspezifika zu erfassen beabsichtigt, ergibt sich daraus die Aufgabe, dem Wörterbuchbenutzer Gemeinsamkeiten und Unterschiede der Lemmazeichen mit den entsprechenden Einheiten der deutschen Standardsprache zu verdeutlichen.¹⁷

Bei der Lemmaauswahl musste für die lexikografisch zusätzlich erfassten Viertel(s)zentren eine strenge und überlegte Auswahl getroffen werden, um einerseits hierfür repräsentative Lexeme aufzunehmen und auch die vorgegebenen Lemmata- bzw. Seitenanzahl bei einem gedruckten Nachschlagewerk, das Voll- Halb- und Viertel(s)zentren ausgewogen berücksichtigt, nicht zu überschreiten.

Die Neuauflage umfasst insgesamt 162 Lemmata der in den jeweiligen Viertel(s)zentren gültigen und typischen Ausdrücke, im VWB als RUM, NAM,

MENN kodifiziert, die fast ausnahmslos Presstexten entnommen wurden und die in keinem anderen Zentrum im Gebrauch sind. Das VWB umfasst folglich *nicht* den gesamten Wortschatz des Standarddeutschen.¹⁸ Es führt nur diejenigen Wörter und Wendungen mit staatlichen oder regionalen Besonderheiten vor, die nicht im gesamten deutschen Sprachgebiet verbreitet sind oder je nach Land oder Region unterschiedliche Bedeutungen aufweisen, verschiedenen Sprachstilen zugerechnet werden, von diversen Sprechergruppen unterschiedlich verwendet werden. Daher nimmt das VWB keine österreichisch-rumäniendeutsche lexikalische Gemeinsamkeiten auf. Diese werden bei Lăzărescu und Scheuringer (2007) als *Rumäno-Austriazismen* ausführlich beschrieben und mit dem hochgestellten Kürzel (RO) markiert, das rechts vom Lemma in runden Klammern steht.¹⁹

Obwohl Deutsch in Rumänien keinen Amtssprachenstatus hat, sind Modelltexte und Normautoritäten vorhanden, in denen eine für Rumänien spezifische Standardvariante des Deutschen erkennbar ist, wodurch dieser Varietät der Anspruch auf Standardsprachlichkeit gegeben ist.²⁰ In Rumänien ist (Hoch-)Deutsch die überregionale und relativ einheitliche Verkehrssprache, zugleich auch Schrift-, Kirchen- und Unterrichtssprache der regional getrennt lebenden deutschsprachigen Minderheiten (Siebenbürger Sachsen, Zipser, Banater und Sathmarer Schwaben, Landler, Bukowinadeutsche, Bessarabiendeutsche, Dobrudschadeutsche, Regatdeutsche).²¹ Die heutige deutsche Minderheit ist geografisch im Zentrum Rumäniens, in Siebenbürgen (rum. Transilvania) um die Städte Hermannstadt (rum. Sibiu), Kronstadt (rum. Braşov) oder Klausenburg (rum. Cluj-Napoca), im westlichen Banat um Temeswar (rum. Timişoara) und Reschitza (rum. Reşiţa), im Nordwesten um Sathmar (rum. Satu Mare) konzentriert.

Schriftsprache der deutschsprachigen Gemeinschaften in Rumänien war das regional gefärbte Hochdeutsche. Die jahrhundertelange Zugehörigkeit deutschsprachiger Gebiete Rumäniens zur Habsburgermonarchie ermöglichte den Erhalt der deutschen Sprache. Anders als in anderen Staaten Mittel- und Osteuropas hat es in Rumänien auch während des Kommunismus und der Ceuşescu-Diktatur ohne Unterbrechung ein deutschsprachiges Schulwesen gegeben.²² Die „deutschen Schulen“ haben in Rumänien Tradition und einen sehr guten Ruf.²³ In den 1960er-Jahren begann die Abwanderung der deutschsprachigen Bevölkerung nach Deutschland.²⁴ Der letzte große Exodus Anfang der 1990er-Jahre hatte eine Minderung der Lehrer- und Schüleranzahl an deutschsprachigen Schulen zur Folge. Nach der Abwanderung der meisten Rumäniendeutschen und des damit verbundenen starken demografischen Rückgangs an muttersprachlichen Schülerinnen und Schülern wurden ab 1990 die traditionellen deutschen Schulen mehrheitlich von Rumänischstämmigen besucht. Auch vor der Wende (1989) galt Deutsch als Prestigevarietät und die deutschsprachigen Schulen waren auch von der rumänischen Mehrheitsbevölkerung begehrt. In den 1990er-Jahren stieg die Nachfrage bei der Mehrheitsbevölkerung kontinuierlich. Trotz Schüler- und Lehrermangel konnte sich das

deutschsprachige Schulsystem und der Unterricht in deutscher (Mutter-)Sprache z.B. in den Städten dadurch erhalten, dass die Mehrzahl der Schüler rumänische oder ungarische Muttersprachler waren und noch ausreichend viele Lehrkräfte vorhanden waren, die als Schüler des deutschsprachigen Schulwesens über adäquate Deutschkenntnisse verfügten. Gehörten Anfang der 1990er-Jahre die Schülerinnen und Schüler deutscher Klassen vorwiegend der deutschen Minderheit an, so sind nach dem starken Rückgang der deutschen Minderheit maximal fünf Prozent aller Schülerinnen und Schüler in „deutschen Schulen“ Angehörige der Minderheit. Rumänische oder ungarische Familien schrieben ihre Kinder in die deutschen Schulen ein, da der Erwerb solider Deutschkenntnisse ihnen bessere Berufschancen sicherten. Für diese ist Deutsch keineswegs die im Alltag gebrauchte Sprache, sondern „Bildungssprache“ und später „Berufssprache“.

Die Überalterung der deutschen Minderheit und das Fehlen einer ausgewogenen Verteilung sozialer Schichten aufgrund der Massenauswanderung gehören zu den größten Schwierigkeiten für den Erhalt des Deutschen als Muttersprache in Rumänien.²⁵ Hinzu kommt, dass ein schwindend geringer Sprecheranteil Deutsch als Muttersprache pflegt und — im Gegensatz zu vielen Ungarn — Rumänisch den Status einer Muttersprache erlangt hat. Keine einzige Region Rumäniens ist heute mehrheitlich von deutschen Muttersprachlern besiedelt, selbst Ortschaften mit vorwiegend deutschsprachiger Bevölkerung gibt es nicht mehr.

Gegenwärtig leben über eine Million Rumäniendeutsche mit ihren Nachkommen in Deutschland. Nach der Rückwanderung der Rumäniendeutschen nach Deutschland (ab 1990) prophezeite man einen bevorstehenden Sprachverlust: Das Rumäniendeutsche würde in die binnendeutsche Standardsprache aufgehen, sodass das Deutsche der aus Rumänien Ausgewanderten als Sprachvarietät an die nächste Generation nicht mehr weitergeben werden könnte. Ein „Sprachtod“ sei nach Lăzărescu (2017: 356), trotz Rückgang deutscher Muttersprachler in Rumänien, allerdings *nicht* zu befürchten. An dem Sozialprestige wie auch an dem hohem gesellschaftlichem Bedarf, das dem Deutschen zukommt, hat sich bis heute in Rumänien nichts geändert. Die deutsche Minderheit hat ein wichtiges kulturelles, sprachliches und geistiges Erbe hinterlassen, das zu erhalten zweifelsohne nicht einfach, jedoch allseits erwünscht und auch möglich ist. Die rumänische Variante der deutschen Standardsprache oder das „rumänische Deutsch“

steht heute nicht mehr ausschließlich für die von L1-Sprechern gebrauchten Varietäten, sondern auch für das Deutsch zahlreicher Sprecherinnen und Sprecher mit rumänischer und ungarischer Erstsprache, die an Schulen mit deutscher Unterrichtssprache lernen und meist eine sehr hohe Kompetenz in der geschriebenen und gesprochenen Standardvarietät erreichen. Für einige dieser L2-Sprecher wird Deutsch zur Berufssprache, für andere bleibt es Schul- und Bildungssprache, für deren Verwendung nach Schulabschluss nur wenige oder keinerlei Kommunikationssituationen bestehen bleiben.²⁶

Der Begriff *Rumäniendeutsch* bezeichnet folglich nicht nur den in Rumänien gepflegten Sprachgebrauch einer historischen deutschen Minderheit, sondern auch den Sprachgebrauch der Deutsch sprechenden Rumänen, die Deutsch vorwiegend als Bildungs- und Berufssprache gebrauchen. Deutsch ist in Rumänien nicht nur Minderheitensprache, sondern — und vor allem — auch Verkehrssprache zwischen Nichtmuttersprachlern. Und schließlich: Deutsch wird neben anderen Minderheitensprachen (z.B. Ungarisch) auch in den Medien und in der Literatur verwendet.²⁷

Als „eigenständige Varietät mit standardsprachlicher Geltung“ (Lăzărescu 2013a: 370) und überregionale Kommunikationsform einer deutschen Minderheit ist das Rumäniendeutsche durch sprachliche Gemeinsamkeiten mit der österreichischen Varietät des Deutschen, den verschiedenen regionalen Mundarten und dem Rumänischen gekennzeichnet. Die rumänische Variante der deutschen Standardsprache ist eine durch eigene Hochsprachlichkeit gekennzeichnete Varietät des Deutschen, die nicht als dialektal oder fehlerhaft einzuschätzen ist und die auch die Kriterien für die Standardsprachlichkeit erfüllt. Daher ist es sehr zu begrüßen, dass das VWB durch die Aufnahme Rumäniens als Viertel(s)zentrums dem recht diffusen Bild der rumäniendeutschen Sprachvarietät schärfere Konturen verleiht und Anerkennung verschafft. Varietäten im deutschen Sprachraum werden auch von den SprecherInnen oft gar nicht als solche wahrgenommen oder als Abweichungen aufgefasst, sodass das Deutsch Deutschlands als das „eigentliche“ Deutsch gilt. Oft begegnet nämlich die Ansicht, rumänisches Standarddeutsch — zumindest in seiner gesprochenen Form — sei „nur eine Variante des österreichischen Deutsch“. Es würde auf Kompetenzdefizite hinweisen und massive Interferenzphänomene aufweisen, folglich als minderwertige Variante des Deutschen einzuschätzen sein. Eine solche Vereinfachung ist nicht vertretbar. Das für das Lemmainventar ausgewertete drucksprachliche Korpus mit diatopischer Verteilung verbietet Pauschalzuordnungen.

Aus synchroner Sicht wird das Rumäniendeutsche als Regionalsprache und Standardvarietät von anderen Sprachen (Rumänisch und Ungarisch) und Varietäten des Deutschen beeinflusst. Das Rumäniendeutsche weist sprachliche Gemeinsamkeiten mit der österreichischen Varietät²⁸, dem Schweizerhochdeutschen und kleineren Varietäten nationaler Halbzentren auf wie auch mit verschiedenen regionalen Mundarten. Hinzu kommen vielfältige lexikalisch-grammatische Interferenzerscheinungen, bedingt durch den Kontakt zum Rumänischen als Sprache der Mehrheitsbevölkerung und anderen autochthonen Minderheitensprachen (z.B. das Ungarische).

Entgegen seinen dialektalen Grundlagen ist das Rumäniendeutsche „dominant *süddeutsch*, genauer gesagt [...] deutlich *österreichisch* gefärbt“ (Lăzărescu 2013a: 375; Hervorhebung im Original). In dem von Ioan Lăzărescu und Hermann Scheuringer 2007 herausgegebenen Wörterbuch zu den österreichisch-rumäniendeutschen lexikalischen Gemeinsamkeiten werden die *Rumäno-Austriazismen* erklärt und beschrieben. Das Wörterbuch erfasst aus-

schließlich Austriazismen und ihre rumänischen Entsprechungen und belegt damit Ähnlichkeiten zwischen dem Rumäniendeutschen und der österreichischen Variante des Deutschen. Im Bereich der Lexik (Küche/Gastronomie, Verwaltung, Beruf, Haushalt, Handwerk), Wortbildung (z.B. Gebrauch der Fugenelemente) und Grammatik (z.B. Bildung des Perfekts mit „sein“ bei den Verben *sitzen, stehen, liegen*) ist der Einfluss der österreichischen Varietät offensichtlich. Vgl. hierzu die Bezeichnungen für Haus-/Einrichtungsgegenstände (z.B. *Gang, Rauchfang, Eiskasten, Mistkübel, Polster, Kleiderhänger*), Lebensmittel (z.B. *Semmelbrösel, Staubzucker, Zuckerwata, Topfen*), Speisen (z.B. *Eierspeise, Kipferl, Knödel*), Gemüse (z.B. *Kren, Karfiol, Fisolen, Paradeis*), Obst (z.B. *Ribisel, Weichsel*). Für das Rumäniendeutsche kann nach dem Zweiten Weltkrieg der zunehmende Einfluss des deutschen Deutsch und die verstärkte Interferenz mit dem Rumänischen ausgemacht werden. Die offiziell existierende Sprachsituation wie auch der Umbruch, der mit dem Aufkommen neuer Medien ausgelöst wurde, ersetzten die Diglossie (Dialekt — Standardsprache) durch den Bilinguismus (Standarddeutsch — Rumänisch) und eröffneten auch einen intensiven Varietätenkontakt, sodass die ehemals stark österreichisch gefärbte rumäniendeutsche Verkehrssprache mit bundesdeutschem Wortgut und deutsch-rumänischen Mischbildungen bereichert wurde. Der an den deutschsprachigen Schulen gebotene Sprachunterricht ist bemüht, Transfer- und Interferenzerscheinungen gezielt zu mindern.²⁹

Als Varietät des Deutschen zeigt das Rumäniendeutsche zudem auch eigene Varianten, d.h. Eigenbildungen — *Rumänismen* —, die in allen deutschsprachigen Regionen Rumäniens im Gebrauch sind. Darunter werden Besonderheiten der deutschen (Hoch-)Sprache in Rumänien erfasst, die jedoch nur teilweise mit den Einflüssen des Rumänischen als Amtssprache verbunden sind. Vgl. hierzu die Sachgebiete Haushalt, Kleidung, Kochkunst, Flora und Fauna, Geselligkeit, Sitten und Bräuche, Beruf, Handel, Wirtschaft, Politik, Verwaltung, Schulwesen u.a. Für Nichteingeweihte muten Wörter oder deren Gebrauchsweisen wie z.B. *Wettbewerb* („Stellenbesetzung“) und insbesondere landestypische Sachbezeichnungen — darunter auch besondere Benennungen für politische, administrative, kulturelle Einrichtungen (z.B. *Erste-Grad-Prüfung, Generalschulinspektor, Inspektorat, Allgemeinschule, Generalschule, Katalog, Klassenkollege, Kulturhaus*) — kurios an. Es handelt sich um typische Lexeme des Rumäniendeutschen als Ergebnis eines multikulturellen und mehrsprachigen Umfeldes, speziell um Einflüsse des Rumänischen als Amtssprache — z.B. rumänische Transferenzen wie *buletin* (dt. Ausweis) oder *stare civilă* (dt. Standesamt), Lehnübersetzungen und hybride Wortformen — wie auch um den Einfluss des Ungarischen als Umgebungssprache auf diese Varietät. Die Eigenbildungen umschreiben ein wichtiges Merkmal dieser Varietät, das sie von denen im geschlossenen deutschen Sprachraum unterscheidet: Es handelt sich um typische Lexeme des Rumäniendeutschen als Ergebnis zahlreicher Sprachkontakte. Kennzeichnend für Siebenbürgen ist, dass mehrere Sprach(varietät)en in direktem Kontakt stehen (z.B. Deutsch und Ungarisch als autochthone Minder-

heitenssprachen), sodass die Prämissen zu vielfältigen Sprachkontakten gegeben sind. Der Einfluss des Rumänischen zeigt sich u.a. in den zahlreichen integrierten Lehnwörtern und Lehnübersetzungen (z.B. *didaktisches Material* ‚Lehrmittel‘ oder *Experiment* ‚Versuch‘), in den lateinbasierten Verben (z.B. *insistieren*, *motivieren*, *inspirieren*), im Gebrauch des Verbs *machen* anstatt semantisch differenzierender Verben (z.B. *Französisch machen*; *das Militär machen*), in den festgeprägte Wendungen (z.B. *jmdn. am Telefon erwischen*; *jmdm. ein Telefon geben*; *eine Prüfung geben*; *eine Prüfung nehmen*; *eine Kontrollarbeit schreiben*), die auf Interferenzen mit dem Rumänischen zurückgehen. Zu den standardsprachlichen Besonderheiten des Deutschen in Rumänien zählen landestypische Sachbezeichnungen wie z.B. der Parade-Rumänismus *Märzchen*. Hierfür schlug 1995 Ulrich Ammon den Begriff *Transsyvanismen* vor, der terminologisch jedoch nicht alle historischen Sprachgebiete deutschsprachiger Siedler abdeckt. *Rumänismus* als Terminus steht daher gleichberechtigt neben *Teutonismus*, *Austriazismus* und *Helvetismus*.

Das Rumäniendeutsche weist Besonderheiten auf allen sprachlichen Ebenen auf, in der gesprochenen Standardsprache wie in der geschriebenen. Über die vielfältigen lexikalisch-grammatischen Interferenzerscheinungen hinaus, die durch den verstärkten Kontakt zum Rumänischen als Sprache der Mehrheitsbevölkerung und anderen autochthonen Minderheitensprachen (z.B. das Ungarische) bedingt sind, kann für das Rumäniendeutsche der Einfluss des DDR-Deutschen nach dem Zweiten Weltkrieg und ab 1990 der zunehmende Einfluss des deutschen Deutsch angenommen werden.

Die im VWB erstmals erfassten Varianten der Viertel(s)zentren — darunter 79 Rumänismen, 37 Namibismen und 46 Lemmata aus den mexikanischen Mennonitensiedlungen — belegen Eigenheiten des hier gesprochenen Deutsch, die inhaltlich verschiedene Bereiche (Verwaltung, Schulwesen, Wirtschaft, Kochkunst, Geselligkeit, Brauchtum) abdecken.³⁰ Das Lemmainventar im VWB weist Eigenbildungen, teilweise nach deutschen Wortbildungsmustern auf, die in anderen Varietäten in dieser Form und/oder Bedeutung unüblich sind. Vgl. z.B.: RUM: *Aufboden* ‚Dachboden‘, *Muskelfieber* ‚Muskelkater‘, *Bierfabrik* ‚Brauerei‘; vgl. auch *Baumstrietzel*, *Lektionsplan*, *Kontrollarbeit*, *ultrazentral*.³¹ Manche Wortvarianten sind denen in anderen Varietäten formgleich und -nahe, haben jedoch in dem betreffenden Viertel(s)zentrum zumindest auch eine spezifische Bedeutung. Vgl. z.B. RUM: *Programm* ‚Stundenplan; Öffnungszeiten eines Geschäfts‘, *Akademiker* ‚Mitglied einer wissenschaftlichen Akademie‘, *Analyse* auch ‚Blut- oder Urintest‘; vgl. hierzu auch die Lemmata *Katalog*, *Notenheft*. Die Besonderheiten dieser Varietät zeigt sich einerseits auch in den Eindeutschungen von rumänischen Wörtern oder Wortteilen — vgl. hierzu u.a. RUM: *Thermozentrale* ‚Wärmeleistungswerk‘, *Mikrobus* ‚Kleinbus, Minibus‘, *Tokane* ‚Gulasch nach rumänischer Art‘, *Vinete* ‚Salat aus gerösteten und zerhackten Auberginen‘, *Amphitheater* ‚Hörsaal‘, *Zuika* ‚[Pflaumen-]Schnaps‘; vgl. auch *Matrikelblatt*, *Turmblock*, *Winterkommando*, *Hydrozentrale* oder *Bo-kantsch* —, jedoch insbesondere in den Realienbezeichnungen, die nur in dem

betreffenden Viertel(s)zentrum gültig sind und oft keine Entsprechung in einem anderen Zentrum aufweisen. Die Sach-Rumänismen (Lehnbildungen, Lehnprägungen, Lehnerschöpfungen oder Lehnübersetzungen) umschreiben landeskundlich relevante Sachbereiche (z.B. Brauchtum, Einrichtungen, Institutionen). Vgl. z.B. RUM: *Märzchen* ‚Glücksbringer, der von Mädchen und Frauen an einer weiß-roten Schnur im Monat März getragen wird‘, *Bakkalaureat* ‚das rumänische Abitur‘, *Allgemeinschule* ‚erste Gymnasialstufe‘, *Lyzeum* ‚zweite Gymnasialstufe‘, *Definitivatsprüfung* ‚erste Lehramtsprüfung‘, *Kulturheim* ‚Kulturhaus [in einem Dorf]‘. Das Lemmainventar verzeichnet auch Wörter und Wendungen, die durch den Kontakt zur Amtssprache Rumänisch entstanden sind oder aus dem Dialekt übernommen wurden. Aus dem Rumänischen unverändert übernommen wurde z.B. *Mititei* ‚gegrillte Röllchen aus Hackfleisch‘, dialektales Wortgut lebt weiter z.B. in RUM *Palukes* ‚Maisbrei‘, *Hanklich* ‚eine siebenbürgisch-sächsische Art Kuchen‘ oder *Ägrisch* ‚Stachelbeere‘ und *Urzeln* ‚maskierte Gestalten (in der Narrenzzeit)‘ bzw. ‚*Hattert* ‚Gemarkung‘. Zu den Grenzfällen des Standards gehören u.a. RUM *Bizikel*, *Bulibasse*, *Motorin*, *Muskelfieber* oder *Sarmale*, da diese Wörter z.B. aus dem familiären Bereich in den überregionalen Sprachgebrauch eingegangen sind. Diese Lemmata werden mit einem Verweis an verschiedenen Stellen im VWB angeführt.

Einer breiten Wörterbuchnutzung kommt der im VWB angebrachte Varianten-Hinweis entgegen. Vgl. hierzu *Kletitten* RUM die; nur Plur. (Küche) < aus rumän. *clătite*, Pl. von *clătită* > (Grenzfall des Standards): → Omelett A, → Palatschinken A, → Omelette A CH, → Eierkuchen D-nordost/mittelost, → Pfannkuchen D-mittelwest, → Pfannkuchen D-süd, → Pfannkuchen D (ohne nordost/mittelost), → Plinse D-mittelost ‚Gericht aus einem dünnen Teig aus Eiern, Milch und Mehl, der in der → Pfanne in Fett gebacken, mit → Marmelade o.Ä. bestrichen wird‘: In der Mitte des Umzuges befand sich der „Pfannkuchenwagen“ mit einem qualmenden Ofen, auf dem symbolisch die Kletitten zubereitet wurden (Hermannstädter Zeitung 24. 2. 2012) — Selten auch in der Form *Kletiten* oder *Klettiten* geschrieben. — Dazu: *Kletitten-Festival* (S. 394).

Das Wörterbuch belegt auch das Vorkommen unterschiedlicher Phraseologismen in den Standardvarietäten des Deutschen, damit auch eigene, gebräuchliche regionalspezifische phraseologische Varianten. Vgl. z.B. RUM *in den Ägrisch gehen* (S. 21) und *sich Rechenschaft geben* (S. 579) bzw. NAM *ein Rivier kommt ab* (‚plötzliches starkes Wasserführen eines ausgetrockneten Gewässerlaufs‘; S. 8). Dabei handelt es sich um schriftsprachliche Phraseologismen, die aus dem aktuellen Sprachgebrauch (z.B. Presse und Internet) exzerpiert worden sind. Die Auswertung verschiedener Quellen belegt, dass obwohl mehrere Phraseologismen strukturell dialektale Merkmale aufweisen, diese Phraseologismen nicht areal begrenzt gültig sein müssen, d.h. Phraseologismen mit einer regional gebundenen Konstituente (z.B. *etw. geht jmdn. einen [feuchten] Kehricht an* = gemeint.; *Kehricht* CH, D-südwest; S. 378) müssen nicht auf das Verbreitungsgebiet der betreffenden Dialektwörter beschränkt sein. Hier sind empirische Untersuchungen erforderlich, um den arealen Geltungsbereich eines Idioms

ermitteln zu können. Unter dem Lemma *Ägrisch* [RUM der; -(e)s, -e < aus rumän. *agrișă* und in A dialektal *Agrasel* > ‚Stachelbeere‘ (S. 21)] wird auch der Phraseologismus **in den Ägrisch gehen* RUM (nur im Imp., salopp, Grenzfall des Standards) aufgeführt in der Bedeutung ‚abfahren CH, putzen A, vertschüssen A, verpissen CH D, dünnemachen D-nord/mittel; sich entfernen; verschwinden, abhauen‘. Der Phraseologismus wird aber auch unter anderen Lemmata (abfahren; S. 5; abschieben; S. 11; abschleichen; S. 12; Fliege; S. 243; Mücke; S. 483) angeführt.

Das VWB bietet in seiner Neuauflage ein empirisch ausgewiesenes Inventar, das auf eine spezifische Realität Bezug nimmt: häusliches und landwirtschaftliches Arbeitsleben, die damit verbundenen Sozialverhältnisse und Gewohnheiten, Einrichtungen, kurzum: Sachliche, soziale und sprachliche Lebensformen, die sich historisch unter besonderen Bedingungen entwickelt haben und mit denen sich die SprecherInnen dieser deutschen Varietät *identifizieren*.

Dieser Überblick zur Lemmaselektion zeigt, dass die kodifizierten systemintegrierten lexikalischen Spezifika der deutschen Standardsprache in Sprachinsellage historisch, wirtschaftlich, politisch, sozial und kulturell bedingt sind. Sicherlich wären auch andere Einträge „wörterbuchreif“ gewesen, doch den (Regional-)Experten sind Einschränkungen auferlegt worden. Allerdings kommen im Wörterbuch auch einige für die Vollzentren spezifischen Wörter vor, die es auch im Rumäniendeutschen gibt und die leider nicht als RUM ausgewiesen wurden.³² Vgl. z.B. *Aufbaustudiengang* (D; S. 55), *Inspektorat* (A CH; S. 349), *Gehschule* (A; S. 268), *Gelse* (A; S. 270), *Präfekt* (CH; S. 506), *Trolleybus* (CH; S. 507). Falls man bestimmte Wörter auch in Rumänien in der vermerkten Bedeutung verwendet, wäre es angebracht, dass man bei den Lemmata aus D, A, CH, LUX usw. auch RUM hinzufügt. Auch wenn viele Wörter bei Lăzărescu und Scheuringer (2007) gebucht sind, wäre der Wörterbuchbenutzer *dieses* Wörterbuchs dankbar gewesen, wenn in der Neuauflage auch die rumänien-spezifischen Lemmata, die vom österreichischen Deutsch geprägt sind, entsprechend markiert sein würden. Dieser Zeit- und Arbeitsaufwand hätte sich gelohnt und hätte auch den einen oder anderen Sprachinselfreund erfreut.

Die im VWB aufgenommenen *Rumänismen* entstammen einem umfangreichen (Zeitungs-)Korpus, wobei für die ausgewählten Lemmata auch Belege gefunden werden mussten, die dem aktuellen Sprachstand entsprechen.³³ Die Beispielsätze stammen ausnahmslos aus Preetexten. Es handelt sich vorwiegend um Preetexte aus der *Allgemeinen Deutschen Zeitung für Rumänien*, die Tageszeitung für die Angehörigen der deutschen Minderheit und deren Regionalbeilagen³⁴ wie auch aus der *Siebenbürgischen Zeitung*, die vom Verband der Siebenbürger Sachsen in Deutschland ab 1950 herausgegeben wird, da Angehörige der rumäniendeutschen Minderheit auch nach ihrer Auswanderung nach Deutschland untereinander diese Varietät noch verwenden. Auch die Sichtung verschiedener Aufsätze zu den als typisch rumäniendeutschen Wörtern erklärten Rumänismen erwies sich als notwendig. Belege aus der Belletristik

blieben unberücksichtigt.

Mit Blick auf die Viertel(s)zentren ist erfreulich, dass eine gute Beleglage, repräsentativ ausgewählt, das Rumäniendeutsche abdeckt im Vergleich zu den spärlichen Quellen anderer Viertel(s)zentren. Hinsichtlich der Viertel(s)zentren entfallen von den 17 Titelangaben jeweils sechs auf Rumänien, acht auf Namibia und drei auf die Mennonitensiedlungen.

Abschließend muss erwähnt werden, dass im Hinblick auf die (konsequente) Erfassung der Regionalspezifität die Kodifizierungspraxis der lexikalischen Varianten in den Viertel(s)zentren Rumänien, Namibia und Mexiko nicht mit der gleichen Sorgfalt erfolgt ist. Insgesamt kann eine sorgfältigere Bearbeitung der unter dem Kürzel RUM aufgenommenen Lemmata registriert werden. Dies zeigt sich zunächst im konsequenten Vorgehen bei der Lemmatisierung und bei der Einhaltung einzelner Artikelpositionen (Angaben zur Herkunft, zum Vorkommensbereich oder zur Markierungspraxis; vgl. hierzu auch Angaben wie „selten auch in der Form *Klettiten* oder *Klettiten* geschrieben“; S. 394). Darüber hinaus begegnen beim RUM-Wortschatzausschnitt mehrere Verweise auf (mindestens zwei) Zusammensetzungen (vgl. z.B. Lemma *Märzchen*; S. 464) im Schlussteil des Wörterbuchartikels. Bei vielen MENN- oder NAM-Lemmata bleiben etymologische Angaben oder Hinweise auf Komposita aus. Vgl. z.B. NAM *Kamp* (‘eingezäunte Fläche’), wo keine Herkunftsangabe erscheint (S. 364). Auch sind die zitierten RUM-Belege zutreffender als bei vielen MENN- oder NAM-Lemmata. Die Beispielsätze sind nützlich, da sie den Gebrauch der Stichwörter erklären und in ein Umfeld einordnen. Andererseits zeigen sie auch, weshalb eine bestimmte Bedeutungsangabe in der einen oder anderen Weise formuliert wurde. Bei manchen Beispielsätzen wie z.B. beim Stichwort *Veld* (NAM; ‘offenes, weites Land, Savanne’; S. 775) wäre anstatt „Blitz entzündet trockenes Veld“ (AZN 26.10.2010) eine aussagekräftigere Beispielangabe benutzerfreundlicher gewesen. Auch bei *Seida* (MENN; ‘Limonade’; S. 665) wäre ein anderer Beleg sicherlich angebrachter gewesen wie auch einige Zusammensetzungen zu diesem Lemma. Bei MENN *Komitee* (→ Magistrat A, Stadtamt → A, → Gemeindeamt A D, → Komunalverwaltung D, → Bürgermeisteramt D LIE, → Munizip MENN ‘Verwaltungsorgan einer mennonitischen → Kolonie; Gemeindeverwaltung’; S. 404) und MENN *Munizip* (‘lokaler Verwaltungsbezirk, Gemeinde’; S. 485) bzw. MENN *Ohm* (S. 512) werden keine Zusammensetzungen vermerkt.

3. Schlussbemerkungen und Ausblick

Die Neuauflage des Wörterbuchs belegt Varianten der deutschen Standardsprache, wobei dies bislang für keine andere Sprache — auch für die großen plurizentrischen Sprachen (z.B. Englisch, Französisch, Spanisch, Portugiesisch)³⁵ nicht — erfolgt ist, sodass mit der Aufnahme der Viertel(s)zentren eine lexikografische Lücke geschlossen werden kann.

Das empirisch erhobene Sprachmaterial verdeutlicht Eigenheiten auch des

in den jeweiligen Viertel(s)zentren gesprochenen Standarddeutschen, die der wissenschaftlichen Fachwelt und auch dem interessierten Nichtfachpublikum zugänglich gemacht werden müssen, da es hier einen großen Bedarf an noch zu erbringenden Forschungen gibt. Sowohl die Viertel(s)zentren wie auch ihre Standardvarianten sind in der Vergangenheit kaum in den wissenschaftlichen Fokus gerückt. Mit der Aufnahme standardsprachlicher Besonderheiten des Deutschen in Rumänien, in Namibia und bei den Mennoniten in Amerika und die Bereitstellung für eine eingehendere linguistische Forschung ist zu erwarten, dass künftig verstärkt Untersuchungen zu den Auffälligkeiten in den Viertel(s)zentren angestellt werden.³⁶

Als aktuelle Forschungsdesiderate sind vorwiegend (sozio-)linguistische Fragestellungen zu erwähnen, die auch andere deutschsprachige Minderheiten anvisieren sollten, um einerseits die Existenz weiterer Viertel(s)zentren zu dokumentieren, andererseits verschiedene Viertel(s)zentren miteinander zu vergleichen. In diesem Zusammenhang wäre auch danach zu fragen, wie die sinkenden Sprecherzahlen in den Viertel(s)zentren einzuschätzen sind und ob damit das Bestehen dieser Viertel(s)zentren gefährdet wäre. Wichtig sind auch Überlegungen zu sprachpolitischen Maßnahmen zum Erhalt und zur Förderung der deutschen Sprache in den jeweiligen Viertel(s)zentren wie auch die Ermittlung neuer Standardvarianten und Themenbereiche durch die Auswertung von Sprachkorpora aus Modelltexten, die Erfassung der geografischen Verteilung der Mennonitensiedlungen oder die Erforschung der Besonderheiten des Sprachgebrauchs weiterer religiöser Minderheiten deutscher Herkunft aus Nordamerika.³⁷

Deutsch am Rande und weit außerhalb des geschlossenen deutschen Sprachgebietes dokumentiert ein Varietätenspektrum, dem das VWB gebührende Beachtung verschafft. Hier werden Sprachzeugnisse deutschsprachiger Minderheiten verschiedener Regionen — darunter auch der *rumäniendeutschen Standardvarietät* — nicht nur adäquat erfasst, für künftige Generationen aufbewahrt und im öffentlichen Bewusstsein verankert. Auch der an historischen Sprachinseln oder an der sprachlichen Heterogenität interessierten Laienschaft, die kompetent Sprachvariation beurteilen möchte, bietet dieses Nachschlagewerk Wissenswertes über das gegenwärtige Deutsch in seiner eigenen standardsprachlichen Charakteristik, die in verschiedenen Ländern und Regionen auszumachen ist. So sind die Neuaufnahmen sowie die neu hinzugekommenen Sprachräume für das VWB ein zusätzlicher Gewinn und die Freude über Gefundenes — längst aus dem Gedächtnis Geschiedenes — groß. Auch damit hat das Autoren- und Expertenteam die im Vorwort der Erstauflage (2004: X) verkündete „weitere Verbesserung“ und „andauernde Aktualisierung“ als Zukunftsaufgabe verwirklicht.

Das Acknowledgement

Dieser Artikel entstand im Rahmen des von der Lucian-Bloga-Universität Sibiu/Hermannstadt geförderten Projekts LBUS-IRG-2018-04 (Laufzeit: 2018–2020).

Anmerkungen

1. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Herausgegeben von Ulrich Ammon, Hans Bickel und Jakob Ebner. Berlin: de Gruyter Mouton. 2004. LXXV, 954 S.
2. Die im Beitrag verwendete Terminologie im Zusammenhang mit den Voll-, Halb- und Viertel(s)zentren geht auf Ammon (1995) zurück.
3. Die Initiative und wissenschaftliche Grundlegung stammen von Ulrich Ammon. Vgl. hierzu das Vorwort zur Erstauflage (2004: VII-X). Dem Wörterbuchteil gehen daher Ausführungen zu den nationalen Voll- und Halbzentren des Deutschen voraus, eine Unterscheidung, die auf dem Konzept der plurizentrischen Sprache beruht.
4. Dazu Scheuringer (1996). Auf den Unterschied, der in der Forschung zwischen den Termini „plurizentrisch“ und „pluriareal“ gemacht wird, wird hier nicht näher eingegangen.
5. Zur Ermittlung der standardsprachlichen Variation wurde ein breites Korpus angelegt, das vorwiegend Periodika (über- und regionale Tages- und Wochenzeitungen, Monatsmagazine, Zeitschriften, Illustrierte, populäre Fachzeitschriften), Belletristik (Prosa, Kinder- und Jugendliteratur), Krimis und Trivialliteratur, populäre Sachtexte verschiedener Themenbereiche (Bildung/Erziehung, Wirtschaft, Gesundheit, Wohnen, Medien, Soziales, Natur, Kultur, Religion, Sport, Tourismus), Informations- und Werbebroschüren, Werbematerialien, Prospekte, Kalender, Kataloge, Formulare, Gesetzestexte, Audio- und Videoquellen und das Internet umfasst. Berücksichtigt wurden neben dem Grundwortschatz auch Sachspezifika, Bezeichnungen für bedeutende und typische Institutionen, geografische Namen und typische Vornamen, Abkürzungen und Kurzwörter sowie Redewendungen, Sprichwörter oder substantivierte Attribute, denen (Bedeutungs-)Erklärungen, Arealangaben und Belege folgen. Vgl. dazu die Ausführungen zur Auswahl der Stichwörter (2004: XI-XII).
6. Die Anordnung der Stichwörter (S. XII) erfolgt alphabetisch, wobei Schreibvarianten nicht gesondert aufgeführt werden. Bei nationsinternen Schreibvarianten wird die häufigere Form angegeben und die Zweitform im Kommentar erwähnt.
7. Bei den Lemmata werden u.a. folgende Abkürzungen als Verweise auf andere Varietäten verwendet: A = Österreich, CH = Schweiz, D = Deutschland.
8. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. Herausgegeben von Ulrich Ammon, Hans Bickel und Alexandra N. Lenz. Berlin/Boston: de Gruyter Mouton. 2016. LXXVIII, 916 S.
9. Eine von mir verfasste Rezension in englischer Sprache des VWB in seiner Neuauflage erscheint im Frühjahr 2018 im *Journal of Germanic Linguistics* 30: 88-96.
10. Dazu Ammon (1995) und den einleitenden Teil des Variantenwörterbuchs (2016: XI-LXXVIII).
11. Terminus nach Ammon (1995: 14). Zum Terminus „Rumäniendeutsch“ vgl. Lăzărescu (2013a: 369-389).
12. Zu den sprachlichen Besonderheiten der in Rumänien gesprochenen deutschen Sprache vgl. insbesondere die Arbeiten von Ioan Lăzărescu und in der älteren Literatur Kelp (z.B. 1982 und 1985). Zur deutschen Sprache in Rumänien aus variationslinguistischer Sicht vgl. auch den Sammelband von Lăzărescu, Scheuringer und Sprenger (2016) und die Publikations-

reihe des *Forschungszentrums Deutsch in Mittel-, Ost- und Südosteuropa* (FZ DiMOS) an der Universität Regensburg. Das Forschungszentrum widmet sich der Erforschung und Dokumentation der historischen und aktuellen Mehrsprachigkeitssituation in diesem Areal unter Einbeziehung der dortigen Nachbarsprachen des Deutschen.

13. Vgl. hierzu die Projekthomepage <http://www.variantenwoerterbuch.net/>; 12.01.2017.
14. Vgl. hierzu die Projekthomepage <http://www.variantenwoerterbuch.net/>; 12.01.2017.
15. Ein Wort wurde dann aufgenommen, wenn es nicht im gesamten deutschen Sprachgebiet vorkommt oder wenn es je nach Land oder Region unterschiedliche Bedeutungen trägt, unterschiedlich verwendet wird oder von unterschiedlichen Sprechergruppen unterschiedlich häufig verwendet wird. Nicht aufgenommen wurden Wörter, die sich nur in der Schreibung und der Aussprache von gemeindeutschen Wörtern unterscheiden.
16. Dazu Ammon (1995: 73-75).
17. Bei den Lemmata erscheinen Abkürzungen als Verweise auf andere Varietäten. Vgl. z.B.: *Bakkalaureat* das; -(e)s, -e: 1. A [...] 2. RUM; → Matura A CH, → Reifeprüfung A D, → Matur CH, → Maturität CH, → Abitur D ,Prüfung oder Schulabschluss zur Erlangung der Hochschulreife'[...]. Dazu: → Bakkalaureatsdiplom, Bakkalaureatskandidat(in), Bakkalaureatsprüfung (S. 82).
18. Wörter und Wendungen der Fach- und Verwaltungssprache, Dialekte, veraltetes oder selten gebrauchtes Wortmaterial und Umgangssprachliches wurden nicht erfasst. Ebenfalls nicht aufgenommen wurden Wörter, die sich nur in der Schreibung und der Aussprache von gemeindeutschen Wörtern unterscheiden.
19. Wenn auch die *Rumäno-Austriazismen* im VWB nicht aufgenommen werden, so verweist jedoch das Vorwort des VWB in seiner Neuauflage auf das Wörterbuch von Lăzărescu und Scheuringer (2007), das mit seinen rund 6.100 Einträgen zu den großen Austriazismen-Wörterbüchern gehört.
20. Für die deutsche Standardsprache in Rumänien fehlt ein Korpus. Überlegungen zur Erstellung eines Korpus lexikalischer Rumänismen und der Nutzung des Internet als Quelle für die Variationslinguistik vgl. Serbac (2017).
21. Einen Überblick über die Entstehung der deutschen Gemeinschaften in Rumänien bietet Bottesch (2008: 329-392).
22. Es handelt sich hierbei um staatlich subventionierte Schulen mit deutschsprachigen Klassenzügen, in denen der Unterricht teils oder gänzlich in deutscher Sprache erfolgt. Diese existieren in Rumänien vom Kindergarten über die Grundschule (Klassen 1–4) und das Gymnasium (Klassen 5–8) bis zum Lyzeum (Klassen 9–12). Näheres dazu bei Bottesch (2014).
23. Vgl. z.B. die deutschen Gymnasien in Temeswar, Hermannstadt, Kronstadt, Schäßburg oder Bukarest.
24. Die Rumäniendeutschen in Zahlen: 384.000 (1956), 360.000 (1977), 120.000 (1992). Bei der Volkszählung (2002) gehörten bei einer Landesbevölkerung von 21.700.000 Personen etwa 0,3 Prozent (60.000 Personen) der deutschen Minderheit an. Etwa 42.000 Personen haben Deutsch als Erstsprache angegeben. Bei der letzten Volkszählung (2011) haben sich 36.000 rumänische Staatsbürger (0,2 Prozent) als Deutsche erklärt und 27.000 Personen Deutsch als ihre Muttersprache angegeben. Vgl. Bottesch (2014).
25. Dazu ausführlicher Ammon (2015: 341-349).
26. Bottesch (2008: 351).

27. Vgl. die deutschsprachigen Zeitungen aus dem Quellenverzeichnis des VWB und die deutschsprachigen Sender des staatlichen rumänischen Rundfunks und Fernsehens. 2009 wurde die aus Rumänien stammende deutschsprachige Schriftstellerin Herta Müller mit den Nobelpreis für Literatur ausgezeichnet.
28. Zu den österreichisch-rumäniendeutschen lexikalischen Gemeinsamkeiten vgl. das von Lăzărescu und Scheuringer 2007 herausgegebene Wörterbuch.
29. Zum Schuldeutsch rumänischer Schüler an deutschen Schulen und zum veränderten Status des Rumäniendeutschen in den letzten Jahren vgl. Lăzărescu (2013b: 171-183).
30. Vgl. hierzu Lăzărescu (2017). Umfangreichere Bedeutungsangaben werden hier teilweise gekürzt wiedergegeben.
31. Wobei viele in den herkömmlichen zweisprachigen Wörterbüchern Deutsch-Rumänisch nicht kodifiziert sind.
32. Davon sind etliche als Grenzfälle des Standards einzustufen.
33. Zu den Auswahlkriterien der Lemmata und zur Arbeitsmethode am Variantenwörterbuch vgl. Lăzărescu (2017).
34. Die deutschsprachige Tageszeitung [*Allgemeine Deutsche Zeitung für Rumänien* (ADZ); erscheint ab 1945 in Bukarest; vor 1993: *Neuer Weg*; vorwiegend für Einheimische und nicht nur für Expats] mit zwei Regionalbeilagen (*Banater Zeitung* und *Karpatenrundschau*) wird seit 2005 vom Minderheitenverband, dem Demokratischen Forum der Deutschen in Rumänien herausgegeben, der auch die *Hermannstädter Zeitung* finanziell unterstützt. Die Wochenzeitung *Hermannstädter Zeitung* in Hermannstadt (vor 1990: *Die Woche*) ist die einzige deutschsprachige eigenständige Zeitung in Rumänien. Herausgeber ist die Stiftung *Hermannstädter Zeitung*, die durch Vermittlung des Demokratischen Forums der Deutschen in Rumänien 50 Prozent der Kosten mit dem Departement für interethnische Beziehungen abrechnet. Die restlichen 50 Prozent kommen aus dem Freiverkauf, aus Spenden und Anzeigen.
35. Der Grad der Plurizentrität ist bei den in Europa existierenden plurizentrischen Sprachen unterschiedlich stark ausgeprägt. Eine Gesamtdarstellung der plurizentrischen Sprachen Europas findet sich in Muhr (2003). Zum Deutschen aus plurizentrischer Sicht vgl. Schmidlin (2011).
36. Es gilt, die Heterogenität des Deutschen zu beschreiben und zu erklären, wobei u.a. Fragen der Variationslinguistik, der Sprachkontakt- und Mehrsprachigkeitsforschung oder Soziolinguistik zu beantworten sind.
37. Vgl. hierzu u.a. Schneider-Wiejowski und Ammon (2013: 113-122).

Literatur

- Ammon, Ulrich.** 1995. *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten.* Berlin/New York: Walter de Gruyter.
- Ammon, Ulrich.** 2015. *Die Stellung der deutschen Sprache in der Welt.* 2. Aufl. Berlin/München/Boston: Walter de Gruyter.
- Ammon, Ulrich, Hans Bickel und Jakob Ebner (Hrsg.).** 2004. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Berlin/Boston: de Gruyter Mouton.

- Ammon, Ulrich, Hans Bickel und Alexandra N. Lenz (Hrsg.).** 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2. Aufl. Berlin/Boston: de Gruyter Mouton.
- Bottesch, Johanna.** 2008. Rumänien. Eichinger, Ludwig M., Albrecht Plewnia und Claudia M. Riehl (Hrsg.). 2008. *Handbuch der deutschen Sprachminderheiten in Mittel- und Osteuropa*: 329-392. Tübingen: Gunter Narr.
- Bottesch, Martin.** 2014. Festvortrag. Hat die deutsche Sprache in Siebenbürgen eine Chance? *Symposium der Deutsch-Rumänischen Akademie*, 3.-4. Oktober 2014. Thema: *Die Sprache: Aspekte des Sprachbegriffes aus der Perspektive unterschiedlicher Disziplinen*: 3-13. Sibiu/Hermannstadt: Global Media.
- Kelp, Helmut.** 1982. Lexikalische Besonderheiten unserer deutschen Schriftsprache. 50 Zeitungsartikel. *Neuer Weg* (Bukarest), 30.01.1982-8.12.1984.
- Kelp, Helmut Martin.** 1985. *Die lexikalischen Besonderheiten der deutschen Schriftsprache in Rumänien*. Heidelberg: Quick.
- Lăzărescu, Ioan.** 2013a. Rumäniendeutsch — eine eigenständige, jedoch besondere Varietät der deutschen Sprache. Schneider-Wiejowski, Karina, Birte Kellermeier-Rehbein und Jakob Haselhuber (Hrsg.). 2013. *Vielfalt, Variation und Stellung der deutschen Sprache*: 369-389. Berlin: de Gruyter Mouton.
- Lăzărescu, Ioan.** 2013b. Heutiges „Schuldeutsch“ in Rumänien, oder wie sich Austriazismen, Austro-Rumänismen, Rumänismen und „Kiritzismen“ zu einem einzigartigen Mosaik fügen. Predoiu, Graziella und Beate Petra Kory (Hrsg.). 2013. *Streifzüge durch Literatur und Sprache. Festschrift für Roxana Nubert*: 171-183. Temeswar/Timișoara: Mirton.
- Lăzărescu, Ioan.** 2017. Wie kommen die Rumänismen in die Neuauflage des Variantenwörterbuchs? Zu den Auswahlkriterien der Lemmata und zur Arbeitsmethode am Variantenwörterbuch-NEU. Mauerer, Christoph (Hrsg.). 2017. *Mehrsprachigkeit in Mittel-, Ost- und Südosteuropa. Gewachsene historische Vielfalt oder belastendes Erbe der Vergangenheit. Beiträge zur 1. Jahrestagung des Forschungszentrums Deutsch in Mittel-, Ost- und Südosteuropa, Regensburg, 2.-4. Oktober 2014*: 341-358. *Forschungen zur deutschen Sprache in Mittel-, Ost- und Südosteuropa FzDiMOS*, 4. Regensburg: Friedrich Pustet.
- Lăzărescu, Ioan und Hermann Scheuringer.** 2007. *Limba germană din Austria. Un dicționar German-Român. Österreichisches Deutsch. Ein deutsch-rumänisches Wörterbuch*. Passau: Karl Stutz/Bukarest: Niculescu.
- Lăzărescu, Ioan, Hermann Scheuringer und Max Sprenginger (Hrsg.).** 2016. *Stabilität, Variation und Kontinuität. Beiträge zur deutschen Sprache in Rumänien aus variationslinguistischer Sicht*. *Forschungen zur deutschen Sprache in Mittel-, Ost- und Südosteuropa FzDiMOS*, 2. Regensburg: Friedrich Pustet.
- Muhr, Rudolf.** 2003. Die plurizentrischen Sprachen Europas — Ein Überblick. Gugenberger, Eva und Mechthild Blumberg (Hrsg.). 2003. *Vielsprachiges Europa. Zur Situation der regionalen Sprachen von der Iberischen Halbinsel bis zum Kaukasus*: 191-233. *Österreichisches Deutsch — Sprache der Gegenwart* 2. Frankfurt [etc.]: Peter Lang.
- Sava, Doris.** 2018. Rezension. Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. Herausgegeben von Ulrich Ammon, Hans

Bickel und Alexandra N. Lenz. Berlin/Boston: de Gruyter Mouton. 2016. *Journal of Germanic Linguistics* 30(1): 88-96.

- Serbac, Patricia.** 2017. Rumänismen und ihre Quellen für die Korpuserstellung des Rumäniendeutschen. Mauerer, Christoph (Hrsg.). 2017. *Mehrsprachigkeit in Mittel-, Ost- und Südosteuropa. Gewachsene historische Vielfalt oder belastendes Erbe der Vergangenheit. Beiträge zur 1. Jahrestagung des Forschungszentrums Deutsch in Mittel-, Ost- und Südosteuropa, Regensburg, 2.–4. Oktober 2014*: 330-340. *Forschungen zur deutschen Sprache in Mittel-, Ost- und Südosteuropa FzDiMOS*, 4. Regensburg: Friedrich Pustet.
- Scheuringer, Hermann.** 1996. Das Deutsche als pluriareale Sprache: Ein Beitrag gegen staatlich begrenzte Horizonte in der Diskussion um die deutsche Sprache in Österreich. *Unterrichtspraxis/Teaching German* 29: 147-153.
- Schmidlin, Regula.** 2011. *Die Vielfalt des Deutschen: Standard und Variation. Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. *Studia Linguistica Germanica* 106. Berlin: de Gruyter.
- Schneider-Wiejowski, Karina und Ulrich Ammon.** 2013. Zu den Viertelszentren der deutschen Sprache. Sava, Doris und Hermann Scheuringer (Hrsg.). 2013. *Im Dienste des Wortes. Lexikologische und lexikografische Streifzüge. Festschrift für Ioan Lăzărescu*: 113-122. *Forschungen zur deutschen Sprache in Mittel-, Ost- und Südosteuropa FzDiMOS*, 3. Passau: Karl Stutz.

Pedro A. Fuertes-Olivera. *The Routledge Handbook of Lexicography*. 2018, 810 pp. ISBN: 978-1-138-94160-1. London/New York: Routledge. Price £165.00.

Theoretically and practically, lexicography has "come to a crossroads" (Bergenholtz, Nielsen and Tarp 2009: 8), and is experiencing "a Cambrian explosion driven by the coming of age of the Internet" (Fuertes-Olivera 2018: 37). However, so far no work has been dedicated to offering a comprehensive overview of lexicography in the Internet era. The appearance of *The Routledge Handbook of Lexicography* is timely. It serves as "a guide to what are the most significant contours in the lexicographical world", and offers "a series of possible developments that might be influencing the near future of the field" (ibid).

This book, with its six parts, an introduction and an index, offers "a balanced view of the main approaches to lexicography in general and to some of its specific aspects" (Fuertes-Olivera 2018: 38). The 47 chapters follow a unified format which consists of an introduction, a historical review of the specific topic, core issues, an indication of future developments, a conclusion, related topics, further reading and bibliographical references. Contributors to this book are scholars with theoretical or practical lexicographic background and insight into lexicography in the Internet era. Readers will find each contribution practical and inspiring, rather than abstract and tedious.

Among the recently published handbooks such as Jackson (2013), Durkin (2016) and Hanks and De Schryver (2016), this book is a comprehensive and most up-to-date contribution to lexicography in the Internet era. It differs from the above-mentioned handbooks in three aspects. Firstly, it focuses on lexicography in the Internet era, thus facilitating an overall understanding of the latest theoretical and practical developments in lexicography. Secondly, most of the contributions adopt function theory as a starting point, enabling the reader to form a better understanding of the theory and its application. Thirdly, "Further Reading", with a brief introduction to the related works, and a self-contained reference list, conveniently guides the interested reader to further exploration.

The introduction gives the reader a panoramic view of the book by summarizing the general idea of each part and each chapter, and organizing the chapters in an organic way. The readers can select what they are interested in for further reading in a convenient and time-saving way.

By establishing the position of lexicography as an independent discipline Part I could be seen as the basis for the whole book. Adopting the function theory of lexicography, the five chapters defend this position using the topics dictionary management, access structure, meaning explanation, and dictionary criticism. The status of lexicography as an independent science is defended in Chapter 1, and its five research areas are established accordingly, namely research into (1) the information-search process; (2) dictionary compilation; (3) dictionary form; (4) usefulness of dictionaries; and (5) history of lexicography, which later serve as the topics and issues for discussion in this book. Chapter 2 focuses on dictionary management which is considered to be a peripheral topic

in many lexicographic works. In comparison with other ideas on dictionary management, the author, Henning Bergenholtz, defends the necessity of appointing a professional lexicographer as dictionary project manager by referring to two dictionary projects. This provides quite convincing examples. Chapter 3 deals with the topic of access structure in a macroscopic way by introducing the interactive relations between access structure and other dictionary structures. It can be regarded as an introduction to further discussions on access structure in Chapters 41, 42, 43, etc. Definition is always regarded as the most important aspect in a dictionary. In Chapter 4, meaning explanation is proposed to substitute for definition, and the author, Heidi Agerbo, provides examples of meaning explanation from a functional perspective and suggests considering USER+SITUATION+LANGUAGE CONSTRUCTION as the determiner for selecting lexicographic data. Chapter 5 is innovative in providing a format for dictionary criticism by defining the object and purpose of criticism, proposing three approaches to the task, and examining the qualitative requirement of scientific criticism.

Part II discusses the relation between lexicography and other disciplines, ranging from the often discussed applied linguistics and terminology, to language policy and culture, and the comparatively newly emergent corpus linguistics, natural language processing, information science and domain ontologies. Practically and theoretically, lexicography is interdisciplinary and different stages of dictionary compilation require involvement from different domains (Chapter 6). To be more specific, the compilation of monolingual learners' dictionaries reflects the strong impact of applied linguistics in lexicography, as demonstrated in Chapter 7. Lexicography and terminology are fuzzy-boundary if their practitioners, tools and techniques, and data are taken into consideration (Chapter 9). A dictionary is also the product of language policy which in turn influences language policy, and it will remain a pronounced issue for countries to pursue national identity, as is the case of the African countries mentioned in Chapter 10. A dictionary is also the reflection of culture information, which should be included, especially in learners' dictionaries (Chapter 11). Practically, local and online corpus query tools can be utilized for lexicographic purposes (Chapter 8). Natural Language Processing techniques (NLP) and ontologies can also be used in lexicographic processes such as NLP techniques for collocation extraction, named entity recognition, word sense disambiguation, etc., and in the ontology editor "Protégé" and search engine "SWOOGLE" (Chapter 12, 14). The interdisciplinary nature of lexicography determines that its development will not only benefit from other disciplines, but will also contribute to the development of other disciplines, as is stated in Chapter 13 that information science and lexicography can learn from each other at the level of theory and practice.

Part III presents different types of dictionaries. Contributors to this part deal with different types of dictionaries within the function theory framework and pay close attention to electronic forms. In Chapter 15, Tarp re-defines the dictionary concept according to criteria of form, content and purpose, and pro-

poses four dictionary categories based on communicative, cognitive, operative and interpretive purposes. The dictionaries discussed later in part III fall within this framework, with dictionaries for text reception (Chapter 16), text production (Chapter 17) and translation (Chapter 18) belonging to the dictionary category based on communicative functions. Dictionaries to assist teaching and learning (Chapter 19) belong to the dictionary category based on cognitive functions; and specialized dictionaries (Chapter 20) belong to the dictionary category based on operative functions. Chapter 16 addresses a number of issues in text reception and some lexical categories which pose as challenges to text reception. Chapter 17 adopts a different approach by analyzing two dictionaries for text production and introducing a lexicographic project consisting of different dictionaries to illustrate how to improve current dictionaries for text production. Chapter 18 points out LSP e-lexicography, tailored monofunctional e-dictionaries and a comprehensive translation-oriented platform as the future for translation dictionaries. Chapter 19 gives a brief history of monolingual learners' dictionaries (MLDs) and discusses some core issues in MLDs, such as defining vocabularies, examples, grammatical information, collocations, etc., and special types of MLDs, e.g. electronic ones, bilingualized ones and those for specialized purposes. The author of this chapter, Reinhard Heuberger, points out that the future of MLDs should be electronic learners' dictionaries characterized by customization and user input. Chapter 20 focuses on the defining criteria for specialized dictionaries and examines some issues in the macro- and microstructure. Anne Condamines approaches the topic of terminological knowledge bases (TKBs) from a linguistic point of view, and gives detailed information of the tool-assisted linguistic methods for building TKBs. However, the aim of knowledge engineering has now evolved in building ontologies, as is discussed in more detail in Chapter 14.

Dictionary work mentioned in Part IV is characterized as innovative, whether it is the revision of a long existing monolingual learners' dictionary or the compilation of a new dictionary. Contributors to this part take different approaches to demonstrate the innovative characteristics of the dictionary or dictionaries studied. Chapter 22 examines the main features of online MLDs and points out that future MLDs will be innovative in incorporating material such as video clips. Teachers need to be aware of the advantages and disadvantages of current online MLDs in teaching dictionary use. Students could, for example, be allowed in class to search for a word and to compare entries in different dictionaries. Chapter 23 reports on the revision of a historical Canadian dictionary. The work is innovative not only in its guiding principles but also in its entry structure, applied typology of Canadian English and especially in its use of frequency charts. The *Online Dictionary of New Zealand Sign Language* is innovative in itself in addressing the deaf community's claim for identity recognition but also in utilising the advantages of linguistic research and the digital medium. It also features the use of video clips, corpus-based contents, users' needs studies and interactive user interfaces. The *Alicante Dictionaries*,

the *Oenolex Wine Dictionary* and the *Accounting Dictionaries* are specialized dictionaries adopting a functional theory approach. The author of Chapter 26, Jose Mateo, reports on issues concerning the compilation of the *Alicante Dictionaries*: the basic principles of relevance, clarity and economy, methodology, and data sources. The user-oriented approach requires the inclusion of the most relevant terms in the dictionaries, instead of the traditionally pursued "the more the better". The *Oenolex Wine Dictionary* is distinguished from others by its genuine purpose and by some elements in the construction process, including data generation and acquisition, the use of written and oral corpora, semi-automatic extraction of examples, improved search functionality and cooperative work and interdisciplinary management. Many of these innovative features are shared by the *Accounting Dictionaries* as mentioned in Chapter 28. The authors of this chapter, Pedro A. Fuertes-Olivera and Marta Niño Amo, propose online dictionaries to be regarded as services and not as products. Chapter 24 gives a brief introduction to *FrameNet* database which is innovative in its theoretical basis, different ways to access data, crowdsourcing to generate resources, etc. *Wordnik*, a bottom-up collaborative lexicographic work, features an innovative business model, data-mining and machine-learning techniques and a different technical system.

Part V, *World Languages, Lexicography and the Internet*, provides a panoramic view of world lexicography. Using the criteria of "world language", "dominant and appreciated lexicographic tradition", and "less-resourced languages" (Fuertes-Olivera 2018: 49), eleven languages, namely African, Arabic, Chinese, English, French, German, Hindi, Indonesian, Portuguese, Russian and Spanish, are selected to demonstrate the current situation of world lexicography and to indicate its future development in the Internet era. Although all the contributors to this part give a brief introduction of their lexicographical histories, they approach the core issue from different perspectives. Amongst the eleven languages, English and German lexicography are listed as the leaders in the innovation of world lexicography, e.g. in the use of corpora, the improvement of processing tools, integration of information science in dictionary compilation, etc. Chapter 33 focuses on the use of corpora and the improvement of processing tools in English online lexicography. After discussing general issues concerning online dictionaries like typology, search features, presentation, the use of multimedia, accessibility and customization, the author, Howard Jackson, explores topics for future development, such as the user center, adaptability, hybridization and collaborative lexicography. Besides paying attention to specific issues, Chapter 35 takes a more macroscopic approach. After a review of the different dictionary types, aspects such as user studies, new linguistic data, the exploiting of the Internet, etc. are discussed. The author, Petra Storzjohann, points out the necessity of interaction between lexicography, corpora and information science, the integration of linguistic theory and lexicographic practice, and the raising of dictionary awareness. Amongst the other languages, Arabic lexicography made significant progress in modern development.

Despite uneven development in Arabic lexicography amongst language varieties, it has taken advantage of new technologies, methods and standards in saving the past and developing the new. This example is worth learning from, especially for those who have lagged behind in recent development, despite poor or rich lexicographic traditions. Danie Prinsloo et al. attribute the underdevelopment of African lexicography to the lack of dictionary culture and modern technology. The focus of this chapter is on examining the quality of online dictionaries for African languages, and the methods for improving them. Chapter 32 gives a brief introduction to the history of Chinese lexicography. The authors, however, overlook the recent development of electronic dictionaries by reasoning that the Internet is underused in Chinese lexicography and that both lexicographic theory and practice in China have followed a very different path from other traditions, especially Western ones. To change the status of Hindi lexicography as being underrepresented, the authors of Chapter 36 propose the raising of lexicographic awareness of both scholars and students, global vision of development in lexicography and information science, and to make the development of language policy and political planning a national strategy. Although Indonesian lexicography may be underrepresented, the author is quite optimistic about the future with joint efforts from the public, individual lexicographers and institutions, and about development in theories on corpora, dictionary typology, search capabilities, presentation and access, and the use of multimedia.

For other languages such as French, Portuguese, Russian and Spanish, the contributors approach from slightly different perspectives. Chapter 34 gives an overview of existing French electronic dictionaries and concludes that most of them are just digital versions of printed lexicographical reference works and calls for the improvement of a retrieval system. A history of Portuguese lexicography demonstrates that a dictionary is a cultural object. Besides issues mentioned by other contributors, the author, Teresa Lino, emphasizes the design and compilation of specialized dictionaries. In Chapters 39 and 40, in reviewing dictionaries, some innovations of global tendencies are offered e.g. Russian lexicographers focus on users' needs, the application of new technologies, the development of theoretical lexicography, etc., while their Spanish counterparts value lexical connectivity, personalization and integration.

Part VI, *Looking to the Future: Lexicography in the Internet Era*, focuses on specific issues on electronic lexicography, namely the use of the Web in the lexicographic process, information retrieval, usage research methods, user participation, dictionary portals and the international directory of lexicography. In Chapter 41, Anna Dziemiánko lays the foundation for the discussion in this part. Based on De Schryver's definition and Tarp's typology of Internet dictionaries, the author reviews and comments on existing and future electronic dictionaries, stating "access" and "quality and usefulness" as core issues, pointing out that the future development of electronic lexicography should include the dictionary as digital assistant, automatic lexicographic compilation,

integration of corpora with dictionaries, usefulness research, user studies, more advanced study methods, etc. Chapter 42 explores the development of dictionaries as lexicographic tools in terms of "user", "data" and "access". Chapter 43 introduces some electronic data sources for lexicographers, namely Sketch Engine, Google NGrams Viewer and WordNets, and provides examples of their applications in accomplishing lexicographic tasks with defining words of different parts of speech. Chapter 44 demonstrates how to conduct empirical usage research with specific methods: questionnaires, eye tracking and log files. Chapter 45 illustrates three major types of user participation methods. After examining the types and functions of dictionary portals, the authors of Chapter 46 provide a list of 37 portals. They investigate the lexicographical features and propose the improvement of dictionary portals by augmented search and search in context. The last chapter of this handbook proposes the necessity of having an international directory of lexicography as a source of information on lexicographers, publishers, conferences, elements of the production process, publication information, etc.

In conclusion, some features are worth mentioning.

Firstly, this handbook distinguishes itself by covering a wide variety of lexicographical topics in the Internet era with each chapter following a unified format. In this way, the intended readers, especially inexperienced lexicographers, will have a full understanding of the topic: its history, present practice and future development. In spite of different backgrounds and nationalities, contributors to this book make their contributions accessible in plain and simple language. Besides, topics shared within different chapters are cross-referenced. It is convenient, especially for the electronic edition, which is just a click away. No doubt, with its comprehensiveness, convenience and reader-friendliness, it will qualify as a guidebook for inexperienced lexicographers to have a general and most up-to-date knowledge of lexicography in the digital age.

Secondly, this informative book is well-organized, from the most general to the most specific, guiding the readers to probe into more complex and concrete issues. The readers are guided gradually from general issues such as the nature of the discipline (Parts I and II), different types of dictionaries (Parts III and IV) and dictionaries in different countries (Part V), to specific issues in lexicography (Part VI) to gain an understanding of modern lexicography in both a detailed and comprehensive way.

Thirdly, another merit is its employment of ample examples in discussion, even if it concerns the practical application of theory or the introduction to cutting-edge technology. Adopting a functional theory approach, practical applications of this lexicographic theory are presented with examples of the *Alicante Dictionaries* (Chapter 26), the *Oenolex Wine Dictionary* (Chapter 27), the *Accounting Dictionaries* (Chapter 28), etc. It proves the feasibility of the function theory, but also serves as a good example of the integration of theory and practice. A wide range of resources is also provided for further research, learning and teaching. For instance, Sketch Engine, Google NGrams Viewer

and WordNets are treated as examples in their use in accomplishing lexicographic tasks (Chapter 43). Corpus tools for lexicography such as ANTConc and Corpus Query Processor (CQP) (Chapter 8) and lexical databases such as DANTE, OWID, Pralex, Cornetto and Aralex (Chapter 12) are also introduced. No book of this kind has ever offered so many up-to-date and cutting-edge resources.

Finally, this handbook also serves as a guide for future research by providing good examples of research topics and methodologies. In addition, the format in which research articles are presented is worth being copied/noticed by the readers, especially inexperienced lexicographers. Besides, some contributors also illustrate how to conduct research with specific methods. For example, Chapter 44 demonstrates how to conduct empirical usage research with questionnaires, eye-tracking and log files. Chapter 45 introduces user participation as a new field for lexicographic research.

It is impossible for such a book to cover the detail of lexicography, but as an introductory work, it would be more comprehensive if it could go beyond the functional approach and also cover other approaches such as the communicative and cognitive approaches, especially when dealing with some general issues. Even though such information could be offered in the historical review part of a specific topic, contributors can prefer to emphasize certain aspects. For example, when dealing with definition (Chapter 4), Heidi Agerbo just mentions the conventional practice of defining according to "necessary and sufficient conditions", omitting some important approaches like defining semantic prototypes.

Some gap exists when contributors take a specific perspective in his/her article, which in some cases gives the wrong impression. For instance, when introducing Chinese lexicography (Chapter 32), Heming Yong and Jing Peng give a panoramic view of its history in the past three millennia, omitting the recent development of electronic dictionaries. Readers will find it a pity if they want to know more about the latest development in China.

To summarize, this handbook is a valuable addition to existing books of this kind. It is a very practical introduction to lexicography. It will be useful not only for students and scholars of lexicography as intended by the editor, but also for anyone interested in this topic.

Scholars have different ideas on the substitution of a paper dictionary with an online-only dictionary, however it is quite clear that the online dictionary already has a great influence on lexicography. We should rather prepare ourselves for more changes in the Internet. Therefore, this is a valuable book worth consulting.

References

- Bergenholtz, H., S. Nielsen and S. Tarp (Eds.). 2009. *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern/New York: Peter Lang.

Durkin, P. (Ed.). 2016. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Hanks, P. and G.-M. de Schryver (Eds.). 2016. *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer.

Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*. London/New York: Bloomsbury Academic.

Dai Lingzhen
College of Foreign Languages and Cultures
Xiamen University
Xiamen
P.R. China
(vivian0915@163.com)

María José Domínguez Vázquez, Fabio Mollica and Martina Nied Curcio (Eds.). *Zweispachige Lexikographie zwischen Translation und Didaktik*. Lexicographica. Series Maior 147. 2014, vi + 334 pp. ISBN 978-3-11-036973-1. e-ISBN 978-3-11-036663-1. Berlin/Boston: Walter de Gruyter. Price: € 109.95.

This volume contains articles and reports on bilingual dictionaries and their use with a view to foreign language learning, translation teaching and dictionary use. The authors present different theoretical and practical perspectives, with different focal points, thus giving an overview of the current status of and contemporary trends in contrastive lexicography with regard to the learning of foreign languages, the teaching of translation and the use of bilingual dictionaries for these purposes. The volume offers the latest insights into online lexicography, new trends, as well as suggestions for new research.

With the exception of two, all the articles are in German. The introductory article, written by the three editors, serves to set the scene for the entire volume: bilingual dictionaries are once again gaining in importance, because they are crucial in the learning of foreign languages and translation teaching. Looking for the most adequate equivalent, learners do not always have the knowledge and skills to perform the correct user actions. As Schafroth (p. 83) notes, students without training in dictionary use often select the first available translation equivalent, without taking into account the context. This problem reminds me of Jonathan Safran Foer's novel, *Everything is Illuminated* (2002), where the character of the Ukrainian tour guide/translator always selects the contextually most absurd and incorrect equivalents when he speaks his self-taught English. Learners seem to ignore the available metalinguistic remarks — and even the grammatical information — and fail to orientate themselves towards the structures of the dictionary articles. More and more frequently, learners use online dictionaries and glossaries, which often do not even include such metalinguistic information, exacerbating the problem. Many learners do not acquaint themselves with the user's guidelines. For all these reasons, it is imperative to once again look at the didactics of dictionary use, and incorporate its principles into the didactics of foreign language learning and translation teaching.

Only very few empirical studies are available within this field. In order to deal with this hiatus, this volume concentrates on three important aspects. Part I deals with "Valency, Constructions and Collocations in Bilingual Lexicography", to establish better links between contrastive linguistics and lexicographical practice. Part II is titled "Dictionaries and their Users". Contributors to this section look at bilingual dictionaries as learner's dictionaries, presenting examples of dictionary projects and offering suggestions for improved learning. Part III contains reports on several lexicographical projects, which are planned to offer more opportunities for learners.

Zsuzsanna Fábíán's contribution deals with the description of the three word classes (verb, adjective and noun) in general bilingual dictionaries

between Italian and Hungarian, and bilingual valency dictionaries with Italian and German as language pair. Fábíán points out that as yet no Hungarian valency dictionary in the strict sense of the word has been published. After a short introduction of three comprehensive Italian–Hungarian general dictionaries and three Italian–German valency dictionaries (treating verbs, adjectives and nouns), she focuses on the analyses of the verb *fidare* (=to trust), the adjective *abile* (=skillful) and the noun *condanna* (=condemnation). By taking a look at the methods used by the authors of valency dictionaries, Fábíán makes recommendations for an adequate and more user-friendly presentation of valency in general Italian–Hungarian dictionaries. For example, she recommends that lexicographers should include sentence-like structures in the examples; semantic valency should be presented in a more comprehensive and more accurate way, to avoid confusing learners. Lexicographers of bilingual dictionaries should take note of what has been done in valency research. Fábíán concludes her contribution with examples of what she considers good examples of valency in a potential Italian–Hungarian bilingual dictionary for learners, using *fidare*, *abile* and *condanna*.

In her contribution, Maria Teresa Bianco discusses the German verb *werden* (=to become) and its synonyms in Italian, and how different grammar books assign this verb to different verb classes. She asserts that this verb is only very seldom described as a main verb in textbooks — usually it has the status of an auxiliary verb. Moreover, it is not always clear whether the verb *werden* is considered a main verb or an auxiliary verb. Bianco lists examples from several monolingual German dictionaries and bilingual Italian–German dictionaries which may or may not have adequate information on the usage of the verb *werden*, and then poses some questions, such as whether *werden* is monovalent, and if so, whether it is an auxiliary verb or a main verb; and whether it is only used in fixed expressions. She also asks what equivalents are available in Italian in case of a monovalent verb *werden*, and what a user-friendly entry should look like in a dictionary. According to Bianco, the *Valenzwörterbuch Deutscher Verben* (=VALBU) and its electronic version (=E-VALBU) could serve as examples. These publications are based on research into the German corpus; they list many meanings of the verb *werden* and give ample usage examples.

Klaus Fischer deals with the usefulness of presenting valency and information on the construction of phrases in learner's dictionaries for second language learning, in order to establish how helpful existing dictionaries are. He maintains that valency dictionaries often define their target audience as linguists, grammarians, lexicographers, lecturers and authors of text books. Some of them state that they are also meant to be used by advanced foreign learners when they need help in the construction of phrases. But it seems that there are no resources available to foreign learners with little or no linguistic knowledge. Furthermore, almost all the bilingual valency dictionaries that Fischer took into account were conceptualised from the perspective of German valency, and not from the perspective of the other language in the pair.¹ This, of course, creates

problems for foreign learners, who proceed from the point of view of their own language. Fischer concludes his contribution with presenting a model for an English learner's dictionary of German, based on valency principles and a didactic selection of valency information, which could also be used by learners who do not have extensive linguistic backgrounds. This includes the simple presentation of example sentences and narrative comments.

Elmar Schafroth presents options for the presentation of idiomatic expressions by using a German–French online dictionary as an example. Linking with Goldberg (1995; 2006) and Croft's (2001) grammar of construction and especially Fillmore's (1982) frame semantics, he develops a model of phrase-frames, aiming to describe idiomatic expressions from a holistic point of view. Schafroth uses the example of the French expression *chercher midi à quatorze heures* (=to complicate things needlessly; to seek a knot in a bulrush) and describes not only its syntactic and semantic-pragmatic aspects, but also morphological, prosodic and discursive aspects. He suggests "phrase templates" by means of which lexicographers could present adequate information on the meanings of idiomatic expressions to foreign learners. According to him, the ideal dictionary would be electronic, and would have two monodirectional sections — one with French expressions and one with German — aimed at learners on both sides of the language pair. The descriptions in the French section would mainly be in German, taking German main meanings into account, and including translation possibilities. The German section would be the other way round. Each section aims to help in reception and in production. The phrase-frames could be linked to others of the same type, or with the same or similar meanings. There could be pop-up windows with additional information.

Zita Hollós introduces the KOLLEX Project, which she describes as a bilingual, polyaccessive and polyfunctional syntagmatic learner's dictionary. It is in the first instance production-oriented and based on corpora and data banks of bilingual collocation lexica in German and Hungarian. Its main target group is students studying German at university level, as well as German teachers. The latter group will benefit from using KOLLEX when they are grading their students' assignments. KOLLEX is a combination of a collocation dictionary and a valency dictionary, integrating the didactics of foreign languages, semantics, corpus linguistics and syntax/morphosyntax. Hollós illustrates her exposition with several examples from KOLLEX.

Dirk Siepmann discusses the EMOLEX project, which deals with fields of emotion in French, German and English. He presents a corpus-based analysis of semantic differences between German, English and French collocations of emotion nouns, aiming to determine the translatability of collocations and to possibly close the gaps between the inter-language differences in collocations. The EMOLEX project works with a classification of eight classes of emotion nouns, based on their collocational and colligational behaviour. Some of the problems he discusses have to do with concepts which are unfamiliar in a particular society, but are freely used in another society, and he calls these "collo-

cational gaps". Siepmann states that he found "few significant differences in the distribution of categories" across the languages he investigated, and that there is "comparatively sparse evidence of collocational gaps or interlingual difference" (p. 139). Nevertheless, the examples he discusses point at interesting differences between the three cultures, despite their closeness to each other. According to Siepmann (p.151), a study of Malayan and English emotion nouns denoting the concept "surprise" revealed "considerable divergences" — which makes his investigation fruitful for language pairs with more divergent cultures.

In the second section of the volume, which is titled "Dictionaries and their Users", Monika Bielińska contributes an article about bilingual learner's dictionaries which can support the learning process, in the sense that these dictionaries do not only give support for lexical problems, but they also support learners in grammatical and phraseological matters. She discusses the use of examples and fixed expressions in bilingual dictionaries and maintains that very little theoretical work has been done on this topic. This has resulted in bilingual dictionaries which often do not have a systematic and metalexically thought-through method of dealing with usage examples and fixed expressions. Often, idiomatic expressions are given as examples, but without adequate explanations of the meaning. In addition, there is often a lack of typographical markers to point out phrases or examples to the user. Diatopic, diachronic and diastratic markers are often missing; and phraseological false friends and partial equivalences are often not marked as such, to name only a few of the problems. Bielińska suggests that these issues could be adequately addressed in online dictionaries, where space will not be a problem.

María José Domínguez Vázquez, Fabio Mollica and Martina Nied Curcio discuss the problems which arise when students use bilingual online dictionaries for translating sentences with polysemous verbs and verbs which combine with prefixes or particles. The differences in valency of such verbs, which exist between Italian and Spanish on the one hand, and German on the other hand, create translation problems for Italian and Spanish learners of German as a foreign language. For example, in the case of the same main verb, in Italian *ascoltare*, in Spanish *escuchar*, and in German *hören*, the context determines which German translation equivalent should be used (e.g. *zuhören* or *anhören* instead of *hören* in certain contexts). In spite of the fact that by far the majority of Italian and Spanish students use online dictionaries, they were not very successful in translating the sentences which were requested in the questionnaire they were given, because of valency discrepancies between their native languages and German. From the students' comments on the survey, it also became clear that many of them ignored the grammatical information presented in the dictionaries, or that they did not read through the entire dictionary article. The authors suggest that user-friendly interfaces can be developed in multimedia online dictionaries, ensuring adequate information on the translation of polysemous verbs and verbs combining with prefixes or parti-

cles. In addition, students will have to receive better instruction on the use of dictionaries, in order to better interpret the metalinguistic markers, and to incorporate the given information in their tasks.

Luisa Giacoma asks what a dictionary written by users themselves would look like, delving into her personal perspectives as user and lexicographer. She has many years of experience as a lexicographer, especially in the bilingual lexicography of Italian and German. She maintains that many lexicographers do not draw on research done in the field of contrastive lexicology. For example, the treatment of collocations is often very inconsistent in that collocations are often presented as examples, and at other times in separate text blocks, without recognizable reasons for this inconsistency. In addition, bilingual dictionaries do not always give information on the contexts in which the different equivalents should be used. The syntactical context is often missing, and fixed expressions are not treated in a satisfactory manner. Giacoma's Italian–German bilingual dictionaries, done in collaboration with Susanne Kolb, are the first dictionaries in this language pair which are based on linguistic principles. These dictionaries provide explicit and systematic information on how exactly the keywords can and should be combined with other language elements. They contain information on word syntax, combined with collocators, and they include so-called "structural formulas", which give users a good idea of how to produce texts in the foreign language. For example, such a structural example will tell the user whether a verb needs to combine with an object (direct or indirect), with which prepositions it can be linked, the case of the keyword, especially after prepositions, et cetera. In the printed versions of the Giacoma/Kolb dictionaries, the collocators are positioned within curly brackets, and printed in small capital letters; in the online and CD-Rom versions of these dictionaries, the collocations are marked in red.

According to Giacoma, especially the morphology of German is neglected in bilingual dictionaries. This often results in errors by foreign language learners. Electronic data processing will enable more complete information on the inflection of each German word used in the dictionary articles. Dictionary users will then also be able to do searches, by typing in variations of a particular word. Giacoma concludes her list of (already established) wishes for an adequate bilingual dictionary with some promises for her future publications (already in print). Over 600 windows with false friends have been added, and tips on usage as well as notes on cultural differences and abbreviations. Even though these additions could not be made to the printed versions, users can get this extra information in the online version with a simple click. An important addition is also the possibility that users could get online access to data banks where they can get more information. Her aim is, in her own words (p. 244), to present users with a type of "map" for each word, by means of which they can move through the "new landscape" of the second language almost as effectively as native speakers.

The third section of the volume deals with specific current and planned

lexicographical projects. Firstly, Rufus Gouws offers several suggestions for the development of bilingual dictionaries, based on Wiegand's approach. Different users may use dictionaries for different purposes. It would be useful if different dictionaries could be derived from one comprehensive data bank or "mother lexicon". Gouws discusses a proposal for a new bilingual dictionary project with German and Afrikaans as language pair. The concept he has in mind, however, can be applied to any language pair. The same metalexicographical principles, which were developed for printed dictionaries, can be adapted and applied to e-dictionaries. Gouws proposes that a single large database, which he calls a "polytypological mother dictionary", can be used to extract different dictionaries, such as dictionaries for secondary school learners, university students, tourists, translators, and people in the field of business. The data, upon being entered, should be marked according to the possibilities of its use in the different dictionaries. For example, a translation equivalent that might be useful for translators (and is marked as such) may be marked differently for use by school learners. Gouws states that research done by Bothma et al. (2012) has identified at least thirty-six fields from which one should make selections per specific dictionary article. One advantage of the existence of such a mother database from which different dictionaries can be extracted is that users could personalise their dictionaries on the basis of their individual needs. They could set up a personal profile, which could be changed according to their specific and changing needs. Gouws states that the planned dictionary will be bidirectional, with two central lists, one in each language, so that both the German and Afrikaans lists can serve as source language and target language. Both the text reception function and the text production function can be addressed. One of the advantages of such an electronic mother dictionary is that one can regularly update terms in the database, devise new types of dictionaries, and change the data presentation in the dictionaries. The examples can have different formats and contents, in accordance with the different usage situations of the dictionaries. Gouws illustrates his exposition of such a mother dictionary with some enlightening examples.

David Lindemann presents an overview of bilingual Basque lexicography from the 19th century up to the present. Although the Basque language is not widely spoken, it has an interesting history of dictionaries and other publications on the language. Basque only became a written language in the middle of the 16th century, and literature and research into the language only came into being in the 17th century on the northern side of the Pyrenees, and only from the middle of the 18th century and especially during the 19th century on the southern side of the Pyrenees. A new electronic dictionary project, called EuDeLex, is currently under way at the University of the Basque Country.

Lindemann discusses several lexicographical products in chronological order, beginning with a Basque–German word list which originated around 1500 for use by pilgrims and authors of glossaries. Some of the dictionaries and publications were between Basque and French, others between Basque and

Spanish. Wilhelm von Humboldt (1767–1835) was very interested in the Basque language, and he had a great impact on the linguistic description of this language.

A couple of German–Basque dictionaries appeared in the second half of the 20th century, but there were problems: the one by Löpelmann (1968) was unreliable, containing many errors. A second one, compiled by Helmut Kühnel (1999), was already outdated when it appeared, because it did not take into account the standardised Basque orthography and morphology.

The first German–Basque dictionary to be really useful is the *Euskara–Alemana Hiztegia* (=EAH, 2007). This is a printed pocket dictionary containing 32 400 lemmas and 4 600 examples and phrases. This dictionary can be seen as the first to save users the trouble to have to consult French–German or Spanish–German dictionaries in order to successfully work in the language pairs German and Basque.

The new EuDeLex electronic dictionary will certainly enhance dictionary use involving the Basque language. Lindemann describes the features of this dictionary, which will also be based on the concept of a "mother lexicon" (Gouws, in this volume). Corpus linguistics is nowadays part and parcel of dictionary compilation. Therefore, this Basque dictionary project will be based on a Basque corpus which is derived parallel to the German corpus that is already available. Lindemann illustrates his discussion with examples of the macrostructure and the microstructure, as well as the treatment of dictionary articles with verbs, adjectives and adverbs. The proposed dictionary will have the advantage that it will be freely available to the public, since it is developed within the framework of a research unit of the University of the Basque Country.

Martin Becker's contribution discusses the smaller Slavic languages and the fact that they often do not have dictionaries, in spite of their extended vocabulary. He starts his discussion by classifying "major" languages as opposed to "medium" and "minor" languages among the Slavic languages. Kashubian and Upper- and Lower-Sorbian are examples of "minor" languages, with only about 50 000 speakers of Kashubian, and 55 000 and 12 000 speakers of Upper-Sorbian and Lower-Sorbian respectively. According to Becker, criteria which determine the "importance" of a language include its status as a language with established literary works, its status as an official language for a state, and the extent to which the language is standardised. The cultural and political significance of the language is also linked to these other factors. The vocabulary of the major languages is usually vast, and one can find a great number of general and special-field dictionaries in these languages.

An interesting phenomenon is the position of a minority language such as Sorbian. The cultivation of minority languages is financially supported in countries such as Germany, Austria and Poland, where lexicographical research is done and dictionaries in these languages are published. Small pockets of speakers of Upper- and Lower-Sorbian live in the federal states of Saxony and Brandenburg in Germany. The Sorbian Institute, based on the Institute for

Research into the Sorbian Nation, which was founded in the German Democratic Republic in 1951, undertakes research and advocates the spread of this minority language. This institute publishes in the area of culture, history and linguistics, and over the years, several dictionaries were also published. The same applies to the lexicography of Kashubian: since the 1990s, several dictionaries have been published, including special-field dictionaries, an author's dictionary, a bilingual German–Kashubian dictionary, and many more.

Becker maintains that electronic and online dictionaries hold many possibilities for minority language dictionaries. A multi-language data bank for the Slavic languages could make comparative studies between the Slavic languages possible. Special-field dictionaries would be possible, and they could contain exact explanations of the terms, since space is not a problem. Becker's concept is, of course, applicable to other sets of minority languages in other countries.

In his contribution, Peter Meyer describes the lexicographical process followed during the construction of the comprehensive portal database for the project "Lehnwortportal Deutsch" (=loan word portal German), which is currently being compiled at the Institute for the German Language (IDS) in Mannheim, Germany. This database portal offers several learner's dictionaries in languages such as Polish, Cieszyn Polish and Slovenian, and it concentrates on loan words from German in these languages. It is open to the public on the internet, and allows for extensive search functions, not only in the dictionaries themselves, but also in the database as such. It also contains a so-called "dictionary of origin" ("Herkunftswörterbuch") or "inverted dictionary of loan words", which gives information on the etymology of the German loan words. The lemmas in this etymological dictionary function as etymological "metalemmas", and are considered to be the *tertium comparationis* of the loan word portal.

The loan word portal is a Java-based web application, developed by the IDS, and the data bank contains individual articles of the different dictionaries of loan words as XML documents. The relationships between the various elements of the dictionary are depicted by means of graphs. The relationships between word forms (for example, between the etymon and the loan word, or between the metalemma and the loan word form) or between different word forms in the different dictionaries are shown by means of arrows, to designate derivations, variations, et cetera. A unique ID number is assigned to each word form before the word form is saved within a nodal chart which shows the relationships.

Carolina Flinz discusses special-field languages as a basis for dictionary compilation by presenting examples from a planned German–Italian online dictionary project called TOURLEX. She states that online special-field dictionaries in the field of tourism only came into being towards the end of the 20th century. Usually they are glossaries or lexicons, without information on morphosyntactic issues or collocations. TOURLEX will be freely available on the internet, and it will give information on pronunciation (by means of an audio

example), syllable division, word class, gender, number, translation equivalents, syntagmatic information such as collocations, valency items, sentence examples, and paradigmatic items (synonyms). Flinz describes the deliberations in the planning of the lexicon, which include determining the target users, the user situations and the needs of the users. Questionnaires, user protocols and a blog forum for discussion will be used in this process.

TOURLEX will use computer-based criteria: a corpus of special-field texts on tourism will be analysed with Word Smith Tools 3.0., after which the lemma lists in both languages will be compiled. User-friendly effects will be used in the layout, such as colours, different buttons, links between the index and the lemmas, links to external resources, easy-to-use search functions, and possibilities to give feedback to the dictionary team. This concept, as presented by Flinz, can, of course, be adapted and applied to other bilingual special-field dictionaries.

All in all, this volume is of great interest to lexicographers who would like to see how other dictionary makers plan and execute their projects, and how they apply the latest research and trends in their dictionaries. Each contribution has a formidable reference list, which shows that the contributors base their projects on theoretical principles and solid research. The problem of valency, as the overall topic in this volume, is addressed in different ways, but with imaginative efforts to find solutions so that users' needs can be fulfilled in the best possible way. This volume can help us all in our future planning of bilingual dictionaries.

Endnote

1. According to Fischer, one exception is Curcio's (1999) Italian–German valency dictionary.

Bibliography

- Bothma, Theo J.D., Henning Bergenholtz and Rufus H. Gouws.** 2012. *Filtering, Searching and Navigating for Fixed Expressions*. Unpublished Paper Presented at the EUROPHRAS 2012 Conference in Maribor, Slovenia, August 27–31, 2012.
- Croft, William.** 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Curcio, Martina Lucia.** 1999. *Kontrastives Valenzwörterbuch der gesprochenen Sprache Italienisch–Deutsch. Grundlagen und Auswertung*. Mannheim: Institut für Deutsche Sprache. CD-ROM.
- EAH = *Euskara–Alemana Hiztegia*. Martínez Rubio, Elena (Ed.). 2007. *Euskara–Alemana Hiztegia*. Donostia: Elkar.
- E-VALBU = <http://hypermedia.ids-mannheim.de/evalbu/index.html> [Accessed 8 August 2018].
- Fillmore, Charles J.** 1982. Frame Semantics. The Linguistic Society of Korea (Ed.). 1982. *Linguistics in the Morning Calm*: 111-137. Seoul: Hanshin.
- Foer, Jonathan Safran.** 2002. *Everything is Illuminated: A Novel*. New York: Houghton Mifflin.

- Goldberg, Adele E.** 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
- Goldberg, Adele E.** 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Kühnel, Helmut (Ed.)**. 1999. *Wörterbuch des Baskischen*. Wiesbaden: Reichert.
- Löpelmann, Martin (Ed.)**. 1968. *Etymologisches Wörterbuch der baskischen Sprache; Dialekte von Labourd, Nieder-Navarra und La Soule*. Berlin: De Gruyter.
- VALBU = *Valenzwörterbuch deutscher Verben*. Schumacher, Helmut, Jacqueline Kubczak, Renate Schmidt and Vera de Ruiter. 2004. *Valenzwörterbuch Deutscher Verben*. Tübingen: Gunter Narr.

Maria Smit
Independent Lexicographer and Language Practitioner
Princeton, New Jersey
United States of America
(ria.eden@gmail.com)

Publikasieaankondigings / Publication Announcements

María José Domínguez Vázquez, Fabio Mollica, Martina Nied Curcio (Herausgeber/Editors). *Zweispachige Lexikographie zwischen Translation und Didaktik*. 2014, vi + 334 pp. ISBN 978-3-11-036973-1, e-ISBN 978-3-11-036663-1. Lexicographica. Series Maior 147. Berlin/Boston: Walter de Gruyter. Price: €109.95. (Review in this issue.)

Franco Frescura, Joyce Myeza. *Illustrated Glossary of Southern African Architectural Terms*. 2016, 224 pp. ISBN 978-1869143497. Scottsville: University of KwaZulu-Natal Press. (UKZN Bilingual Glossary Series.) Price R255.00.

Pedro A. Fuertes-Olivera. *The Routledge Handbook of Lexicography*. 2018, 810 pp. ISBN: 978-1-138-94160-1. London/New York: Routledge. Price £165.00. (Review in this issue.)

Vida Jesenšek, Milka Enceva (Herausgeber/Editors). *Wörterbuchstrukturen zwischen Theorie und Praxis*. 2014, vi + 261 pp. ISBN 978-3-11-059630-4, e-ISBN 978-3-11-059432-4. Lexicographica. Series Maior 154. Berlin/Boston: Walter de Gruyter. Price: €99.95.

Roderick McConchie, Jukka Tyrkkö (Editors). *Historical Dictionaries in Their Paratextual Context*. 2018, xii + 331 pp. ISBN 978-3-11-057286-5, ISSN 0175-9264. Lexicographica. Series Maior 153. Berlin: Walter de Gruyter. <https://www.degruyter.com/view/product/497321>. Price: €99.95.

Pharos Junior Tweetalige Skoolwoordeboek/Pharos Junior Bilingual School Dictionary. 2018, 512 pp. ISBN 978-86890-206-4. Cape Town: Pharos. Price R140.00.

Fred Pheiffer (Editor-in-Chief). *Oxford Afrikaans–Engels / English–Afrikaans Skoolwoordeboek / School Dictionary*. Second edition. 2017, xii + 728 pp. ISBN 978 0 19 905468 8 (Paperback), ISBN 978 0 19 072106 0 (Hardback). Cape Town: Oxford University Press Southern Africa. Price R149.95.

Clive Roos, Michael Wilter. *Oxford South African Dictionary of School Terminology*. 2018, viii + 168 pp. ISBN 978 0 19 044106 7. Cape Town: Oxford University Press Southern Africa. Price R170.00.

Albert Venter, Susan Botha, Louis du Plessis, Mariëtta Alberts. *Explanatory Dictionary of Politics: Bilingual Core Terms and Definitions in Political Science / Verklarende Politieke Woordeboek: Tweetalige Kernterme en -definisies in Politieke Wetenskap*. 2017, xvi + 403 pp. ISBN 9781485119944 (Soft Cover). Cape Town: Juta. www.jutalaw.co.za. Price R540.00.

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 engeen 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 nuusvystellings van leksikografiese verenigings wat veral vir die praktiserende leksikograaf van waarde sal wees.

(7) **Ander:** Van tyd tot tyd kan ander rubrieke deur die redaksie ingevoeg word, soos Leksikoprogrammatuur, Leksiko-opname, Leksikobibliografie, Leksikonuus, Lexikofokus, Leksiko-eerbewys, Leksikohuldeblyk, Verslae van konferensies en werksessies.

Bydraes in kategorieë (3)-(7) moet almal aan die eise van akademiese geskrifte voldoen en word met die oog hierop deur die redaksie gekeur.

2. Wetenskaplike standaard en keuringsprosedure

Lexikos is deur die Departement van Hoër Onderwys van die Suid-Afrikaanse Regering as 'n gesubsidieerde, d.w.s. inkomstegenererende navorsingstydskrif goedgekeur. Dit verskyn ook op die *Institute of Science Index (ISI)*.

Artikels sal op grond van die volgende aspekte beoordeel word: taal en styl; saaklikheid en verstaanbaarheid; probleemstelling, beredenering en gevolgtrekking; verwysing na die belangrikste en jongste literatuur; wesenlike bydrae tot die spesifieke vakgebied.

Manuskripte word vir publikasie oorweeg met dien verstande dat die redaksie die reg voorbehou om veranderinge aan te bring om die styl en aanbieding in ooreenstemming met die redaksionele beleid te bring. Outeurs moet toesien dat hulle bydraes taalkundig en stilisties geredigeer word voordat dit ingelewer word.

3. Taal van bydraes

Afrikaans, Duits, Engels, Frans of Nederlands.

4. Kopiereg

Nóg die Buro van die WAT nóg die African Association for Lexicography (AFRILEX) aanvaar enige aanspreeklikheid vir

eise wat uit meewerkende skrywers se gebruik van materiaal uit ander bronne mag spruit.

Outeursreg op alle materiaal wat in *Lexikos* gepubliseer is, berus by die Direksie van die Woordeboek van die Afrikaanse Taal. Dit staan skrywers egter vry om hulle materiaal elders te gebruik mits *Lexikos* (AFRILEX-reeks) erken word as die oorspronklike publikasiebron.

5. Oorspronklikheid

Slegs oorspronklike werk sal vir opname oorweeg word. Skrywers dra die volle verantwoordelikheid vir die oorspronklikheid en feitelike inhoud van hulle publikasies. Indien van toepassing, moet besonderhede van die oorsprong van die artikel (byvoorbeeld 'n referaat by 'n kongres) verskaf word.

6. Gratis oordrukke en eksemplare

Lexikos is sedert volume 28 slegs elektronies beskikbaar op <http://lexikos.journals.ac.za>. Geen oordrukke of eksemplare is dus beskikbaar nie.

7. Uitnodiging en redaksionele adres

Alle belangstellende skrywers is welkom om bydraes vir opname in *Lexikos* te lewer en verkieslik in elektroniese formaat aan die volgende adres te stuur: lexikos@sun.ac.za, of Die Redakteur: LEXIKOS, Buro van die WAT, Postbus 245, 7599 STELENBOSCH, 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 sageware 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 engeen 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.

4. Aantekeninge/voetnote/eindnote

Aantekeninge moet deurlopend in die vorm van boskritef genommer en aan die einde van die manuskrip onder die opskrif **Eindnote** gelys word.

INSTRUCTIONS TO AUTHORS

(For a more detailed version of these instructions, please contact us (lexikos@sun.ac.za) or refer to our website: <http://lexikos.journals.ac.za/>)

A. EDITORIAL POLICY

1. Type and content of articles

Articles may treat pure lexicography or the implications that related fields such as linguistics, general linguistics, terminology, computer science and management have for lexicography.

Contributions may be classified in any one of the following categories:

(1) **Articles:** Fundamentally original scientific research done and the results obtained, or existing research results and other facts reflected in an original, synoptic, interpretative, comparative or critically evaluative manner.

(2) **Review articles:** Research articles presented in the form of a critical review of one or more published scientific sources.

Contributions in categories (1) and (2) are subjected to strict anonymous evaluation by independent academic peers in order to ensure the international research quality thereof.

(3) **Reviews:** An analysis and critical evaluation of published scientific sources and products, such as books and computer software.

(4) **Projects:** Discussions of lexicographical projects.

(5) **Lexiconotes:** Any article containing practice-oriented information, suggestions, problems, questions, commentary and solutions regarding lexicography.

(6) **Lexicovaria:** Any of a large variety of articles containing announcements and press releases by lexicographic societies which are of particular value to the practising lexicographer.

(7) **Other:** From time to time other categories may be inserted by the editors, such as Lexicosoftware, Lexicosurvey, Lexicobibliography, Lexiconews, Lexicofocus, Lexicohonour, Lexicotribute, Reports on conferences and workshops.

Contributions in categories (3)-(7) must all meet the requirements of academic writing and are evaluated by the editors with this in mind.

2. Academic standard and evaluation procedure

The Department of Higher Education of the South African Government has approved *Lexikos* as a subsidized, i.e. income-generating research journal. It is also included in the *Institute of Science Index* (ISI).

Articles will be evaluated on the following aspects: language and style; conciseness and comprehensibility; problem formulation, reasoning and conclusion; references to the most important and most recent literature; substantial contribution to the specific discipline.

Manuscripts are considered for publication on the understanding that the editors reserve the right to effect changes to the style and presentation in conformance with editorial policy. Authors are responsible for the linguistic and stylistic editing of their contributions prior their submission.

3. Language of contributions

Afrikaans, Dutch, English, French or German.

4. Copyright

Neither the Bureau of the WAT nor the African Association for Lexicography (AFRILEX) accepts any responsibility for claims which may arise from contributing authors' use of material from other sources.

Copyright of all material published in *Lexikos* will be vested in the Board of Directors of the Woordboek 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.

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**.