No. 28 | July 2020
https://lexicala.com/kln

K LexicalNews

The new KLN: Lexicalbound. **Ilan Kernerman.** p2

Tom McArthur 1938-2020
The legacy of Tom McArthur. **Lan Li.** p3
What is 'reference science'? **Tom McArthur.** p5
Asian Lexicography: Past, Present, and Prospective. **Tom McArthur.** p10

NexusLinguarum: European network for Web-centred linguistic data science. **Jorge Gracia.** p21

Applying the OntoLex-*lemon* lexicography module to K Dictionaries' multilingual data.
Dorielle Lonke and Julia Bosque-Gil. p30

English WordNet: A new open-source wordnet for English.
John P. McCrae, Ewa Rudnicka and Francis Bond. p37

ELEXIS: Technical and social infrastructure for lexicography.
Anna Woldrich, Teja Goli, Iztok Kosem, Ondřej Matuška and Tanja Wissik. p45

Internet lexicography at the Leibniz-Institute for the German Language.
Stefan Engelberg, Annette Klosa-Kückelhaus and Carolin Müller-Spitzer. p54

***Understanding English Dictionaries*:**
The first MOOC about lexicography. **Michael Rundell.** p78

SuperMemo. p81

GlobaLex update. p82
Tom McArthur and Globalex. **Ilan Kernerman.** p84

KDICTIONARIES

K DICTIONARIES LTD
6 Nahum Hanavi Street, Tel Aviv 6350310 Israel
+972-3-5468102
kdl@kdictionaries.com
https://lexicala.com

Editor:
Ilan Kernerman

ISSN 1565-4745

The new KLN: Lexicalbound

The first four-page issue of this publication appeared in July 1994 under the title **Password News** as "[a] forum for discussion about the semi-bilingual dictionary". Issue No. 2, published in January 1995, was renamed **Kernerman Dictionary News** and has since appeared each July, gradually expanding coverage to all dictionary-related topics and eventually also linking multiple language technology domains, while hosting an ever-growing variety of authors.

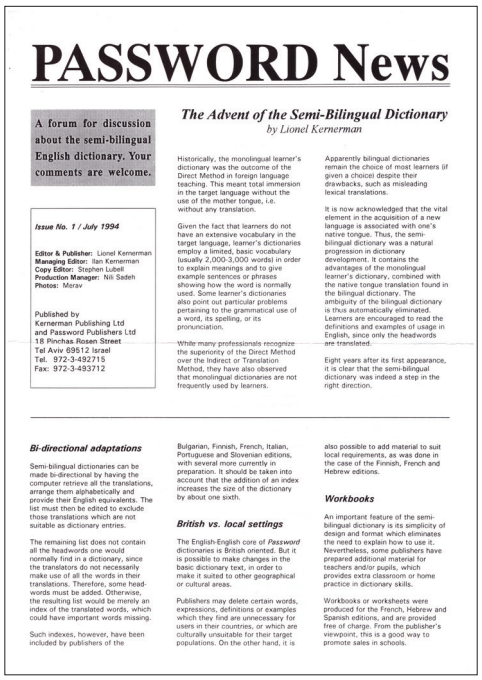
Over the years the newsletter has thus transformed from a single-focus promotion tool to serving a broad and diverse global community. Since the early 2000s, at least one thousand copies of each KDN issue were printed and freely distributed every year, as well as being accessible online, making it possibly the most widely disseminated publication on lexicography (and more) worldwide.

Change being a constant factor in life, time has come to update 😊 The current issue, No. 28, appears for the first time digitally only, in PDF and HTML, and the new name – **K Lexical News** – makes explicit our interest beyond dictionaries and lexicography to everything lexical, complemented by a new look and feel.

I wish to thank the numerous colleagues and friends for their contribution in writing, consulting, reading and all other forms of support, with special thanks to the designer, Orna Cohen.

Ilan Kernerman

KLexicalNews



PASSWORD News

A forum for discussion about the semi-bilingual English dictionary. Your comments are welcome.

Issue No. 1 / July 1994

Editor & Publisher: Lionel Kernerman
Managing Editor: Ilan Kernerman
Copy Editor: Stephen Lubell
Production Manager: Nili Sadeh
Photos: Marav

Published by
Kernerman Publishing Ltd
and Password Publishers Ltd
18 Pinchas Rosen Street
Tel Aviv 69512 Israel
Tel. 972-3-492715
Fax: 972-3-493712

The Advent of the Semi-Bilingual Dictionary
by Lionel Kernerman

Historically, the monolingual learner's dictionary was the outcome of the Direct Method in foreign language teaching. This meant total immersion in the target language without the use of the mother tongue, i.e. without any translation.

Given the fact that learners do not have an extensive vocabulary in the target language, learner's dictionaries employ a limited, basic vocabulary (usually 2,000-3,000 words) in order to explain meanings and to give example sentences or phrases showing how the word is normally used. Some learner's dictionaries also point out particular problems pertaining to the grammatical use of a word, its spelling, or its pronunciation.

While many professionals recognize the superiority of the Direct Method over the Indirect or Translation Method, they have also observed that monolingual dictionaries are not frequently used by learners.

Apparently bilingual dictionaries remain the choice of most learners (if given a choice) despite their drawbacks, such as misleading lexical translations.

It is now acknowledged that the vital element in the acquisition of a new language is associated with one's native tongue. Thus, the semi-bilingual dictionary was a natural progression in dictionary development. It contains the advantages of the monolingual learner's dictionary, combined with the native tongue translation found in the bilingual dictionary. The ambiguity of the bilingual dictionary is thus automatically eliminated. Learners are encouraged to read the definitions and examples of usage in English, since only the headwords are translated.

Eight years after its first appearance, it is clear that the semi-bilingual dictionary was indeed a step in the right direction.

Bi-directional adaptations

Semi-bilingual dictionaries can be made bi-directional by having the computer retrieve all the translations, arrange them alphabetically and provide their English equivalents. The list must then be edited to exclude those translations which are not suitable as dictionary entries.

The remaining list does not contain all the headwords one would normally find in a dictionary, since the translators do not necessarily make use of all the words in their translations. Therefore, some headwords must be added. Otherwise, the resulting list would be merely an index of the translated words, which could have important words missing.

Such indexes, however, have been included by publishers of the Bulgarian, Finnish, French, Italian, Portuguese and Slovenian editions, with several more currently in preparation. It should be taken into account that the addition of an index increases the size of the dictionary by about one sixth.

British vs. local settings

The English-English core of *Password* dictionaries is British oriented. But it is possible to make changes in the basic dictionary text, in order to make it suited to other geographical or cultural areas.

Publishers may delete certain words, expressions, definitions or examples which they find are unnecessary for users in their countries, or which are culturally unsuitable for their target populations. On the other hand, it is also possible to add material to suit local requirements, as was done in the case of the Finnish, French and Hebrew editions.

Workbooks

An important feature of the semi-bilingual dictionary is its simplicity of design and format which eliminates the need to explain how to use it. Nevertheless, some publishers have prepared additional material for teachers and/or pupils, which provides extra classroom or home practice in dictionary skills.

Workbooks or worksheets were produced for the French, Hebrew and Spanish editions, and are provided free of charge. From the publisher's viewpoint, this is a good way to promote sales in schools.

Kernerman
DICTIONARY News

Issue No. 2 / January 1995

Editor & Publisher: Lionel Kernerman
Managing Editor: Ilan Kernerman
Production Manager: Nili Sadeh

Published by
Kernerman Publishing Ltd
and Password Publishers Ltd
18 Pinchas Rosen Street
Tel Aviv 69512 Israel
Tel. 972-3-492715
Fax: 972-3-493712
Email: kern_or@netvision.net.il

CONTENTS:

Forward Ilan Kernerman 1
Monolingual or Bilingual, that is *not* the Question: the 'Bilingualised' Dictionary Kyohei Nakamoto 2-7
Answers to Open Questions Joseph A. Reif 8

The second edition of Oxford Student's Dictionary for Hebrew Speakers *was published by Kernerman Publishing and Lonnie Kahn in January 1993.*

Forward

A number of articles on bilingualised dictionaries appeared in various publications last year. Also, there are plans by publishers to bilingualise existing monolingual EFL dictionaries. Thus, in light of the rising attention towards this genre it was decided to broaden the scope of our newsletter to the wider domain of learner's dictionaries, retitle it, and devote this issue to some of the recent research in bilingualised dictionaries.

First, there was the paper presented by Batya Laufer and Linor Melamed at the EURALEX 1994 Congress in Amsterdam, and published in its *Proceedings*: **Monolingual, Bilingual and 'Bilingualised' Dictionaries: Which are More Effective, for What and for Whom?** It compares the effectiveness of these three types of dictionaries in the comprehension and production of new words by EFL learners, as tested on 123 high-school and university students in Israel. "The practical conclusion of the study seems to be that a good 'bilingualised' dictionary is suitable for all types of learners. When the learner is still unskilled in dictionary use, s/he may rely on the bilingual information. With progress in these skills, the monolingual information will gain relevance and importance... Even when the monolingual part of the entry is used to its full potential ... the translation will always be helpful in reassuring and reinforcing the learner's decisions about the meaning [and use] of new words...".

Another paper, **Translated! A new breed of bilingual dictionaries**, by Mona Baker and Robert Kaplan in the International Journal of Translation *babel* (40: 1), describes the ongoing bilingualisation program of Collins Cobuild dictionaries. Called *bridge bilinguals*, their main feature is that the explanations of the monolingual English learner's dictionary are translated into the user's native-tongue. Under preparation are versions for speakers of Brazilian-Portuguese, French, German, Italian and Spanish. Without entering into the debate about the merits of this particular feature, on the whole, it appears that the rationale for inserting L1 translations is being increasingly adopted by educators, lexicographers and publishers of learner's dictionaries worldwide.

Two recent papers by Reinhard Hartmann enhance his studious research on the subject: **Bilingualised versions of learners' dictionaries** appeared in *Fremdsprachen Lehren und Lernen* (FLuL 23), published by Gunter Narr Verlag in Tübingen. It discusses the bilingualised learner's dictionary in terms of four divisions of dictionary research (history, typology, criticism, and use) and in the light of a project carried out at the University of Exeter. The project involved a reading comprehension task that assessed the reaction of 28 informants to seven bilingualised dictionaries intended for Arabic, French, German, Greek, Korean and Spanish learners of English. The paper concludes that combining L2 definitions and L1 translation equivalents is an attractive feature of such dictionaries, especially for decoding activities.

Dr Hartmann's other paper, **The Bilingualised Learner's Dictionary (A Transcontinental Trialogue on a Relatively New Genre)**, was published by the Language Centre of the Hong Kong University of Science and Technology and the Division of Foreign Languages of Jilin University of Technology, and presents an entertaining and illuminating discussion on the topic.

Finally, the paper by Kyohei Nakamoto - that was originally published by Iwasaki Linguistic Circle in its annual periodical *LEXICON*, and which follows here nearly in its entirety. I wish to express our acknowledgement and gratitude to Mr Nakamoto and to the ILC Chairman, Professor Shigeru Takebayashi, for granting us permission to reprint it.

This paper is supplemented by a brief response from Joseph Reif. I hope our work contributes to generating further discussion on learner's dictionaries.

Ilan Kernerman

(left) **Password News** No. 1. July 1994

(right) **Kernerman Dictionary News** No. 2. January 1995

Tom McArthur 1938-2020

The legacy of Tom McArthur

Lan Li

Tom McArthur was born in the city of Glasgow, Scotland, and studied at the University of Glasgow (MA) and University of Edinburgh (PhD). He had a rich international career, starting as an officer-instructor in the British Army, and subsequently as Head of English at the Cathedral School in Bombay (Mumbai), lecturer and Director of Studies at Extra Mural English Language Courses at the University of Edinburgh, Associate Professor of English at Université du Québec à Trois-Rivières, and Visiting Professor at the University of Exeter's Dictionary Research Centre, Chinese University of Hong Kong, Lingnan University, and Xiamen University.

Tom was a world-renowned linguist, fluent in English, Scots and French, with an academic knowledge of Latin, Ancient Greek and Sanskrit. He could also converse to varying degrees in Spanish, Italian, Greek, Russian, German, Persian/Farsi, and Cantonese. He contributed to the field of linguistics with passion and love for world culture and languages, and shed light in particular on English studies, world Englishes and lexicography.

Contribution to lexicography

Tom was a lexicographer. He proposed the term 'reference science' for works providing lexical, grammatical, encyclopedic and other referential information. Defying the A-Z convention of lexicographic practice, he compiled the thematic dictionary, *Longman Lexicon of Contemporary English* (1981), complementing *Longman Dictionary of Contemporary English*. The Lexicon is an admixture of cognitive science and reference science, containing over 15,000 entries in 130 topics, from life and animals to war and peace. It illuminates word differences in the same semantic field, such as *hotel, motel* and *inn*, and is especially useful for non-native learners of English to

enlarge their vocabulary. The book has had 22 printings and has been translated to different languages.

Alongside Reinhard Hartmann, Tom co-organized 14 sessions of Interlex (International Lexicography Course) at the Dictionary Research Centre at the University of Exeter from 1987 to 2000. They also initiated training lexicographers in the MA and PhD Lexicography programme from 1993 to 2000. Many of their students became practicing lexicographers or university professors in different parts of the world.

Contribution to English language research and education

Tom was an inspiring professor, doing independent academic research. His doctoral thesis was entitled *The English Word?* and his study into the English language covers a wide range of topics, including lexis, syntax, phonetics and sociolinguistics. He was the founding editor of the journal *English Today*, by Cambridge University Press, leading it from 1985 to 2008, and a walking encyclopedia recharging students with not only linguistic knowledge but also culture and history. With English teachers and learners in mind, his books were wittily written and easy to engage with. The peak of his linguistic achievements was in the editorship of *The Oxford Companion to the English Language* (1992), which constitutes an immense, complex and detailed survey of the English language, including extensive facts and sharp opinions from scholars worldwide, describing local, regional and international usages of the language in detail and illustrating standard and non-standard varieties of English that present readers with a full picture of the world lingua franca. Another masterpiece Tom took much pride in was *The Oxford Guide to World English* (2003), which exemplifies how English has been used all over the world by more non-native than native speakers – a stark comparison with Latin in the Middle Ages.

Close link to Asia

Tom was a global citizen, interested in different languages and cultures, with a particular interest in Asia, an early proof being his condensed translation from Sanskrit of *An Easy-to-Read Bhagavad Gita* that appeared in 1978. He worked in India, loved Singapore and lived in Hong Kong, was one of the founders of the Asian Association for Lexicography in 1997 (together with Gregory James and Reinhard Hartmann) and participated in the Asialex conferences of 2003 in Japan and 2005 in Singapore.

Inspiring and sharing world ideas

Tom was a great tutor. He never gave up the thought of nurturing young teachers. While working as the editor of *English Today*, he created a hub bringing together famous experts as well as young scholars worldwide, presenting a comprehensive picture of English yesterday, English today and English tomorrow. His enlightened thinking, open-mindedness, consideration, generosity and encouragement stimulated many minds. He will be remembered forever.

Lan Li was a student of Tom McArthur at the University of Exeter. Currently she is Director of the Centre for Learning Enhancement and Research and an Associate Professor at the Chinese University of Hong Kong (Shenzhen) and Review Editor of *Lexicography – Journal of Asialex*.

Obituary by Roshan McArthur, The Guardian, 12 April 2020.

Tom McArthur's *English Today* by Kingsley Bolton, David Graddol and Rajed Mesthrie, *English Today* 100, Vol. 25, No. 4: 3-8. December 2009.

What is 'reference science'?

Tom McArthur

Reprinted by permission from *Lexicon*, 28: 135-140. 1998. Tokyo: Iwasaki Linguistic Circle.

It was born at a one-day conference at the University of Exeter in England in the spring of 1996. The birth was on time, the baby was small but in excellent health, and hardly made any noise. As a result, very few people knew that it had arrived. At the same time, however, there has been a steadily increasing interest in the new arrival, and in September this year [1997] I talked to the Iwasaki Linguistic Circle about it in Tokyo. I believe it is a subject whose time has come, but it will take a little more time before the precise nature and relevance of 'reference science' become clear.

Before I go on I'd like to look at a rather basic issue — the actual matter of inventing a science. Can one just invent a science when one feels like it? And if you do, how does it stay invented? Does a new science occupy new semantic or conceptual space, does it 'steal' space from other sciences, or does it overlap, flowing in and out of them? Or are these the wrong metaphors? And if you do invent a science, when and how do you know if you've succeeded — ten, twenty, a hundred years later? I would argue that these questions are not just interesting in general terms; they are questions for which reference science could itself provide a framework for answers — and further questions.

Looking back over the year since we launched our fledgling science, four things particularly stand out for me:

- Reinhard Hartmann creating the Dictionary Research Centre, which has proved successful in getting lexicographers and other interested people to talk to each other.
- Study programmes at Exeter, from the doctoral level to the one-week InterLex course, that allow open-ended consideration of everything relating to lexicography. Nothing referential was arbitrarily excluded, and minds could extend themselves.
- The formulation over time of first EuraLex then AfriLex, then this year, AsiaLex. These organizations, alongside the Dictionary Society of North America, provide a firm base for lexicographical debate, without which one could not contemplate anything more fundamental.
- The publication by Cambridge in 1986 of my *Worlds of Reference: Language, lexicography and learning from the clay tablet to the computer* [WoR]. The book was widely and constructively reviewed, and the most enthusiastic reviewers were not lexicographers but librarians and computer people who seemed to feel that it gave them a history and even a charter. Lexicographers generally responded well, but some considered that I did not give enough attention to 'proper' lexicography. But then, the book wasn't about any single art, craft or science. It was about how we refer and inform, how we communicate, and how we know.

One of the most powerful developments since WoR was published has been our understanding of DNA. In a few short years humankind has uncovered and begun to map a referential software system that is built into us and all other life known to us. It seems to me that we need a framework within which we can ask such questions as 'How similar are human language and DNA?' and 'How similar to and different from DNA are our systems of information storage and retrieval?' It is not enough to talk about 'the language of the genes' and 'genetic letters'. Are these simply metaphors, or do language systems and gene systems share a basic pattern that could also underlie some third system that we have not yet encountered? This is just one of the possible areas that reference scientists might in due course look at.

We can consider next something not quite so cosmic, but nonetheless large: what at the end of WoR I described as a 'global nervous system'. In just ten years, that nervous system has immensely, almost incalculably, increased — a vast multiplex of old copper cable and new fibre-optics, older ground TV and newer satellite TV, and

many other things. Technology is one thing; however, content and use another, and part of that content and use relates to asking for information either from other humans by e-mail or from the system itself on, say, the World-Wide Web. Reference science has a place in observing and reporting on this largest and most integrated reference service humanity has ever known, into which many of the resources of the world's great libraries are currently being woven, to form the largest work of reference that has ever existed.

When pushed, users and observers of works of reference will concede that both the dictionary and the telephone directory have much in common, as do indexes, concordances, atlases, manuals, and catalogues (whether the mail-order kind or in libraries). It is hard, however, to conceive of the circumstances in which the compiler of a telephone directory, an atlas, a computer manual, or a catalogue would be accepted as members of Euralex or the DSNA. Yet these varied products are linked by their reference function and a range of common techniques and technologies. The current computerization of all such materials only serves more fully to emphasize this point.

Indeed, they belong within something larger than, but closely associated with, traditional lexicography, have never had any generic names, and at the close of this century they need such names. On offer since at least 1986 have been, for the practical business of producing artifacts, such terms as reference art and reference technology, and since 1996 the term for their assessment has been reference science, the study of all aspects of organizing data, information, and knowledge in any format whatever, for any purpose whatever, using any materials whatever. The lack of such a level of study may be due in part at least to a historical current which, in the terminology of postmodernist literary theory, has 'privileged' the position of dictionaries and to some extent also encyclopedias, gazetteers, chronologies, concordances, and indexes (all in archetypal A-Z order) and along with them privileged the position of lexicography and its practitioners.

Lexicographers might, in Johnson's term, be 'harmless drudges', but their drudgery has for centuries been held in higher esteem than that of makers of catalogues, directories, time-tables, ready-reckoners, and travel guides. It might be wise in McLuhan's age of information overload to seek greater egalitarianism in the worlds of reference, by focusing on reference itself rather than on language and alphabeticism (significant as these are), and to examine and exploit all techniques and insights associated with all works of reference from any time, place, language, and writing system.

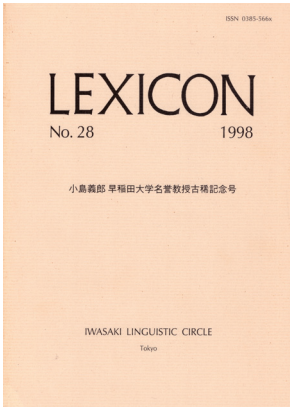
Of course, it is only relatively recently that lexicography has been systematically critiqued, a development that has however proved both successful and useful. Nowadays, lexicographers no longer simply compile dictionaries according to formulas that seldom change but are liable as they work to develop theories about what they do and novel practices tied to those theories. Given this advance, is it asking too much to say now: Look beyond this recently-raised consciousness and recognise a greater link with other professionals and products.

It is not surprising that the academic world has paid little or no attention to the making of directories and catalogues. So crucial, however, is the business of organizing information in our time, and on a global basis, that it may soon be difficult — impossible — to avoid bringing all the tools and vehicles of reference together within one subject area with one name. This will happen, I suspect, if for no other reason than that anything informative and referential, when stored in a computer, becomes quite simply a database, regardless of whatever name or function or prestige or lack of prestige it might traditionally have had. The electronic revolution is a leveller.

At the moment, however, I feel that we can identify three areas of immediate concern to reference science, the first with a traditional name, the second with a new name, and the third with no name at all:

- The first is lexicography, that aspect of reference art and technology which deals wholly or mainly with language and pre-eminently with words, regardless of the format used (in the main alphabetic, thematic, or a hybrid of the two).
- The second is encyclopedics, that aspect of reference art and technology which deals with information about the world, and for me includes atlases, gazetteers, almanacs, and manuals (and ties in with textbooks).
- The third covers tabulations (such as time-tables), directories (as for telephone subscribers), and catalogues (among other things). It may prove to be several areas and require us to conclude that certain divisions of reference science necessarily overlap with other disciplines and activities, such as library science and social and business life, because they have common concerns.

Fairly obviously, the bulk of research and commentary in reference science in the immediate future will concern dictionaries and probably also encyclopedics. I anticipate, however, that increased interest in databases, hypertext, multimedia, and information structures at large — from satellite linkups to DNA — will ensure that more attention is paid to my third, unnamed element, which to date has been the part of the iceberg below the referential waterline.



It seems to me that there are all sorts of fertile possibilities within the framework made possible by the concept reference science. I will close by looking at only one of these, a contrast that has become important in lexicography in recent years: macrostructure and microstructure. This dichotomy is usually interpreted as covering on the one hand the overall ('macro') organization of a dictionary and on the other any single entry within such a work (the 'micro' organization). I would argue here, however, that the contrast is valuable not only in terms of dictionaries and their entries (and by extension library catalogues and whatever their constituent units may be) but also in other levels of organization among information, knowledge, and communication structures.

Thus, just as an entry is microstructural within the macrostructure of a dictionary, so such a dictionary is microstructural within a publisher's list of dictionaries. Such a reference list is in its turn microstructural within the macrostructure of all publisher's reference lists everywhere. The same is true with each bibliographical catalogue in a library, which is microstructural within the macrostructure of all bibliographical collections within all libraries and similar institutions in a city, state, or the world — especially if such resources are linked electronically. Again, within such a system as the World-Wide Web, each website is microstructural within the WWW at large.

Such matters can become discussable if we have such a framework as reference science, whose findings and postulations can feed back into the practical business of making books and other artifacts. Reference science could be a liberating and integrating discipline, in which lexicography would not be eclipsed but strengthened, not downgraded but upgraded, in intriguing theoretical and practical ways. The term proposed is, I suggest, neither a cute neologism nor a novelty for its own sake, but at the close of this century a necessity.

Asian Lexicography: Past, Present, and Prospective

Tom McArthur

Introduction to *Lexicography in Asia. Selected papers from the Dictionaries in Asia Conference, Hong Kong University of Science and Technology, 1997, and other papers*. Editors: Tom McArthur and Ilan Kernerman. 1998: 9-20. Tel Aviv: Password Publishers.

Introduction

In 1997, I had the good fortune to attend two international conferences held in East Asia, the first in Hong Kong in March, the second in Tokyo in August. Both were concerned with lexicography but, although a number of people attended both, there was no intended link between them, and their approaches to lexicography were markedly different. They were:

- *Dictionaries in Asia*. A gathering organized by the Language Centre of the Hong Kong University of Science and Technology, and held at its campus at Clearwater Bay in Kowloon. During the conference proper, attention focused in the main on alphabetic lexicography and analogous formats, and on the closing day members inaugurated the Asian Association for Lexicography (ASIALEX). In addition to a large attendance from many parts of Asia, representatives and other well-wishers were present from four already established continental organizations: the Dictionary Society of North America (DSNA), the European Association for Lexicography (EURALEX), the African Association for Lexicography (AFRILEX), and the Australian Association for Lexicography (AUSTRALEX). I attended as publications consultant.
- *Language Study and the Thesaurus in the World*. This gathering, organized by the Kokuritu Kokugo Kenkyuzyo (National Language Research Institute) in Tokyo, was held at the National Olympics Memorial Youth Center and focused mainly on thematic lexicography – and is as far as I know the first conference in the world to do so. I was present as a guest speaker, invited to describe the nature, origin, and compilation of my *Longman Lexicon* (1981; see also 1986a, 1998b).

Despite the differences between the two (or rather because of them), the conferences proved to be valuable complementary events for those able to attend both. Because of such meetings, in Asia as elsewhere, it has now become possible to look forward to a conference devoted to 'world lexicography' (on whatever continent it may be held), that will seek to cover as wide a sampling as possible from our immense international heritage of reference materials, in all their formats,

genres, rationales, writing systems, technologies, languages of origin, and languages of translation. It would be particularly good if the four continental -lexes and the DSNA could consider jointly sponsoring such a 'Globalex' development.

Asia and its Languages

Hong Kong and Tokyo, the venues of the conferences in question, are relatively close together, in a part of the world once Eurocentrically known in English as 'the Far East' and in French as l'Extrême-Orient. Two decades ago such terms were internationally commonplace, and they are certainly still with us, but on the edge of a new century they have an archaic feel about them, especially as the region is now more commonly and straightforwardly referred to, in English and especially in the media, as 'East Asia'.

It is intriguing to consider what the participants might have thought and felt if the conferences had been held not in 'East Asia' but, say, in Ankara and Beirut (located in the former 'Near East': a label now virtually extinct), or in Damascus and Teheran (both still located in the 'Middle East' but increasingly also in 'West Asia'), or in Tashkent and Samarkand (formerly and still safe in 'Central Asia'), or in Karachi and Calcutta (formerly in 'the Indian subcontinent' but more recently in 'South Asia' or, on occasion, simply in 'the Subcontinent'), or in Saigon and Manila (both located in a hyphenated 'South-East Asia'). But wherever the conferences might have been situated and however they might have been nuanced in geocultural terms, they are significant for one reason above all others: that until now, Arabs, Iranians, and Indians, for example, have not been in the habit of discussing lexicography with Chinese, Koreans, and Japanese – except perhaps in such venues as the Dictionary Research Centre of the University of Exeter in England, where for years lexicographers from many backgrounds have been meeting. But if they have been talking to each other in such places, it has been more as lexicographers at large than as Asian lexicographers.

Asia is old and immense, but this lexical club is very new, and its members are so thin on the ground and many of the issues that concern them are so novel that much of the continent may remain unrepresented in their ranks for some time to come. To see why this is so, it may make sense here to consider the origins and nature of some of the names and concepts involved and at least raise the question of whether lexicography in Asia is – or can be? – based on any kind of unified – or unifiable? – sociolinguistic culture.

In looking for the origins of 'Asia' as both word and concept, one must turn to the Greeks, a people who have been squeezed for

several millennia between two cultural tectonic plates – so much so indeed that Herodotus wrote the first universal 'history' in terms of war between East and West: first between the Greeks and Trojans (who were in fact close neighbours), then between the Greeks and Persians (who were much more widely separated). The Greeks had a word for both the subject of this book (*lexikographia*) and the region in question (Asia), but they also had two – now largely forgotten – original senses for Asia, one of them mythological the other geographical. In mythology, Asia was a titan and the mother of titans. One of her sons was Atlas (who has served as an eponym three times over: for an everyday work of reference, for a range of mountains in North Africa, and for the Atlantic Ocean), another was Prometheus (a symbol of human, and later Western, arrogance in challenging the fundamental forces of nature and being punished for it). In geographical terms, however, Asia had more modest beginnings, as a small city on the eastern shore of the Aegean Sea, inland from which lay an uncertainly large region known as Anatolia ('Land of the Rising Sun'). The later Latin equivalent of this name, *oriens* ('rising'), is the literal root of the mysterious 'Orient'.

By the time the Romans took over the eastern Mediterranean, the area of coverage of 'Asia' had become properly titanic. Both the city of Asia and Anatolia had by then been lumped together in a west-facing peninsula which the Romans called in Latin Asia Minor ('Lesser Asia'), in contrast to a vast and conceptually shapeless Asia Major ('Greater Asia') that was now known to stretch all the way to Sinae and Serica (their names for parts of China). In later centuries, perhaps under pressure from inquisitive Europeans, the inhabitants of this huge expanse came to perceive themselves as inhabiting a single region from Mediterranean to Pacific, although in strictly geographical terms the landmass in question is a single 'Eurasia' rather than a smaller 'Europe' to the west and a larger 'Asia' to the east, Europe being in effect an Atlantic equivalent of the Indian subcontinent. The division of this single hard-to-encompass landmass into two such unequal continents is topographically illogical, but the distinction does make a kind of psychological sense. As the Palestinian-American literary critic Edward Said (1978:2-3) has observed, regarding European views of what lies to the east:

> Orientalism is a style of thought based upon an ontological and epistemological distinction between "the Orient" and (most of the time) "the Occident". Thus, a very large mass of [European] writers, among whom are poets, novelists, philosophers, political theorists, economists, and imperial administrators, have accepted the basic distinction of East and West as the starting point for

elaborate theories, epics, novels, social descriptions, and political accounts concerning the Orient, its peoples, customs, "mind", destiny, and so on.

The whole matter is both culturally and emotionally charged, as a consequence of which a range of European expressions that include the English terms Asiatic, Oriental, and Eastern have acquired over time certain suspect connotations, as a consequence of which the phrases 'Oriental lexicography', 'Asiatic lexicography', and 'Eastern lexicography' are impossible. At the end of the twentieth century, the only viable term to match such phrases as 'European lexicography' and '(North) American lexicography' is 'Asian lexicography', because out of the set of relevant adjectives only Asian is neutral in terms of international pride and prejudice.

However, if denomination is odd, delimitation is odder, for where do Asia, its languages, and its lexicography begin and end? Arabia, India, China, and Japan (among other territories) are unequivocally 'Asian' and so therefore are their languages, but what does one do with Russia, an entity that extends over vast tracts of North-Eastern Europe and North and East Asia? Even makers of post-Soviet atlases are chary about the geopolitics of Russia, as for example the editors of the Reader's Digest Illustrated Atlas of the World (UK: 1997), who divide the 'old world' into: Northern Europe; Southern Europe; Central Europe; Russia and its Western Neighbours; Central and Eastern Asia; South-East Asia, the Middle East and the Gulf, the Indian Subcontinent and its Neighbours; and Oceania.

The Digest may dodge this issue, but we should not, and can reasonably ask: Is Russian to be classed as an Asian language and, if so, should there have been a place for it and its lexicography both at the Hong Kong conference and in a book whose content derives largely from that conference? Or should Russian and its dictionaries be considered no more than the overland extension of a European culture into Asia, much as Dutch and its lexicography for a time extended by sea to what is now Indonesia (as Soekemi notes in his paper) and to Japan (as Yamada and Komuro point out in theirs)? One might say 'yes', categorizing Russian as alien despite the size of the territory involved and the obvious need to list indigenous Siberian languages that co-exist with Russian as unassailably Asian – along with any work done on them by Russian-speaking lexicographers.

There are also thought-provoking parallels elsewhere. Arabic, for example, is manifestly an Asian language, but is every bit as bicontinental as Russian, having ancient extensions into North and East Africa. It would be impossible to exclude Arabic from any

comprehensive lexicographical discussion of 'languages of Africa' (as opposed to, say, 'African languages', if that formulation is to be reserved for the ultimately indigenous). But the time is likely to come – and probably quite soon – when Russian cannot be excluded from discussions of language and lexicography in Asia; it is after all as firmly established to the north of India and China and the west of Japan as Arabic is established south of the Mediterranean.

If the Russian and Arabic languages are bicontinental (and therefore the concern alike of EURALEX, AFRILEX, and ASIALEX), what can one say about omnicontinental English? Its inroads into Asia are so marked that no fewer than five papers in this volume relate to its Asian roles and to Asian dictionaries and dictionary research associated with teaching, learning, and using it: Lu Gusun on bilingual Chinese/English lexicography, Li Lan on dictionaries as aids to the learning of English in China; Jacqueline Lam Kam-mei on a glossary to help (especially Hong Kong) students with computer science texts in English; Ilan Kernerman on semi-bilingualized English learners' dictionaries in Asia and elsewhere; and Shigeru Yamada and Yuri Komuro on the origin and immense educational and commercial success of Japanese English learners' dictionaries. Reiko Takeda even turns the tables entirely, and as an Asian researcher into European lexicography reports on lesser-known aspects of the lexicography of English not in Asia at all but in England in the fifteenth century. Sauce for the goose....

In addition, English enters obliquely into other papers, as for example where Lee Sangsup, discussing the Dictionary of Korean, indicates the key role played by the *Oxford English Dictionary* as a model, and where Arvind Kumar compares two Indian thesauruses (one ancient and in Sanskrit, the other recent and in Hindi) with Roget, an originally nineteenth-century English-language work which he treats as a touchstone for the genre.

Finally, the medium of the present collection of papers is uniformly English, and it is hard to imagine any other language that could have served to weave together such varied strands as these. [It is noteworthy, however, that at the Hong Kong conference papers could be and were delivered in Mandarin or English, and at the Tokyo conference in Japanese, Mandarin, or English. How many other languages might be deemed to merit the same treatment at a comprehensively pan-Asian gathering?] English is here at least 'a language of Asia' if not (yet) 'an Asian language', although already these days – safely beyond lexicographical circles – it is often referred to as just that, for at least the following five reasons (see also McArthur, 1998a):

- English has been used widely in Asia for as long as it has been used in the Americas (that is, since the seventeenth century), and by considerable numbers of people, especially in South and South-East Asia.
- In recent years (much to the surprise of many of its own inhabitants), Australia has been 're-branded' as Asian rather than Australasian (in origin a Latinate term meaning 'South Asian'), and is often so listed in international periodicals (especially for economic and financial purposes). Thus, Philip Bowring comments in the article 'Australia: Regional Leader or Orphan Adrift?' (*International Herald Tribune*, 1 October 1992): "Australia and its neighbors have to recognize that Asia is simply a geographical definition, and for practical purposes Australia is part of it." The national language of Australia is English, and many East Asians send their children there for educational reasons that pre-eminently include improving their English – in the process of course Asianizing it further.
- It is the language that Asians need not only for purposes of communicating with other continents and engaging in worldwide scientific and other activities whose dominant medium is English, but also (pre-eminently?) for intra-Asian communication: Thais with Japanese, Koreans with Indonesians, Filipinos with Asian Russians, Chinese with Pakistanis, Gulf Arabs with Indians.
- It has highly significant and long-standing official roles within Asia. Thus, in the Philippines it is co-official with Filipino (Pilipino, Tagalog); in Singapore it is one of four official languages, alongside Mandarin, Malay, and Tamil; in Hong Kong (now integrated into China as a special administrative region) it is a key everyday language of business and education alongside Cantonese and increasingly Mandarin/Putonghua; and, momentously, it has in India three distinct legislated roles, as the associate official language (Hindi being official), as a national language (alongside Bengali, Gujerati, Tamil, and other state languages), and as the sole official language of eight Union territories (including Delhi, Nagaland, and Pondicherry) – all additional to its use as a medium of education, business, and – famously – 'a window on the world'.
- It is the working language of ASEAN (the Association of South-East Asian Nations), a regional organization founded in 1967 for economic, social, and cultural co-operation, whose members are currently Brunei, Burma/Myanmar, Indonesia, Laos, Malaysia, the Philippines, Singapore, Thailand, and Vietnam.

There have always been world languages, in the sense that the language of culturally, economically, and militarily powerful communities have impacted on the known worlds of their time and place. Asia has had its share of such languages, which include Sanskrit (brought to our attention here by Arvind Kumar), Persian (whose lexicography is discussed by Ahmad Taherian), Malay (covered by both Nur Ida Ramli of Malaysia and Soekemi of Indonesia), and Classical Chinese (with its influence not only in the Middle Kingdom but also in Korea, Japan, and Indo-China, and the concern here particularly of Lu Gusun and Li Lan). English differs from other world languages only – yet it is an overwhelming 'only' – is that its world is the entire planet, its speakers are the most widely distributed and the most ethnoculturally varied ever, and their numbers increase by the year. Demographically the only Asian rival to English – and it is a powerful 'only' – is Mandarin/Putonghua, which may not be spoken an written by all Chinese but is for all of them the touchstone of linguistic excellence. Inevitably, these two giants among languages will have much to do with each other in the coming century, including in lexicographical terms.

Asia and its Lexicographies

The word *lexicography* has the same Greco-Latin pedigree and structure as *biology*, *astronomy*, *osteopathy*, *phylogeny*, and other widely-used names for academic activities and subjects. As such, it is part of what the American dictionary editor Philip Gove (1963:7a) has called *International Scientific Vocabulary* (ISV). Although Gove has for his purposes treated such words as restricted to English, they are in reality 'translinguistic': they operate (with appropriate phonological and orthographic adaptations) in many languages that serve as mediums for education, culture, science, and technology: not only in, say, Russian, Spanish, Swedish, or English (European languages traditionally receptive to Classical word elements and patterns) but also in Japanese, Malay, Tagalog/Pilipino, and other Asian languages (to which they are often transmitted through modern European languages). In effect, such words have no ultimate canonical forms: their embodiments in any language are all equally valid as citation forms. Because no language-specific version of such a term has primacy, an ISV word is truly international, transcending individual languages, a point which lexicographers worldwide have yet to come to terms with. ISV words would appear to be – both in their own right and through any loan translations that may have been made from them – the most universal set of lexical items on earth.

Not all such Greco-Latinisms are however equally 'scientific'. On the one hand, such terms as *biology* and *physics*, which serve to label branches of science itself, are manifestly part of an originally European endeavour that has in the last century or so become fully cosmopolitan, but on the other hand terms such as *lexicography* and *psychotherapy* refer to social and professional activities, not to 'hard' sciences, and other terms still, such as *necromancy* and *anthropophagy*, label activities that are not at all scientific – although scientists and scholars may take an interest in them, and are likely to be prominent among the few who use the terms. All such words are however at their very least specialist terms, for which reason (*pace* Gove) I prefer to interpret 'ISV' as 'International Specialist Vocabulary' (cf. Kirkness, 1997, who identifies them more particularly as 'Euroclassicisms').

Because the strictly scientific ISV terms are unitarian and now cosmopolitan, one cannot treat a 'biology in Europe' and a 'biology in Asia' as being different in kind: they are the same thing pursued in different locales. Matters are not so clear, however, for such items as 'lexicography' and 'psychotherapy'. Do such terms mean something essentially European that is spreading throughout the world, as biology has done, and may at length have the same comprehensive status as biology, or do they – actually or potentially – refer to more general, more culturally varied matters, so that for example traditional, millennia-old Chinese lexicography might differ markedly from centuries-old British, American, and French lexicography yet be recognised everywhere instantly and fully as equally lexicographical? Indeed, are we seeing a kind of hybridization under way, where aspects of Western lexicography combine usefully with aspects of Eastern lexicography? An example might be present-day bilingual English-Chinese dictionaries such as Lu Gusun and Li Lan discuss, where the English-Chinese section has an A-Z ordering of lemmata and the Chinese-English section is traditionally ordered according to a conventional listing of the strokes of which Chinese characters are composed.

The discussion need not however end there. The condition of lexicography in Asia may be closer to that of a comparably culture-laden activity that has travelled the other way, from East to West, as for example yoga in Europe and America. Such a comparison leaps to my mind because intermittently over some thirty years I have attended (and spoken at) conventions of yoga teachers and students in the United Kingdom, have written two books about India, yoga, Indian philosophy, and their spread worldwide (McArthur 1986b/c), and at one time, for several years, edited the journal of an association

which was concerned (in effect) with indigenizing yoga in Scotland: a process that included the accreditation of local teachers of yoga by the Scottish Sports Council – an example of culture clash if ever there was one. During that period such concepts as *asana* (a physical pose), *dhyana* (meditation), and *mantra* (a repeated sound serving to focus the mind) have gone from being generally regarded in the West as eccentrically and exotically Eastern to being about as common and virtually as unremarked as the terminology of golf.

The organization of conferences about dictionaries in Asia and conventions for yoga in Europe can be perceived as a vast process of cultural exchange. In such an exchange, questions like the following arise: In their encounter with yoga in Europe and other non-Asian locales, should non-Asians regard it as 'essentially' Eastern and therefore forever 'other', no matter how strong the effort to naturalize it, or do they absorb and extend the subject so as to incorporate comparable practices among Europeans and others into a more inclusive view of yoga (that may also include such other Asian philosophical-cum-physical systems as tai-chi, Zen, and Sufism)? Comparably, in their encounter with lexicography, should Asians (and others) regard it as 'essentially' Western and focused on 'dictionaries' (understood in an A-Z sense), and so forever to some degree 'other', or do they absorb and extend the subject so as to include comparable practices among Asians within what can become a more inclusive view of lexicography?

There may be no neat and tidy answer to such questions, but the papers in this volume, it seems to me, in addition to their valuable immediate aims contain the seeds of studies, both diachronic and synchronic, that could be immensely helpful in placing lexicography in a geographically wider and chronologically deeper frame of reference. Let me mention here only three areas that belong very much to Asia, about which one day I hope to know more:

(1) Lexicophony

At present I can think of no better name for something which Arvind Kumar discusses in his paper: a tradition probably over three millennia old in South Asia, in which the brahmins of Vedic India orally and aurally encoded in Sanskrit verse not only religious but also lexical information, to be recited as the need for consultation and instruction arose. Such pre-literate artifacts have been the lexicographical equivalents of Homer's Iliad or, in more local terms, of Vyasa's Mahabharata.

(2) Bilingual word lists

Such lists, which recur throughout this collection in relation to the present-day bilingual-dictionary industry, had their origins in West Asia. Some three millennia ago in Mesopotamia, Semitic-speaking scribes in the city state of Akkad (and later in Babylon and Nineveh), borrowed cuneiform writing from their southern neighbours in Sumer, the creators of the world's earliest known writing system (cf. McArthur 1986a, Chs. 4-5). In the process, the formulated Semitic equivalents for Sumerian originals, creating the first lists of language equivalents set side by side in columns on clay tablets.

(3) Ideographic lexicography

First formulated in China over two millennia ago, the signs in such a system in the main represent concepts rather than sounds and words as such: that is, they are ideographic rather than phonographic and logographic. As such, they are in principle as detachable from the language to which they initially relate as alphabetic letters have been, as demonstrated for example by their adoption to serve Japanese, which is structurally entirely different from Chinese. In essence, such a system is a (successful and extensive) ancient cousin of the (failed and more limited) philosophical language with which Bishop John Wilkins experimented in seventeenth-century England, a quest for a conceptual 'language' that in due course inspired Roget when he created his *Thesaurus* in the mid-nineteenth century.

The prospects are endless and enticing, and the present collection of papers already provides a varied spread of approaches, perspectives, descriptions, and proposals ranging from the remotest times to the day after tomorrow, contributing significantly to an academic discipline which Reinhard Hartmann and I call 'reference science' (see McArthur, 1998c). It is refreshing that the collection covers several generations of scholars, all of whom I wish to thank here for their collaboration in making the volume possible; I am immensely pleased to have been part of its creation. *Lexicography in Asia*, it seems to me, is a noteworthy step towards the collaborative formulation of a single over-arching typology for all works of lexical reference, wherever and whenever compiled, by whomever and in whatever language, and through whatever compiling, recording, and presentational technology.

References

Gove, Philip (ed.). 1963. *Webster's Third International Dictionary*. Springfield, Mass: G. & C. Merriam.

Kirkness, Alan. 1997. Eurolatin and English today. *English Today*, 13.1:3-8.

McArthur, Tom. 1981. *Longman Lexicon of Contemporary English*. Harlow: Longman.

McArthur, Tom. 1986a. *Worlds of Reference: Language, Lexicography and Learning from the Clay Tablet to the Computer*. Cambridge: Cambridge University Press.

McArthur, Tom. 1986b. *Yoga and the Bhagavad-Gita*. Wellingborough: Thorsons/HarperCollins.

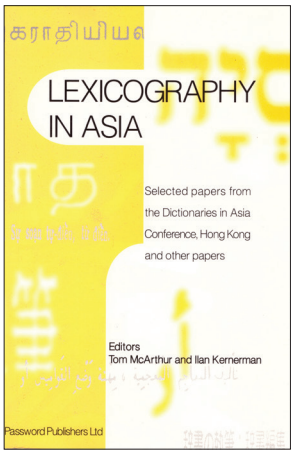
McArthur, Tom. 1986c. *Understanding Yoga: A Thematic Companion to Yoga and Indian Philosophy*. Wellingborough: Thorsons/HarperCollins.

McArthur, Tom. 1998a. *The English Languages*. Cambridge: Cambridge University Press.

McArthur, Tom. 1998b. A mutually defining circle of words: some reflections on the making of the *Longman Lexicon of Contemporary English*. Ch. 14 in *Living Words: Language, Lexicography, and the Knowledge Revolution*. Exeter: University of Exeter Press.

McArthur, Tom. 1998c. What is 'reference science'? *Lexicon*, 28. Tokyo: Iwasaki Linguistic Circle. 135-140.

Said, Edward. 1978. *Orientalism: Western Conceptions of the Orient*. London: Routledge & Kegan Paul.



NexusLinguarum: European network for Web-centred linguistic data science

Jorge Gracia

The kickoff meeting of the newly created 'European network for Web-centred linguistic data science' (*NexusLinguarum* in its short name) took place in Brussels, Belgium, on 28 October 2019. The meeting brought together representatives of the 33 countries constituting the initial network to discuss the objectives, the plans for implementing the different networking tools, and the scope and goals of the different working groups, as well as to elect the action's management board. The initial group consisted of a broad network of experts from different areas, like computer science, semantic web, artificial intelligence, linguistics, humanities, etc. That was the beginning of an exciting journey towards building a common ecosystem to support research on *linguistic data science* in a Web-centred context.

We understand linguistic data science as a subfield of the growing data science field that focuses on the systematic analysis and study of the structure and properties of linguistic data at a large scale, along with methods and techniques to extract new knowledge and insights from it. Linguistic data science is concerned with providing a formal basis to the analysis, representation, integration and exploitation of linguistic data for language analysis (e.g. syntax, morphology, terminology, etc.) and language applications (e.g. machine translation, speech recognition, sentiment analysis, etc.).

NexusLinguarum will last for four years and is funded by the European Cooperation in Science and Technology (COST) organization, which supports such highly competitive projects (COST Actions) by financing research networks on emerging issues, through mechanisms such as research visits, organization of congresses, scientific meetings, summer schools, etc.

To enable the study of linguistic data in the most productive and efficient ways, the NexusLinguarum COST Action is set to enhance the construction of an ecosystem of multilingual and semantically interoperable linguistic data at the scale of the Web. To this end, methods and techniques of the Semantic Web, Natural Language Processing and Language Resources are studied and combined. Such an ecosystem could reduce language barriers in Europe (and eventually beyond) and favour both electronic commerce and cultural

Jorge Gracia is the Chair of the NexusLinguarum 'European network for Web-centred linguistic data science' COST Action. He works as assistant professor at the Department of Computer Science and Systems Engineering (University of Zaragoza, Spain) as a member of the Aragon Institute of Engineering Research (I3A) and of the Distributed Information Systems research group. His main research interests are Semantic Web, Ontology Matching, Multilingual Web of Data, Query Interpretation, and Web Intelligence, and his recent work focuses on linked data-based lexicography as well as on methods and techniques for cross-lingual linking and information access. http://jogracia.url.ph/web/

Participants at NexusLinguarum kickoff meeting, Brussels, 28 October 2019

exchange between countries with different languages. Another objective is to support minority languages whose technological support is currently limited.

Through the study of Web-centred linguistic data science, we will be able to better understand the nature of language, through innovative methods for the representation, integration and comparison of linguistic data. Furthermore, since language is the medium in which human knowledge is transmitted, this field has the potential to decisively influence studies that use natural language for knowledge sharing, as is the case of the humanities, the legal domain, journalism, social sciences, etc.

Some of the main research coordination objectives of NexusLinguarum are to:

- propose, agree upon and disseminate best practices and standards for linking data and services across languages;
- organise activities to foster collaboration and communication across communities, such as scientific workshops involving broader communities to reach agreement on best practices;
- collect and analyse relevant use cases for linguistic data science and develop prototypes and demonstrators that will address some prototypical cases.

Furthermore, we plan to work out a curriculum for a Europe-wide master degree that the participating institutions could adopt to train a new generation of researchers in the area, thus introducing linguistic data science in a cross-discipline academic infrastructure.

Currently we count on participants from 42 countries (37 COST

Countries, 3 Near Neighbour Countries, and 2 International Partner Countries). So far, 137 members have joined the different working groups (WGs), a number which is steadily growing since the network is still open to new participants.

NexusLinguarum is organised in five working groups, four technical ones and a one for management activities:

WG1 – Linked data-based language resources. This WG lays the foundations to develop best practices for the evolution, creation, improvement, diagnosis, repair and enrichment of linguistic linked open data (LLOD) resources and value chains.

WG2 – Linked data-aware NLP services. This WG focuses on the application of linguistic data science methods including linked data to enrich NLP tasks in order to take advantage of the growing amount of linguistic (open) data available on the Web.

WG3 – Support for linguistic data science. This WG aims to foster the study of linguistic data by following data analytic techniques at a large scale in combination with LLOD and linked data-aware NLP techniques

WG4 – Use cases and applications. This WG focuses on studying use cases and practical applications of the relevant technologies involved in the Action.

WG5 – Management and dissemination. This WG takes care of the measures to be taken to ensure the creation of added value of the Action as a whole, to ensure its maximum visibility, and to monitor the cross-WG activities.

All these WGs have already started their activities, although they are still in initial phases. One of the fist outcomes to be delivered by NexusLinguarum is a study of use case definitions, which is currently under development by WG4. The following use cases are being analysed currently: Humanities and Social Sciences, Linguistics (Media and Social Media, and Language Acquisition), Life Sciences, and Technology (Cybersecurity and FinTech). The idea is to analyse the current state-of-the-art on each of these topics, analyse their needs and challenges, and determine the techniques and ideas of linguistic data science that might improve them, in close collaboration with the other WGs.

The other three technical WGs are also conducting initial surveys and analysing related projects and initiatives to set the ground for further development. Collaboration with other projects and initiatives are already on course, for instance with the W3C Linked Data for

Language Technologies community group in relation to the ongoing discussion towards a consolidated Linked Open Data vocabulary for linguistic annotations, in the context of WG1.

NexusLinguarum has already organised two face-to-face meetings of its Management Committee (MC): in Brussels in October 2019 (kickoff meeting), and in Prague (Czech Republic) in January 2020 collocated with the first WG meetings. The next MC + WGs meeting is due to take place in October 2020 in Lisbon (Portugal). In addition, regular teleconferences take place to enable and monitor the scientific progresses of the different WGs, and a number of training schools and scientific events are planned for 2021.

More information can be found at https://nexuslinguarum.eu/. New participants can join the network through this registration form https://forms.gle/ZML87XLHnxXdPrbh6.

NexusLinguarum Core group

Chair. Jorge Gracia
University of Zaragoza, Spain

Vice-chair. John McCrae
National University of Ireland, Galway, Ireland

Grant Holder Scientific representative. Elena Montiel-Ponsoda
Universidad Politécnica de Madrid, Spain

Science Communication manager. Thierry Declerck
DFKI, Germany

Short Term Scientific Missions coordinator. Penny Labropoulou
Athena Research Center, Greece

Inclusiveness Target Countries Conference Grant coordinator. Vojtech Svatek
University of Economics, Prague, Czech Republic

WG1 leader. Milan Dojchinovski
Czech Technical University in Prague, Czech Republic

WG1 co-leader. Julia Bosque-Gil
University of Zaragoza, Spain

WG2 leader. Marieke van Erp
KNAW Humanities Cluster, The Netherlands

WG3 leader. Dagmar Gromann
University of Vienna, Austria

WG3 co-leader. Amaryllis Mavragani
University of Stirling, UK

WG4 leader. Sara Carvalho
University of Aveiro, Portugal

WG4 co-leader. Ilan Kernerman
K Dictionaries, Israel

An overview of NexusLinguarum Working Groups

The Action is composed of five working groups (WGs) interoperating and providing mutual feedback. They cover, in a bottom-up approach, the technical and infrastructural groundings needed to attain the objectives of the Action along with a range of use cases and applications. In addition to their own tasks, all WGs participate in preparing cross-group dissemination activities. The scientific work is carried out over four years through workshops and other meetings as well as remote cooperation through electronic communication means (email, teleconference, etc).

* Short-Term Scientific Missions (STSMs) organized within each WG, to promote synergies and maximize cooperation, and International Training Schools (ITCs), are not included in this overview.

WG1 – Linked data-based language resources

Objective. Lay the foundations and develop best practices for the evolution, creation, improvement, diagnosis, repair and enrichment of linguistic linked open data (LLOD) resources and value chains.

Tasks

Task 1.1: LLOD modelling. Update, extension and improvement of existing models for representing linguistic information as linked data (LD, e.g. lemon-ontolex, LexInfo, OLiA, NIF, etc).

Task 1.2: Creation and evolution of LLOD resources in a distributed and collaborative setting. Analysis of new approaches for the distribution and collaborative creation and extension of linguistic resources to facilitate the extension of existing resources and their publication as LD.

Task 1.3: Cross-lingual data interlinking, access and retrieval in LLOD. Studying novel (semi-)automatic methods aimed to increase the interlinking across LLOD datasets, and methods and techniques for accessing and exploiting data on the Web across different languages, based on the use of linguistic linked data (LLD).

Task 1.4: Improving and monitoring quality of LLOD sources. New techniques to monitor and improve the quality of LLOD sources by novel approaches for diagnosis and repair and new measures allowing to monitor and assess the quality of such sources, as well as analyzing semi-automatic and automatic methods for validating LD and cross-resource links via collaborative strategies.

Task 1.5: Development of the LLOD cloud for under-resourced languages and domains. Analysis and development of language technologies serving under-resourced languages and domains in the LLOD cloud.

Deliverables

- Scientific papers on linguistic linked data and language resources
- Training school on linguistic linked data
- Guidelines and best practices on the generation, interlinking, publication and validation of LLOD (new and update of existing ones)
- Policy brief about the inclusion of data from under-resourced languages
- Intermediate and final activity reports

WG2 – LD-aware Natural Language Processing services

Objective. Applying LLD to enrich NLP tasks taking advantage of the growing amount of linguistic linked (open) data available on the Web.

Tasks

Task 2.1: LLD in Knowledge Extraction. Analysis of large-scale integrated linguistic and semantic knowledge for multiple domains and languages to open up new possibilities in taxonomy and ontology-based information extraction.

Task 2.2: LLD in Machine Translation. Incorporating multilingual LLD in machine translation (MT), both syntactically (e.g. using dependency relations) and semantically (using lexical semantics), as well as exploring LD for expressing translation workflow metadata to improve MT output.

Task 2.3: LLD in Multilingual Question Answering. Examining how lexical knowledge required by QA systems can be extracted from LLD.

Task 2.4: LLD in Word Sense Disambiguation and Entity Linking. Studying the impact of LLD on disambiguation in multilingual content processing, such as for the translation of terms and idioms in user-generated content to detect words or phrases used in a potentially offensive manner.

Task 2.5: LLD in Terminology and Knowledge Management. Cross-disciplinary research on applying LLD in multilingual terminology and knowledge resource management, including their linking, merging with enterprise (proprietary) resources, and publishing on the Web as part of a global ecosystem of multilingual data.

Deliverables

- Scientific papers on linked data-aware NLP
- Guidelines and best practices on LLOD and NLP (new and update of existing ones)
- Intermediate and final activity reports

WG3 – Support for linguistic data science

Objective. Understand linguistic data by following data analytic techniques at a large scale in combination with LLOD and LD-aware NLP techniques, covering scalability issues in the study of multilingual linguistic data given the fact that datasets are rapidly growing in size, leading to huge amounts of data on the Web (*big data*).

Tasks

Task 3.1: Big data and linguistic information. Studying big data sources and state of the art statistical analysis in combination with LLOD to better understand language, also considering visual analytics, having an impact on the linguistics aspect in all sub-domains, from typology to syntax to comparative linguistics.

Task 3.2: Deep learning and neural approaches for linguistic data. Study the effective use of deep learning in understanding the specificities of linguistic data in a big data context, to be better exploited and combined with LD mechanisms.

Task 3.3: Linking structured multilingual language data across linguistic description levels. Explore how diverse data regarding phonology, morphology and lexicon that is spread across datasets of varying extent, quality and format, can be described, stored and accessed uniformly.

Task 3.4: Multidimensional linguistic data. Link language resources across various dimensions (such as time axis, style, genre, media, etc) to facilitate diachronic and sociolinguistic interoperable research.

Task 3.5: Education in linked data science. Develop a curriculum for linguistic data science in a cross-discipline academic infrastructure for a Europe-wide master's degree for training a new generation of researchers.

Deliverables

- Scientific papers on techniques that support linguistic data science
- Academic curriculum for the studies on linguistic data science
- Training school on linguistic data science
- Intermediate and final activity reports

WG4 – Use cases and applications

Objective. Exploring practical use cases and applications of the relevant methodologies and technologies involved in the Action.

Tasks

Task 4.1: Use cases in legal domain. Explore legal terminology and translation, and identify use cases ranging from unique identification and re-use of licenses at a Web-scale to assisted translation based on semantic annotations.

Task 4.2: Use cases in humanities and social sciences. Study how linguistic data science can deeply influence studies in the humanities and social sciences, allowing us to trace the history of the peoples of the world, understand literature and culture in new ways. and predict and analyze social trends.

UC4.2.1 – Use Case in Humanities

UC4.2.2 – Use Case in Social Sciences

Task 4.3: Use cases in linguistics. Investigate how linguistic data science and richer understanding of language can benefit research in linguistics (lexicography, typology, syntax, comparative linguistics, etc).

UC4.3.1 – Use Case in Media and Social Media

UC4.3.2 – Use Case in Language Acquisition

Task 4.4: Use cases in life sciences. Exploit structured linguistic information in the process of discovering hidden facts out of textual data, expanding text analytics techniques applied in biomedicine and other life sciences.

Task 4.5: Use cases in technology. Look into incorporating text analytics into advanced technological systems, such as for sentiment analysis and fake news detection, by developing customized domain-specific models.

UC4.5.1 – Use Case in Cybersecurity

UC4.5.2 – Use Case in Fintech

Deliverables

- Scientific papers on in-use applications of LLOD, NLP and linguistic big data
- Report describing requirements elicitation and use cases definitions
- Intermediate and final activity reports

WG 5 - Management and dissemination

Objective. Manage the measures taken to ensure the creation of added value of the Action as a whole and its optimal visibility, and monitor the cross-WG activities.

Tasks

Task 5.1: Management. Day-to-day management and administrative coordination.

Task 5.2: Cross-working group communication. Monitor communication across the different WG activities, especially in the MC and SC meetings.

Task 5.3: Capacity building. Coordinate the Short-Term Scientific Missions (STSMs) and elaborate a plan for the training schools, datathons and hackathons to be developed by the different WGs.

Task 5.4: External communication. Coordinate the Action's external communication including its website, social media streams, press releases, and the organization of (and participation in) events and workshops, etc.

Task 5.5: Scientific publications strategy. Monitor the scholarly publications produced in the Action, define a strategy for journal special issues, and provide the means to document them in a central repository (e.g. Zenodo).

Deliverables

- Roadmap document containing a common research agenda for linguistic data science
- Generated dissemination materials (blog entries, short reports of the different Action's events, press releases, etc.)
- Policy brief about the social and technological interest of linguistic data science

Applying the OntoLex-*lemon* lexicography module to K Dictionaries' multilingual data

Dorielle Lonke and Julia Bosque-Gil

Introduction

In recent years, K Dictionaries (KD) has been working on representing its multilingual lexicographic resources in Linked Data (LD) format (Bizer et al. 2011), adhering to the RDF Ontolex-*lemon* model for lexical resources (cf. McCrae et al. 2017). The latest iteration, which began in 2019, follows two previous rounds, and focuses on adapting the newly added *lexicog* module[1], which addresses the need to preserve the original structure of a lexicographic dataset. In cooperation with Julia Bosque-Gil and Jorge Gracia from University of Zaragoza, the KD team has been involved in implementing *lexicog* in the Global series of multi-language, multi-layer resources (cf. Bosque-Gil et al. 2019). The contributions of this effort are twofold: primarily, offering the first use case of real-world lexicographic data represented entirely in LD format; secondly, addressing previously reported limitations of the Ontolex-*lemon* model and offering new solutions, allowing a more agnostic approach to a graph representation of lexicographic data.

The first two iterations of this data conversion were carried out under a strict principle of 'round-tripping', wherein every element in the original XML structure must be accounted for within the new corresponding RDF structure in terms of hierarchy and order, thus enabling one-to-one reconstruction of the lexicographic resource from RDF back to XML (Klimek and Brümmer 2015, Bosque-Gil et al. 2016). In addition to obtaining perfect matching back and forth between the two data forms, this principle served to assure perfect validation of the data conversion from either format to the other. However, this principle was abandoned in the latest iteration due to the understanding that each type of data format should be applied freely and fully in accordance with its own nature and not be restricted by characteristics of the other. In consequence, the alternative validation process provided in this iteration introduced a new incremental approach, which actually proved to be more efficient in validation and error catching. In section 2 we describe the revised

1 https://www.w3.org/2019/09/lexicog

Dorielle Lonke has been working at K Dictionaries for the past three years, involved in different projects pertaining to linguistic data design and managment, and leading conversion from XML to RDF. Currently she is graduating in Linguistics and completing her Philosophy degree at Tel Aviv University. Her main fields of interest include language technologies, computational processes in natural languages and ontological representation of data.
dorielle@kdictionaries.com

pipeline which features the incremental approach. The modelling improvements that were performed as part of the new *lexicog* module adaptation are presented in section 3, and section 4 reports on the ensuing validation process, including queries and results, followed by a brief summary of the process in section 5.

The pipeline

The incremental approach saw a step-by-step modelling process, in which each component was first modelled to fit the *lexicog* module, and then manually described as RDF triples serialized in Turtle format (TTL)[2]. Based on the manual description, a generalization was applied in the automatic conversion process, allowing the simultaneous conversion of numerous entries in the dataset. The final stage of this process was to upload the dataset onto a triple store, enabling querying and detection of errors, and fixing such errors in the conversion. The process was repeated for each cluster of components. Instead of committing to a restrictive structure that requires a one-to-one conversion, this has enabled a looser, more flexible workflow that facilitated casting off excessive information that encumbers the model, while still fitting each lexicographic component with its ontology counterpart and retaining the original hierarchy and order of the lexical data where necessary.

Progressing incrementally has not only enabled constant validation and error management, but also allowed for an adaptation period, during which the process of writing queries for validation shed light on the model and methods of improvement. Taking into account input from partners and collaborators who have been experimenting with the RDF data, particularly as part of our work in the H2020 Lynx project[3], we were able to improve the queries and iteratively modify the model so that the results optimally represent the needs of the users. This working method has proved efficient, not only in the sense of illuminating problems that would have otherwise remained unknown, but also due to the involvement of practical users who make use of the RDF data, resulting in a model that is both theoretically and empirically sound.

Advancements in modelling

The ultimate goal of the latest iteration was to retain generalizability and universality of the model, while still representing the richness and complexity of the Global series. To that end, the iteration involved numerous updates and improvements to the 2016 model which

2 https://www.w3.org/TR/turtle/
3 http://lynx-project.eu/

Julia Bosque-Gil is a postdoctoral researcher at the Distributed Information Systems Group at University of Zaragoza. She has recently obtained her PhD at the Ontology Engineering Group (Universidad Politécnica de Madrid) for her thesis investigating the use of linguistic linked data for lexicography, and collaborated with K Dictionaries and Semantic Web Company in the representation of multilingual lexicographic data as RDF in the LD4HELTA project. She is currently working in the representation, transformation and linking of multilingual resources as linguistic linked data as part of the Prêt-à-LLOD project and the NexusLinguarum COST Action.
jbosque@unizar.es

was proposed in the framework of the round-tripping condition (Bosque-Gil et al. 2016). One facet of improvement was a thorough revision of the mappings proposed for the different XML paths, the KD ad-hoc vocabulary that was developed for internal use to bridge the gaps between the OntoLex and LexInfo RDF vocabularies, and the KD XML Schema. In addition to applying the *lexicog* module illustrated in Bosque-Gil et al. (2019), this revision saw an update of the KD XML Schema (DTD) in terms of its collection of predetermined semantic or syntactic cues and their values. In the 2016 model, the KD vocabulary included individuals, classes and properties that could not be directly mapped to the LexInfo vocabulary, primarily for two reasons: (a) mismatches between the DTD values of a tag and LexInfo classes, and (b) a different level of granularity in the predefined values in the DTD and the individuals in the linguistic category registry of LexInfo[4]. Given that in these cases a one-to-one mapping from KD into LexInfo was not viable, new elements had to be created under the KD namespace, for example, *kd:prepositionalCase*. The 2019 revision attempted to align the KD DTD values with LexInfo's most recent version[5] as much as possible, to avoid the less desired solution of adding ad hoc ontology elements to represent elements unique to KD, and thus limiting the possibility of linking to external resources. In general, the conversion strives to be as universal as possible, to allow more extensive cross-linking to different resources and consequently expanding the graph. The 2019 conversion has extended the outreach of LexInfo elements, covering significantly more data in KD versus the previous iterations. However, the source data annotation does not only pertain to lists of DTD tags and predefined values; part of the lexicographic workflow takes into account the free values that editors suggest for a given tag, especially in cases in which the predefined list of attribute values does not offer an adequate annotation in the editor's eyes and hence a nuance or further detail is provided. Since this is valuable content for both the data description as well as the day-to-day operations of KD aimed at schema improvement, the 2019 revision is systematically treating free values provided by the editors as individuals in the KD namespace. By dynamically adding these values to the namespace every time the pipeline runs, we allow for future inference of their types thanks to restrictions on properties range, as well as future careful revision of them and even consideration as a new (predefined) value in the Schema, reflecting semantic and pragmatic shifts in the language, or as a potential replacement for a predefined value of which the usage is gradually in decline.

4 http://www.lexinfo.net/ontology/2.0/lexinfo.owl
5 http://www.lexinfo.net/ontology/3.0/lexinfo

```
PREFIX lexicog: <http://www.w3.org/ns/lemon/lexicog#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX skos: <http://www.w3.org/2004/02/skos#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX vartrans: <http://www.w3.org/ns/lemon/vartrans#>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?mainEntry ?nestedEntry ?lexicalEntry ?originalResource
WHERE{
  ?mainEntry rdfs:member ?nestedEntry .
  ?nestedEntry lexicog:describes  ?lexicalEntry .
  ?originalResource lexicog:entry ?nestedEntry .
} LIMIT 500
```

Query 1. Retrieving a list of nested dictionary entries, along with their containers and the lexical entries they refer to, according to the OntoLex lexicog module

mainEntry	nestedEntry	lexicalEntry	originalResource
:ES-00000024-00000025-nested	:ES_DE00000024	:LexiconES/abajar-vb	:mlds-ES3
:ES-00000024-00000025-nested	:ES_DE00000025	:LexiconES/abajar-vb	:mlds-ES3
:ES-00000026-00000027-nested	:ES_DE00000026	:LexiconES/abajo-adv	:mlds-ES3
:ES-00000026-00000027-nested	:ES_DE00000027	:LexiconES/abajo-interj	:mlds-ES3
:ES-00000028-00000029-nested	:ES_DE00000028	:LexiconES/abalanzar-vb	:mlds-ES3
:ES-00000028-00000029-nested	:ES_DE00000029	:LexiconES/abalanzarse-vb	:mlds-ES3

Table 1. Extract from the results of Query 1

Validation and querying

Following the development process described in section 3, the validation step consists initially of a JSON Schema (detailed in Bosque-Gil et al. 2019), followed by performing a series of queries on the KD SPARQL endpoint, with specific queries designed for each step in the conversion. These predefined queries permit inspecting the modelling tag by tag in the DTD, as its OntoLex-*lemon* counterpart. For example, Query 1 allows to validate the representation of containers of nested entries and their links to the lexical entries they describe.

An extract of Query 1 results is shown in Table 1. The same dictionary entry container (e.g. 024-025-nested) contains two nested entries (24 and 25), which both describe the Spanish verb *abajar* [to lower, decrease], having two different forms (transitive vs intransitive) hence originally separated into two entries. The container 026-027 groups together two dictionary entries, *abajo* [down, downstairs] (adverb) and *abajo* [down!] (interjection), and 028-029 represents the container that in the original resource gathered the transitive and reflexive uses of *abalanzar* [to leap on, jump, throw].

```
SELECT DISTINCT ?entry
{
  ?entry ontolex:lexicalForm [ontolex:writtenRep "bow"@en ] .
}
```

Query 2. (same prefixes as in Query 1 apply) Retrieving all lexical entries with the lemma *bow* in English

```
SELECT DISTINCT ?sense
{
  :LexiconEN/bow-n ontolex:sense ?sense .
}
```

Query 3. (same prefixes as in Query 1 apply) Retrieving all senses linked to the artificial entry :LexiconEN/bow-n, created to act as a "container" of senses and to allow linkage with other resources if the homograph number is unknown

Queries 2 and 3 delve into the modelling of homographs. Query 2 retrieves the list of lexical entries with the lemma *bow* (noun) in English, which will include separate lexical entries for homographs.

Query 3 retrieves all lexical senses linked to the artificial entry :LexiconEN/bow-n. The artificial entry *bow* enables gathering the information originating from the different homographs (i.e. from :LexiconEN/bow-n-1, and :LexiconEN/bow-n-3), as well as from other dictionaries in which *bow* is given as a translation (without specifying to which homograph it applies). Thanks to this method of clustering, the query currently results in 56 possible senses in different

sense
:LexiconEN/bow-n-arco-n-ES_SE00006877-sense
:LexiconEN/bow-n-arco-n-ES_SE00006878-sense
:LexiconEN/bow-n-inclinaci%C3%B3n-n-ES_SE00041443-sense
:LexiconEN/bow-n-lazo-n-ES_SE00045188-sense
:LexiconEN/bow-n-mo%C3%B1o-n-ES_SE00050680-sense
:LexiconEN/bow-n-proa-n-ES_SE00060090-sense
:LexiconEN/bow-n-reverencia-n-ES_SE00065273-sense
:LexiconEN/bow-n-arco-n-IT_SE00002942-sense
:LexiconEN/bow-n-arco-n-IT_SE00002945-sense
:LexiconEN/bow-n-fiocco-n-IT_SE00016220-sense
:LexiconEN/bow-n-gala-n-2-IT_SE00017474-sense
:LexiconEN/bow-n-prora-n-IT_SE00033071-sense
:LexiconEN/bow-n-riverenza-n-IT_SE00036412-sense
:LexiconEN/bow-n-Bogen-n-DE_SE00006470-sense
:LexiconEN/bow-n-Bogen-n-DE_SE00006471-sense

Table 2. Extract of the results from Query 3

languages of the word *bow* as a noun in English across the Global series.

An extract of the list of results for Query 3 is shown in Table 2. In this way, the artificial entry :LexiconEN/bow-n, which was absent in the original resource, serves now as a linking point with other multilingual resources in the Global series as well as an entry point for the different senses of the homographs in the English dataset (cf. Image 1).

These query samples exemplify the way in which uploading the data onto a triple store and querying it for results enable validation of the conversion. By counting instances in the original data and matching with the triple store results, we can make sure that every XML component underwent proper conversion to an RDF counterpart. By examining the results of goal-oriented queries, such as obtaining nested entries or all related instances of a noun (as in the case of *bow*), we can see clearly how the cross-linking operates de facto. By determining an incremental conversion and querying of the results as our *modus operandi*, we were able to obtain a clear visual and structural representation of the model. The opposite also applies: by querying a particular instance and retrieving an unexpected result, we were able to handle errors and fix them in the origin.

Summary

Following recent publications regarding the theoretical aspects of KD's initiative and collaboration with partners to convert its data

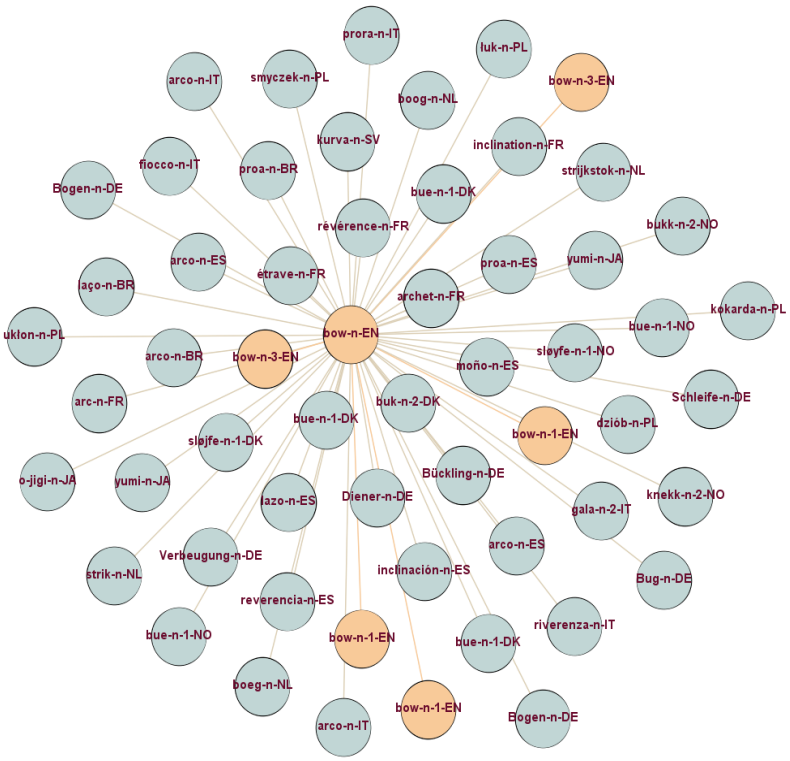


Image 1. The artificial entry :bow-n and its links to senses; senses stemming from translations are in green/blue and those from English in orange

to an LD format, this short paper demonstrates the application of the modelling in terms of pipeline, practical advancements to the model, and the validation and querying process. This endeavor features a real-world example of RDF representation of lexicographic data, demonstrating how a model should account for the structural constraints of a lexical resource, as well as the linguistics shifts and changes to the semantic and syntactic information that is represented therein. The dynamic vocabulary is just one example, proving that in a constantly changing environment, the theoretic representation should be able to develop accordingly. We exemplified how a good model that takes into account such constraints while retaining as much information as possible and remaining flexible, will yield impressive and expansive results. Such is the case of the artificial entry *bow* (noun), which gathers information across all lexical resources of the Global series, creating a de facto graph of cross-linked information within one larger context. Finally, through mutual consultation and exchange we were able to design the queries to match our partners' needs and obtain the best results for real-word applications.

References

Bizer, C., Heath, T., and Berners-Lee, T. 2011. Linked data: The story so far. *Semantic services, interoperability and web applications: emerging concepts*: 205-227. IGI Global.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de-Cea, G. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology*: 65-72. http://www.lrec-conf.org/proceedings/lrec2016/workshops.html

Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. 2019. Validating the OntoLexlemon Lexicography Module with K Dictionaries' Multilingual Data. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*: 726-746. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_41.pdf

Klimek, B., and Brümmer, M. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23, 5-10. https://www.kdictionaries.com/kdn/kdn23_2015.pdf

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. 2017. The Ontolex-Lemon model: Development and applications. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*: 19-21. http://john.mccr.ae/papers/mccrae2017ontolex.pdf

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the Lynx project (grant agreement No 780602). It has been also partially supported by the Spanish projects TIN2016-78011-C4-3-R (AEI/FEDER, UE) and DGA/FEDER 2014-2020 "Construyendo Europa desde Aragón". http://lynx-project.eu/

English WordNet:
A new open-source wordnet for English

John P. McCrae, Ewa Rudnicka and Francis Bond

Introduction

Wordnets have become one of the most popular dictionaries for use in natural language processing (NLP) and other areas of language technologies. This is primarily due to their structure as a graph of words, that is much easier for computers to understand than the traditional form of a dictionary. The first wordnet was introduced by Arthur Miller (Miller 1995) and later extended by Christiane Fellbaum (Fellbaum 1998) at Princeton University before finally being released in its definitive form as Princeton WordNet 3.0 in 2006. However, since then there has only been a single maintenance release of the resource (3.1) in 2011, that actually reduced the number of words it covered. Meanwhile, interest and use of wordnets have grown with many projects around the world creating new wordnets for languages other than English as well as projects adding extensions to Princeton WordNet such as extending it with sentiment information (Esuli and Sebastiani 2006), encyclopedic information (Navigli and Ponzetto 2012) and pronouns and exclamatives (Da Costa and Bond 2016), and providing domain-specific terminology (McCrae, Wood, and Hicks 2017). Furthermore, it is clear that the English language has changed in the last 14 years and Princeton WordNet does not cover recent neologisms and other language usage changes, which are important for many of the social media analytics tasks that we wish to apply wordnets to. Moreover, perhaps one of the biggest criticisms of Princeton WordNet has been that it contains many errors (McCrae and Prangnawarat 2016) and has, at times, overly fine or coarse sense distinctions (Hovy et al. 2006).

Given the lack of change in Princeton WordNet, in spite of the abundant criticisms, we decided to make a 'fork' of the Princeton WordNet, to create a new open-source project called English WordNet (EWN). This project aims to produce the highest quality and most complete wordnet for English and to do so in an open manner. This is implemented by means of a GitHub repository with a collection of XML files that are clear and can easily be edited by anyone. The project accepts suggestions from any parties and so far has been very active with over 650 commits over 500 issues over the course of two years. This has lead to over 18,500 individual improvements over the

John P. McCrae is a research lecturer at the Data Science Institute at the National University of Ireland Galway and a member of the SFI Insight Research Centre for Data Analytics. He holds a PhD from SOKENDAI University (National Institute of Informatics, Tokyo). He is the coordinator of the H2020 project Prêt-à-LLOD on linguistic linked open data and leads the task on linked data in the ELEXIS infrastructure H2020 project, and holds an IRC (Irish Research Council) consolidator laureate award on NLP for minority and historical languages and is a board member of the Global WordNet Association.
john.mccrae@insight-centre.org

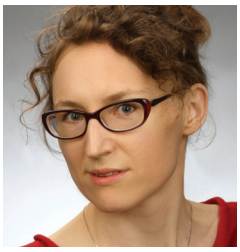
Princeton WordNet, producing a resource that is clearly of much better quality and more comprehensive than the previous releases that have been available to date.

In this article, we provide a brief description of the idea of wordnets and how they are frequently used in natural language processing for readers who may not be familiar with this form of dictionary. Then, we describe the development methodology we have for this dictionary and how we have built and adapted to the growing community of English WordNet users. We then describe the resource of English WordNet and the changes over Princeton WordNet in its two releases so far. Finally, we detail our future plans for this wordnet and make some concluding remarks.

Wordnets are a form of dictionary that aim to make information more easily processable for computers. The primary unit of a wordnet is a set of synonyms or a *synset*, consisting of a list of words that in some context can be substituted for each other. These synsets then form the nodes of a graph, which is connected by edges, consisting of relations such as *hypernym*, indicating a broader/narrower relation, *antonym*, indicating opposition, and *meronym*, indicating a part/whole relation. A word may be part of multiple synsets and as such, we refer to the word within a given synset as a *sense* of the word. An example of such a graph is shown in Figure 1.

Princeton WordNet and most other wordnets cover only four parts-of-speech: noun, verb, adjective and adverb. The nouns are grouped into a hierarchy, where every term is ultimately a hyponym of a single word 'entity'. Verbs similarly are grouped into hierarchies, however, there is no overall supreme concept for verbs and the graph is more disconnected. For adjectives, the structure is generally based around a 'dumbbell' model, where adjectives are grouped into pairs of antonyms, such as 'hot'-'cold', and then 'satellite' adjectives that are related to the meaning of these adjectives, such as 'scorching' or 'frosty', are connected to one end of the dumbbell with a *similar* relation. Alternatively, adjectives may be classified as *pertainyms*, whose meaning is defined by 'of or relating to' a noun, such as 'French' to 'France'. For adverbs, there is little structure and many adverb synsets have no connections in the graph.

The graph-based nature of wordnets has made them highly amenable to NLP applications and a number of methods have been developed that exploit this. For example, word similarity can be computed by simply calculating how many edges must be followed to connect two words (Wu and Palmer 1994) and more sophisticated methods have been built on this principle (Lin and Sandkuhl 2008). Moreover,



Ewa Rudnicka is a research associate at the Department of Computational Intelligence, Wroclaw University of Science and Technology, Poland, and a member of CLARIN-PL Language Technology Centre. She holds a PhD and an MA in comparative linguistics from the University of Wroclaw (Faculty of Languages, Department of English). She is the coordinator of a team of lexicographers working on the mapping between plWordNet and Princeton WordNet, and building an extension to the latter, enWordNet, and a member of the Global Wordnet Association. Her research interests cover lexicography, semantics, theory of equivalence, and natural language processing.
eva.rudnicka@pwr.edu.pl

Princeton WordNet is still the most widely used resource for *word sense disambiguation*, the task of deciding which sense of a word is used in a given context, and wordnets are still the basis of most evaluations in this area (Navigli, Jurgens and Vannella 2013). Even with the recent developments in the field of NLP, relating to the use of neural networks and other methods, there has been interest in exploiting the graph structure of wordnets to develop neural networks (Kutuzov et al. 2018) and embeddings (Rothe and Schütze 2015).

Open-source methodology of EWN

English WordNet has adopted an open-source methodology for the development of the wordnet, meaning that anyone can comment and suggest changes, although these changes are implemented by a core team of developers. There are principally two ways to contribute to EWN, either directly by suggesting changes to the XML through a method called a *pull request*, that is a standard part of open-source development, or by making an *issue*, that is a report of a bug. We have found that the vast majority of suggested changes are made by opening an issue. These suggestions are then categorized by the type of change that is requested, for example adding/removing a relation, updating a definition or example of usage, adding, removing, merging or splitting a synset, or another technical issue. We find that most issues refer to merging synsets, perhaps because wordnet tends to split word senses too much, but can often be resolved by making the definitions more distinct. In addition, there have been many requests for new synsets to be added and to accommodate this we have developed guidelines that determine when a new term should be added to the wordnet.

- Concepts should be significant and represent general English usage. English WordNet does not need to include the name of every place, person and organization in the world. Such things are better handled by other projects, such as Wikidata.
- Terms should not be compositional, that is the meaning of a multiword expression cannot be inferred from its words, or a single word is not derived by the obvious use of a prefix or suffix.
- The word (or sense) should be distinct from other synsets already in the wordnet.
- It should be possible to give a clear textual definition of the concept and to link it to at least one other concept already in the wordnet.

Francis Bond is an Associate Professor in Linguistics and Multilingual Studies at Nanyang Technological University, Singapore and the co-coordinator of NTU's Digital Humanities Cluster. He holds a BA, BEng and PhD from the University of Queensland. He is an active member of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN) and the Global WordNet Association, has developed and released wordnets for Chinese, Japanese, Malay and Indonesian, and coordinates the Open Multilingual Wordnet. His main research interest is in natural language understanding.
bond@ieee.org

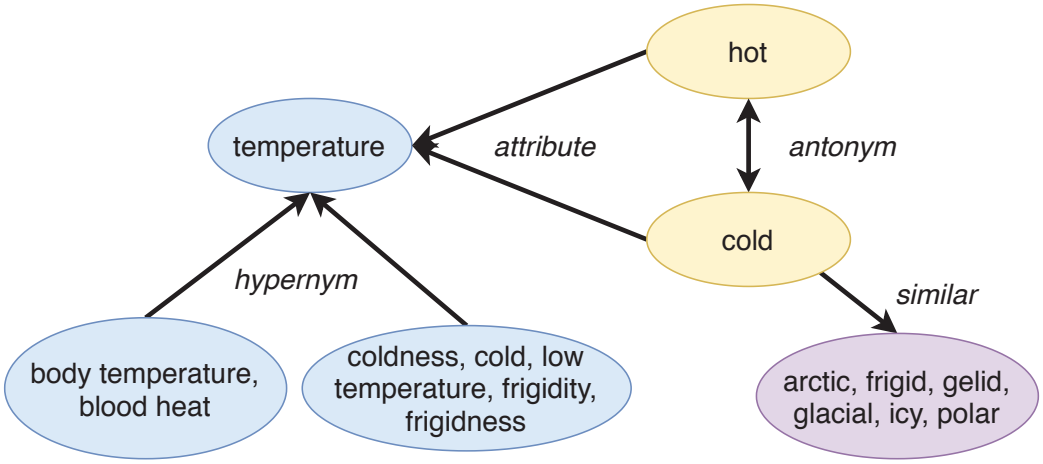


Figure 1. An example of a wordnet graph, showing 'temperature' and its hypernyms, the dumbbell of 'hot' and 'cold' and a satellite adjective

- In difficult cases, we look for clear distinctions in the hypernym to distinguish similar concepts, such as for 'wood' by consistently distinguishing between a tree (an organism) and its wood (a material) or by finding collocations that clearly distinguish this sense.

A key goal is to ensure that there is backwards compatibility between these releases of EWN and the previous Princeton releases. We achieve this by also releasing the data in the form of the database files that are used by the Princeton WordNet tools. This can create some issues in that this format uses the *offset*, that is the number of bytes in the file that need to be read to reach the start of an entry, to identify synsets. For English WordNet, we have fixed the identifiers to be the offsets of the Princeton WordNet 3.1 release and in fact, use random numbers for new synsets so it would be impossible and impractical to keep these in-sync with the release. Otherwise, we try to keep all of the features of Princeton WordNet's structure as is, even if some aspects may be unnecessary, complex or scientifically questionable.

To date, there have been two releases of English WordNet, the 2019 and 2020 edition. These have expanded the scope of the project and

	Princeton 3.1	EWN 2019	EWN 2020
Synsets	117,791	117,791	120,054
Lemmas	159,015	159,789	163,079
Senses	207,272	208,353	211,864
Relations	378,203	378,201	383,825

Table 1. Size and coverage of Princeton WordNet 3.1 and the two releases of English WordNet

while they have obviously introduced many new changes, the focus of the work has been on improving the quality of the resource. In fact, we found nearly 2,000 typos in the text of PWN, and even in one case a misspelt lemma!

Another major source of changes was the inclusion of external data from other sources. We directly included other English wordnets, including enWordNet developed as part of plWordNet (Rudnicka, Witkowski and Kaliński 2015) and Colloquial WordNet (McCrae, Wood and Hicks 2017), with some modification to better fit the structure of this wordnet. Secondly, we incorporated updated definitions from the Open Multilingual WordNet project (Bond and Foster 2013) and also used the linking to Wikipedia (McCrae 2018) to add extra lemmas for many concepts. Finally, we fixed many minor errors related to issues such as examples which do not use any lemma in the synset.

Future plans

English WordNet is an expanding project and we intend to continue to develop the resource through the open-source methodology. There have been several areas that have been identified as key long-term areas to improve the resources. Firstly, the modelling of adjectives and adverbs is, as discussed above, quite unusual and adjectives and adverbs have far fewer links and more disconnected nodes in the graph than for nouns or verbs. Adopting a new structure would be much more preferable (Mendes 2006), and finding similar ways to define adverbs and their relations to the noun and verb hierarchies would enhance usability for NLP applications that depend on these links. Secondly, we are working to improve the methodology for developing the wordnet, in particular, there has been much discussion in the community about moving on from the XML model to something less verbose and more readable and the use of YAML markup is likely to be adopted. As an example we compare the current XML markup with the proposed YAML form, which significantly reduces the size of the file.

In addition, we have a browsing interface available at https://en-word.net/ which provides a searchable interface to the most recent interface and provides a linked data version of the data in RDF using the OntoLex-Lemon model (Cimiano, McCrae and Buitelaar 2016). As such, the YAML format is intended to be an internal working format with releases still made according to the standards such as the LMF XML format, and OntoLex-Lemon. An example of the data encoding is available in Figure 2.

```
<LexicalEntry id="ewn-dictionary-n">
 <Lemma writtenForm="dictionary"
    partOfSpeech="n"/>
 <Sense id="ewn-dictionary-n-06430544-01"
    n="0" synset="ewn-06430544-n"
    dc:identifier="dictionary%1:10:00::"/>
</LexicalEntry>
<Synset id="ewn-06430544-n" ili="i70226"
   partOfSpeech="n"
   dc:subject="noun.communication">
 <Definition>a reference book containing an alphabetical list of
words with information about them</Definition>
 <SynsetRelation relType="hypernym"
       target="ewn-06430336-n"/>
 <SynsetRelation relType="mero_part"
       target="ewn-06311813-n"/>
</Synset>
```

```
06430544-n:
 definitions:
 - a reference book containing an
  alphabetical list of words
  with information about them
 entries:
 - dictionary%1_10_00
 - lexicon%1_10_00
 hypernym:
 - 06430336-n
 ili: i70226
 mero_part:
 - 06311813-n
 pos: n
```

Figure 2. An example of the XML and YAML encoding of the English WordNet data as available on Github

Furthermore, tools for supporting changes in English WordNet and validating the consistency are already deployed and continue to be developed. Finally, we would like to move on from the model of a single monolithic dictionary and support a network of wordnets, including domain-specific wordnets or large-scale encyclopedic resources that could be of use to a wide range of tasks, although this would create further issues with maintaining and integrating such a wide range of tasks.

Conclusion

English WordNet is an open-source fork of the Princeton WordNet, whose aim is principally to ensure that there is an English wordnet which is up-to-date and can be of the highest quality, as the many users of wordnets can easily contribute changes and improvements back to the project. We have done this in a simple way, by providing a GitHub repository for simple XML documents. This has proven successful with over 18,500 changes and many contributions from all sides. We plan to continue to develop this resource and hope that it continues to be one of the core dictionaries for NLP applications. Further, while this project is intended to be limited to the English language we hope that this methodology can be adopted by wordnets for other languages and support linking and connecting to create multilingual resources such as through the Open Multilingual WordNet.

References

Bond, Francis and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1352–62. Sofia, Bulgaria: Association for Computational Linguistics.

Cimiano, Philipp, John P. McCrae and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report. W3C. https://www.w3.org/2016/05/ontolex/.

Da Costa, Luis Morgado and Francis Bond. 2016. Wow! What a Useful Extension! Introducing Non-Referential Concepts to WordNet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4323–28.

Esuli andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. In *LREC*, 6:417–22. Citeseer.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Hovy, Eduard, Mitch Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. "OntoNotes: The 90% Solution." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60.

Kutuzov, Andrey, Mohammad Dorgham, Oleksiy Oliynyk, Chris Biemann and Alexander Panchenko. 2018. Learning Graph Embeddings from WordNet-Based Similarity Measures. *arXiv [cs. CL]*. arXiv. http://arxiv.org/abs/1808.05611.

Lin, Feiyu and Kurt Sandkuhl. 2008. A Survey of Exploiting WordNet in Ontology Matching. In *Artificial Intelligence in Theory and Practice II*, 341–50. Springer US.

McCrae, John P. 2018. Mapping WordNet Instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.

McCrae, John P. and Narumol Prangnawarat. 2016. Identifying Poorly-Defined Concepts in WordNet with Graph Metrics. In *Proceedings of the First Workshop on Knowledge Extraction and Knowledge Integration (KEKI-2016)*.

McCrae, John P., Ian Wood and Amanda Hicks. 2017. The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*, 194–202.

Mendes, Sara. 2006. Adjectives in WordNet.PT. In *Proceedings of the Global WordNet Conference*, edited by Petr Sojka, Key-Sun Choi, Christiane Fellbaum and Piek Vossen, 225–30. Citeseer.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39–41.

Navigli, Roberto, David Jurgens and Daniele Vannella. 2013. Semeval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 222–31.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193 (Supplement C): 217–50.

Rothe, Sascha and Hinrich Schütze. 2015. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1793–1803.

Rudnicka, Ewa Katarzyna, Wojciech Witkowski and Michał Kaliński. 2015. Towards the Methodology for Extending Princeton Wordnet. *Cognitive Studies| Études Cognitives*, no. 15: 335–51.

Wu, Zhibiao and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, 133–38. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics.

ELEXIS: Technical and social infrastructure for lexicography

Anna Woldrich, Teja Goli, Iztok Kosem, Ondřej Matuška and Tanja Wissik

Since the European lexicographic community was brought together by the European Network of e-Lexicography (ENeL) COST action (http://elexicography.eu/) in 2013–2017, the following needs have become apparent: the flow of broader and more systematic exchange of expertise; the establishment of common standards and solutions; the development and integration of lexicographic resources; and the wide-scale application of these quality resources to wider research communities. This has resulted in launching the four-year H2020 infrastructure project ELEXIS, European Lexicographic Infrastructure in February 2018 (extended for six months until July 2022).

ELEXIS brings together research and industrial partners from various fields, such as the Semantic Web, Artificial Intelligence, Natural Language Processing (NLP) and Digital Humanities, thus supporting developments in (e-)lexicography in order to open up dictionary data and enable access to lexicographic standards, methods, data and tools.

Among the most obvious outputs of the project are the tools and services it offers. In its first two years, ELEXIS has been enriched by seven different tools which were either developed as part of the project or made freely accessible through its infrastructure, and by the end of the project, the ELEXIS infrastructure is planned to enable and support the whole dictionary creation process. The tools and services already available include:

Sketch Engine. This corpus query system, which existed prior to the project, was one of the first tools made freely accessible to academics and observer institutions in ELEXIS. It includes over 500 preloaded corpora and analysis functions, such as concordancing, building wordlists, compiling word sketches, thesauri and automatic dictionary drafting. https://sketchengine.eu/elexis/

Lexonomy. Another infrastructure component which already existed before ELEXIS but whose further comprehensive development continues within the project. This is a cloud-based dictionary-writing and online-publishing system that interacts closely with Sketch Engine. For example, Sketch Engine can push lexicographic data into Lexonomy to create automatically generated dictionary drafts and Lexonomy can pull data from Sketch Engine's corpora during the entry editing process. https://lexonomy.eu/

Anna Woldrich is communication expert at the Austrian Academy of Sciences (ACDH-CH) and has worked previously as social editor and campaign manager. She graduated at Universität Wien/Universitá degli Studi di Siena in mass media and communication studies, focusing on radio, broadcasting, marketing, communication research and communication theory. As a part of ELEXIS project, she is responsible for planning, managing and monitoring on- and offline communication activities, and manages the social media channels and content-management on the ELEXIS website.
anna.woldrich@oeaw.ac.at

Elexifier. A brand new cloud-based dictionary conversion service, using advanced XML parsing and machine learning techniques to help convert PDF and XML dictionary data into a standardized machine-readable format. Users can upload PDF and custom XML dictionaries, define mapping rules for XML transformation or create a machine learning training set for PDF conversion and download the transformed XML or PDF dictionary in a TEI-compliant file format based on the Elexis Data Model[1]. https://elexifier.elex.is/

VerbAtlas. A novel large-scale manually-crafted semantic resource for wide-coverage, intelligible and scalable semantic role labeling. The goal of VerbAtlas is to manually cluster WordNet synsets that share similar meanings into sets of semantically-coherent frames, available both for download and via a RESTful API, featuring resources such as PropBank and BabelNet. http://verbatlas.org/

SyntagNet. A manually-curated large-scale lexical-semantic combination database which associates pairs of concepts with pairs of co-occurring words. The goal of SyntagNet is to capture sense distinctions evoked by syntagmatic relations, hence providing information which complements the essentially paradigmatic knowledge shared by currently available lexical knowledge settings such as WordNet. http://syntagnet.org/

NAISC. A tool for linking datasets. NAISC takes as input 2 RDF documents (referred to as 'left' and 'right) and outputs an alignment (set of RDF triples) between these two documents. It typically relies on a configuration, which is a JSON document. https://youtube.com/watch?v=maYEv8rG0_k

Elexifinder. A search tool dedicated to helping lexicographers and researchers find scientific output in lexicography and related fields. Elexifinder enables users to search through papers and videos, using concepts, that is words or sets of words with a Wikipedia page, and various other conditions, for example source (conference, etc.), author, language, etc. Each paper/video is linked to its page where the user can download or view it. https://elex.is/tools-and-services/elexifinder/; http://er.elex.is/.

Lexicographic news feed. A service using the Event Registry API to extract recent news articles related to lexicography. Articles are extracted from 30,000 news sources, supporting over. 35 languages. https://elex.is/tools-and-services/lexicographic-news/.

1 Information on the ELEXIS Data Model is available in the recordings of the ELEXIS Observer Event 2019: http://videolectures.net/elexisobserver2019_tiberius_data_model/

Teja Goli is an assistant at the Artificial Intelligence Laboratory at Jožef Stefan Institute and at the University of Ljubljana, where she has finished her master's degree in Translation at the Faculty of Arts. Her research interests include translation, corpus linguistics and lexicography. In the ELEXIS project, she is mainly responsible for contact with observers and managing website content.
teja.goli@ijs.si

Ondřej Matuška oversees sales and marketing activities and external communication at Lexical Computing, and is the main point of contact for information about and user support for Sketch Engine.
ondrej.matuska@sketchengine.co.uk

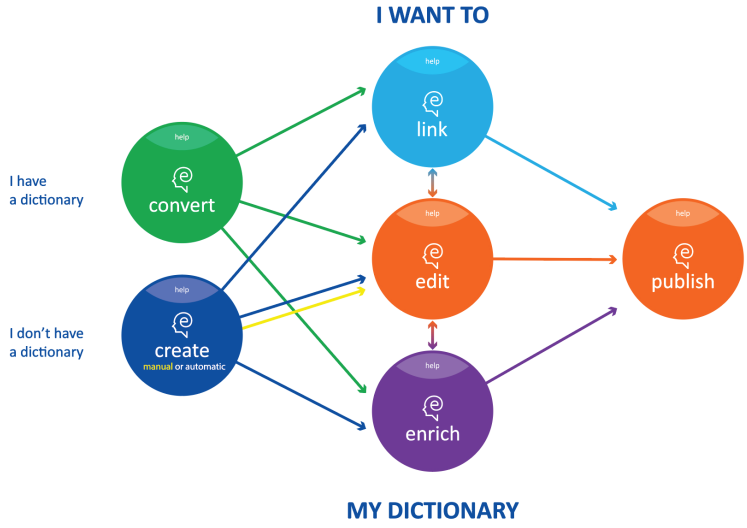


Image 1. ELEXIS offers a user-friendly way to create dictionaries or edit and publish existing ones

More tools and services as well as instruction manuals will be added during the lifetime of the project, accumulating to a full-scale service for a user-friendly dictionary publication process (cf. Image 1).

Besides this extensive technical infrastructure, ELEXIS provides a social infrastructure to foster cooperation and support knowledge exchange among lexicographic communities. Additionally, it is bridging the gap between lesser-resourced languages and those with higher e-lexicographic expertise. One aspect of this social infrastructure is organizing training sessions and workshops at conferences as well as summer/spring schools all over Europe (https://elex.is/all-events/). Due to COVID-19 restrictions, several events had to be canceled this year, but we have managed to overcome the obstacle prohibiting face-to-face interaction for community building by moving several activities online. As part of the GlobaLex 2020 Workshop on Linked Lexicography at the Language Resource and Evaluation Conference (LREC 2020), ELEXIS organized the first shared task on monolingual word-sense alignment (MWSA). While the workshop itself had to be cancelled, the papers and the results are available as part of the proceedings (https://aclweb.org/anthology/volumes/2020.globalex-1/). The goal was to find senses in two monolingual dictionaries (in the same language), that describe the same concept. The MWSA task made use of data in 15 languages from ELEXIS partners and observers. The participants developed strong systems with the overall best system scoring 84% accuracy in sense alignment. https://elex.is/mwsa2020/.

Furthermore, ELEXIS supports individual researchers and research teams via trans-national access, enabling them to reach facilities

Iztok Kosem (PhD) is Research Associate at Jožef Stefan Institute and at the University of Ljubljana. His main areas of research are lexicography and lexicogrammar, corpus linguistics, crowdsourcing, and computer-aided language learning and teaching. In ELEXIS, he has the role of Community Manager, and he is heavily involved in the development of Elexifinder, Lexonomy and games with a purpose (gamification).
iztok.kosem@ijs.si

and lexicographic resources which are not fully or easily accessible online or where professional on-site expertise is needed. Researchers, scholars and students are invited to apply for a fully-funded short- or long-term research visit to leading lexicographic institution partners (https://elex.is/grants-for-research-visits/). Calls for visiting grants are launched twice a year, in summer and in winter, amounting to seven calls in total during the project period. The travel grant reports as well as mini-interviews with the respective winners from various countries all over Europe are available at https://elex.is/travel-grant-reports/.

While individual researchers can participate through travel grants, institutions are invited to join the ELEXIS network via observer status (https://elex.is/join-as-observer). Observing institutions may request new customized lexicographic data or have their existing data enriched and expanded with both monolingual and multilingual information, Moreover, they can access the ELEXIS cloud, tools and open-access resources as well as resources in the partner and observer's area of the cloud. Observers are notified about newly developed tools, services and activities (e.g. hackathons, tool demo sessions, etc.) aimed at improving and enriching their own lexicographic data. To keep up

The first Lexonomy Hackaton took place in Brno on 23-25 April 2019

a sustainable infrastructure after the end of the project in 2022, the observer status guarantees the possibility to participate actively in the post-project stage.

To this end, ELEXIS organized an Observer Event in early 2019, dedicated to inform representatives of various lexicographic institutions on its activities (https://elex.is/observer-event/).

Tanja Wissik is a senior scientist and project leader at the Austrian Centre for Digital Humanities and Cultural Heritage of the Austrian Academy of Sciences and she teaches at the University of Graz and University of Vienna. She holds a PhD in Translation Studies from the University of Vienna and in the last couple of years she has been working in a number of European and national research projects in the field of language resources, text technologies and DH methods.
tanja.wissik@oeaw.ac.at

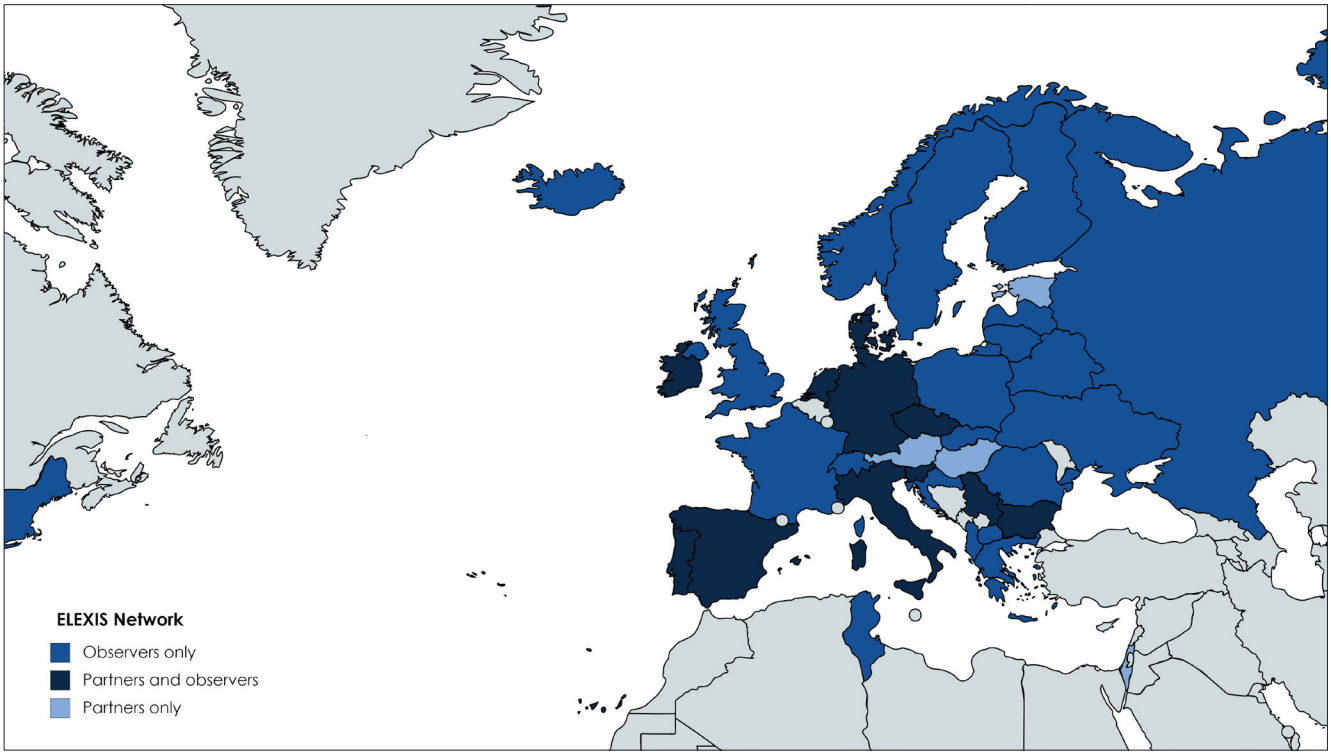


Image 2. Overview of the ELEXIS network in June 2020

Institutions from all over Europe (and beyond) have been joining the network: as of June 2020, the ELEXIS community is made up of 17 partner and 50 observer institutions from 35 different countries (cf. Image 2; https://elex.is/observers/). In addition, ELEXIS is running a campaign on social media, describing the characteristics of each observing institution – all portraits are collected in the #elexisobserver moment on Twitter.

Since community building is a key factor for ELEXIS, it is important to assess the experience and opinions regarding the project's intermediate outcomes. This is a way to reflect on the work done so far as well as to fine-tune the final outcomes to respond best to the needs of the community. Thus, the ELEXIS impact survey was launched in May 2020, containing 16 questions on different aspects of the technical and social infrastructures. The results have shown that 79% (n=123) of the respondents already knew ELEXIS or were following its activities actively. For most respondents the most important aspects of ELEXIS are the tools and services as well as open access and open data, followed by training and education, knowledge exchange and community building (cf. Image 3).

elexis
european lexicographic infrastructure

Horizon 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015.

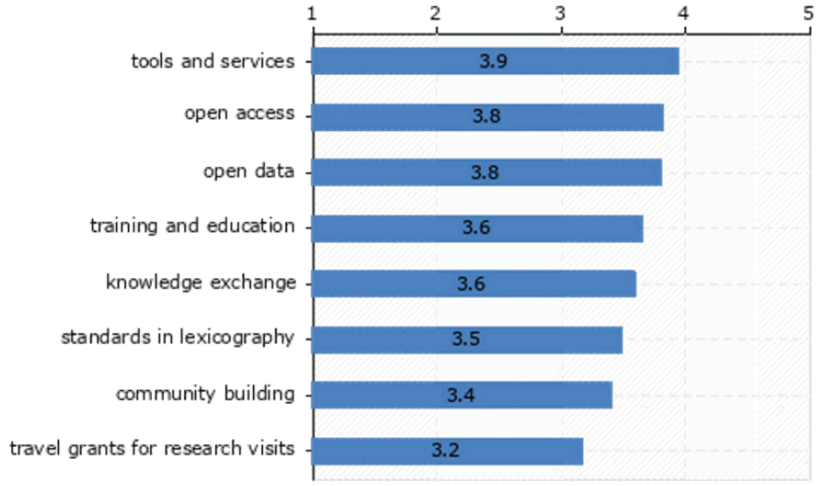


Image 3. Usefulness of ELEXIS services for those who are familiar with the network (Q12, N=97)

Although some respondents did not know ELEXIS before, we were interested to find out how useful specific aspects of the infrastructure might be to them. These turned out to include access to the corpus query tool Sketch Engine, open data and open access, as well as knowledge exchange, training and education (cf. Image 4).

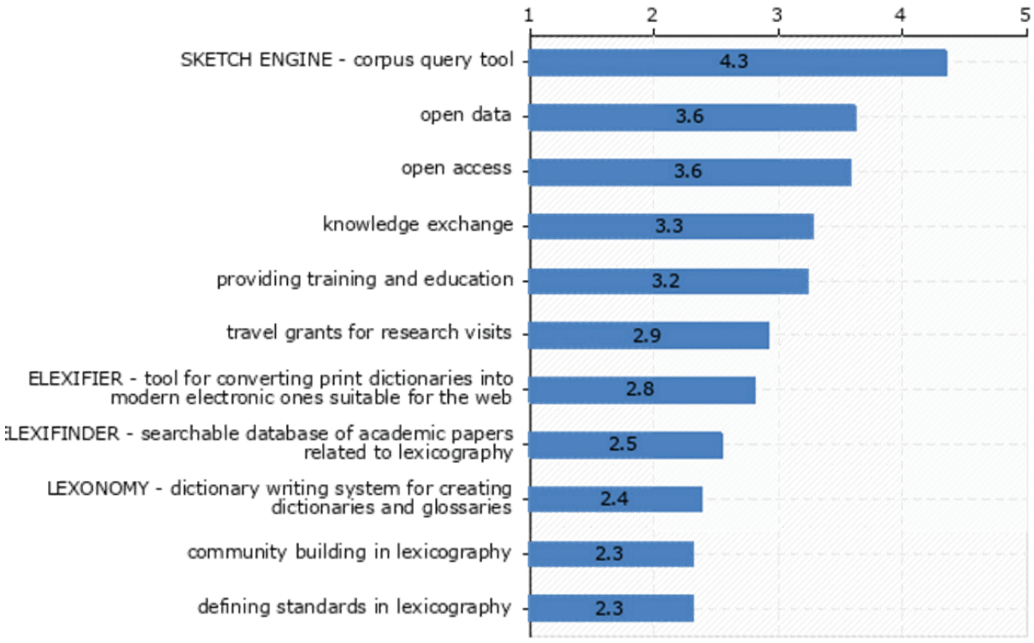


Image 4. Potential usefulness of ELEXIS services for those who don't know the network (Q6, N=26)

The full survey as well as other project reports are available at https://elex.is/deliverables/.

Additionally, all conference papers, peer-reviewed articles and journal articles published in the course of the project with ELEXIS are available on Zenodo.

Partners

Original name	English name	Country
Österreichische Akademie der Wissenschaften	The Austrian Academy of Sciences	Austria
ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК	Institute for Bulgarian Language Prof Lyubomir Andreychin	Bulgaria
Lexical Computing CZ sro	Lexical Computing CZ sro	Czech Republic
Det Danske Sprog- og Litteraturselskab	The Society for Danish Language and Literature	Denmark
Center for Sprogteknologi (CST) Institut for Nordiske Studier og Sprogvidenskab	The Centre for Language Technology at the Department of Nordic Research, University of Copenhagen	Denmark
Eesti Keele Instituut	Institute of the Estonian Language	Estonia
Universität Trier	The Trier Center for Digital Humanities	Germany
Magyar Tudományos Akadémia Nyelvtudományi Intézetének	Research Institute for Linguistics at the Hungarian Academy of Sciences	Hungary
The National University of Ireland, Galway/ OÉ Gaillimh	The National University of Ireland, Galway	Ireland
ק מילונים בע"מ	K Dictionaries Ltd	Israel
Sapienza Università di Roma	The Sapienza University of Rome	Italy
Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale "A Zampolli"	The Institute for Computational Linguistics "A Zampolli"	Italy
Universidade Nova de Lisboa – Faculdade de Ciências Sociais e Humanas	Universidade NOVA de Lisboa – The NOVA School of Social Sciences and Humanitie	Portugal
ентар за дигиталне хуманистичке науке	The Belgrade Center for Digital Humanities	Serbia
Inštitut "Jožef Stefan"	"Jožef Stefan" Institute	Slovenia
Real Acamedia Espanola	The Royal Spanish Academy	Spain
Instituut voor de Nederlandse Taal	Dutch Language Institute	The Netherlands

ELEXIS on GitHub

https://github.com/elexis-eu

Lexonomy
https://github.com/elexis-eu/lexonomy
Elexifinder
https://github.com/elexis-eu/elexifinder
elexifier-api
https://github.com/elexis-eu/elexifier-api
elexifier
https://github.com/elexis-eu/elexifier
dictionary service
https://github.com/elexis-eu/dictionary-service
MWSA
https://github.com/elexis-eu/MWSA
NAISC
https://github.com/insight-centre/naisc
word games
https://github.com/elexis-eu/word-games
elexis-rest
https://github.com/elexis-eu/elexis-rest
elexifier-pdf
https://github.com/elexis-eu/elexifier-pdf
tei2ontolex
https://github.com/elexis-eu/tei2ontolex
CrossTheWord
https://github.com/elexis-eu/CrossTheWord
ocd
https://github.com/elexis-eu/ocd
D3.1
https://github.com/elexis-eu/D3.1

Observers (June 2020)

	Original name / English name	Country
1	Institut za hrvatski jezik i jezikoslovlje / Institute for the Croatian Language and Linguistics	Croatia
2	Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE) / Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE)	Italy
3	Univerzitet u Beogradu, Rudarsko-geološki fakultet / Univesity of Belgrade, Faculty of Mining and Geology	Serbia
4	SIL International / SIL International	International (US)
5	Лексикографски центар при Македонската академија на науките и уметностите / Lexicographic Centre at the Macedonian Academy of Sciences and Arts	North Macedonia
6	Kotimaisten kielten keskus / Institute for the Languages of Finland	Finland
7	Київський університет імені Бориса Грінченка / Borys Grinchenko Kyiv University	Ukraine
8	Институт по информационни и комуникационни технологии към Българската академия на науките / Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences (IICT-BAS)	Bulgaria
9	Институт лингвистических исследований Российской академии наук / Institute for Linguistic Studies, Russian Academy of Sciences	Russia
10	Universitatea de Vest din Timisoara / West University of Timisoara	Romania
11	Universitetsbiblioteket ved Universitetet i Bergen / University of Bergen Library	Norway
12	Sveučilište Jurja Dobrile u Puli / Juraj Dobrila University of Pula	Croatia
13	Filozofski fakultet, Sveučilište u Rijeci / Faculty of Humanities and Social Sciences, University of Rijeka	Croatia
14	Institut de Recherche et d'Histoire des Textes (CNRS) / Institut de Recherche et d'Histoire des Textes (CNRS)	France
15	Dipartimento di filologia, letteratura, linguistica / Department of Philology, Literature and Linguistics	Italy
16	Vytauto Didžiojo universitetas, Kompiuterinės lingvistikos centras / Vytautas Magnus University, Centre of Computational Linguistics	Lithuania
17	Uniwersytet im. Adama Mickiewicza w Poznaniu / Adam Mickiewicz University in Poznań	Poland
18	Institutul de Filologie Română „A. Philippide" / "A. Philippide" Institute of Romanian Philology	Romania
19	Lietuvių kalbos institutas / The Institute of the Lithuanian Language	Lithuania
20	ZRC SAZU Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	Slovenia
21	Stofnun Árna Magnússonar í íslenskum fræðum / The Árni Magnússon Institute for Icelandic Studies	Iceland
22	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, v. v. i. / Ľudovit Štúr Institute of Linguistics of the Slovak Academy Sciences	Slovakia
23	Sveučilište u Zagrebu, Filozofski fakultet / University of Zagreb, Faculty of Humanities and Social Sciences	Croatia
24	Schweizerisches Idiotikon / Swiss Idiotikon	Switzerland

	Original name / English name	Country
25	Universidad de Castilla-La Mancha / University of Castilla–La Mancha	Spain
26	Institute for Applied Linguistics, Eurac Research / Institute for Applied Linguistics, Eurac Research	Italy
27	UPV/EHU University of the Basque Country / UPV/EHU University of the Basque Country	Spain
28	Instytut Neofilologii – Państwowa Wyższa Szkoła Zawodowa w Raciborzu / Institute of Modern Language Studies – State University of Applied Sciences in Racibórz	Poland
29	Institut Superior d'Investigació Cooperativa – IVITRA (Universitat d'Alacant) / Higher Institute Of Cooperative Research – IVITRA (University of Alicante)	Spain
30	Foras na Gaeilge / Foras na Gaeilge	Ireland
31	Stiechting Limbörgse Academie / Limburgish Academy Foundation	The Netherlands
32	Zavod za lingvistička istraživanja Hrvatske akademije znanosti i umjetnosti / Linguistics Research Institute of the Croatian Academy of Sciences	Croatia
33	Latvijas Universitātes Matemātikas un informātikas institūts / Institute of Mathematics and Computer Science, University of Latvia	Latvia
34	Center za jezikovne vire in tehnologije, Univerza v Ljubljani / Centre for Language Resources and Technologies, University of Ljubljana	Slovenia
35	Academia das Ciências de Lisboa / Lisbon Academy of Sciences	Portugal
36	Canolfan Uwchefrydiau Cymreig a Cheltaidd Prifysgol Cymru / University of Wales Centre for Advanced Welsh & Celtic Studies	United Kingdom
37	Institut für Deutsche Sprache / Institute for the German Language	Germany
38	Svenska Akademien / Swedish Academy	Sweden
39	Cologne Center for Humanities / Cologne Center for Humanities	Germany
40	Ústav Českého národního korpusu / Institute of the Czech National Corpus	Czech Republic
41	Mykolo Romerio Universitetas / Mykolas Romeris University	Lithuania
42	Institute for Language and Speech Processing, ATHENA R.C. / Ινστιτούτο Επεξεργασίας του Λόγου, Ε.Κ. ΑΘΗΝΑ	Greece
43	Universitatea de Medicină, Farmacie, Științe și Tehnologie „George Emil Palade" din Târgu Mureș / George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures	Romania
44	Гродзенскі дзяржаўны ўніверсітэт імя Янкі Купалы / Гродненский государственный университет имени Янки Купалы / Yanka Kupala State University of Grodno	Belarus
45	Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) / Berlin-Brandenburg Academy of Sciences and Humanities	Germany
46	Univerzitet u Kragujevcu, Filološko-umetnički fakultet / University of Kragujevac, Faculty of Philology and Arts	Serbia
47	Fakulteti i Historisë dhe i Filologjisë / Faculty of History and Philology	Albania
48	Instituti i Gjuhësisë dhe i Letërsisë / Institute of Linguistics and Literature	Albania
49	Universidad de Alcala / University of Alcala	Spain
50	Dansk Sprognævn / The Danish Language Council	Denmark

Internet lexicography at the Leibniz-Institute for the German Language

Stefan Engelberg, Annette Klosa-Kückelhaus and Carolin Müller-Spitzer

1. Introduction

Over ten years ago, Ilan Kernerman has invited us to write an article about lexicography at the Department of Lexical Studies at the Leibniz-Institute for the German Language in Mannheim (IDS). We wrote about the tenets of Internet lexicography at our institute, about the dictionaries we made, about the lexicographic processes, and about the demands with respect to IT competences and staff recruitment (Engelberg, Klosa and Müller-Spitzer 2009). Back then, we emphasized that the "ability to handle and analyse mass data and the need for Internet-adequate lexicographic concepts is changing the profession", that Internet lexicography is driven by "new possibilities of data integration and crosslinking" and that "the demands made on the competences that have to be gathered within lexicographic projects, reaching from the development of corpus analysis methods to text technology and web technology", were constantly increasing (Engelberg, Klosa and Müller-Spitzer 2009: 16). Admittedly, it did not take a prophet to predict the relevance of these parameters for future Internet lexicography. However, it should be interesting to see how – in more detail – these parameters shaped the lexicographic praxis in our institute within the last ten years.

In 2009, we had just implemented the *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online Lexical Information System on German) as "a lexicographic Internet portal for various electronic dictionary resources that are being compiled at the Institute for German Language" (Engelberg, Klosa and Müller-Spitzer 2009: 16). OWID has proven to be a successful concept and it still constitutes the backbone of lexicography at our institute. In 2007, it was released with four lexicographic resources: a general dictionary on contemporary German, a dictionary of neologisms, a dictionary of idioms, and a discourse dictionary. Since then, six more dictionaries have been integrated into OWID (cf. section 2).

Stefan Engelberg is head of the department 'Lexik' at the Leibniz-Institute for the German Language (IDS) in Mannheim, professor for German linguistics at the University of Mannheim, and honorary professor at the University of Tübingen.
engelberg@ids-mannheim.de

On re-reading the article from 2009, we found that we have followed the path pointed out by OWID quite consistently. The focus is now explicitly on specific-domain dictionaries, that cover areas of the lexicon that have so far been neglected by lexicography and lexicology. Despite its focus on resources for special vocabulary areas, OWID reaches users in many different countries around the world (cf. Figure 1).

However, since 2009, we have also conceptualized and implemented new types of lexicological-lexicographic platforms, and OWID and its dictionaries have developed new kinds of access structures and modes of presentation (cf. section 2). This was partly necessary because it turned out that a single portal with a particular concept of dictionary integration cannot serve all purposes, either because dictionaries develop new forms of presentation that cannot easily be integrated into OWID (such as the dictionary of paronyms, cf. section 2, which is only linked to OWID via its lemma list) or because particular uses of lexical data, especially those by scientific communities, require very different kinds of data aggregation, access to data, and lexicographic collaboration. These demands led to the development of new lexicographic portals (cf. sections 3 and 4).

Annette Klosa-Kückelhaus heads the 'Lexicography and Language Documentation' program area in the Department of Lexicology at the Leibniz-Institute for the German Language (IDS) in Mannheim, and is chief editor of its online dictionary of neologisms.
klosa@ids-mannheim.de

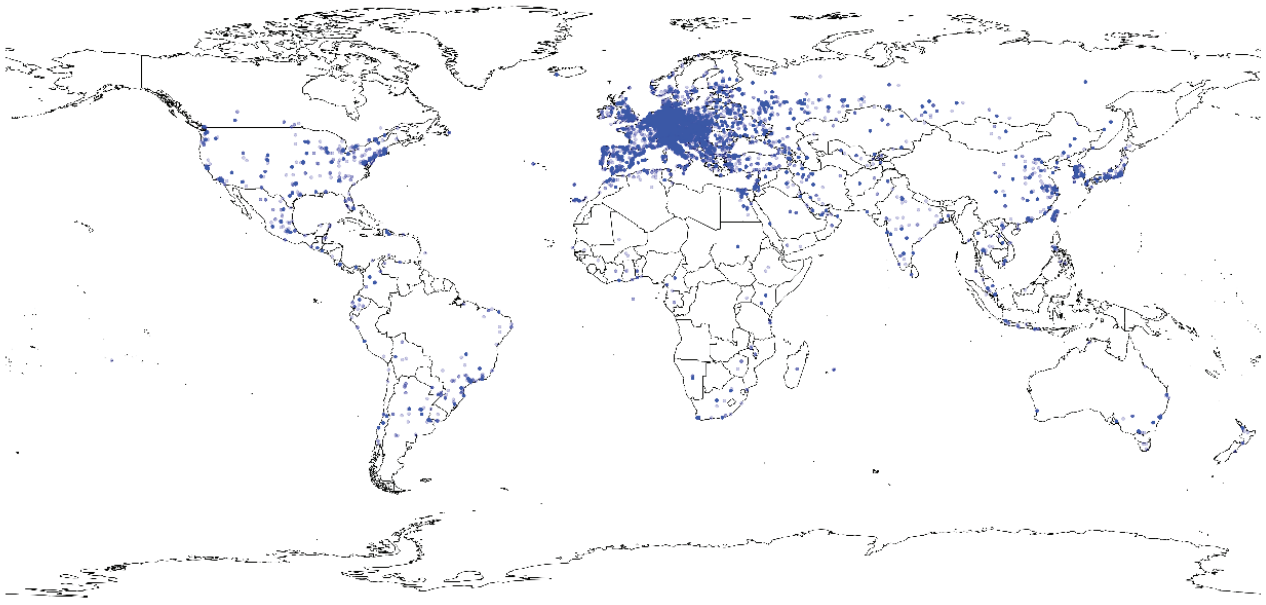


Figure 1.
Geographically located accesses to OWID from the years 2018 and 2019

2. OWID

A. Dictionaries in OWID and OWID^{plus}

As of April 2020, eleven different lexicographic resources are offered in the OWID dictionary portal and in OWID^{plus}:

(1) *elexiko* – Online-Wörterbuch zur deutschen Gegenwartssprache[1] (online since 2003, no further editing). *elexiko* is an online information system for the contemporary German language, which documents, explains and scientifically comments on the vocabulary based on current language data in individual modules. It comprises almost 1,900 systematically compiled, detailed entries for individual words, over 50 word group articles on meaning-relational groups (e.g. 'Defizit – Mangel' [deficit – deficiency], 'Kindheit – Jugend – Alter' [childhood – youth – old age]), thematic fields (e.g. 'Beruf und Familie' [job and family], 'Jahreswechsel' [turn of the year]) and word fields (e.g. 'Getränke' [drinks], 'Speisen' [food]) as well as over 250,000 entries that offer only automatically generated information on the headwords (orthographic information, corpus citations, frequency information).

(2) Paronymwörterbuch[2] (online since 2018, work in progress). This paronym dictionary documents easily confusable expressions in their current public usage. It contains expressions that, for example, have strong similarities in spelling, pronunciation and meaning, such as 'farbig – farblich' [colored], 'kindlich – kindisch – kindhaft' [childlike – childish], 'universell – universal' [universal]. The paronyms are presented together in new, contrasting dictionary articles. Their similarities and their differences can be seen at a glance, with the users deciding for themselves which sections or comparative views they want to get presented. The paronym dictionary clearly illustrates the focus of the dictionaries in OWID: on the one hand, to pick out certain vocabulary sections, for which lexicographic work has not been carried out to date, and on the other hand, to go new ways for the user interface. As such new visualizations are not easy to understand for users, a short video was created as a tutorial.[3] This is also new for scholarly lexicography, but could certainly serve as a model.

1 https://www.owid.de/docs/elex/start.jsp
2 https://www.owid.de/parowb/
3 https://www.owid.de/parowb/docs/hilfe.html

Carolin Müller-Spitzer is head of the program area 'Lexic Empirical and Digital' in the Department of Lexicology at the Leibniz-Institute for the German Language (IDS) in Mannheim, and professor of German Linguistics at the University of Mannheim. Her research focuses on user research, empirical gender linguistics and online lexicography.
mueller-spitzer@ids-mannheim.de

(3) Sprichwörterbuch[4] (online since 2012, no further editing). This proverb and slogan dictionary illustrates that proverbs are still alive and an object of constant variation (e.g. 'Harte Schale, weicher Kern' [hard shell, soft core] with the variants 'Rauhe Schale, weicher Kern' [rough shell, soft core] and 'Harte Schale, kaputter/genialer/leckerer Kern' [hard shell, broken/ingenious/delicious core]), how speakers use them and how new ones are created, for example in advertising (e.g. 'Nicht immer, aber immer öfter' [not always, but more and more often]). It is also the first empirically validated documentation of currently used fixed phrases in the German language based on criteria of scholarly lexicography (created as part of the multilingual EU project SprichWort. An Internet Platform for Language Learning[5], 2008-2010).

(4) Kommunikationsverben[6] (online since 2013, finished). This dictionary is the electronic version of a handbook of German speech act verbs, the *Handbuch deutscher Kommunikationverben* (edited by G. Harras, S. Erb, K. Proost and E. Winkler. Berlin/New York: de Gruyter 2004-2007). It contains 241 entries on German verbs that refer to communicative actions, focusing on speech act verbs (e.g. 'schimpfen' [to scold], 'resümieren' [to summarize]). A specialty of this dictionary is that the semantic description takes place on two levels: the conceptual one (represented by the reference situation types) and the lexical one (represented in the word articles).

(5) Kleines Wörterbuch der Verlaufsformen im Deutschen[7] (online since 2013, finished). This small dictionary of aspectual forms in German presents German verbs with regard to their occurrence in three aspectual forms, the *am*-progressive, the absentive and the *beim*-progressive. The aim is to provide researchers, students and teachers with the largest and most easily searchable collection of evidence on over 900 verbs that illustrate the use of these forms. This dictionary will be moved to the OWID^{plus} platform in the future (cf. section 3).

4 https://www.owid.de/wb/sprw/start.html
5 http://www.sprichwort-plattform.org/sp/Projekt
6 https://www.owid.de/docs/komvb/start.jsp
7 https://www.owid.de/wb/progdb/start.html

(6) Deutsches Fremdwörterbuch – Neubearbeitung[8] (letters A-H of the revision online since 2016, work in progress). This German dictionary of foreign words (DFWB) describes and documents the vocabulary of today's learned everyday language, both in its current use and in its historical development from the respective date of borrowing to date. It currently comprises around 1,700 entries (with approximately 25,000 main and secondary headwords as well as approximately 130,000 corpus citations), e.g. 'Charakter' [character] with its numerous composites ('Charakterkopf' [striking head]' etc., as well as 'Nationalcharakter' [national character], etc.) and derivatives ('charakterlos' [unprincipled], etc.). The entire first edition (letters I-K edited by Hans Schulz, published 1913; letters L-Q edited by Otto Basler, published 1942; letters R-Z, edited by IDS, published 1977-1983[9]) has been retro-digitized and is fully available online. The DFWB is the most comprehensive dictionary in OWID. The new edition contains the above-mentioned number of entries from the more than 5,500 pages of the previously published volumes 1 to 7 (1995-2010). The data released in April 2019 contains 3,354 entries from approximately 3,400 pages of volumes 1 to 6 of the first edition (1913-1983).[10]

(7) Neologismenwörterbuch[11] (online since 2004, work in progress). This dictionary presents over 2,000 new words, new phrase units and new meanings of established words that were incorporated into the general part of the vocabulary of the German standard language between 1991 and today. The new vocabulary from the three decades – 1990s (e.g. 'Handy' [cellphone]), the first (e.g. 'skypen' [to skype]) and the second decade of the 21st century (e.g. 'Influencer' [influencer]) – can be searched using various access routes (by subject groups, via the advanced search) (cf. section B).

(8) Schulddiskurs 1945-55[12] (online since 2008, finished). This dictionary is the online version of Heidrun Kämper's printed reference work *Opfer – Täter – Nichttäter. Ein Wörterbuch zum Schulddiskurs 1945-1955* [Victim – Culprit – Non-Culprit. A

8 https://www.owid.de/wb/dfwb/start.html
9 For more detailed information see https://pub.ids-mannheim.de/laufend/fremdwort/auflage1.html?loop=2
10 For information on the different parts of „Deutsches Fremdwörterbuch" see: https://www.owid.de/wb/dfwb/uebersicht.html
11 https://www.owid.de/docs/neo/start.jsp
12 https://www.owid.de/wb/disk45/einleitung.html

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

The Leibniz Institute for the German Language (Leibniz-Institut für Deutsche Sprache, IDS) is the central non-university institution for the study and documentation of the contemporary usage and recent history of the German language. Together with more than 90 research and service institutions, it belongs to the Leibniz Association, one of the four major research organizations in Germany. At present, the IDS has 227 employees including 105 researchers. In the Department of Lexical Studies, more than 30 full and part time researchers and about 20 student researchers work in three so-called program areas on lexicography and language documentation (head: Annette Klosa-Kückelhaus), on syntagmatic aspects of the lexicon (head: Stefan Engelberg), and on empirical methods and the digital foundation of lexical studies (head: Carolin Müller-Spitzer). https://ids-mannheim.de

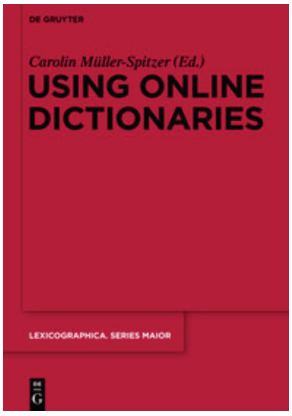
dictionary on the discourse of guilt 1945-1955] (Berlin & New York: de Gruyter 2007), which summarizes the lexical-semantic results of this investigation in the form of lexicographic entries. The entries describe those words according to lexicographic principles which represent the lexical framework of the discourse on guilt (having in the center the lexemes 'Pflicht' [duty] and 'Schuld' [guilt]).

(9) Protestdiskurs 1967/68[13] (online since 2012, finished). This dictionary is the online reference work on Heidrun Kämper's monograph *Aspekte des Demokratiediskurses der späten 1960er Jahre. Konstellationen – Kontexte – Konzepte* [Aspects of the Discourse on Democracy in the late 1960s. Constellations – Contexts – Concepts] (Berlin and Boston: de Gruyter 2012). Conceptually, it is a further development of the *Wörterbuch zum Schulddiskurs 1945-55*, containing around 90 articles on approximately 210 lemmas, each of which represent the discourse of left-wing students and intellectuals in a specific way.

(10) Schlüsselwörter der Wendezeit 1989/90[14] (online since 2014, finished). This dictionary is the online version of the reference book by Dieter Herberg, Doris Steffens and Elke Tellenbach *Schlüsselwörter der Wendezeit. Wörter-Buch zum öffentlichen Sprachgebrauch 1989/90* [Keywords of the time of the German reunification. Wordbook of words in public language 1989/90] (Berlin and New York: de Gruyter 2007). It represents the public language around 1989-1990, is consistently corpus-based, and was prepared for the online version on the occasion of the 25th anniversary of the German reunification. It gives information on the more than 1,000 words and phrases that are relevant to that time, ordered according to 150 keywords (e.g. 'Akte' [file] and 'Mauer' [wall]) and thematic groups (e.g. 'Die Deutschen vor und nach der Wiedervereinigung' [The Germans before and after reunification], 'Die politischen Veränderungen in der DDR und deren Vorboten' [The political changes in the GDR and its heralds]).

(11) LeGeDe – Lexik des gesprochenen Deutsch[15] (online since 2019, finished). This prototype of a dictionary of spoken German (conceptualized and implemented in a third-party funded, three year research project) offers a limited number of entries on nouns, adverbs, adjectives, verbs, etc. as well as multiword expressions which are characteristic for discourse in spoken language, e.g. the adjective

13 https://www.owid.de/wb/disk68/start.html
14 https://www.owid.de/wb/swwz/start.html
15 https://www.owid.de/legede/



Research into dictionary use at the IDS

The IDS has a strong focus on research into dictionary use. Besides numerous journal papers, the team also published the first book-length work on empirical research on *Using Online Dictionaries*. A detailed review is: Lew, Robert. 2015. Research into the use of online dictionaries. *International Journal of Lexicography* 28.2, 232–253.

'gut' [good], that can indicate the closure of a communicative task, or the expression 'keine Ahnung' [no idea], which may be used as a manifestation of uncertainty. For the first time in German lexicography, the lexicographic focus was not only placed exclusively on spoken language, but the dictionary entries are the first ones based entirely on corpora of spoken language, showing and explaining examples with (aligned) transcripts and audio files. Lexicographic information comprises details on the function of each word or expression in discourse, its function for signaling turns, its prosodic integration, etc., but also on how the use specifically for spoken language is connected to the respective word in written language (e.g. the expression 'keine Ahnung' is connected to the meaning 'knowledge' of the noun 'Ahnung' [idea, knowledge]). Our work on lexicographic information on spoken German is continued in a new project, which will cooperate with the new online research platform for spoken varieties of German outside the closed German language area in Central Europa (cf. section 4.B).

To summarize, in addition to retro-digitized online dictionaries (e.g., *Kommunikationsverben*), there are also dictionaries in OWID that were developed directly for online publication, e.g. *Sprichwörterbuch*. Besides completed dictionaries (e.g. *Schlüsselwörter der Wendezeit 1989/90*), there are some that are constantly worked on and are published dynamically (e.g. *Paronymwörterbuch*), and there are diachronic (e.g. *Deutsches Fremdwörterbuch*) as well as synchronic dictionaries

Recent journal articles on dictionary use

Wolfer, Sascha; Kosem, Iztok; Lew, Robert; Müller-Spitzer, Carolin and Ribeiro Silveira, Maria. 2018. Web-based exploration of results from a large European survey on dictionary use and culture: ESDexplorer. *Lexikos* 28, 440-447.

Müller-Spitzer, Carolin; Nied Curcio, Martina; Domínguez Vázquez, María José; Silva Dias, Idalete Maria and Wolfer, Sascha. 2018. Correct hypotheses and careful reading are essential: Results of an observational study on learners using online language resources. *Lexikos* 28, 287-315.

Kosem, Iztok; Lew, Robert; Müller-Spitzer, Carolin; Ribeiro Silveira, Maria and Wolfer, Sascha. 2018. The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32.1, 92-114.

Wolfer, Sascha; Bartz, Thomas; Weber, Tassja; Abel, Andrea; Meyer, Christian M.; Müller-Spitzer, Carolin and Storrer, Angelika. 2018. The effectiveness of lexicographic tools for optimising written L1 texts. *International Journal of Lexicography* 31.1. 1-28.

Wolfer, Sascha and Müller-Spitzer, Carolin. 2016. How many people constitute a crowd and what do they do? Quantitative analyses of revisions in the English and German Wiktionary editions 26/16. *Lexikos* 26, 347-371.

(e.g. *Neologismenwörterbuch*). A special feature of OWID are the two discourse dictionaries *Schulddiskurs 1945-55* and *Protestdiskurs 1967/68*, a type of dictionary that was developed at the IDS.

B. Access structures: OWID and *Neologismenwörterbuch*

The main function of OWID is to provide a common access structure in the form of search options across the individual dictionaries. This is the typical function of lexicographic portals (e.g. YourDictionary.com, Wörterbuch-Portal) (cf. Müller-Spitzer and Engelberg 2013).

In OWID, there is a clear distinction between the level of the portal and the level of an individual dictionary. The search box of the portal is always accessible on the top of the webpage, while for each of the dictionaries, specific access structures are offered, which are shown once a user selects a certain dictionary by clicking on the dictionary button. With this distinction, we address two different user needs: firstly, searching for one word in any dictionary or, secondly, searching within one specific dictionary only.

Some of the OWID dictionaries offer advanced search options. An outstanding feature of OWID is that we try to develop appropriate advanced searches for each dictionary and use very diverse technologies to do so. For the German dictionary of foreign words (DFWB), for example, a rather narrative dictionary, we have invested primarily in developing effective full-text search. Since the lexicographic information in this dictionary is not granularly structured, a good full-text search is the most interesting way to access it. The full-text search can be narrowed down according to different text levels (keywords, article text, examples). On the other hand, in the neologism dictionary the situation is quite different, as the entries are structured in a fine-grained way. Therefore, we have applied another search technology and developed a user interface that aims to provide exactly these fine-grained structures to the end user with very distinguished search options. The advanced search[16] here enables users to find all keywords that have a common feature (e.g. all new lexemes that were borrowed from languages other than English in the 1990s, as shown in Figure 2).

In addition, users can select the page 'Wortartikel'[17] [entries] as a starting point for further exploration of the dictionary content. Different lists group all entries according to different criteria:

Type of headword (single word entries, multiword units, new elements

16 https://www.owid.de/docs/neo/suche/index.jsp
17 https://www.owid.de/docs/neo/wortartikel.jsp

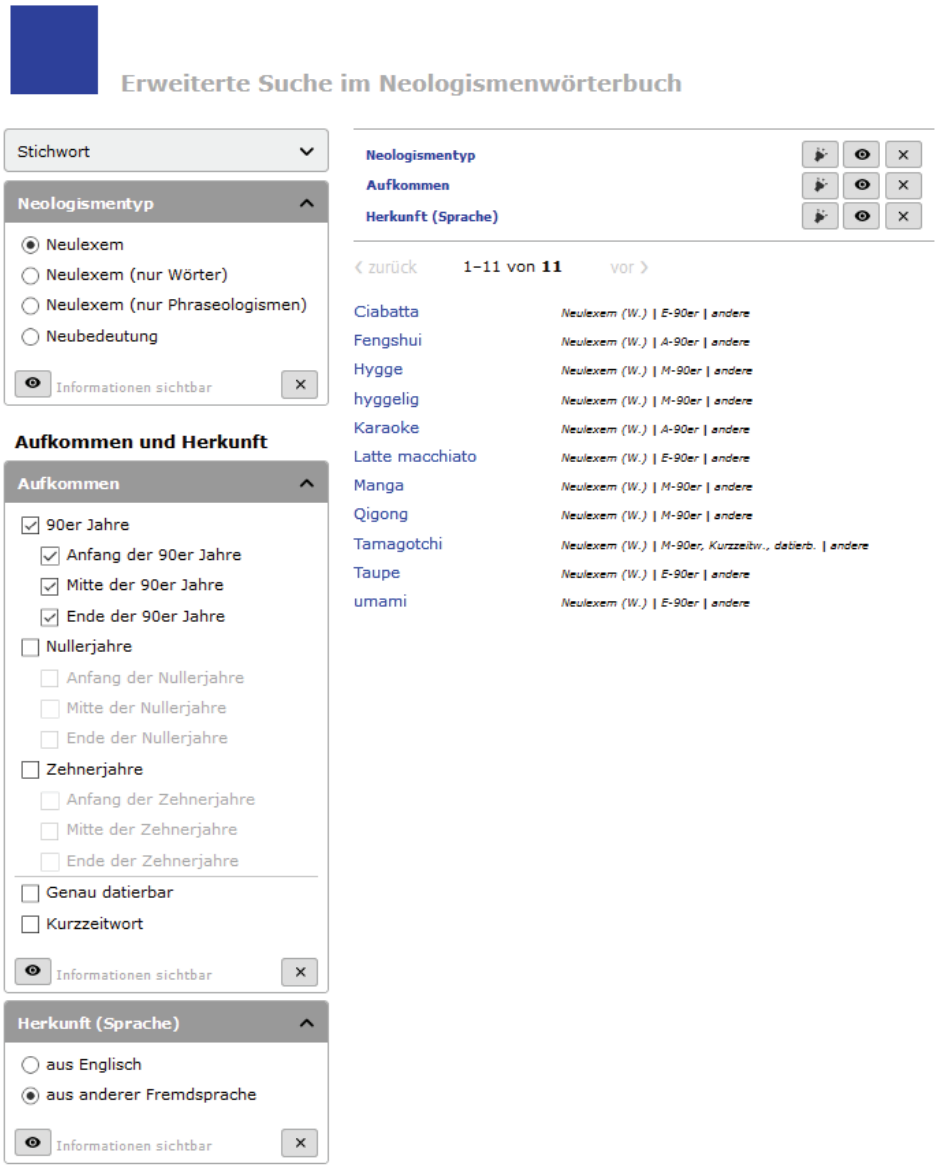


Figure 2.
Advanced search in *Neologismenwörterbuch* showing new lexemes from the 1990s borrowed from languages other than English

of word formation plus neologisms which are not headword but are contained in entries, e.g. lesser-used synonyms, derivations and compounds with the lemma).

Decades (neologisms from 1991-2000, 2001-2010, and 2011-2020).

Different types of entries such as the latest full entries (published always at the end of a calendar year), groups of short entries (published monthly, each related to a single topic, e.g. Europe, education), a list of neologisms that might enter the dictionary at some point and which are still monitored, or the most recent list of neologisms relating to the coronavirus pandemic (as shown in Figure 3)

Subject groups, for example sports, media, health and wellbeing, economics.

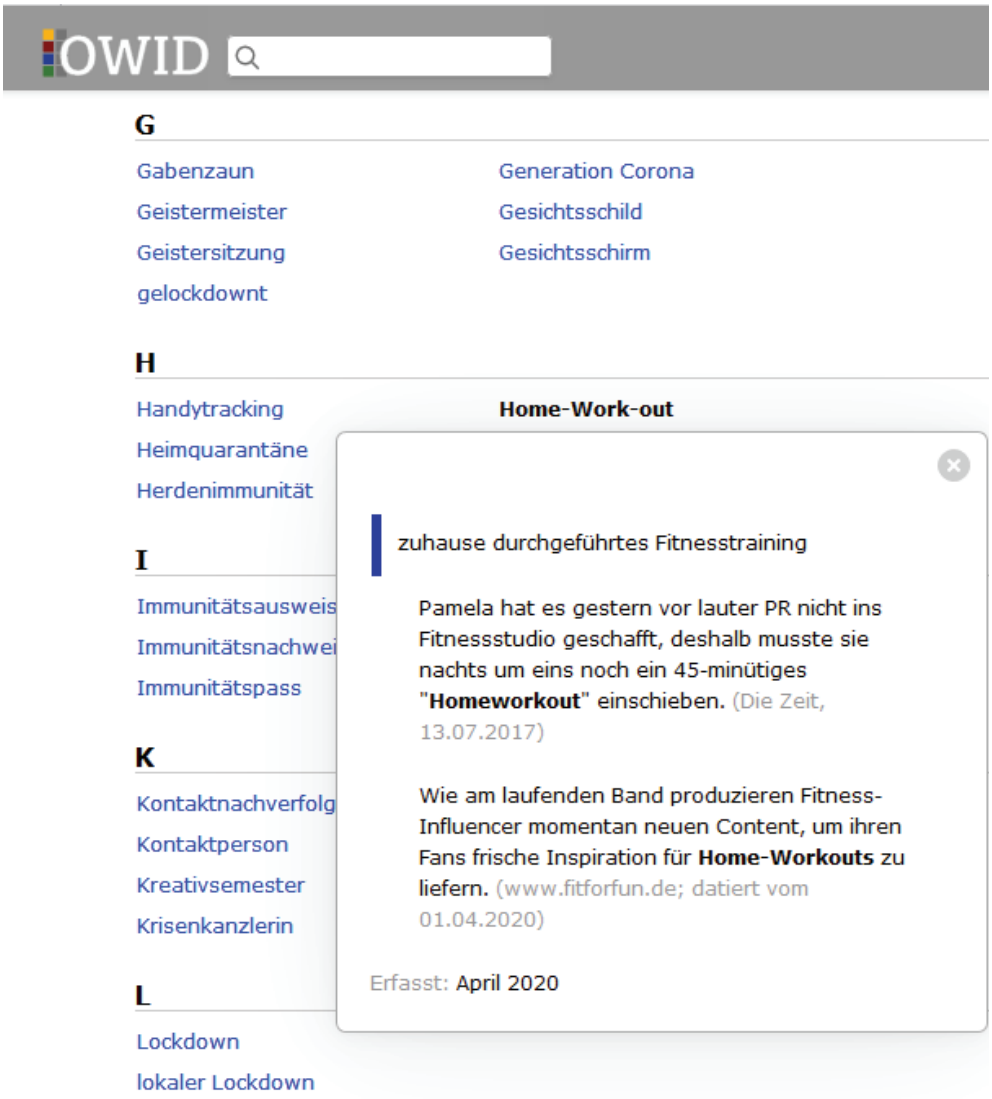


Figure 3.
Extract from the list of monitored neologisms relating to the coronavirus pandemic

The interactive listing of entries according to their subject group[18] is combined with the presentation of all relevant entries according to decades, thus enabling users to gain information on the social, cultural, technical, economical, etc., developments over the last 30 years. In addition, the *Neologismenwörterbuch* offers various possibilities for accessing the dictionary content on its homepage[19] (e.g. direct links to lists of latest additions, the advanced search and entries in subject groups plus a link to suggest a new word), where also a very short introduction into the content is given.

Over all, users find many different options to explore the *Neologismenwörterbuch*. Once an entry has been selected and is presented on screen, exploration inside and outside of the dictionary

18 https://www.owid.de/docs/neo/gruppen.jsp
19 https://www.owid.de/docs/neo/start.jsp

may continue by following one or the other of numerous (dictionary internal and external) hyperlinks contained in the entries.

C. Information types: Examples from different dictionaries in OWID

All dictionaries in OWID analyse and describe the entries on the basis of extensive empirical, mostly corpus-derived, linguistic data. These dictionaries are products of scholarly lexicography and are the result of lexicological-lexicographic and metalexicographic research. Although most of them focus on specific parts of the vocabulary and not the general language, through their connection in OWID they offer fascinating insights into the German vocabulary, as shown in the following examples.

A search for 'smart' in OWID yields several results (cf. Figure 4) from three different dictionaries, namely *Deutsches Fremdwörterbuch*, *elexiko* and *Neologismenwörterbuch*. While *Deutsches Fremdwörterbuch* focuses on etymology and historical sense development of the adjective and noun derivations 'Smartness' and 'Smartheit' [being smart] as sub-lemmas, in *elexiko* three senses are disambiguated (clever, chic, technically highly developed) and explained from a synchronic view, focusing especially on judgmental usages specific for the first two senses. The third sense is the one found in compounds like 'Smart-TV' or 'Smartboard', neologisms of this millennium which are recorded in *Neologismenwörterbuch*. Here, users learn, among other things, that both loanwords stem from English and are productive in the formation of new

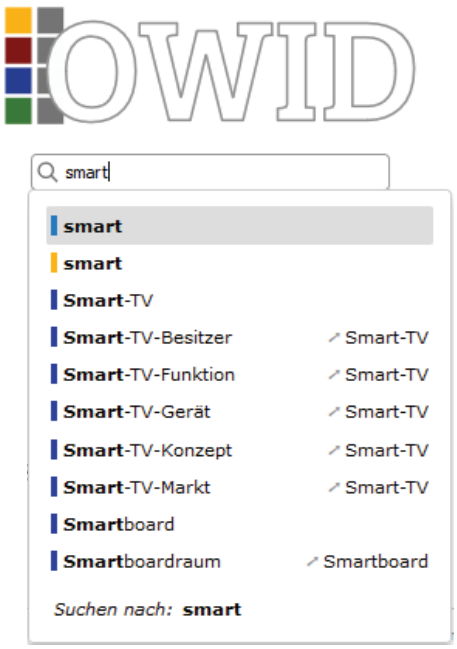


Figure 4.
Search results for 'smart' in OWID

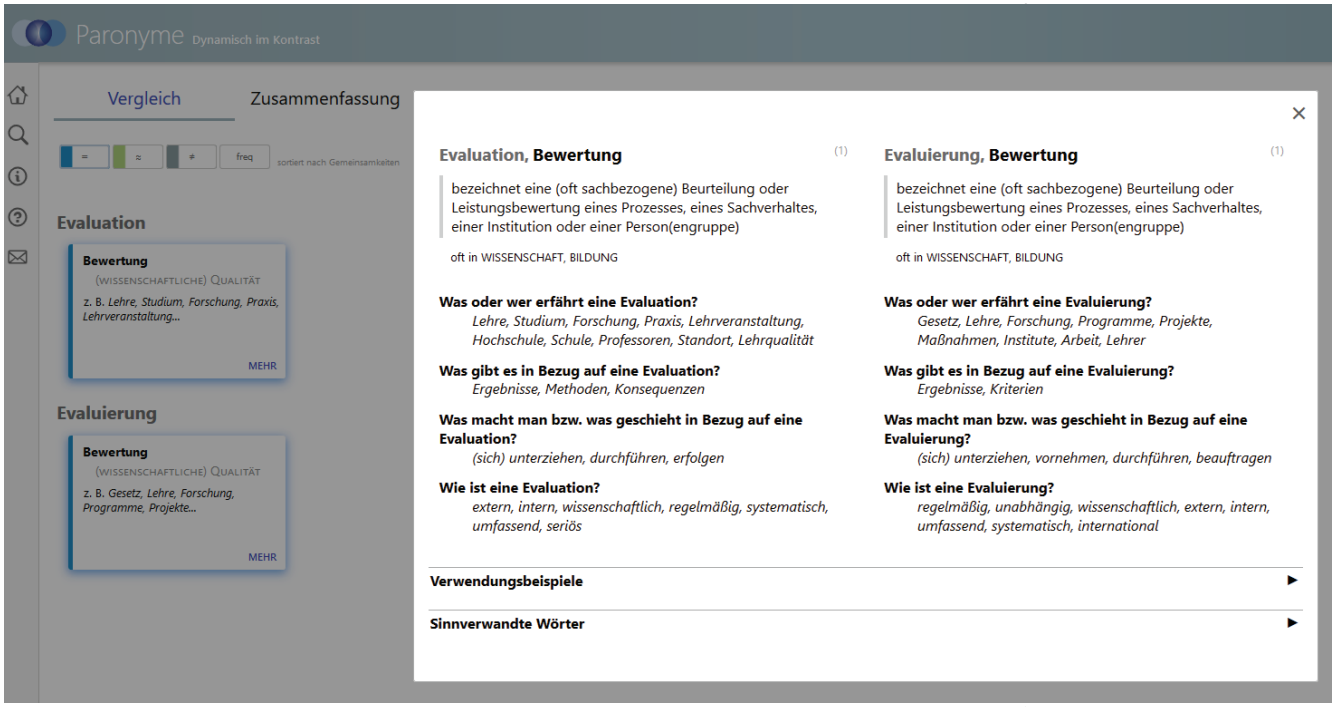


Figure 5.
The entry 'Evaluation – Evaluierung' from *Paronymwörterbuch* presenting information on both words simultaneously, enabling users to compare them

compounds such as 'Smartboardraum' [room with a smart board] or 'Smart-TV-Konzept' [smart TV concept].

A search for the string 'evalu' yields the following results (among others): the entry 'Evaluation – Evaluierung' [evaluation] from *Paronymwörterbuch* and the verb 'evaluieren' (to evaluate) in *Kleines Wörterbuch der Verlaufsformen im Deutschen*. Here, users find the information that 'evaluieren' is used in the progressive form with 'am' (as in 'Die Forscher sind noch am evaluieren' [The researchers are still evaluating]), which is illustrated in many corpus citations. In *Paronymwörterbuch*, the two nouns derived from this verb are compared with regard to similarities or differences in their meaning and use. In a highly innovative way of presenting the data distributed on different tiles in a horizontal order, users are able to understand that Evaluation and Evaluierung are used synonymously in most contexts, especially in scholarly language (cf. Figure 5).

When looking for the string 'leben', OWID yields (among others) the entry 'Leben' [life] in *Schulddiskurs 1945-55* and a number of proverbs from *Sprichwörterbuch*: 'Leben und leben lassen' [live and let live], 'Ordnung ist das halbe Leben' [a tidy house, a tidy mind], 'Totgesagte leben länger' [there's life in the old dog yet], and 'Wer zu spät kommt, den bestraft das Leben' [life punishes the latecomer]. For each proverb, not only are the meaning and usage explained and many corpus examples given, but also information on variability, for example 'X ist das halbe Leben' [X is half the life] and 'X und X lassen' [live and let X live], are provided along with many examples

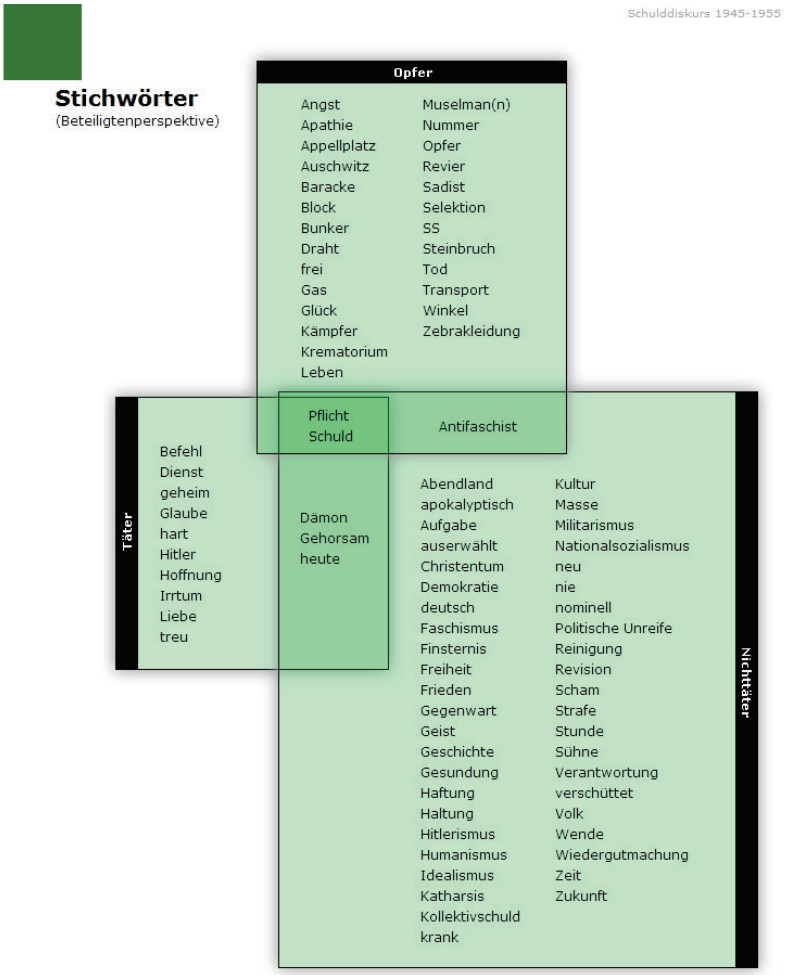


Figure 6.
Access structure in *Schulddiskurs 1945-1955* for the lexical items belonging to the discourse on guilt by victims (Opfer), culprits (Täter), and non-culprits (Nichttäter)

for each variant. In *Schulddiskurs 1945-55*, the noun 'Leben' [life] is described as one of a number of concepts as used by Holocaust victims (cf. Figure 6 showing the words used by victims, culprits and non-culprits). After being liberated, 'Leben' for them acquired the specific meaning of "life given back" and it is used often in contrast to those, who lost their lives, as shown in the essayistic entries in this dictionary.

Over all, most of the dictionaries contained in OWID are not only innovative in choosing specific parts of German vocabulary as dictionary matter, but also in developing new types of lexicographic information, by consistently linking between lexicographic information and corpus data, and in presenting information to users in new ways adapted to each dictionary type.

3. OWID^{plus}

With OWID^{plus}, a new experimental platform was established to complement the dictionary portal OWID. The background for creating OWID^{plus} was that the variety of lexicological-lexicographic data of

interest for publication now extends far beyond digital dictionaries. From a data perspective, digital dictionaries are resources prepared for end users. As a general rule, they are presented in such a way that they can be used without prior knowledge. Therefore, resources that are more likely to appeal to a specialist audience fit less into a general dictionary portal. In addition, it is essential for a dictionary portal such as OWID that all the dictionaries included have at least a few central similarities. Only in this way, the portal is able to offer a uniform user concept and common access structures. With OWID, these lowest common denominators are the access by words or by word units and the restriction to the German language as the dictionary matter.

In the course of our research work and the contact with external colleagues, however, it became increasingly clear that internal and external research projects often produce data sets that are not prepared for end users, but which are too valuable for the professional community not to be published. In OWID^{plus}, we provide space for a wide variety of resources, also multilingual ones. The individual resources are modularly implemented as independent interactive applications. Whether it might be useful in the long run to create a common index for all resources in OWID^{plus} is still open at the moment. Currently we are working on a common faceted search option of OWID and OWID^{plus}, because with the growing number of resources, the existing interface of OWID^{plus} becomes too heterogenous.

At present, 15 different resources can be found in OWID^{plus}, for example (i) the *Lexical Explorer*[20], with which quantitative corpus data on spoken German can be researched on the basis of frequency tables regarding the distribution over word forms, co-occurrences and metadata, (ii) the *ZAS Database of Clause-Embedding Predicates*[21] of the Leibniz-Centre for General Linguistics (Leibniz-Zentrum Allgemeine Sprachwissenschaft) in Berlin, which contains clausal complementation patterns of lexical predicates in several languages, including multiple historical stages of German, (iii) a resource for log file statistics of six Wiktionary language editions[22], and (iv) various visualizations of lexical change, for example Lexical change in *Der Spiegel*[23]. With the last resource we would like to encourage users to move from passive consumption of linguistic knowledge to active exploration of a limited textual basis (all texts of the news magazine *Der Spiegel* from 1947 to 2016) and a limited phenomenon (here,

20 https://www.owid.de/lexex/
21 https://www.owid.de/plus/zasembed/
22 https://www.owid.de/plus/wikivi2015/index.html
23 https://www.owid.de/plus/wwspiegel2018/

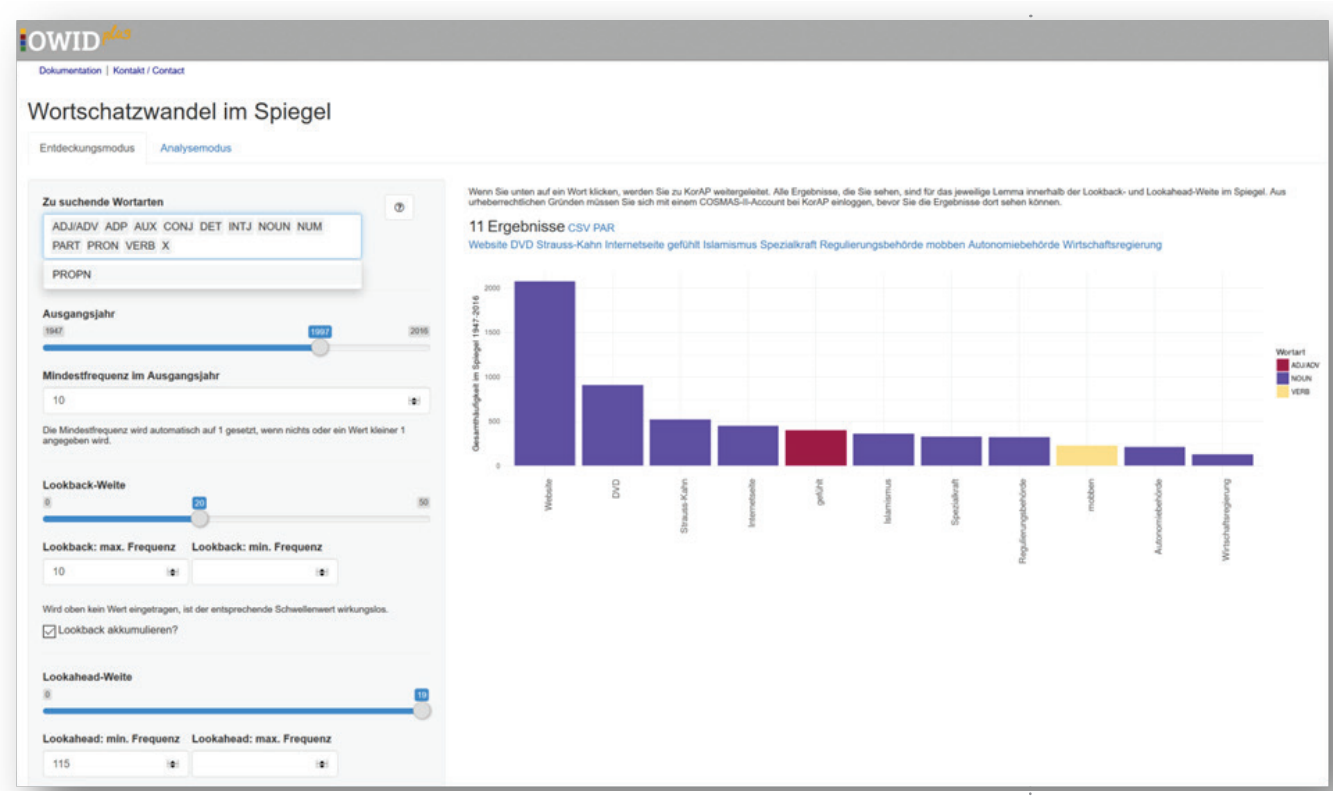


Figure 7.
Results from 'Wortschatzwandel im *Spiegel*' from 1997

lexical innovations and archaisms). With a few settings, lists and frequency diagrams of emerging or dying word forms can be created. For example, it is possible to see that the word forms 'website', 'DVD', 'internet page' and 'mobbing' appeared in 1997 and continue to be frequently used since then (cf. Figure 7). The tool can be usefully applied in teaching to demonstrate the quantitative research of lexical innovation, or generally used to pursue one's own linguistic urge to explore more information.

The architecture of OWID^{plus} enables us to publish new resources very quickly. We have recently used this opportunity to upload two brand new resources on the occasion of the corona pandemic: cOWID^{plus} Analysis[24] and cOWID^{plus} Viewer[25]. The motivation for these resources was that all around the globe, the coronavirus pandemic has been affecting almost every part of public life. Consequently, the pandemic is the subject of discussion not only in private face-to-face conversation (whenever this kind of talk has been possible), but also in the news. With lots of daily life activities like sports and cultural events coming to a stop, corresponding newspaper desks might very well run out of events to report on and shift their focus to pandemic-related topics. This gives rise to the assumption that the vocabulary used in articles,

24 https://www.owid.de/plus/cowidplus2020/
25 https://www.owid.de/plus/cowidplusviewer2020/

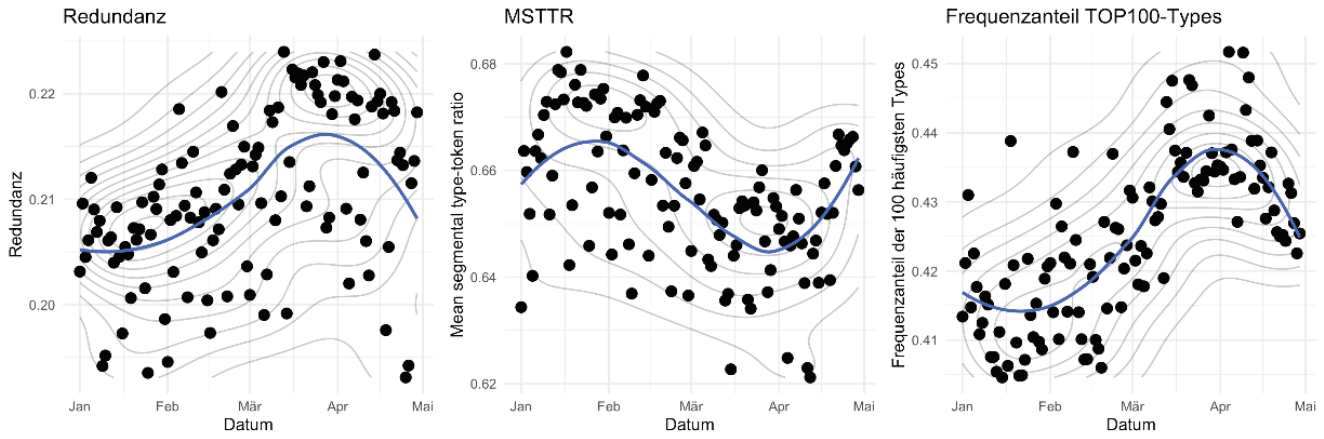


Figure 8.
Development of various lexical diversity measures from January to May 2020 (Redundanz = redundancy, MSTTR = Mean segmental type-token ratio, Frequenzanteil Top100-Types = frequency proportion of Top100 Types)

not only in printed but also in online media, is changing. To be precise, we would assume a concentration of (content word) vocabulary on concepts that are associated with the coronavirus pandemic. This does not necessarily mean that fewer word types are used. Rather, it suggests a shift in frequency distribution over types in a way that such distributions are more heavily skewed towards (temporarily) important types. Such a change is detectable by quantitative measures. In cOWID^{plus} Analysis we can indeed see a measurable narrowing of topics (and hence of the vocabulary) during the corona pandemic in selected (online) news outlets in German language, especially in mid-March 2020 (cf. Figure 8).

The cOWID^{plus} Viewer enables users to explore the data of cOWID^{plus} Analysis. An easy-to-use interface enables visualisation of the frequency curves of word forms. Data and illustrations are also available for download and are updated weekly. It is especially the topicality of the data that allows us to quickly discover and follow interesting lexical trends of the corona crisis. Figure 9 gives an example: at the beginning of the corona crisis, panic shopping (in German 'Hamsterkäufe', literally "hamster purchases") were a big issue. All over the world different things were "hoarded", in Germany it was mainly toilet paper ('Klopapier') and flour ('Mehl'). This is apparent in the news feeds as well: mid-March was the climax of the "hamster topic".

All resources in OWID^{plus} are primarily aimed at the linguistic community. Accordingly, we are always interested in receiving feedback on the existing offers as well as in offers to provide further data.

4. Two Internet portals for the lexicography of language contact

A. The Loanword Portal for German

The Loanword Portal for German ('Lehnwortportal Deutsch') documents words that have been borrowed from German into other languages. It is mostly not based on original etymological research but integrates already-existing loanword dictionaries or ongoing dictionary projects. Methodologically, it differs from OWID in

Lehnwortportal Deutsch

Published within the Loanword Portal for German are the following lexical resources:

Belarusian, Russian, and Ukrainian. Hentschel, Gerd et al. *Dictionary of German loanwords in the East Slavic languages with a parallel in Written and Standard Polish*. [compiled for the portal]

Czech and Slovak. Newerkla, Stefan Michael (2nd ed., 2011): *Sprachkontakte Deutsch – Tschechisch – Slowakisch. Wörterbuch der deutschen Lehnwörter im Tschechischen und Slowakischen*. Frankfurt am Main: Peter Lang. [digitized from a PDF text file]

Dutch. Veen, Pieter Arie Ferdinand van and van der Sijs, Nicoline (1997): *Etymologisch woordenboek: de herkomst van onze woorden*. Utrecht: Van Dale Lexicografie. [retro-digitalized, excerpts]

English. Pfeffer, J. Alan and Cannon, Garland (1994): *German Loanwords in English. A Historical Dictionary*. Cambridge, New York, Melbourne: Cambridge University Press. [retro-digitalized]

French. (1) *Le Trésor de la Langue Française Informatisé*, online: http://atilf.atilf.fr [relevant excerpts converted from another digital format]; (2) Sarcher, Walburga (2001): *Das deutsche Lehngut im Französischen als Zeugnis für den Wissenstransfer im 20. Jahrhundert*. Hamburg: Verlag Dr. Kovač [retro-digitalized]

Hebrew. Adiv, Uriel and Mendel, Jakob (2015ff.), *Deutsche Lehnwörter im Hebräischen*. [compiled for the portal]

Hungarian. Benkő, Loránd and Büky, Béla (1993-1997): *Etymologisches Wörterbuch des Ungarischen*. Budapest: Akademiai Kiadó. 3 vols. [retro-digitalized, excerpts]

Polish. de Vincenz, Andrzej and Hentschel, Gerd (2010), *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache*. [converted from another digital format]

Portuguese. Schmidt-Radefeldt, Jürgen and Schurig, Dorothea (1997): *Dicionário dos Anglicismos e Germanismos na Língua Portuguesa*. Frankfurt/M.: Teo Ferrer de Mesquita. [retro-digitalized, entries on German loanwords]

Slovene. Striedter-Temps, Hildegard (1963), *Deutsche Lehnwörter im Slovenischen*. [retro-digitalized]

Teschen dialect of Polish. Menzel, Thomas and Hentschel, Gerd (2005), *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen*. [converted from another digital format]

Tok Pisin. Engelberg, Stefan; Möhrs, Christine and Stolberg, Doris (2017ff.), *Wortschatz deutschen Ursprungs im Tok Pisin*. [compiled for the portal]

The following resources are going to be added in 2021 (some licenses pending):

Uzbek. Jumaniyozov, Atabay and Vohidova, Nofiza (in preparation): *Dictionary of German loanwords in Uzbek* [retro-digitalized, edited and translated version of an Uzbek monograph]

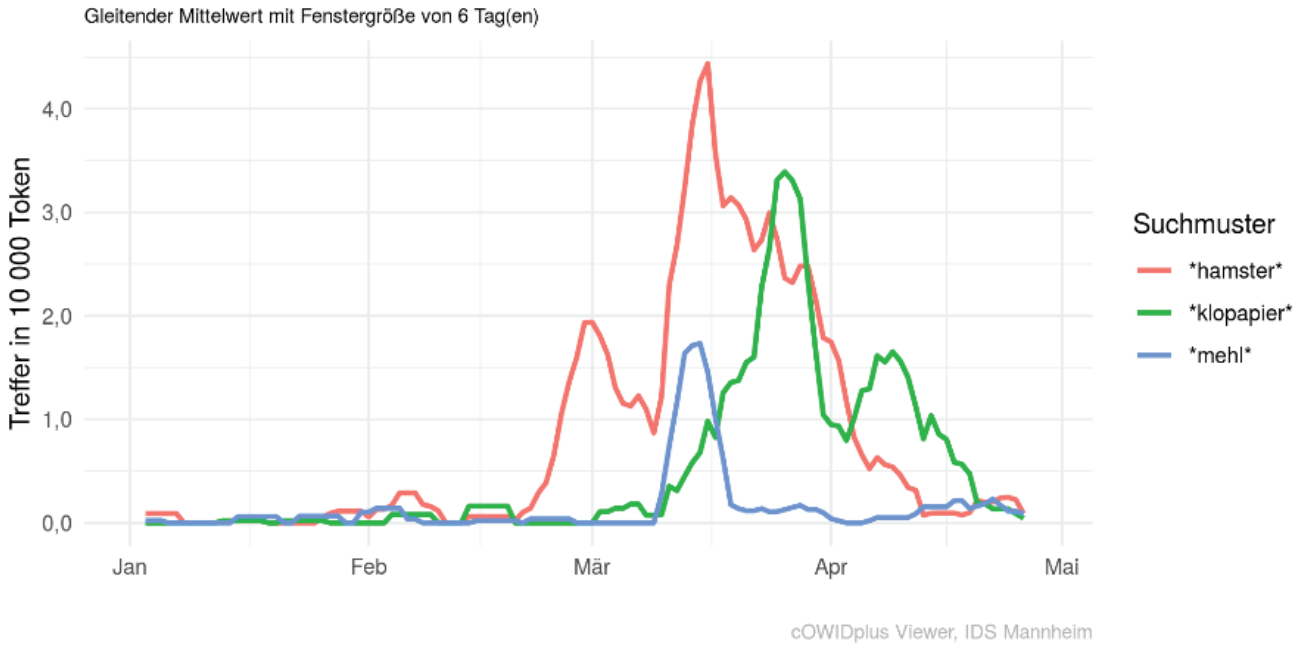


Figure 9.
hamster, *klopapier* and *mehl* in cOWID[plus] Viewer

two respects, as it does not primarily include dictionaries that were compiled in-house and (therefore) does not have a thorough corpus-lexicographic basis. The integrated dictionaries, which usually have lemmatized words in the target language of the borrowing process, can be consulted as individual works; however, in the database of the portal all information is additionally represented as a complex, cross-dictionary network of loanwords and words of origin. Thus, the portal can be used as a "reverse loanword dictionary": in the dictionary of words of origin, generated from a German meta-lemma list, it is possible to search for related loanwords in other languages.

The loanword portal in its current version contains dictionaries of German loanwords in Hebrew, Polish, Slovene, the Teschen dialect of Polish, and Tok Pisin. A new, entirely revised, version, to be released in mid-2021, will also include dictionaries of German loans in Czech, Dutch, the East Slavic languages Belarusian, Russian, and Ukrainian (restricted to cases with a corresponding loan in Polish), English, French, Hungarian, Portuguese, Slovak, and Uzbek.

The portal stands out due to its complex search options that enable expert searches of lexical items in the database network based on arbitrary combinations of search features (target language, time of borrowing, dialect of source item, semantic components, grammatical category, relationship to other words in the network, etc.).

B. A language variety platform for German languages enclaves

A new project at our institute is in its early stages of development. Its aim is the implementation of a lexicological-lexicographic

Figure 10.
Search for 'Adresse' [address] in the German meta-lemma list of the *Lehnwortportal Deutsch*

online research platform for spoken varieties of German outside the closed German language area in Central Europe. Varieties of this sort are spoken on all five continents and include (i) varieties that show a large degree of dialect levelling and therefore often exhibit a considerable overlap with the lexicon of standard German, for example Barossa German in Australia, Namibian German, or Paraguayan Laguna German, (ii) varieties that retain strong features of German dialects and therefore show a considerable difference from the standard German lexicon, for example Puhoi Bohemian German in New Zealand, or Siberian Mennonite East Low German, and (iii) German-based pidgin and creole languages such as Kiche Duits in Namibia or Unserdeutsch from New Guinea.

We plan to develop a modular online platform for variational and contact lexicology that dissolves the distinction between dictionary and corpus and accommodates the needs and interests of lexicographers, linguists and non-expert speakers of the included varieties. The platform builds on the increasing number of digitally searchable corpora of the spoken language of these varieties that adhere to current standards of transcription and annotation. The envisaged system will allow lexicologists to supplement standard metadata and token annotations of such corpora by custom annotation layers for specific research questions and additional variational parameters. Researchers and end users will be provided with advanced tools and presentation options for lexicographic documentation, quantitative analyses and information extraction beyond standard corpus query technology. The integrated toolset will make it possible

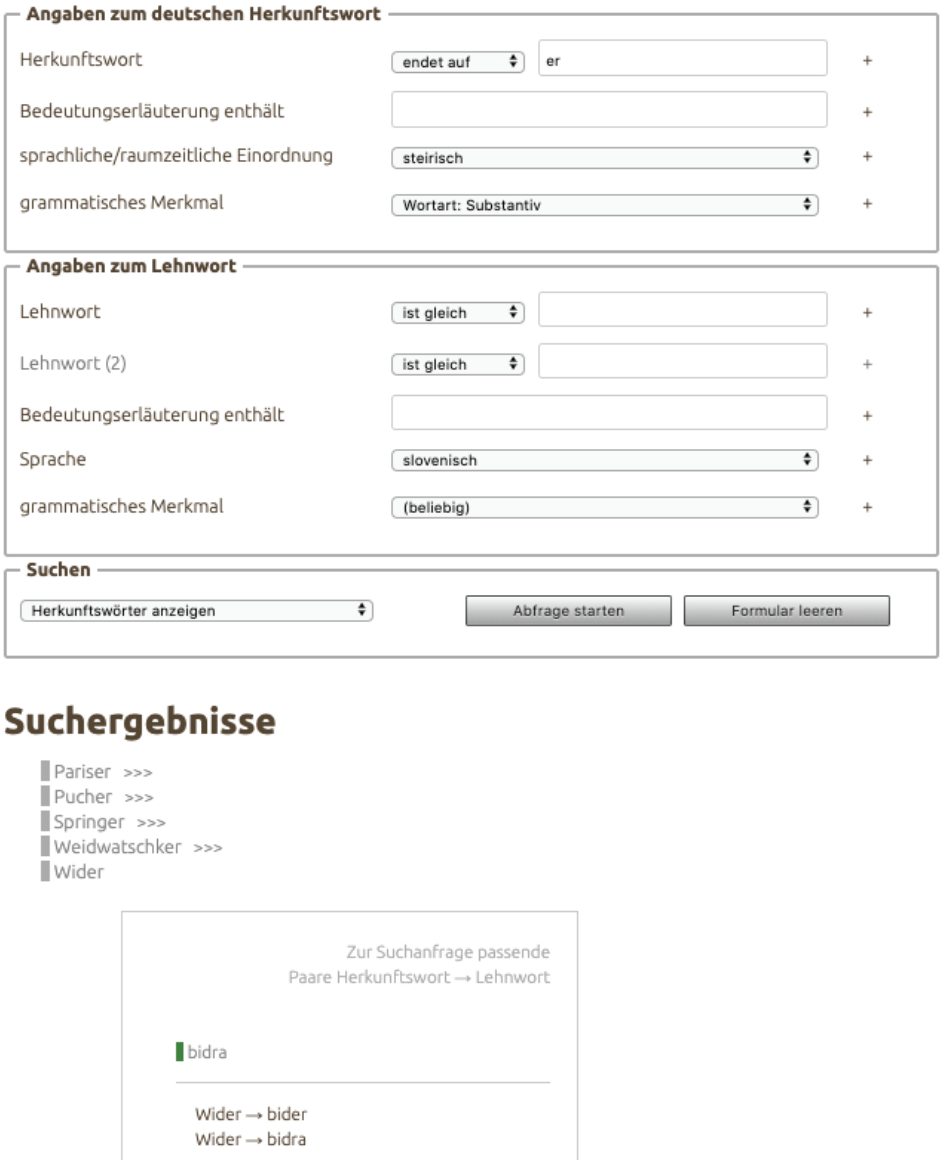


Figure 11.
Complex search in the *Lehnwortportal Deutsch*: nouns ending in '-er' (an instrumental suffix) borrowed from the Styrian dialect of German into Slovene

to systematically address questions of intra-variety linguistic variation in domains such as register-specific collocations, phenomena at the interface between lexicon and grammar, lexical specifics of spoken language, or lexical contact phenomena. The first test case will probably be Namibian German, and as soon as the number of varieties integrated into the platform increases, the portal will also allow for comparative studies, both between these post-migration varieties and in relation to spoken varieties of German as a majority language in Germany. The details of the development of the platform will depend on both funding and the interest of linguistic colleagues in cooperating with us.

5. Competences and workflow

A: Consequences of changing lexicographic processes

When we began to develop OWID, the overall Internet lexicographic workflow was fairly straightforward: modelling XML structures, writing entries in an XML editor, storing entries in an Oracle database, creating stylesheets for HTML-based data presentation, and programming access structures. With more diversity in Internet dictionaries and portals and an increasing number of external cooperation partners, the workflow becomes more complex. Standardizing processes and structures is a strong demand on the one hand, while on the other hand, we want to develop new forms of lexicographic data presentation, new lexicographic methods, and new steps within lexicographic processes adapted to these methods. In the following, we sketch some developments in our department which determine work processes and staff recruitment.

Corpus analysis. The Leibniz-Institute of the German Language has a department for research infrastructure and corpus linguistics. The corpora and the corpus analysis systems developed in this department are used fairly intensively by our lexicographers. However, with our specific lexicological research questions and lexicographic projects, we often need certain corpora and methods of analysis that cannot all be provided by our in-house corpus linguistics resources. Therefore, we have to rely more and more on expertise on corpus linguistics within our own department and a larger part of the staff budget has to be shifted to meet this demand.

Quantitative methods. Quantitative analysis of corpora has become more important for lexicological and lexicographic purposes, which becomes particularly obvious in some of the resources in OWID^{plus}. Ten years ago, quantitative linguistics hardly played a role in the department; meanwhile, we have hired several persons who work mostly or partly with quantitative methods.

Collaboration. Lexicography is a labor-intensive process. Sooner or later it turns out that you cannot achieve everything on your own. During the last decade we have established a number of lexicographic co-operations, in particular with universities in Germany and several European countries. This pertains partly to OWID^{plus} and moreover to the loanword portal and the emerging variation platform, where the lexicographic content is almost exclusively produced with our external partners. These co-operations, in turn, require the development of specific lexicological and lexicographic tools.

New platforms. Our new Internet platforms require technical solutions

that go beyond what we had implemented for OWID. Besides the relational database (Oracle) used for OWID, we now employ also a full-text search engine (Elasticsearch) in OWID and NoSQL storage and retrieval (Neo4J for graphs, BaseX for XML) in the loanword portal and projects therein. The collaboration with lexicographers outside the institute led to developments affecting the lexicographic process. While many of our in-house lexicographers compile dictionary entries using the commercial XML editor Oxygen, the special needs of our collaborative projects required the programming of dedicated online editorial tools based on open-source XML and rich text editing components. In general, web development has become much more frontend-centered in recent years, increasingly replacing server-side generation of static browser content by interactive presentations and visualisations driven by modern reactive technology that runs in the browser.

Sustainability. The more lexicographic resources we produce, and the more diverse they are, the more work we have to allocate to measures of maintenance and sustainability. This includes trying to standardize procedures and formats where it is possible to develop software of a more generic sort. In spite of these efforts, the number of lexicographic products and services that have to be accompanied through the change of technical surroundings is growing, which creates a rising demand for computational lexicographic support.

B. Staff recruitment

The last ten years at our department have seen lexicological studies and lexicographic practice growing closer together. Many of our employees now conceive and compile dictionaries on the one hand and publish on empirical lexicological research on the other. In our paper from 2009, we estimated that the work of about 10-12 full-time equivalents of staff distributed over 18 members was allocated to Internet lexicography. While the overall time devoted to lexicography has not changed much (ca. 12 FTE), almost everybody in the Department of Lexical Studies is now involved part-time in one way or the other in lexicographic activities.

More than we did ten years ago, we advertise vacancies that combine expertise from various fields: in linguistics and lexicography, in corpus linguistics and computational lexicography, in lexicography and text technology, etc. We still do not hire IT specialists without a thorough background in the humanities, since the close collaboration among these domains in our department necessitates a common background in the conception and implementation of linguistic and lexicographic projects. A particular challenge arises from the

fact that IT expertise is in high demand in the private sector. As a publicly-funded institute, being restricted to standard wages, we have to offer a particularly interesting and inspiring work environment to attract highly qualified and motivated applicants with competences in these domains.

6. Outlook

We have shown how lexicography has changed at the Institute for the German Language within the last ten years. We have further developed our main dictionary portal OWID into a dictionary platform for specific-domain dictionaries in areas that have not been covered adequately by lexicological research and lexicography. Beyond that, new platforms have been designed and implemented: OWID^{plus} as an experimental platform for different kinds of lexical resources and tools, the Loanword Portal for German as a dictionary network for the publication of resources on German loanwords in other languages, and – forthcoming – a lexicological and lexicographic platform for spoken varieties of German outside the closed German language area in Central Europe. With these developments in progress, there is still an increasing demand for expertise in corpus analysis methods, text technology and web design but also for competences in the creation of new lexicographic formats and the lexicographic integration of current lexicological research.

The lexicographic resources at our department now reflect two different basic types of lexicography: communication-oriented dictionaries are directed more towards lay people and serve to solve communication problems or support language acquisition, whereas knowledge-oriented dictionaries mainly address a linguistically informed audience and document linguistic knowledge about the lexical system of a language (cf. for similar distinctions Tarp 2008; Engelberg 2014; Engelberg, Klosa-Kückelhaus and Müller-Spitzer 2019). However, there are of course over-arching principles of scholarly lexicography that both types of dictionaries have to adhere to: a thorough empirical basis of all lexicographic information, a concept of the nature and breadth of lexical knowledge, and a user-orientation with respect to access structures and information presentation.

References

Engelberg, S. 2014. Gegenwart und Zukunft der Abteilung Lexik am IDS: Plädoyer für eine Lexikographie der Sprachdynamik. In *Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, 243-253.

Engelberg, S., Klosa, A, and Müller-Spitzer, M. 2009. Challenges to Internet lexicography: The Internet dictionary portal at the Institute for German Language. *Kernerman Dictionary News*, 17: 16-25. https://kdictionaries.com/kdn/kdn17_2009.pdf

Engelberg, S., Klosa-Kückelhaus, A. and Müller-Spitzer, C. 2019. Lexikographie zwischen Grimm und Google? *Sprachreport* 35.2, 30-34.

Engelberg, S. and Müller-Spitzer, C. 2013. Dictionary portals. In Gouws, R., Heid, U., Schweickard, W. and Wiegand, H.E. (eds.), *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin and Boston: de Gruyter, 1023-1035.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer.

Understanding English Dictionaries: The first MOOC about lexicography

Michael Rundell

For those who aren't familiar with the subject, a MOOC (Massive Open Online Course) is a way of delivering educational content via the web to an unlimited number of participants, and courses are typically free. MOOCs have been around for some time. Lancaster University's well-regarded corpus linguistics MOOC, for example, was launched as far back as 2014. But the growth in online educational resources has dramatically accelerated, as the COVID-19 pandemic has – for the time being – made face-to-face learning impracticable. We had little idea of what was to come when we began creating the first dictionary-related MOOC in late 2017, but recent developments have made courses like this even more relevant.

Understanding English Dictionaries is a MOOC produced through a collaboration between experts at Coventry University (Hilary Nesi and Sharon Creese), The Alan Turing Institute (Barbara McGillivray), and Macmillan Education (Katalin Süle and Michael Rundell). The six-week course is hosted on the FutureLearn platform, and is free to join. It first ran in September/October 2019, and its success led to a second run in January/February 2020. At the time of writing, it is in the middle of a third iteration. Over its first two outings, the Dictionaries MOOC attracted a total of almost 5,000 learners from all over the world. (We don't yet have participant figures for the third iteration.) The word *English* in our MOOC's title shouldn't be interpreted in too restrictive a way. This simply reflects the background and experience of the course creators. In reality, most of the MOOC's content is equally relevant to dictionaries in any language.

The background to the course is described in Creese et al. 2018, where we make the point that it "is emphatically not a training course in dictionary-making … [but] … is intended to provide a lively introduction to a topic about which non-experts often have strong opinions but little real knowledge". Our working assumption was that, although almost everyone uses dictionaries, most people know very little about them. This turned out to be broadly accurate. A minority of participants had some prior knowledge of the field and at least a basic understanding of language corpora and their use in lexicography, but most of our participants came to the MOOC without much grasp of how dictionaries are created. It is easier than ever for people to consult

a dictionary, on their phone or laptop, without having any awareness of which dictionary the entry came from, why the word was included in the dictionary, what kind of evidence the entry was based on, or how far they could rely on the information provided. One of our primary goals, therefore, was to challenge common misconceptions about dictionaries and dictionary-making. The format of the MOOC makes it easy to track participants' changing attitudes as the course progresses.

In the course of six weeks, the MOOC explores the following topics:

- Why use dictionaries (when you could just use a search engine)?
- What's in a dictionary entry: the range of information types which a dictionary may include
- Evidence and method: where the information in dictionaries comes from and how lexicographers process it, including the role of language technology
- What goes into a dictionary, and who decides
- Meanings and definitions: how people create meanings, why some words have multiple meanings, and how dictionaries explain them
- What the future holds: trends in research and publishing (including the automation of lexicographic tasks, and the potential of crowdsourcing), and the likely future role of dictionaries

Each teaching 'Week' typically begins by asking participants to think about and respond to a fundamental question. Thus, in Week 3, we begin by asking students where they think the information found in dictionaries actually comes from. An end-of-week summary offers an opportunity for students to reflect on what they have learned, and to report on how this has affected their earlier preconceptions. Between these two points, the topic of the week is developed through a diverse range of short learning activities, including tasks and quizzes, opinion polls, short articles, and video interviews with well-known experts (including Michael Proffitt, Jane Solomon, Ramesh Krishnamurthy, and Patrick Hanks). At various points, participants work with language data (in the form of concordances, Word Sketches, etc) in order to complete tasks they have been set.

Like all FutureLearn courses, the dictionaries MOOC is highly interactive. At every stage in the process, learners are invited to contribute their thoughts and ideas or ask questions. There is typically a lively exchange of views among participants, and we as 'educators' will frequently jump in to add a comment, answer a question, or just give encouragement. It is a key feature of introductory courses like

this that each learning component is short and easily digested. Video clips generally last only three or four minutes, while the articles that summarize key ideas rarely run to more than 500 words. (For the course creators, who are all used to discussing these subjects in longer journal papers, this represents an interesting challenge.) Finally, for each of the main topics we cover, links are provided to a wide range of online resources, such as longer, more scholarly articles for those who feel motivated to explore an issue in greater depth.

At this stage, we don't know whether the MOOC will run for another iteration, but the experience so far has been a fascinating one for those of us who developed the course and participated as educators. In a follow-up paper (McGillivray et al. forthcoming), we reflect in greater detail on our experience and on our interactions with participants, as well as analysing data on learners' engagement at different stages in the course. One representative comment was: "The mentor and instructor participation on the course forums was outstanding as was the content. Really fascinating course and I'll be sure to recommend it widely. A HUGE thanks!"

Many other comments supported our initial hypothesis that participants would see their preconceived notions about dictionaries called into question or completely overturned. This end-of-course observation is typical:

"I have never imagined there was so much work behind dictionaries! I have learned about different dictionaries, different uses, the parts of a dictionary, the latest technology, their future... Now I see things more clearly and I appreciate dictionaries much more. It has been a fantastic trip".

References

Creese, S., McGillivray, B., Nesi, H., Rundell, M. and Sule, K. 2018. Everything You Always Wanted to Know about Dictionaries (But were Afraid to Ask): A Massive Open Online Course. In Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (eds.), *Proceedings of the XVIII EURALEX International Congress – Lexicography in Global Contexts, 17-21 July 2018, Ljubljana, Slovenia* https://videolectures.net/euralex2018_mcgillivray_rundell_online_course/

McGillivray, B., Creese, S. Nesi, H., Rundell, M. and Süle, K. forthcoming. *Understanding English Dictionaries*: The experience from a massive open online course. *Proceedings of EURALEX 2020*.

SuperMemo

Based in Poznań, Poland, and part of the PWN Group, SuperMemo World is an e-learning publisher with over 25 years of experience in algorithm-based services and content for language learning. It has been at the forefront of long-term human memory research and it specializes in designing vocabulary courses that optimize language acquisition and word retention.

As part of a new-type, wide-scope cooperation project with SuperMemo, K Dictionaries will apply innovative methods to automatically generate new data by merging existing lexicographic resources, producing lexical sets for 19 language cores with translation equivalents in 14-15 languages each, accumulating to 276 language pairs in total!

In the early 1990s, based on its own research, SuperMemo World was the first company ever to implement advanced spaced repetition algorithms in computer programs with the aim of supporting effective learning. The so-called SuperMemo method consists of optimizing intervals between repetitions of the knowledge acquired, thus minimizing the number of repetitions and at the same time achieving the desired level of knowledge retention in each learner's memory. As a result, learners are provided with a tool that helps them memorize any amount of information, for example, words in a foreign language, dates, names, rules, facts or hierarchies, with the retention level close to 100%. Today, SuperMemo is still the world leader in research into human long-term memory.

Since 2017, SuperMemo.com offers an ecosystem for studying and creating courses for learning powered by the SuperMemo method. Apart from free user-generated content, the Web service and apps for iOS and Android provide over 200 high-quality premium courses to study 19 different languages, catering for users at various levels having different objectives and styles of learning. Notably, the collection features the award-winning Olive Green course that includes a full-feature interactive action film produced to teach English and the Extreme series of learner's dictionaries for 4 languages that contain 100,000 entries altogether.

https://supermemo.com

GlobaLex update

globaLex

Status

The Global Alliance for Lexicography (GlobaLex) is undergoing registration as a non-profit organization in Leiden, the Netherlands by the representatives of its five founding continental lexicography associations who currently serve on the Management Committee (MC), as follows:

- African Association for Lexicography (AFRILEX). Dion Nkomo
- Asian Association for Lexicography (ASIALEX). Ilan Kernerman, MC Chair
- Australasian Association for Lexicography (AUSTRALEX). Julia Miller, MC Vice-chair
- Dictionary Society of North America (DSNA). Edward Finegan
- European Association for Lexicography (EURALEX). Lars Trap-Jensen

The MC includes also Simon Krek from the European Lexicographic Infrastructure ELEXIS, which hosts and maintains the GlobaLex website and will add the new Elexifinder search tool for lexicographic publications.

The members have been holding monthly virtual meetings since mid-2018, with reports posted regularly on the GlobaLex website.

Edward Finegan will be replaced in August 2020 by the new DSNA representative, Sarah Ogilvie.

Workshops

Two GlobaLex workshops were canceled in 2020 because of the COVID-19 pandemic:

- ***Globalex Workshop on Linked Lexicography.*** The third workshop organized in conjunction with the LREC conference series was due to be held in Marseille, France in May. Focusing on cross-linking different lexicographic resources as well as other lexical data, it featured monolingual and bilingual/multilingual shared-task tracks supported by ELEXIS, the TIAD (Translation Inference Across Dictionaries) workshops, and K Dictionaries.
  - **workshop**
    https://globalex2020.globalex.link/globalex-workshop-lrec2020-about-globalex-lrec2020/
  - **proceedings**
    https://aclweb.org/anthology/2020.globalex-1.0.pdf

- ***Globalex Workshop on Lexicography and Neologism.*** The second iteration of GWLN was scheduled as part of the Euralex conference in Alexandroupolis, Greece in September. Selected papers will be published as a special issue of the *International Journal of Lexicography* in 2021 and others as part of the conference proceedings.
  - **workshop** https://globalex2020.globalex.link/gw-euralex2020/

In addition, ***Globalex Seminar on Learner's Dictionaries*** that was planned to be held in this year's conference of AFRILEX has been rescheduled to 2021.

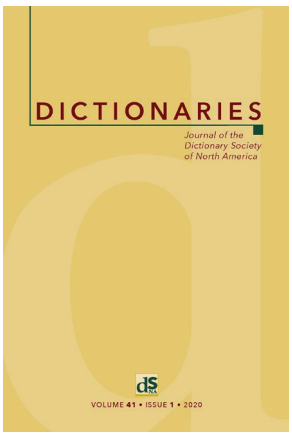
Conferences

The three conferences of the continental lexicography associations planned for 2020 were rescheduled to 2021, when the biennial Australex and DSNA meetings are scheduled to take place too:

- ***DSNA.*** Boulder, Colorado, June 2-5
- ***ASIALEX.*** Yogyakarta, Indonesia, June 12-14
- ***AFRILEX.*** Stellenbosch, South Africa, June/July
- ***AUSTRALEX***. New Zealand, September 1-2
- ***EURALEX.*** Alexandroupolis, Greece, September 7-12

The next **eLex Conference** due in Brno, Czech Republic in 2021 may be postponed to 2022.

Publications



Dictionaries. The Journal of the Dictionary Society of North America devoted a special issue to papers from the first Globalex Workshop on Lexicography and Neologism held as part of the DSNA conference at Bloomington, Indiana in May 2019. Volume 41, Issue 1, 2020 includes eight of the 13 papers from GWLN 2019.

- https://muse.jhu.edu/issue/42292

Lexicon. The journal of the Iwasaki Linguistic Circle in Tokyo is accessible on the Globalex website starting with Issue 25 from 1995.

- https://globalex.link/publications/lexicon/

Lexicography – Journal of ASIALEX. The publication of LJA is taken over from Springer by Equinox Publishing, from Sheffield, UK. The journal will be published online-only from 2021 in a hybrid open access model combining ASIALEX membership. LJA is now indexed also by Scopus.

- https://journal.equinoxpub.com/lexi/index

Tom McArthur and Globalex

In the end of 1997 and much of 1998, I had the good fortune to collaborate with Tom McArthur on editing *Lexicography in Asia*. Tom offered to co-edit when I asked for his paper from the Dictionaries in Asia Conference, which he didn't have but proposed to rewrite as an introduction to the volume. He was kind, bright and visionary – early in developing interest in language radiating into new information science, from his *reference science* to (*linguistic*) *data science* today – and reciprocally and wholeheartedly both local and global.

Earlier in 1997 I wrote that "[a] future Globalex (or Unilex, in the words of Tom McArthur) concerns globalization and co-existence in multilingual societies" (cf. Towards PEOPLEX, KDN 5, 1997). Tom felt uneasy by the possible connotation of the word *people* in the title. In the introduction to *Lexicography in Asia* he wrote:

> … it has now become possible to look forward to a conference devoted to 'world lexicography' … that will seek to cover as wide a sampling as possible from our immense international heritage of reference materials, in all their formats, genres, rationales, writing systems, technologies, languages of origin, and languages of translation. It would be particularly good if the four continental -lexes and the DSNA could consider jointly sponsoring such a 'Globalex' development.

The wish for Globalex is answered and pursued.

Ilan Kernerman