Recent developments in German lexicography

Alexander Geyken

Alexander Geyken works at the Berlin-Brandenburg Academy of Sciences and the Humanities since 1999, where he directs the long-term research project Digital Dictionary of the German Language (http://dwds.de/). He received his Ph.D. in Computational Linguistics at the University of Munich in 1998. His main research interests are computational lexicography, corpus linguistics, and the use of syntactic and semantic resources for the mining of large textual data.

geyken@bbaw.de

> The digital revolution is changing the way readers consume news and search for information. People are moving away from printed reference books and going online where, generally, they expect to get their information for free. (Press release by Chambers Harrap, 15 September 2009.)

This declaration by Chambers Harrap Publishers in 2009 was one of the rare public statements by a publishing house before closing its business. It points to the fact that the technological change is a decisive factor for the crisis of traditional dictionary production that has led to numerous staff reductions or insolvencies of dictionary publishers on an international level. Similarly, the national German dictionary market has been confronted with dramatic changes in the past years. Traditional dictionary publishers shrank dramatically (Duden, Langenscheidt) or even disappeared completely (Wahrig), and the largest academic dictionary, the *Deutsches Wörterbuch* (DWB, German Dictionary, by the Grimm brothers), compiled by the two Academies in Berlin and Göttingen, will cease its work in 2016. Except for Langenscheidt where the decline has a longer history, all this was announced to the public in one and the same year: 2013. The timing was pure coincidence since the momentous decisions were taken much earlier. To begin with, in 2009, the publishing house Langenscheidt with a tradition of more than 150 years of business in bilingual dictionaries, sold the prestigious Duden department to Cornelsen, a large company known for its text books in the field of education. This happened just a few months after Langenscheidt sold the Brockhaus encyclopedia to the Bertelsmann group. These sales were not the end of Langenscheidt's decline. In 2011 Langenscheidt also separated from Polyglott and in 2012 the 'adult education and school' department was taken over by its competitor Klett. Langenscheidt ended up as "Langenscheidt light"[1]. Instead of investing into a technological reorganization the company surprised the public by announcing a recent strategy shift that foresees a stronger focus on print products[2]. Another challenge for Langenscheidt was the advent of collaborative internet platforms. Based on large initial vocabularies donated by third parties and on a very active user community, the two most popular German dictionary internet platforms, Leo[3] and dict.cc[4], managed to compile very large translation databases of currently nine (Leo) and 26 (dict.cc) language pairs with German as the pivot language and became much more popular on the internet than Langenscheidt's rather traditional website. In addition Linguee[5] – a large database of paragraph-aligned translations where the translation quality of words or phrases in the sentence context can be rated by contributors – is becoming increasingly popular among users. Finally, Klett, who has invested very early in technological products, occupied a large market share. Under its brand PONS it has published a diligently curated set of bilingual dictionaries that have been made available free of charge on the internet since 2001 and continually extended to 13 language pairs at present with access to over 10 million words and phrases[6]. In 2009 Klett has also published a monolingual German dictionary that challenged Duden's spelling dictionary[7].

As a reaction to the stronger pressure from the competitors, Duden was also working on a more powerful internet platform. In April 2011 a website was launched where free access was given to the complete edition of Duden's flagship product, the *Großes Wörterbuch der Deutschen Sprache* (GWDS, Great Dictionary of the German Language, 1999[8]). GWDS is the largest dictionary of contemporary German. It was published as a print edition in 10 volumes with a total of 7,200 pages and 200,000 entries in 1999, and one year later in a CD-ROM version. In order to appreciate the full impact of the free internet version on the market strategy of Duden one has to remember that the CD-ROM was initially

[1] http:// buchreport.de/nachrichten/verlage/verlage_nachricht/datum/2012/11/08/langenscheidt-light.htm/

[2] http:// boersenblatt.net/949754/

[3] http://dict.leo.org/

[4] http://dict.cc/

[5] http://linguee.de/

[6] http://pons.com/

[7] http://text-gold.de/praxistipps-fuer-online-redakteure/pons-online-woerterbuch-macht-dem-duden-konkurrenz-ein-praxistest/

[8] http://duden.de/woerterbuch/

sold for an equivalent of 500 Euros. According to experts in the field, the entry of Duden into the market came too late. The sharp decrease in sales of the spelling dictionary, previously the no. 1 selling work of Duden, could not be counterbalanced. Only two years later, in 2013, Duden announced a dramatic reduction of staff from 190 to 30 employees[9]. Of course, plans for a complete revision of the GWDS were unrealistic under these conditions. Duden now concentrates on its one volume works, including the spelling dictionary, the grammar and the idiom dictionary.

Wahrig, the number two in the monolingual German dictionary market never managed to obtain a significant brand visibility on the internet. Being almost hidden among many other resources in Bertelsmann's large knowledge platform[10], it does not come as a surprise that Wahrig's dictionary was buried together with the Brockhaus encyclopedia: it was also in the year 2013 that Bertelsmann announced the discontinuation of their knowledge platform. The entire lexicographic staff was made redundant and since then, work on Wahrig's dictionary came to its end.

This crisis of lexicography in Germany is more than only an economic one. It is a well known fact among publishing houses that revenues of the large flagship dictionaries do not exceed their expenses. However, in the past these expenses could be cross-financed, for example by the revenues of print products derived from a flagship dictionary. This somewhat comfortable scenario stopped with the sharp decrease in sales of printed books and the triumph of the internet. Users do no longer rely on print products or, for that matter, on classical browser interfaces. Access via smartphones or tablets has become more and more common, and users are not willing to pay for these services. German dictionary producers have not been prepared for compiling tailored products for these new devices. The good old times when dictionaries were produced in three consecutive phases, i.e. planning the dictionary, compiling the dictionary and producing the dictionary (Landau, 1984), are definitively over. Nowadays dictionaries are not produced sequentially anymore but the various phases run in parallel or in cycles. Protagonists of dictionary production are no longer restricted to a team of lexicographers alone but prefer to work with an interdisciplinary team consisting of corpus linguists, computational linguists, IT specialists and lexicographers. Concerns that have to be addressed nowadays include the appropriateness of corpus compilation and its dynamic adaptation to new needs rather than the compilation of citation slips. Also, the automatic extraction of lexicographic information from corpora via statistics or machine learning techniques plays a major role in the dictionary production process today. Numerous papers on lexicography bear witness to these new challenges (e.g. Gouws 2011, Rundell 2012).

With the decreasing lexicographic staffs in publishing houses, further development of lexicography relies predominantly on institutional funding, namely the *Union der deutschen Akademien der Wissenschaften* (Union of German Academies) and the *Institut für Deutsche Sprache* (IDS, Institute for German Language). Both have a long tradition of compiling monolingual dictionaries. Currently there are more than 20 different dictionary projects funded by the Academies. However, the majority of these projects were started a long time ago with traditional methods and will run out of funding in the coming ten years. And given the above-mentioned technological changes it is not likely or desirable that new projects will start in the traditional way that is currently still typical of almost all these projects.

By contrast, there are currently two larger projects in Germany that recognize and implement the principles of the new era of e-lexicography. Both can hope for a sustainable funding: *elexiko* and DWDS.

Elexiko[11] started in 2000 as a long-term project of the IDS. The goal is to describe the German language from the end of the 1940's to the present in all its national variants. Practically, the focus in *elexiko* is set on the description since the 1990's corresponding to the text representation in the underlying corpus base, i.e. the DEREKO-corpus, a continually growing corpus of currently more than 25 billion words. A list of 300,000 lemmas has been selected for *elexiko*. Until the end of 2014, approximately 2,000 entries with high frequency in the corpora were manually edited by the lexicographers. Most lemmas consist of semi-automatically generated minimal articles with information about the spelling, the morphology and corpus examples. The hypertextual structure of the lexicon in *elexiko* played a role right from the beginning. Therefore particular emphasis is put on cross-referencing individual articles and providing links to external resources (Meyer 2014). The online presentation of *elexiko* is embedded into the

[9] http://boersenblatt.net/543236/

[10] http://wissen.de/

[11] http://owid.de/wb/elexiko/start.html/

lexical information system OWID[12]. OWID grants access to a set of lexical modules including the lexicon of neologisms, the lexicon of paronyms and its core module *elexiko*.

The DWDS (*Digitales Wörterbuch der Deutschen Sprache*, Digital Dictionary of the German Language) began in 2007 as a long term academic project at the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW). The motivation to launch this project was threefold: firstly, there is no satisfactory account for the history of the German vocabulary since the end of the 19th century. Secondly, the *Grimmsches Wörterbuch* will remain outdated for the letters G-Z even after the completion of the second edition of the DWB that ends in 2016 after the completion of the letters A-F (by the way, 'Frucht' (fruit) was the last word compiled by the brothers Grimm). And thirdly, existing dictionaries at that time did not draw on large corpus data and computational methods right from the outset. Given the comparatively small project size of ten specialists, the goal of the DWDS project cannot be to compile a full historical dictionary. Instead it was decided to compile a large synchronic dictionary, to which diachronic modules could be added if such work will be funded in the future. More precisely, the aim of DWDS is to build an aggregated information system that draws on several complementary lexical resources, word statistics and corpora. The DWDS can make use of several lexical resources that are part of the heritage of the BBAW: the *Wörterbuch der Gegenwartssprache* (WDG), a synchronic dictionary of 4,800 pages with 90,000 keywords, compiled between 1961 and 1977, the *Etymologisches Wörterbuch des Deutschen* (Etymological Dictionary of German)) and the *Grimmsches Wörterbuch*. Moreover, some 60,000 dictionary articles were licensed from the Duden-GWDS for cases where the WDG articles are missing or outdated. The platform integrates an automatic collocation extractor and a good example finder (Didakowski and Geyken 2012, Didakowski et al 2012). Finally, the DWDS draws on large corpora with a size of 4 billion running words that cover the period between 1600 to the present. The results of this project are accessible under http://dwds.de/.

To sum up, the past decade has brought a shift in German lexicography away from private publishing houses to publicly funded institutions and collaborative internet platforms. The next years will show in what way the two institutional key players in Germany, namely the IDS and the Academies, are able to keep pace with the rapidly developing technology, thus being able to bring academic lexicographic knowledge to the public of the 21st century.

Acknowledgement
The author would like to thank Lothar Lemnitzer for his comments on an earlier draft of this manuscript.

References

Didakowski, J. and Geyken, A. 2012. From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Arbeiten zur Linguistik* 2/2012 (OPAL), Mannheim: Institut für Deutsche Sprache, 43–52.

Didakowski, J., Lemnitzer, L. and Geyken, A. 2012. Automatic example sentence extraction for a contemporary German dictionary. In Fjeld, R.V. and Torjusen, J.M. (eds.), *Proceedings of the XVth EURALEX Congress*. Oslo: University of Oslo, 343-349.

Gouws, R.H. 2011. Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In Fuertes-Olivera, P.A. and Bergenholtz, H. (eds.), *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London: Continuum International Publishing Group, 17–29.

Landau. S. 1984. *Dictionaries. The Art and Craft of Lexicography*. New York: Charles Scribner's Sons.

Meyer, P. 2014. Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *elexiko*. In Abel A. and Lemnitzer, L. (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Arbeiten zur Linguistik* 2/2014 (OPAL). Mannheim: Institut für deutsche Sprache.

Rundell, M. 2012. It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical (Hornby Lecture). In Fjeld, R.V. and Torjusen, J.M. (eds.), *Proceedings of the XVth EURALEX Congress*. Oslo: University of Oslo, 47–92.

[12] http://owid.de/