The lexicographical process of *elexiko*

Christine Möhrs
Institut für Deutsche Sprache
E-mail: moehrs@ids-mannheim.de

1 Introduction

The dictionary *elexiko* is a lexicological-lexicographic project which was developed at the INSTITUT FÜR DEUTSCHE SPRACHE (IDS) in Mannheim (cf. Haß 2005a, Klosa et al. 2006, Klosa 2011a). This dictionary was specifically designed for online publication and is one of the dictionary projects included in the dictionary portal OWID (Online Wortschatz Informationssystem Deutsch). The dictionary *elexiko* can be characterized as an online dictionary under construction, a so called "Ausbauwörterbuch" (cf. Schröder 1997: 60; Klosa 2013).

The lexicographers of *elexiko* compile a reference work that explains and documents contemporary German. After publishing a complete list of headwords on the Internet in 2003, the dictionary was filled with sense independent information for each headword which was automatically or semi-automatically generated from the underlying corpus (for example the frequencies of the headwords), the *elexiko*-corpus[1] (cf. Storjohann 2005). The next step was the publication of 250 headwords, which were defined as the demonstration module (Demonstrationswortschatz) that has been fully lexicographically described. Following the latter, we are currently working in the module "Lexikon zum öffentlichen Sprachgebrauch" (dictionary on public discourse). It contains entries selected mainly according to their (high) frequency in the *elexiko*-corpus, such as *demonstrieren*, *Arbeit* or *Staat*.

The technical background enables the project to offer a list of headwords and fully lexicographically described headwords on the Internet along with specific search features. Dictionary consultants can find single headwords but can also look up groups of lexemes with the same semantic, syntactic, or morphological characteristics (the search options will be extended continually). Furthermore, *elexiko* offers auditory examples for selected groups of headwords and illustrations connected to a meaning of a word.

2 Lexicographical workflow

The dictionary *elexiko* can be described as a dictionary under construction ("Ausbauwörterbuch" according to Schröder 1997: 60). In the following the different phases of the lexicographical workflow will be specified. Contrary to printed dictionaries, writing dictionary entries in *elexiko* do not follow an alphabetic order. In *elexiko* dictionary entries are published in modules, i.e., batches of headwords grouped according to specific criteria, for example their frequency. The first module was called the demonstration module (Demonstrationswortschatz[2]), the second was called the module on public discourse (Lexikon zum öffentlichen Sprachgebrauch[3]). The concept behind this project was to plan and to realize a dictionary for publication on the Internet. The computer (regarding to the workplace and all relevant tools) is the main working tool during the process of compilation and publication of the dictionary *elexiko*. According to Klosa (2013: 519) or rather Wiegand (1998: 233ff.) this process is a computer-lexicographical process. As described in Klosa (2013: 519) "working on a dictionary usually (but especially for long-term academic projects) means that while still writing the dictionary, new material may be added to the corpus, corrections for entries already published are gathered, headwords and cross-references will be supplemented, and modifications in the original concept will become necessary." In the project *elexiko* the phases of planning, writing and producing (and also the phases of reconception, extension, revision [cf. Hahn et al. 2008]) of the dictionary or of specific lexicographic information are parallel.

2.1 Preparation

The idea of compiling a lexicological-lexicographical project was developed in the middle of the nineties. First outlines of the project were written at the end of the nineties (cf. Haß 2005: 13f.). A phase of evaluating the outlines followed which included pilot studies – especially by the lexicographers – on the list of headwords, possible search options, the concept of compiling dictionary entries, the software and all computational tools, and the linguistic-lexicographical

[1] http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html

[2] „Der Demonstrationswortschatz besteht aus zwei Teilmengen: einmal aus den signifikanten Kontextpartnern des willkürlich nach gesellschaftlicher Relevanz ausgewählten Zentral-Lexems *Mobilität*, und zum anderen aus systematisch ergänzten Lexemen. Letztere sollten garantieren, dass alle Wortarten, Wortbildungstypen, Alphabetstrecken und dass hoch- und niedrigfrequente Wörter in diesem Wortschatzausschnitt in realistischer Verteilung vorkommen." (Haß 2005b: 15)

[3] „Das ***Lexikon zum öffentlichen Sprachgebrauch*** ist das erste Modul nach dem Demonstrationswortschatz, das als Bearbeitungsteilwortschatz festgelegt wurde. Dieser Wortschatz deckt sowohl Themen aus Politik und Gesellschaft als auch speziellere Sachverhalte ab. Gut die Hälfte der insgesamt rund 2.700 hochfrequenten Wörter (jeweils zwischen 10.000- und 500.000-mal im *elexiko*-Korpus belegt) sind Nomen, die häufig die zentralen politischen und gesellschaftlichen Diskurse, sie sie im *elexiko*-Korpus erscheinen, eingebettet sind. Sie werden hauptsächlich durch Verben und Adjektive ergänzt, die zu einem geringeren Teil selbst diskursgebunden sind (z.B. *reformieren*, *global*, *sozialverträglich*), die aber auch zur Versprachlichung der Diskurse benötigt werden (z.B. *feststellen*, *abstimmen*)." (Klosa 2011b: 17f.)

components. After evaluating all aspects of the concept, the lexicographers began to compile dictionary entries to test the concept.

2.2 Data acquisition

The *elexiko*-corpus forms the foundation of the compilation of all lexicographical information which the user finds in the dictionary. The *elexiko*-corpus is based upon the "Deutsches Referenzkorpus" (DEREKO) from the IDS which generates the primary sources. In addition to the corpus, secondary sources like other paper or electronic dictionaries and tertiary sources like grammar books were collected (cf. Klosa 2013: 520). The acquisition and incorporation of multimedia elements, namely illustrations and audio files, did not start during the phase of data acquisition. These data were supplied and implemented during the phase of data analysis.

2.3 Computerisation

For writing a corpus-based dictionary the lexicographers of *elexiko* can resort to COSMAS II – the corpus analysis tool of the IDS – and also to a data base of co-occurrences (Kookkurrenzdatenbank CCDB, cf. Belica 2011ff.). ORACLE is used as a data base (it is the data base which is also used for a grammatical tool (grammis) at the IDS). Furthermore, the project first used XMetaL as an XML editor and in the meantime Oxygen. In addition to that, computer linguists wrote a software program to connect the different data bases (EDAS = "Electronic Dictionary Administration System"[4] by Roman Schneider) and a program to control all linking between semantic related partners ("Vernetziko"[5] by Peter Meyer).

2.4 Data processing

On the basis of the *elexiko*-corpus a list of candidates for entries was identified. After a manual analysis of these candidates the lexicographers created a list of headwords for *elexiko* which included approximately 300.000 entries. In collaboration with corpus linguists the frequencies of the headwords were identified and the frequency layers were determined. This information is important for the definition of the modules. A further aspect (among others) was the specification of the data base structure on the basis of DTD-standards (cf. Müller-Spitzer 2005: 26ff.).

2.5 Data analysis

In this phase the lexicographers of the dictionary project *elexiko*, corpus linguists and computational linguists worked on the application of automatic data acquisition for example to specify the aspect of syllabication. Automatic methods are used to determine the co-occurrence profile of a headword. This is the most important method in the process phase of data analysis. In contrast to the other phases the phase of data analysis is the longest one. The compilation of dictionary entries and the conception and the maintenance of links were and are part of this phase.

2.6 Preparation for online release

The phase of the preparation for online release paralled with the process phase of data analysis. This phase included the content review and proofreading of the compiled headwords, many tests of the online application and presentation of the data on the Internet, and the evaluation and extension of the search options. Directions for use and a glossary were compiled. During 2009 and 2011, a team of lexicographers, computational linguists and social scientists worked together in a third party funded project, which was called "Benutzeradaptive Zugänge und Vernetzungen in elexiko (BZVelexiko)"[6]. One of this project's main aims was to investigate questions about usage research of dictionaries (cf. Müller-Spitzer et al. 2011, Klosa et al. 2011, Koplenig 2011).

2.7 Afterlife

On completion of the project module "Lexikon zum öffentlichen Sprachgebrauch" the lexicographers of the project *elexiko* plan further modules. One of the following projects will be a module about the phenomenon of paronymy. This project will discuss the phenomenon of paronymy in a lexicological, corpus linguistic and lexicographic way (cf. for example Storjohann 2013). The aim of the new *elexiko* module(s) will be to present a comparatively small section of a specified vocabulary and to edit the data for the interests of special user groups. Besides publishing new modules, existing lexicographical entries in *elexiko* will be maintained and updated. Long term archiving of the data also needs to be taken care of.

[4] Cf. Müller-Spitzer / Schneider (2009).

[5] "vernetziko has primarily been developed as a software tool for the automated insertion, correction and checking of cross-references in an extensible set of XML-based electronic dictionaries." (Meyer 2011: 192)

[6] http://www1.ids-mannheim.de/lexik/bzvelexiko.html

3 Time span of the different phases

Phase	Duration																	
	Ende 1990	2000	‘01	‘02	‘03	‘04	‘05	‘06	‘07	‘08	‘09	‘10	‘11	‘12	‘13	‘14	‘15	…
Preparation																		
Data acquisition																		
Computerisation																		
Data processing																		
Data analysis																		
Preparation for online release																		
Afterlife																		

Table 1 Process phases of the dictionary project *elexiko* and their time span

4 References

Belica, Cyril: Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. © 2001ff., Institut für Deutsche Sprache, Mannheim.

COSMAS I/II (Corpus Search, Management and Analysis System), http://www.ids-mannheim.de/cosmas2/, © 1991-2010 Institut für Deutsche Sprache, Mannheim.

DEREKO – Deutsches Referenzkorpus des IDS. Internet: http://www1.ids-mannheim.de/kl/projekte/korpora/.

EDAS – Electronic Dictionary Administration System / Lexikographisches Redaktions- und Recherchesystem für digitale Wörterbücher. Internet: http://www1.ids-mannheim.de/gra/projekte/grammis2.html.

grammis – das grammatische Informationssystem des Instituts für Deutsche Sprache. Internet: http://hypermedia.ids-mannheim.de/.

Hahn, Marion / Klosa, Annette / Müller-Spitzer, Carolin / Schnörch, Ulrich / Storjohann, Petra (2008): elexiko – das elektronische, lexikografisch-lexikologische korpusbasierte Wortschatzinformationssystem. Zur Neukonzeption, Erweiterung und Revision einzelner Angabebereiche (934 KB). In: Klosa, Annette (ed.): Lexikografische Portale im Internet. Mannheim: Institut für Deutsche Sprache, 2008, p. 57-86. (= OPAL Sonderheft 1/2008; OPAL - Online publizierte Arbeiten zur Linguistik 1/2008).

Haß, Ulrike (ed.) (2005a): Grundfragen der elektronischen Lexikographie. elexiko - das Online-Informationssystem zum deutschen Wortschatz. Berlin / New York: de Gruyter. (Schriften des Instituts für Deutsche Sprache, Bd. 12).

Haß, Ulrike (2005b): elexiko – das Projekt. In: Haß (ed.) (2005a), p. 1-17.

Klosa, Annette (2013): The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H. / Heid, Ulrich / Schweickard, Wolfgang / Wiegand, Herbert Ernst (ed.): Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume. Recent developments with focus on electronic and computational lexicography. Berlin / Boston: de Gruyter, p. 517-524. (= Handbücher zur Sprach- und Kommunikationswissenschaft, Bd. 5.4).

Klosa, Annette (ed.) (2011a): elexiko. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Gunter Narr. (= Studien zur Deutschen Sprache. Forschungen des Instituts für Deutsche Sprache, B. 55).

Klosa, Annette (2011b): Einleitung. In: Klosa (ed.) (2011a), p. 9-26.

Klosa, Annette / Schnörch, Ulrich / Storjohann, Petra (2006): ELEXIKO - A Lexical and Lexicological, Corpus-based Hypertext Information System at the Institut für Deutsche Sprache, Mannheim. Alessandria. Edizioni dell'Orso. In: Corino, Elisa/Marello, Carla/Onesti, Cristina (ed.): Proceedings of the Twelfth EURALEX International Congress, Torino, Italia, 6 - 9 September 2006. Alessandria: Edizioni dell'Orso, p. 425-429.

Klosa, Annette / Koplenig, Alexander / Töpel, Antje (2011): Benutzerwünsche und Meinungen zu einer optimierten Wörterbuchpräsentation – Ergebnisse einer Onlinebefragung zu elexiko. *35 p. - Mannheim: Institut für Deutsche Sprache, 2011. (OPAL - Online publizierte Arbeiten zur Linguistik 3/2011).*

Koplenig, Alexander (2011): Understanding How Users Evaluate Innovative Features of Online Dictionaries – An Experimental Approach. In: Kosem, Iztok/Kosem, Karmen (Hgg.): Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10 - 12 November 2011 (http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-18.pdf), p. 147-150.

Meyer, Peter (2011): vernetziko: A Cross-Reference Management Tool for the Lexicographer's Workbench. In: Kosem, Iztok / Kosem, Karmen (ed.): Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10 – 12 November 2011. Ljubljana: Trojina, Institute for Applied Slovene Studies, p. 191-198.

Müller-Spitzer, Carolin (2005): Die Modellierung lexikografischer Daten und ihre Rolle im lexikografischen Prozess. Berlin/New York. de Gruyter. In: Haß (ed.) (2005a), p. 21-54.

Müller-Spitzer, Carolin / Koplenig, Alexander/Töpel, Antje (2011): What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. In: Kosem, Iztok / Kosem, Karmen (Hgg.): Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10 - 12 November 2011 (http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-27.pdf), p. 203-208.

Müller-Spitzer, Carolin / Schneider, Roman (2009): Ein XML-basiertes Datenbanksystem für digitale Wörterbücher - Ein

Werkstattbericht aus dem Institut für Deutsche Sprache. In: Information Technology & Multimedia in English Language Teaching 51/4, p. 197-205.

Schröder, Martin (1997): Brauchen wir ein neues Wörterbuchkartell? Zu den Perspektiven einer computerunterstützten Dialektlexikographie und eines Projektes „Deutsches Dialektwörterbuch". In: Zeitschrift für Dialektologie und Linguistik 64/1, p. 57-65.

Storjohann, Petra (2005): Das elexiko-Korpus: Aufbau und Zusammensetzung. In: Haß (ed.) (2005a), p. 55-70.

Storjohann, Petra (2013): Korpuslinguistische und lexikografische Ansätze zur Beschreibung deutscher Paronyme. In: Hermann Scheuringer / Doris Sava (ed.): Im Dienste des Wortes. Lexikologische und lexikografische Streifzüge. Festschrift für Ioan Lazarescu. Passau: Stutz-Verlag, p. 401-418. (= Forschung zur deutschen Sprache in Mittel-, Ost- und Südosteuropa, Bd.3).

Wiegand, Herbert Ernst (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin / New York: de Gruyter.

(Internet resources: last retrieved on 10.06.2014)