Annette Klosa

Der lexikographische Prozess im Projekt *elexiko*

1. Einführung

elexiko,[1] ein Onlinewörterbuch zur deutschen Gegenwartssprache, wurde von Anfang an für die Publikation im Internet geplant und realisiert. Es ist also im wahrsten Sinne des Wortes ein Internetwörterbuch, für das außerdem charakteristisch ist, dass es noch nicht vollständig vorliegt, dennoch aber schon publiziert wird. Man kann bei *elexiko* daher von einem „Ausbauwörterbuch" (Schröder 1997, S. 60) oder einem „dynamischen Wörterbuch" (Lemberg 2001, S. 81) sprechen, nach dem Kriterium der Vollständigkeit würden Storrer/Freese (1996) sowie Storrer (1998 und 2001) es als „Wörterbuch im Aufbau" einordnen. Dieses Merkmal hat einen großen Einfluss auf den lexikographischen Prozess von *elexiko*, indem bei solchen Wörterbüchern wie *elexiko* dieser im Grunde nicht abgeschlossen ist bzw. zu Ende geführt werden kann, sondern so lange andauert, wie das Wörterbuch publiziert wird.

Da *elexiko* sowohl durchgehend mit Unterstützung von Computern erarbeitet wird als auch auf/in Computern publiziert wird, ist der lexikographische Prozess in diesem Fall genauer als computerlexikographischer Prozess (nach Wiegand 1998, S. 233ff.) zu bestimmen. In Fortführung von Wiegands Vorschlägen werden für computerlexikographische Prozesse nach Klosa (2013) die Phasen der Vorbereitung und Planung, der Datenbeschaffung, der Computerisierung, der Datenaufbereitung, der Datenauswertung und der Vorbereitung für den Onlinerelease unterschieden. Charakteristisch für Wörterbücher wie *elexiko* ist dabei, dass alle Phasen nicht notwendigerweise hintereinander stattfinden, sondern dass sie auch parallel zueinander ablaufen können. Eine klare Grenzziehung zwischen den einzelnen Phasen ist daher nicht immer möglich.

Ein zweites Charakteristikum für solche Wörterbücher im Aufbau ist, dass in der Phase der Datenauswertung nicht von A bis Z vorgegangen werden muss, sondern dass die Stichwörter des Wörterbuches nach anderen Kriterien als ihrer alphabetischen Abfolge für die Bearbeitung und sukzessive Publikation ausgewählt werden können. Die Erarbeitung des Wörterbuches erfolgt dann in sogenannten Modulen (Haß 2005, S. 13ff.; Klein 2004, S. 300f.). Verschiedene Module eines Onlinewörterbuches können sich in verschiedenen Phasen des lexikographischen Prozesses befinden, in dem für ein Modul beispielsweise noch Daten beschafft werden müssen, während die Stichwörter eines anderen Moduls schon online erscheinen. Und sogar innerhalb eines Moduls können sich verschiedene Stichwörter in verschiedenen Phasen der lexikographischen Prozesse befinden, z.B. im folgenden Fall, der auch für *elexiko* zutrifft:

> Modules not only as a whole but also single entries within a module can be situated in different phases if the dictionary allows online release of single entries. Thus, a number of entries in the same module may already be online while still others are in the phase of data analysis. (Klosa 2013, S. 522).

[1] Vgl. www.elexiko.de (zuletzt eingesehen: 5.2.2015). Zur Konzeption von *elexiko* vgl. Haß (Hg.) (2005). Zur praktischen Umsetzung dieser Konzeption vgl. Klosa (Hg.) (2011).

2. Die Phasen des computerlexikographischen Prozesses von *elexiko*

2.1 Vorbereitungsphase

In der Vorbereitungsphase für *elexiko*, dessen Projektlaufzeit im Jahr 1997 begonnen hat, wurde viel Zeit in die Entwicklung der inhaltlichen Konzeption und in etliche Pilotstudien investiert. Die Ergebnisse dieser Bemühungen sind im Band „Grundfragen der elektronischen Lexikografie. *elexiko* – Das online-Informationssystem zum deutschen Wortschatz" (Haß (Hg.) 2005) publiziert, in dem sowohl die Makrostruktur des Wörterbuches wie auch die Mikrostruktur ausführlich dargelegt werden. Fragen der Datenmodellierung, der Korpuszusammenstellung sowie der Mediostruktur werden ebenso erläutert wie erste Überlegungen für die Onlineoberfläche u.Ä. Erst später wurde allerdings mit der Umsetzung der inhaltlichen Konzeption in ein lexikographisches Redaktionshandbuch begonnen (vgl. Klosa 2011a, S. 15).

Das Konzept für die Wörterbuchinhalte wurde erprobt, indem zahlreiche Musterartikel (im sogenannten *elexiko*-Demonstrationswortschatz) verfasst wurden. Dieser Demonstrationswortschatz enthält knapp 250 Stichwörter, bei deren Auswahl zwei Kriterien zur Anwendung kamen: Ausgewählt wurden erstens die mithilfe des Verfahrens der Kookkurrenzanalyse[2] ermittelten statistisch signifikanten Kookkurrenzpartner, also Kollokatoren des Nomens *Mobilität*[3] Das zentrale Wort *Mobilität* wurde gewählt, weil die Diskussion hierüber, vor allem im beruflichen Kontext, im *elexiko*-Korpus einen breiten Raum einnimmt und auch allgemein wichtig und aktuell ist, ohne zu sehr religiös oder weltanschaulich gefärbt zu sein. Zweitens wurde diese Menge systematisch ergänzt nach den aus anderen bzw. älteren Wörterbüchern bekannten Proportionen von Wortarten, Alphabetstrecken und diversen Wortbildungstypen. Durch diese Auswahl bzw. ihre systematische Ergänzung war es möglich, Schätzungen für die Bearbeitungsdauer des später begonnenen Moduls „Lexikon zum öffentlichen Sprachgebrauch"[4] anzufertigen. Die Erarbeitung des Demonstrationswortschatzes gehört inhaltlich zwar noch in die Vorbereitungsphase, weil im Grunde erst nach Abschluss dieser Arbeiten am systematischen weiteren Ausbau von *elexiko* gearbeitet werden konnte bzw. wurde, greift natürlich aber auch schon auf die später beschriebenen Phasen im lexikographischen Prozess aus.

Die Arbeit am Projekt *elexiko* wurde zunächst ohne vollständige Durchdringung computerlexikographischer Prozesse, dafür mit viel Pioniergeist und opportunistischer Ausnutzung der am IDS gegebenen Möglichkeiten durch die beteiligten lexikographischen Mitarbeiter(innen) aufgenommen. Dies lässt sich u. a. daran ablesen, wie die Projektschritte von 1997 bis 2004 in Haß (2005, S. 13ff.) beschrieben werden:

> Erste Skizzen zu einem lexikalischen Informationssystem entstanden Ende 1997. Der Aufbau einer Arbeitsgruppe, die Exploration der computertechnischen Möglichkeiten und die Konkretisierung eines den personellen wie technischen Möglichkeiten angepassten Konzepts mündeten in die linguistische Entwicklung und texttechnologische Realisierung einer Wortartikelstruktur in Form einer ungewöhnlich komplexen DTD-Struktur. [...]
> In der darauffolgenden Phase wurde diese DTD getestet, evaluiert und insgesamt dreimal überarbeitet. [...] In diese Konzeptionsphase fallen sämtliche Bemühungen um eine mittelfristig stabile Lösung der gesamten Softwarearchitektur eines solches [sic!] Vorhabens, die auch heute nicht und vermutlich nie ganz wunschgemäß gegeben ist. Anfang 2003 war die Konzeptionsphase sowohl linguistisch-lexikografisch als auch computertechnisch so weit abgeschlossen, dass mit der Realisierung erster Wortschatzbereiche (Module [...]), d.h. mit dem Verfassen von Wortartikeln und ihrer Speicherung in der Datenbank begonnen werden konnte.

[2] Zur Methode der Kookkurrenzanalyse vgl. www.ids-mannheim.de/kl/projekte/methoden.html (zuletzt eingesehen: 5.2.2015).

[3] Vgl. das Stichwort *Mobilität* online unter www.owid.de/artikel/62377 (zuletzt eingesehen: 5.2.2015).

[4] Vgl. www.owid.de/wb/elexiko/projekt/modSprachgeb.html (zuletzt eingesehen: 5.2.2015).

> In die Realisierungsphase ist auch die Erarbeitung der knapp 300.000 Wörter umfassenden Stichwortliste aus dem IDS-Korpus einbezogen, [...].
> Seit Januar 2004 sind diese 300.000 gebräuchlichen Einwortausdrücke mit orthografischen und morphologischen Angaben öffentlich zugänglich. [...]
> *elexiko* wird auch in Zukunft modular weiterentwickelt und ausgebaut, d. h. nicht entlang dem Alphabet, sondern vertikal nach Teilwortschätzen bzw. horizontal nach bestimmten Angabearten. [...] Das erste Beispiel für einen vertikal ausgearbeiteten Teilwortschatz ist der 240 Lexeme umfassende Demonstrationswortschatz. [...]
> Seit Juni 2004 stehen die 240 Artikel des sog. Demonstrationswortschatzes online zur Verfügung; [...]. (Haß 2005, S. 13ff.)

Zu Ende der 90er Jahre des vorigen Jahrhunderts wurde mit dem geplanten Wörterbuch im Bereich der wissenschaftlichen Lexikographie Neuland betreten. Es gab keine Vorbilder für vollständig neu zu erarbeitende, kontinuierlich zu erweiternde Onlinewörterbücher, sodass es im Grunde keine Erfahrungswerte gab, auf denen das Projekt hätte aufbauen können. Allerdings lagen damals durchaus schon Veröffentlichungen vor, die sich mit dem lexikographischen Prozess an sich befassen (z.B. Dubois 1990; Landau 1984), und es gab Ansätze zur Analyse der Einsatzmöglichkeiten des Computers in der Lexikographie (z.B. Knowles 1990). Trotzdem wurde im Projekt *elexiko* in der Vorbereitungsphase auch einiges versäumt, vor allem die Erstellung eines Organisationsplans mit Angaben zu Finanzierung, Personal (auch technischem Personal!), Zeitplan und Workflow, der als zweiter wichtiger Bestandteil des Wörterbuchplans das inhaltliche Konzept für *elexiko* hätte ergänzen müssen. Ebenfalls versäumt wurde eine gründliche Marktanalyse der Ende der 1990er Jahre existierenden lexikographischen Redaktionssysteme. Beide Versäumnisse haben in Folge die Erarbeitung von *elexiko* zeitlich und finanziell belastet, indem es zu eigentlich unnötigen Investitionen von Zeit und Geld kam, um eingangs schlecht oder gar nicht geplante Umstände zu korrigieren (z.B. kein Rückgriff auf ein käufliches Redaktionssystem, sodass später erhebliche Personalmittel und Zeit aufgewandt werden mussten, um eine eigene Redaktionsumgebung zu schaffen (vgl. Abel/Klosa 2012, S. 417f.)). Schließlich wurde die Planung des Onlinedesigns des Wörterbuches ohne Unterstützung von Webdesignspezialisten durchgeführt und verschiedene Benutzungsmöglichkeiten von *elexiko* (z.B. erweiterte Suche, Angebot der reinen Stichwortliste in links- und rechtsalphabetischer Sortierung) wurden ohne die Durchführung von Benutzungsstudien[5] entworfen.

2.2 Datenbeschaffung

Die Phase der Datenbeschaffung war im Projekt *elexiko* vor allem der Zusammenstellung des *elexiko*-Korpus als virtuelles Korpus aus DEREKO, dem Deutschen Referenzkorpus des IDS Mannheim,[6] gewidmet. Das Korpus besteht aus Texten unter anderem aus folgenden Zeitungen und Zeitschriften: Berliner Morgenpost, Berliner Zeitung, Der Spiegel, Die Presse, die tageszeitung, Die Welt, Die Zeit, Frankfurter Allgemeine Zeitung, Frankfurter Rundschau, Hamburger Morgenpost, Kleine Zeitung, Mannheimer Morgen, Neue Kronen-Zeitung, Oberösterreichische Nachrichten, Salzburger Nachrichten, St. Galler Tagblatt, Tiroler Tageszeitung, Vorarlberger Nachrichten und Züricher Tagesanzeiger. Es verbindet also bundesdeutsche, österreichische und Schweizer Quellen in einer den jeweiligen Sprecherzahlen in etwa entsprechenden Menge. Da es sich bei dem *elexiko*-Korpus aber um ein dynamisches Korpus (bzw. ein sogenanntes Monitorkorpus) handelt, welches regelmäßig erweitert und ausgebaut

[5] Einige Elemente der Onlineoberfläche von *elexiko* wurden zwischenzeitlich in zwei Benutzungsstudien untersucht, vgl. hierzu Klosa/Koplenig/Töpel (2011).

[6] Vgl. www.ids-mannheim.de/kl/projekte/korpora.html (zuletzt eingesehen: 5.2.2015).

wurde, ist das hier vorgestellte Wörterbuchprojekt immer wieder (zuletzt 2013) in die Phase der Datenbeschaffung zurückgekehrt, um neue Quellen zu ergänzen.

Parallel wurden sekundäre Quellen (andere gedruckte wie elektronische Wörterbücher) und tertiäre Quellen (z.B. Grammatiken)[7] gesichtet und beschafft, die allerdings im Gesamtkontext des Projektes eine wesentlich unbedeutendere Rolle als das Korpus selbst einnehmen, weil die lexikographische Beschreibung in *elexiko* generell einem korpusgestützten Ansatz verpflichtet ist (vgl. Klosa 2011a, S. 16f.). Auch diese Teilaufgabe in der Phase der Datenbeschaffung war nicht zu einem bestimmten Zeitpunkt abgeschlossen, sondern der inzwischen über 10-jährigen Projektlaufzeit ist geschuldet, dass immer wieder neu erschienene sekundäre und tertiäre Quellen ergänzt wurden. An diesem kleinen Beispiel zeigt sich (und dies gilt vermutlich für generell jeden lexikographischen Prozess), dass die theoretisch angenommenen Phasen nicht wirklich eine nach der anderen ablaufen, sondern sich immer wieder überlappen können.

In einigen Wortartikeln des *elexiko*-Demonstrationswortschatzes waren schon Illustrationen enthalten; das entsprechende Bildmaterial wurde in der Phase der Datenbeschaffung besorgt. Seitdem hat sich allerdings gerade in diesem Bereich viel getan: Zwischenzeitlich wurden weitere Phasen der Datenbeschaffung (zuletzt 2010 bis 2013) eingeschoben, und zwar beispielsweise zur systematischen Gewinnung von Bildmaterial und zur Auswahl von Hördateien, welche das nach Abschluss des Demonstrationswortschatzes begonnene und vorangetriebene Modul „Lexikon zum öffentlichen Sprachgebrauch“ bereichern sollten. Zur Sammlung des Bildmaterials ist anzumerken:

> Daneben wurde überprüft, welche Möglichkeiten es zur kostenlosen Gewinnung von Illustrationen gibt. Hierbei sind verschiedene Online-Datenbanken mit Zeichnungen, Fotografien, Comics und Videos ausgewertet worden (z.B. pixelio.de, Wikimedia Commons). Das Ergebnis ermutigt, was die Menge kostenlos zur Verfügung stehender Illustrationen (vor allem Fotos) betrifft, wirft aber auch einige neue Probleme auf: So sind die Fotos hinsichtlich ihrer Art, Qualität, Auflösung und Größe sehr unterschiedlich. Ebenso unterschiedlich sind die Vorgaben dazu, wie die Quellenangabe erfolgen muss. (Klosa 2011b, S. 164)

Die Hörbelege, die in *elexiko* die natürlichsprachige Aussprache eines Stichwortes im Kontext demonstrieren, wurden zur gleichen Zeit aus der „Datenbank für gesprochenes Deutsch“ im „Archiv für gesprochenes Deutsch (2009)“[8] des IDS ermittelt. Mit dem Jahr 2013 wurde die nachträglich eingeschobene Phase der Beschaffung von Bild- und Tondaten abgeschlossen.

2.3 Computerisierung

Zwei besonders zeitaufwändige Arbeitsschritte in der Phase der Computerisierung, nämlich die digitale Aufbereitung von Korpustexten (Annotierung, Lemmatisierung) sowie die Programmierung eines Korpusrecherchetools, konnten im Projekt *elexiko* insofern entfallen, als das Projekt ein virtuelles Korpus aus dem „Deutschen Referenzkorpus (DEREKO)“ des IDS zusammenstellen konnte und für die Auswertung des Korpus auf COSMAS II,[9] das Korpusrecherche- und -analysewerkzeug des IDS zurückgreifen kann. Im Projekt werden außerdem die in der Kookkurrenzdatenbank CCDB[10] des IDS dokumentierten Ergebnisse berücksichtigt.[11]

---

[7] Zu primären, sekundären und tertiären Quellen in der Lexikographie vgl. Wiegand (1998, S. 140).

[8] Vgl. http://agd.ids-mannheim.de/datenbanken.shtml (zuletzt eingesehen: 5.2.2015).

[9] Vgl. https://cosmas2.ids-mannheim.de/cosmas2-web/ (zuletzt eingesehen: 5.2.2015).

[10] Vgl. http://corpora.ids-mannheim.de/ccdb/ (zuletzt eingesehen: 5.2.2015) und Belica (2011).

[11] Vgl. hierzu genauer Storjohann (2011).

Zur redaktionellen Arbeitsumgebung im Projekt gehören daneben die folgenden Komponenten (vgl. hierzu Abel/Klosa 2012, S. 416f.):

- ein XML-Editor (zunächst XMetal, jetzt Oxygen), in dem die lexikographischen Daten ediert werden,
- ein Verweismanager (*vernetziko*),[12] der die Erstellung und Pflege konsistenter Vernetzungen im Wörterbuch ermöglicht und außerdem die lexikographische Arbeit in anderer Hinsicht unterstützt (z.B. durch die Möglichkeit, Textbausteine anzulegen oder durch Abfragemöglichkeiten nach dem Bearbeitungsstand einzelner Einträge),
- eine Schnittstelle (EDAS – Electronic Dictionary Administration System)[13] zur Datenbank (Oracle), in der die Daten gespeichert werden, in der sie durchsuchbar sind und aus der sie für die Online-Präsentation exportiert werden,
- über EDAS außerdem Zuweisung eines bestimmten Bearbeitungsstatus zu Wortartikeln und Durchführung komplexer Suchanfragen im Wortartikelbestand,
- interne Online-Ansicht der Wörterbuchoberfläche innerhalb des Portals OWID zur Überprüfung vor der Freischaltung,
- ein elektronisches Redaktionshandbuch und eine DTD-Dokumentation.

Im Projekt *elexiko* wird also kein lexikographisches Redaktionssystem im eigentlichen Sinn eingesetzt, doch funktioniert die Arbeitsumgebung nach einiger Zeit der Entwicklung und zwischenzeitlichen Optimierungen (hier ist insbesondere das Vernetzungstool zu nennen) trotzdem zufriedenstellend.

2.4 Datenaufbereitung

Ein großer Arbeitsbereich in der Phase der Datenaufbereitung war die Erstellung der Stichwortkandidatenliste aus dem *elexiko*-Korpus. Dabei wurde folgendermaßen vorgegangen:[14]

1. Mithilfe eines automatischen Lemmatisierers wurden die Flexions- bzw. Paradigmenformen von Wörtern aus den Texten der IDS-Korpora geschriebener Gegenwartssprache Wortformen zugewiesen, die gemäß der Annahmen des automatischen Lemmatisierers als 'wörterbuchübliche Grund- oder Nennformen' gedeutet werden können, z.B. dem Infinitiv bei Verben, dem Nominativ Singular bei Nomen, dem Positiv bei Adjektiven usw.; das sind die sog. Stichwortkandidaten. Zusätzlich wurden auch lexikografisch relevante Flexionsformen eigens erfasst, beispielsweise Partizipien, unregelmäßige Pluralbildungen, die erste Person Präteritum der starken Verben u.Ä. [...].
2. Um von vornherein die Zahl insbesondere formal fehlerhaft angesetzter Stichwortkandidaten einzugrenzen, erfolgte ein sichernder Korrekturabgleich der Stichwortkandidatenliste mit Listen anderer Wörterbücher.

Die Stichwortkandidatenliste ist das Ergebnis eines Zusammenspiels der beiden Methoden 'Korpusextraktion mit automatischer (Grund-) Formzusammenführung' und 'korrektiver formaler Abgleich mit vorhandenen Wörterbuchlisten' im Anschluss daran. Das Resultat wäre allerdings noch um ein Vielfaches umfangreicher, hätte man auf dieser Stufe nicht eine dritte, quantitativ einschränkende, also auswählende Größe berücksichtigt: die Frequenz.

Die weitgehend automatisch generierte Liste enthält deshalb 'nur' knapp 320.000 Stichwortkandidaten, die mit einer Mindestfrequenz von acht im Korpus vorkommen und in anderen Wörterbüchern enthalten sind. Die Frequenz von acht bedeutet, dass das betreffende Wort in seiner Grund- bzw. Nennform und/oder in den dazugehörigen Flexionsparadigmenformen mindestens achtmal im Korpus belegt ist. Die Zahl acht ist dabei zufällig jene Größe, mit deren Hilfe ein Schnitt bei der Extraktion aus dem Korpus vorgenommen werden konnte, durch den sich der Umfang der Stichwortkandidatenliste in idealer Weise beschränken ließ: [...]. (vgl. Schnörch 2005, S. 75f.)

[12] Vgl. Meyer/Müller-Spitzer (2010) und Meyer (2011).

[13] Entwickelt im früheren Projekt Texttechnologie, jetzt Grammis II: Grammatische Datenbanken und Informationssysteme, vgl. www.ids-mannheim.de/gra/projekte/grammis2.html (zuletzt eingesehen: 5.2.2015) und Müller-Spitzer/Schneider (2009).

[14] Verantwortlich für diesen Arbeitsschritt war der Programmbereich Korpuslinguistik (damals noch „AG Korpustechnologie") des IDS, siehe www.ids-mannheim.de/kl/projekte.html (zuletzt eingesehen: 5.2.2015).

Ein wichtiger weiterer Schritt war die (automatische) Ermittlung von Frequenzen der Stichwörter, und zwar die Summe aller Vorkommen der jeweiligen Grundform und die Summe aller dieser Grundform zugeordneten Flexionsformen. Zwar können diese Frequenzangaben „bei einem dynamischen Monitorkorpus wie dem *elexiko*-Korpus immer nur eine Momentaufnahme sein" (Klosa 2011b, S. 158), doch wurden sie genutzt, um sogenannte Frequenzschichten zu definieren, die wiederum die Grundlage für die Definition von zu bearbeitenden Stichwortmengen (Modulen) in *elexiko* bilden (z.B. sind im „Lexikon zum öffentlichen Sprachgebrauch" Stichwörter mit einer Frequenz zwischen 10.000 und 500.000 im Korpus enthalten).

Schließlich gehört in diese Phase des lexikographischen Prozesses auch die Umsetzung der Modellierung der lexikographischen Daten in eine entsprechende Datenbankstruktur (vgl. Wiegand 1998, S. 238). Testweise wurde hierbei zunächst mit der Datenbank Tamino,[15] später dann mit Oracle[16] gearbeitet. Bei der Entwicklung des XML-basierten Datenbanksystems konnte das Projekt von den für grammis (das grammatische Informationssystem des IDS) gewonnenen Erfahrungen des Projektes „Texttechnologie und Datenbanken" profitieren (vgl. Müller-Spitzer/Schneider 2009).

2.5 Datenauswertung

In der Phase der Datenauswertung können sich automatisch-korpuslinguistische sowie redaktionell-lexikographische Arbeiten verbinden. Im Projekt *elexiko* wurden z.B. Belege aus dem *elexiko*-Korpus ermittelt, die bei all jenen Stichwörtern in *elexiko* angezeigt werden, die noch nicht redaktionell bearbeitet sind.[17] Hierbei, wie auch bei weiteren automatisch erzeugten lexikographischen Angaben, wird aber auf völlig automatische Integration der analysierten Daten in die Wortartikel verzichtet, sondern wo immer möglich wird ein redaktioneller Prüfgang (wenigstens stichprobenartig) durchgeführt und entsprechende Korrekturen werden vorgenommen.[18]

In die Phase der Datenauswertung fallen insbesondere redaktionelle Arbeitsgänge. Hierzu zählte im Projekt *elexiko* die redaktionelle Prüfung der automatisch erstellten Stichwortkandidatenliste (s.o.).[19] Hauptaufgabe in dieser Phase eines jeden lexikographischen Prozesses ist das Verfassen neuer Wortartikel, und dieser Arbeitsschritt umfasst auch im Projekt *elexiko* den längsten Zeitraum in der Projektlaufzeit (nämlich kontinuierlich seit 2003).[20] Dabei werden die Korpusdaten (in Form von Kookkurrenzlisten als Ergebnis der Kookkurrenzanalyse,[21] KWIC-Zeilen, Frequenzlisten usw.) analysiert und interpretiert, die Lesarten werden disambiguiert, Definitionen geschrieben, Korpusbelege zur Integration in den Wortartikel ausgesucht, Sekundärquellen geprüft usw. Diese Arbeiten laufen dabei im Grundsatz nicht anders ab, als sie es für ein Printwörterbuch tun würden:

---

[15] Vgl. www.softwareag.com/corporate/products/az/webmethods/default.asp (zuletzt eingesehen: 5.2.2015).

[16] Vgl. www.oracle.com/de/products/database/overview/index.html (zuletzt eingesehen: 5.2.2015).

[17] Vgl hierzu genauer Klosa (2011a, S. 18ff.).

[18] Zur Verbindung redaktionell erarbeiteter wie automatisch ermittelter Angaben in Wörterbüchern vgl. genauer Klosa (2010).

[19] Vgl. genauer Schnörch (2005, S. 77f.).

[20] Einen genauen Einblick in die redaktionelle Arbeit im Projekt *elexiko* geben die in Klosa (Hg.) (2011) enthaltenen Beiträge.

[21] Vgl. www.ids-mannheim.de/kl/projekte/methoden/ka.html (zuletzt eingesehen: 5.2.2015).

> Alle genuin lexikographischen, philologischen und linguistischen Kenntnisse und Fertigkeiten, welche nötig sind, um ohne Computereinsatz in der Phase der Materialauswertung Handlungen vom Typ EINEN WÖRTERBUCHARTIKEL SCHREIBEN erfolgreich ausführen zu können, werden auch benötigt, um in der Phase der Datenauswertung Handlungen vollziehen zu können, die zum Typ EINEN WÖRTERBUCHARTIKEL COMPUTERUNTERSTÜTZT SCHREIBEN gehören. (Wiegand 1998, S. 239)

Der Unterschied bei der Erarbeitung eines Internetwörterbuches zeigt sich eher dann, wenn für ein Wörterbuch wie *elexiko* nicht nur (elektronische) Vernetzungen zwischen einzelnen Wortartikeln oder den Wortartikeln mit den lexikographischem Umtexten angelegt werden, sondern auch zu externen Quellen (z.B. Enzyklopädien) verlinkt wird.

In die Phase der Datenauswertung fällt auch ein Arbeitsgang, in dem die während der Datenbeschaffungsphase zusammengetragenen multimedialen Elemente (Illustrationen, Hördateien) gesichtet bzw. angehört und getestet werden, damit die für die Wortartikel geeigneten ausgesucht werden können. Bild- und Tondaten werden außerdem analysiert, um die für die Wortartikel gewünschten Bild- und Tonausschnitte festlegen zu können.

2.6 Vorbereitung für Onlinerelease

Vor der Freischaltung von Wörterbüchern im Internet sind eine Menge technischer wie inhaltlicher Arbeitsschritte zu leisten, damit das Wörterbuch in möglichst überzeugender Qualität genutzt werden kann. So ist natürlich nötig, dass die Präsentation der Wortartikel im Internet technisch durch entsprechende Programmierung ermöglicht wird. Im Projekt *elexiko* wird dies mithilfe von Stylesheets, welche die Umwandlung der XML-Inhalte für die Ansicht in HTML steuern, realisiert:

> Der redaktionelle Input bestand hier im Entwurf der Wortartikelansichten und der Festlegungen dazu, welche Angaben online wie erscheinen sollen, sowie im gründlichen Testen der Umsetzung dieser Vorgaben. (Klosa 2011a, S. 22)

Umfänglich müssen auch alle Wortartikel vor ihrer Freischaltung im Internet getestet sowie inhaltlich und formal Korrektur gelesen werden. In *elexiko* werden alle Wortartikel zunächst einer doppelten inhaltlichen Korrektur sowie anschließend einer formalen Korrektur unterzogen. Alle vorgenommenen Korrekturen werden nochmals kontrolliert, und vor dem Erscheinen des fertigen Artikels im Internet werden generell alle Hyperlinks im Wortartikel überprüft, Illustrationen werden testweise geöffnet, Tondateien testweise abgespielt usw.

Während dies bei einem Ausbauwörterbuch wie *elexiko* ein fortwährender Prozess ist, war das Verfassen der Benutzungshinweise, des Glossars und weiterer Wörterbuchaußentexte ein geschlossenes Arbeitspaket, das vor der Veröffentlichung der ersten Wortartikel des „Lexikons zum öffentlichen Sprachgebrauch“ eingeschoben wurde. Zwischenzeitlich (2007) sind diese Texte allerdings schon wieder gründlich überarbeitet und ergänzt worden.

3. Kritische Betrachtung des lexikographischen Prozesses in *elexiko*

Wie eingangs schon erwähnt wurde, wurde mit der Arbeit am Projekt *elexiko* begonnen, ohne den für die Realisierung notwendigen computerlexikographischen Prozess vollständig zu durchdenken, sodass es zu Versäumnissen hinsichtlich wichtiger planerischer Schritte kam. Zugleich ist aber festzuhalten, dass erst in der praktischen Arbeit an diesem vollständig neu aus Korpusdaten erarbeiteten, ausschließlich für das Medium Internet konzipierten und erst-

mals als Wörterbuch im Aufbau publizierten Nachschlagewerk die nötigen Erfahrungen gesammelt werden konnten, um schließlich den lexikographischen Prozess in seiner Gänze erfahren und beschreiben zu können.

Da die Arbeit am Projekt *elexiko* nicht isoliert, sondern in einem ganz bestimmten Umfeld am IDS stattgefunden hat bzw. stattfindet, muss sich die Ausgestaltung des lexikographischen Prozesses unter Umständen an die hier vorhandenen Gegebenheiten anpassen. Dies hatte zum einen Auswirkungen auf die Inhalte (die grammatischen Angaben in *elexiko* sind beispielsweise konform mit den Angaben in grammis), zum anderen auf die eingesetzte Technik (z.B. auf die letztendliche Entscheidung für die Datenbank Oracle, die bereits für grammis eingesetzt wurde).

Generell ist der lexikographische Prozess von *elexiko* weniger dadurch zu beschreiben, dass seine Phasen hintereinander ablaufen, sondern eher dadurch, dass stets bestimmte Phasen des Prozesses parallel ablaufen und sich bestimmte Phasen auch wiederholen können. Der Prozess lässt sich im Grunde weniger linear, als in gewisser Weise zirkulär beschreiben. 2014 befand sich etwa ein Bearbeitungsteilwortschatz zu Wortbildungsmitteln noch in der Planungsphase, während sich das Modul „Lexikon zum öffentlichen Sprachgebrauch" bereits in der Phase der Datenauswertung befand bzw. teilweise bereits veröffentlicht war. Auch einzelne Stichwörter in *elexiko* können sich in unterschiedlichen Phasen des Prozesses befinden: Ein und dasselbe Stichwort kann z.B. bereits mit automatisch generierten lexikographischen Angaben publiziert sein, für die nachgelagerte redaktionelle Bearbeitung setzt aber erst die Phase der Datenauswertung ein. Ebenso können sich Mengen von Stichwörtern in einem Bearbeitungsteilwortschatz in unterschiedlichen Phasen befinden. Derzeit läuft für alle redaktionell bearbeiteten und online freigeschalteten Stichwörter, die illustrierbar sind, teilweise noch die Phase der Datenbeschaffung, indem gezielt nach neuen oder weiteren Illustrationen gesucht wird. In dem Moment, in dem die Stichwörter nachträglich mit Illustrationen versehen werden, durchlaufen sie wieder die Phase der Datenauswertung (Prüfung der Abbildungen, Auswahl des Bildausschnittes, Erfassen der nötigen Informationen im Wortartikel) und der Vorbereitung für den Onlinerelease (Kontrolle der Darstellung, Überprüfung des Links zur Bildquelle).

Generell lässt sich aus der beschriebenen Komplexität des lexikographischen Prozesses im Projekt *elexiko* Folgendes schließen, was sicherlich auch auf andere Wörterbuchprojekte übertragen werden kann:

> An online dictionary under construction is an open system (see Schröder 1997: 16). This also implies that organizing (calculating, segmenting, regulating) and controlling this process (see Wiegand 1998: 134) is of the utmost importance. Training the staff to work simultaneously on many different things in different stages is also helpful (as discussed by Landau 1984: 264). It is also very important to render this process transparent for users [...]. (Klosa 2013, S. 506).

4. Literaturverzeichnis

4.1 Fachliteratur

Abel, Andrea/Klosa, Annette (2012): Der lexikographische Arbeitsplatz – Theorie und Praxis. In: Vatvedt Fjeld, Ruth/Torjusen, Julie Matilde (Hg.): Proceedings of the 15th EURALEX International Congress in Oslo 2012, S. 413-421.

Belica, Cyril (2011): Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In: Abel, Andrea/Zanin, Renata (Hg.): Korpora in Lehre und Forschung. Bozen-Bolzano: Freie Universität, S. 155-178.

Dubois, Claude (1990): Considérations générales sur l'organisation du travail lexicographique. In: Hausmann et al. (Hg.), S. 1574-1588.

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

Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst/Zgusta, Ladislav (Hg.) (1990): Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie. 2. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 5.2). Berlin/New York: de Gruyter

Klein, Wolfgang (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: Scharnhorst, Jürgen (Hg.): Sprachkultur und Lexikographie. (= Sprache – System und Tätigkeit 50). Frankfurt a.M. u.a.: Lang, S. 281-308.

Klosa, Annette (2013): The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus Hjalmar/Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herbert Ernst (Hg.): Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. (= Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 5.4). Berlin/Boston: de Gruyter, S. 517-524.

Klosa, Annette (2011a): Einleitung. In: Klosa (Hg.), S. 9-26

Klosa, Annette (2011b): Von Abbildung bis Wortelement: Weitere Ergänzungen und Änderungen in *elexiko*. In: Klosa (Hg.), S. 157-172.

Klosa, Annette (Hg.) (2011): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur Deutschen Sprache 55). Tübingen: Narr.

Klosa, Annette (2010): On the combination of automated information and lexicographically interpreted information in two German online dictionaires. In: Granger, Sylviane/Paquot, Magali (Hg.): eLexicography in the 21st century: New challenges, new applications, Louvain-la-Neuve, 22-24 October 2009, Centre for English Corpus Linguistics, Université catholique de Louvain. Louvain-la-Neuve: CECL, S. 157-163.

Klosa, Annette/Koplenig, Alexander/Töpel, Antje (2011): Benutzerwünsche und Meinungen zu einer optimierten Wörterbuchpräsentation – Ergebnisse einer Onlinebefragung zu *elexiko*. (= OPAL – Online publizierte Arbeiten zur Linguistik 3/2011). Mannheim: Institut für Deutsche Sprache. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2011-3.pdf.

Knowles, Francis E. (1990): The Computer in Lexicography. In: Hausmann et al. (Hg.), S. 1645-1672.

Landau, Sidney L. (1984): Dictionaries. The Art and Craft of Lexicography. New York: Cambridge University Press.

Lemberg, Ingrid (2001): Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In: Lemberg/Schröder/Storrer (Hg.), S. 71-91.

Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.) (2001): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. (= Lexikographica. Series Maior 107). Tübingen: Niemeyer.

Meyer, Peter (2011): *vernetziko*: A Cross-Reference Management Tool for the Lexicographer's Workbench. In: Kosem, Iztok/Kosem, Karmen (Hg.): Electronic lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex2011, Bled, Slowenien, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovene Studies, S. 191-198. www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-25.pdf.

Meyer, Peter/Müller-Spitzer, Carolin (2010): Consistency of sense relations in a lexicographic context. In: Barbu Mititelu, Verginica/Pekar, Viktor/Barbu, Eduard (Hg.): Proceedings of the Workshop 'Semantic Relations. Theory and Applications', 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta. www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf.

Müller-Spitzer, Carolin/Schneider, Roman (2009): Ein XML-basiertes Datenbanksystem für digitale Wörterbücher – Ein Werkstattbericht aus dem Institut für Deutsche Sprache. In: it-Information Technology 51.4, S. 197-206.

Schnörch, Ulrich (2005): Die *elexiko*-Stichwortliste. In: Haß (Hg.), S. 71-90.

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, S. 57-65.

Storjohann, Petra (2011): Paradigmatische Konstruktionen in Theorie, lexikografischer Praxis und im Korpus. In: Klosa (Hg.), S. 99-129.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion III. (= Lexikographica. Series Maior 84). Tübingen: Niemeyer, S. 106-131.

Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg/Schröder/Storrer (Hg.), S. 54-69.

Storrer, Angelika/Freese, Katrin (1996): Wörterbücher im Internet. In: Deutsche Sprache 24, S. 97-136.

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

4.2 Internetquellen (alle zuletzt eingesehen am 5.2.2015)

COSMAS II: https://cosmas2.ids-mannheim.de/cosmas2-web/

elexiko – ein Onlinewörterbuch zur deutschen Gegenwartssprache: www.elexiko.de

Datenbank gesprochenes Deutsch im Archiv für gesprochenes Deutsch: http://agd.ids-mannheim.de/datenbanken.shtml

Das Deutsche Referenzkorpus DEREKO: www.ids-mannheim.de/kl/projekte/korpora.html

grammis – das grammatische Informationssystem des IDS: http://hypermedia.ids-mannheim.de

Kookkurrenzdatenbank CCDB – V3.3: http://corpora.ids-mannheim.de/ccdb/

Lexikon zum öffentlichen Sprachgebrauch: www.owid.de/wb/elexiko/projekt/modSprachgeb.html

Oracle: www.oracle.com/de/products/database/overview/index.html

Projekt „Kookkurrenzanalyse und deren Erschließung": www.ids-mannheim.de/kl/projekte/methoden/ka.html

Projekt „Methoden der Korpusanalyse und -erschließung", Teilprojekt „Kookkurrenzanalyse und deren Erschließung": www.ids-mannheim.de/kl/projekte/methoden.html

Projekt „Texttechnologie und Datenbanken": www.ids-mannheim.de/gra/projekte/grammis2.html

Tamino: www.softwareag.com/corporate/products/az/webmethods/default.asp

Wortartikel *Mobilität* in *elexiko*: www.owid.de/artikel/62377