Erschienen in: Storrer, Angelika/Geyken, Alexander/Siebert, Alexander/Würzner, Kay-Michael (Hrsg.): KONVENS 2008 – Ergänzungsband. Textressourcen und lexikalisches Wissen. - Berlin: BBAW, 2008. S. 29-35.

# *elexiko* – ein Online-Wörterbuch zum Gegenwartsdeutschen

Annette Klosa

**Zusammenfassung.** *elexiko* ist ein Online-Wörterbuch zum Gegenwartsdeutschen, das korpusbasiert und modular erarbeitet wird. Ein Schwerpunkt liegt dabei auf der ausführlichen korpusbasierten Beschreibung der Bedeutung und Verwendung sprachlicher Ausdrücke sowie ihrer Vernetzung untereinander. Die Präsentation des Wörterbuchs soll insbesondere zeigen, wie Korpusdaten in den Wortartikeln aufbereitet werden und wie *elexiko* genutzt werden kann, um lexikalisches Wissen in verschiedenen Benutzungssituationen aus den Wortartikeln zu gewinnen.

## 1 Das Projekt *elexiko*

*elexiko* ist ein Online-Wörterbuch zur deutschen Gegenwartssprache, das am Mannheimer Institut für Deutsche Sprache erarbeitet wird.[1] Der Name *elexiko* verweist dabei darauf, dass ein elektronisches, lexikalisch-lexikologisches und korpusbasiertes Informationssystem entwickelt wird, in dem Bedeutung und Verwendung, Grammatik und Rechtschreibung der einzelnen Wörter beschrieben werden sollen.

Das Projekt stellt insbesondere den aktuellen Sprachgebrauch dar, weshalb alle Angaben aus dem zugrunde liegenden, regelmäßig aktualisierten *elexiko*-Korpus mit über 1,3 Milliarden laufenden Wortformen gewonnen werden. In diesem Korpus sind hauptsächlich Texte überregionaler Zeitungen und Zeitschriften aus Deutschland, Österreich und der Schweiz enthalten, die es ermöglichen sollen, den allgemeinen Sprachgebrauch zu untersuchen.

Auf dem Korpus basiert auch die Stichwortliste des Projektes (knapp 300.000 Einträge), die in zwei Schritten erstellt wurde: Zunächst wurden die im Korpus vorkommenden Wortformen auf entsprechende Grundformen zurückgeführt. Diese wurden dann ab einer bestimmten Vorkommenshäufigkeit in die Liste der relevanten Stichwortkandidaten aufgenommen. Die Stichwortkandidatenliste wurde kompetenzgestützt nachbearbeitet, sodass eine bereinigte Stichwortliste als Ausgangspunkt für die Artikelbearbeitung zur Verfügung steht und online unter `www.elexiko.de` (normal- und rechtsalphabetisch sortiert) publiziert ist.

1. Einen kurzen Einblick in das Projekt bietet Klosa et al. (2006). Zu aktuellen Informationen (z.B. Projektstand, Projektmitarbeiter, Zahl der bearbeiteten Stichwörter) vgl. `www.elexiko.de`.

Der elektronische Publikationsweg von *elexiko* erlaubt es, die Beschreibung der ermittelten Lemmata nicht in alphabetischer Reihenfolge vorzunehmen. Für die Bearbeitung der einzelnen Stichwörter werden deshalb zwei Methoden kombiniert: Zum einen können (teil)automatisch erzeugte Informationen über große Teile der Stichwortstrecke hinzugefügt werden. Dies sind z.B. die Angaben zur Rechtschreibung oder (zukünftig) zufällig ausgewählte Textbelege. Zum anderen können innerhalb eingegrenzter Wortschatzbereiche die Stichwörter mit detaillierten, in die Tiefe gehenden Informationen versehen werden. Dies kann z.B. die ausführliche Bedeutungs- und Verwendungsbeschreibung eines Wortes sein. Der Artikelbestand wird also modular ausgebaut, wobei die Module Mengen gleicher oder ähnlicher Arten von Wörtern umfassen, z.B. alle Wörter, die eine bestimmte Frequenz im *elexiko*-Korpus erreichen oder alle Sprechaktverben.

Projektziel ist aber nicht nur, die korpusbasierte Sprachuntersuchung zur lexikografischen Praxis zu machen, sondern auch, neue Benutzungsmöglichkeiten für die Nachschlagenden zu schaffen, die dem Medium Internet angemessen sind. Die Vernetzung der Artikel untereinander, die in einem elektronischen Wörterbuch weit über das hinausgehen kann, was ein gedrucktes Wörterbuch bieten kann, ist ein weiteres wichtiges Ziel. Schließlich ist der Wortschatz auf verschiedene Weise, nämlich durch Begriffsbeziehungen, Wortfamilien und Themen, in sich vernetzt. Diese Netze soll *elexiko* sichtbar machen, indem die "Fäden" zwischen den Wörtern systematisch z.B. in Form von Hyperlinks, zukünftig aber möglichst auch in grafischer Form dokumentiert werden.[2]

## 2 Lexikalisches Wissen in *elexiko*

Korpusdaten sind die Grundlage der Beschreibung von Bedeutung und Verwendung der Stichwörter in *elexiko*; *elexiko* ist also ein korpusgestütztes Wörterbuch.[3] Das Korpus ist als solches erkennbar vor allem in den vielen Textbelegen, welche die lexikografischen Angaben in den Wortartikeln begleiten. Belege sind in *elexiko* natürlich der Ausgangspunkt der Bedeutungsbeschreibung, dienen aber auch dem wissenschaftlichen Nachweis der lexikologischen Aussagen. Sie veranschaulichen und beweisen darüber hinaus die im Wortartikel festgehaltenen Gebrauchsregeln für ein Stichwort anhand konkreter sprachlicher Ausschnitte.

Das Prinzip der Korpusbasiertheit zeigt sich ebenfalls in der Art, wie (bei den zurzeit bearbeiteten hochfrequenten Stichwörtern) die redaktionelle Arbeit an einem Lemma beginnt: Grundsätzlich wird zunächst nach allen Treffern im *elexiko*-Korpus

2. Müller-Spitzer (2007) gibt einen ersten Einblick in die Vernetzungsstrukturen von *elexiko* und ihre Modellierung.
3. Zu korpusgestützter Lexikografie vgl. Klosa (2007).

gesucht; die durch Zufallsauswahl auf 10.000 Treffer reduzierte Belegmenge wird dann einer Kookkurrenzanalyse (Belica 1995) unterzogen. Die Lexikografen bearbeiten die Liste der statistisch signifikanten Kookkurrenzpartner und der jeweiligen syntagmatischen Muster und erfassen die hier schon greifbaren Lesarten des Stichwortes.[4] Diese werden zunächst beschrieben, bevor in einem zusätzlichen Schritt Rückprüfungen an anderen Wörterbüchern vorgenommen werden. Gegebenenfalls beginnen dann weitere Korpusrecherchen, um wenig frequente Lesarten im *elexiko*-Korpus zu belegen.

Die umfassende Dokumentation des Lesartenspektrums eines Stichwortes ist nur eine Art, wie lexikalisches Wissen in (wörterbuchtypischer) Form in *elexiko* präsentiert wird. *elexiko* bietet darüber hinaus einen relativ neuen, möglicherweise noch ungewohnten Angabetyp, der Korpusdaten in besonderer Form aufarbeitet: die Angaben zur semantischen Umgebung und den lexikalischen Mitspielern. Hinter dieser Angabeart stehen im Prinzip die (semantische, nicht syntaktische) Argumentstruktur eines Wortes und die entsprechenden semantischen Rollen. Deshalb wird diese Angabe online auch in direkter Nachbarschaft zur Bedeutungserläuterung angeboten, während Angaben zur Valenz und den möglichen Satzbauplänen innerhalb der grammatischen Angaben dokumentiert sind. Man kann diese Angabeart "linguistisch also auch oder eher unter dem Aspekt von Frames sehen" (Haß 2005: 228), wobei die möglichst leicht verständlich formulierten Fragen sozusagen die Slots repräsentieren und die jeweiligen Antwortwörter die Filler oder Partizipanten.

Wichtig ist dabei, dass die als Antworten auf eine Frage erfassten Mitspieler aus dem *elexiko*-Korpus gewonnen werden. Bei frequenten Wörtern handelt es sich um statistisch signifikante Lexeme, die vom Lexikografen gesammelt und klassifiziert wurden. In Einzelfällen und bei niedrig frequenten Stichwörtern werden daneben exemplarische Mitspieler aus Textbelegen aufgenommen, was entsprechend kommentiert wird. Dieses Vorgehen ist am Beispiel des Verbs dringen gut nachzuvollziehen; Abbildung 1 zeigt einen Ausschnitt aus der Liste der Kookkurrenzpartner[5] mit den jeweiligen syntagmatischen Mustern. Hierin sind umrahmt Mitspieler, die benennen, wer oder was dringen kann, kursiv sind Mitspieler, die anzeigen, wohin etwas dringt, und unterstrichen sind Mitspieler, die anzeigen, wodurch etwas dringt. Im Wortartikel aufbereitet finden sich als Antworten auf die Frage "Was oder wer dringt?" z.B. Einbrecher, Gas, Geräusch, Lärm, Licht, als Antworten auf die Frage "Wodurch dringt etwas?" z.B. Fenster, Haut, Mauer, Ritze und auf die Frage "Wohin dringt etwas?" finden sich als Antworten z.B. Bewusstsein, Flur, Seele, Tageslicht.

4. Zu korpusbasierter Lesartendisambiguierung vgl. Storjohann (2003).
5. Der Ausschnitt stammt aus der Kookkurrenzliste, wie sie innerhalb von COSMAS II (vgl. `http://www.ids-mannheim.de/cosmas2/uebersicht.html`) angeboten wird. Zum Analysemodul "Statistische Kollokationsanalyse und Clustering" (Belica 1995) vgl. `http://www.ids-mannheim.de/kl/projekte/methoden/ka.html`.

| | | |
|---|---|---|
| ***Bewusstsein*** öffentliche | 66% | ins öffentliche Bewusstsein gedrungen |
| Bewusstsein | 46% | in das\|ins Bewusstsein [der ...] gedrungen |
| **Ritzen** | 41% | dringt **durch** die alle Ritzen |
| **Lärm** Haus | 33% | Lärm der ... dringt ... Haus |
| Lärm | 37% | Der Lärm [der ...] dringt aus ... |
| **Licht** Fenster spärlich | 66% | Fenster drang spärlich Licht |
| Licht | 38% | dringt [...] Licht |
| **Rauchschwaden** dichte | 70% | drangen [...] dichte Rauchschwaden aus ... |
| Rauchschwaden | 44% | Rauchschwaden [...] drangen |
| **Geräusche** | 40% | dringen [die und ...] Geräusche |
| ***aussen*** | 29% | **nach** aussen ... dringen |
| **Fenster** spärlich | **100%** | Fenster dringt ... spärlich |
| Fenster | 27% | das\|die ... Fenster [...] dringt ... der |

*Abbildung 1.* Kookkurrenzpartner zu dringen

Nicht immer sind die Korpusdaten so leicht zu interpretieren und für den Wortartikel aufzubereiten wie im Beispiel dringen. Zwar gibt es redaktionelle Richtlinien dafür, welche Fragen zu welchen semantischen Klassen formuliert werden müssen (im Beispiel dringen handelt es sich um einen Handlungsprädikator, zu dem auf jeden Fall ein Frageset die möglichen Handlungsträger benennen muss), doch finden sich auch statistisch signifikante Mitspieler, die eher thematisch / diskursiv mit dem Stichwort verbunden sind. Solche Mitspieler werden unter der Frage "Was wird in Zusammenhang mit X thematisiert?" gruppiert, wie die Beispiele aus den Wortartikeln Frau und Mann zeigen:

> Was wird im Zusammenhang mit *Frau* thematisiert? Alter, Beruf, Brustkrebs, Emanzipation, Erwerbstätigkeit, Geburt, Kinder, Sex, Wechseljahre

> Was wird im Zusammenhang mit *Mann* thematisiert? Auto, Erektionsstörung, Feuerwehr, Fußball, Gleichberechtigung, Gleichstellung, Handball

Aus der Fülle der Kookkurrenzpartner ist für die Wortartikel immer eine Auswahl zu treffen, besonders dann, wenn wie im ausschließlich zeitungs- und zeitschriftensprachlichen *elexiko*-Korpus bestimmte Themen dominieren (z.B. die Sportberichterstattung, Berichte über Verbrechen und ihre Aufklärung, politische Berichterstattung) oder die Texte bestimmte Klischees widerspiegeln und zugleich transportieren. So ermorden, missbrauchen, prügeln, töten, vergewaltigen Väter ihre Kinder im Normalfall nicht (alles Beispiele aus der Kookkurrenzliste zu Vater), sondern sie lieben und erziehen sie, sie spielen mit ihnen oder erzählen ihnen etwas. In solchen Fällen muss der Lexikograf die Korpusdaten sorgfältig interpretieren, um im Wortartikel nicht korpusbedingt verzerrtes lexikalisches Wissen zu präsentieren.

Ein weiterer Angabebereich in den *elexiko*-Wortartikeln ermöglicht es den Lexikografen deshalb daneben, die Beoachtungen zu auffälligen Thematisierungen im *elexiko*-Korpus festzuhalten: die Angaben zu Besonderheiten des Gebrauchs. Hier können neben Angaben zur Einstellung des Sprechers, zum Situationsbezug, zur Textbindung oder zum Sachgebiet auch themengebundene Verwendungen dokumentiert werden, hier am Beispiel Vater (Lesart 'Mann mit Kind'):

> **Vater** findet im *elexiko*-Korpus häufig Erwähnung, wenn es um die Rechte und Pflichten eines Vater nach der Scheidung (z.B. das Besuchs- und Sorgerecht, Unterhaltszahlungen) geht. [...]

## 3 Benutzungssituationen von *elexiko*

An wen wenden sich Angaben wie die eben gezeigten und in welchen Benutzungssituationen können sie von Interesse sein? Grundsätzlich gilt, dass *elexiko* so konzipiert wurde, "dass es auf viele verschiedene Nutzungsinteressen antworten und damit mehr Wörterbuchfunktionen abdecken kann, als es bei einem gedruckten Wörterbuch sinnvoll ist. [...] Die Autoren erarbeiten ein Informationspotenzial, das unterschiedliche Funktionen des einsprachigen Wörterbuchs und unterschiedliche Nutzungsinteressen unterschiedlicher Adressatengruppen abdecken kann." (Haß 2005: 3).

Wie erste Nutzerrückmeldungen zeigen, werden die Angaben zur lexikalischen Umgebung und den semantischen Mitspielern von Dozenten des Deutschen als Fremdsprache beispielweise zur Vorbereitung des Wortschatzunterrichts genutzt. Die Angaben können darüber hinaus sogar landeskundliche Informationen vermitteln und Einblicke nicht nur in die Sprache, sondern auch in Natur und Kultur im deutschen Sprachraum Mitteleuropas geben, wie etwa die Mitspielerangaben zu den Jahreszeitenbezeichnungen (hier das Beispiel Herbst), zu den Bezeichnungen für die Wochentage, den Monatsnamen oder den Bezeichnungen für die Tageszeiten zeigen:

Was wird in Zusammenhang mit der Natur im Herbst thematisiert? Bäume, Blätter, Blumenzwiebel, Ernte, Laub, Pflanzzeit, Sträucher, Trauben, Weinlese, Zugvögel

Was wird bezogen auf den Herbst (als Termin) thematisiert? Ausbildungsplatz, Börsengang, Buchmesse, Bundestagswahl, [...], Oktoberfest, [...] Regierungswechsel, Tournee, Wahl

Damit erfüllen diese Angaben eher den Zweck enzyklopädischer Nachschlagewerke, nämlich, sachbezogene Informationen zu liefern. Benutzt man die Informationen in *elexiko*, um Sprachunterricht vorzubereiten oder um sich Sachinformationen zu verschaffen, entspricht dies keiner der klassischen Wörterbuchbenutzungssituationen (Textrezeption oder Textproduktion). Natürlich können die Angaben zur semantischen Umgebung und den lexikalischen Mitspielern aber auch bei der Textproduktion benutzt werden, beispielsweise, um den passenden nominalen Mitspieler zu einem Verb oder das treffende Adjektivattribut zu einem Nomen zu finden.

Ein Angabebereich, der ebenfalls für Situationen der Textproduktion (hier genauer für die Wortfindung beim Verfassen eines Textes) von Interesse ist, umfasst die Angaben zu den sinnverwandten Wörtern.[6] Neben Synonymen oder Antonymen werden in *elexiko* auch Relationen der Über- und Unterordnung oder Teil-Ganzes-Beziehungen erfasst. Hinzu kommen Relationspartner, die etwa Ursache und Folge (z.B. Ursache – Wirkung) ausdrücken oder eine responsive Folgerelation bezeichnen (z.B. abverlangen mit den möglichen responsiven Handlungen geben, gewähren, schenken).

Die Erfassung und Beschreibung dieser Relationspartner nimmt nicht nur für die Lexikografen bei der Wortartikelerarbeitung eine zentrale Rolle ein, indem die sinnverwandten Wörter Bedeutung und Gebrauch des Stichwortes zu verdeutlichen helfen, Verwendungspräferenzen oder -beschränkungen illustrieren und nicht zuletzt bei der Disambiguierung einzelner Lesarten herangezogen werden können. Dieser Angabebereich dient dem Nutzer auch in besonderer Weise der Erkenntnis über die Vernetztheit eines Stichwortes mit anderen lexikalischen Ausdrücken, vermittelt also in der Zusammenschau im Wortartikel lexikalisches Wissen rund um das jeweilige Stichwort.

## Literaturverzeichnis

Belica, Cyril (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode*. Mannheim: Institut für Deutsche Sprache.

6. Die korpusbasierte Erfassung paradigmatischer Partner wird in Storjohann (2005) erläutert.

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

Klosa, Annette (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In: Werner Kallmeyer und Gisela Zifonun (Hrsg.), *Sprachkorpora - Datenmengen und Erkenntnisfortschritt*, Band 2006 , *Jahrbuch des Instituts für Deutsche Sprache*, 105–122, Berlin/New York: Mouton de Gruyter.

Klosa, Annette, Ulrich Schnörch, und Petra Storjohann (2006). A lexical and lexicological, corpus-based hypertext information system at the Institut für Deutsche Sprache, Mannheim. In: Carla Marello et al. (Hrsg.), *Proceedings of the 12th EURALEX International Congress (Atti del XII Congresso Internazionale di Lessicografia)*, Band 1, 425–430, EURALEX, Turin: Edizioni dell'Orso Alessandria.

Müller-Spitzer, Carolin (2007). Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. *Hermes* 38:137–171.

Storjohann, Petra (2003). Computergestützte Lesartendisambiguierung. *Deutsche Sprache* 2003(1):3–28.

Storjohann, Petra (2005). Corpus-driven vs. corpus-based approach to the study of relational patterns. In: *Proceedings of the Corpus Linguistics Conference*, Birmingham, URL `http://www.corpus.bham.ac.uk/PCLC/`.