Erschienen in: Kratochvílová, Iva/Wolf, Norbert Richard (Hrsg.): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. - Heidelberg: Winter, 2010. S. 103-115. (Germanistische Bibliothek 38)

Chancen und Probleme korpusgestützter Lexikografie

Am Beispiel deutschsprachiger Online-Wörterbücher

Annette Klosa, Mannheim

Eine der linguistischen Teildisziplinen, in der schon seit vielen Jahren korpusgestützt gearbeitet wird, ist die Lexikografie. Wörterbücher sind lange vor der Entstehung großer elektronischer Textsammlungen mit den entsprechenden Korpusrecherche- und -analysewerkzeugen auf der Basis von umfänglichen Belegsammlungen entstanden, die nach dem Verständnis vieler Lexikografen das Korpus (bzw. die Primärquelle) des Wörterbuches darstellen. Noch heute arbeiten verschiedene Großwörterbücher (z. B. das Oxford English Dictionary) am Ausbau ihrer Belegsammlungen und benutzen diese neben zum Teil eigens aufgebauten elektronischen Wörterbuchkorpora im engeren Sinn. Welche Chancen und Probleme sich bei korpusgestützter Arbeit an Wörterbücher ergeben, wird in diesem Beitrag allerdings nicht am Beispiel traditioneller, gedruckter Wörterbücher erläutert (vgl. hierzu Klosa 2007), sondern ausschließlich an deutschsprachigen Online-Wörterbüchern aufgezeigt, wobei zunächst einige Definitionen erarbeitet werden müssen. Ein kurzer Ausblick auf die Auswirkungen korpusgestützter Arbeit an Online-Nachschlagewerken auf den lexikografischen Prozess schließt diesen Beitrag ab.

1. Definitionen

Nach Herbert Ernst Wiegand (1998) versteht man unter einem Sprachwörterbuch (im Folgenden nur noch: ‚Wörterbuch') ein „Nachschlagewerk, dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann" (Wiegand 1998, 58). Der Benutzer ist dabei immer ein menschlicher Benutzer. Offen bleibt in dieser Definition die Frage, vor allem im Bereich der korpusgestützten Lexikografie, ob die Daten automatisch oder von menschlicher Hand erstellt sind. Deshalb sollte man (mit Müller-Spitzer 2003) differenzieren zwischen automatisch erstellten „Wortschatzinformationssystemen" und lexikografisch bearbeiteten Wörterbüchern. Unter lexikografischer Bearbeitung wird dabei „jede Art der reflektierten menschlichen Bearbeitung der automatisch erstellten Daten verstanden, vom Überprüfen über das Umsortieren bis hin zum Kommentieren" (Müller-Spitzer 2003, 150). Daneben gibt es Nachschlagewerke, in denen beide Verfahren, also automatische und lexikografische Erarbeitung wortbezogener Informationen, kombiniert werden. Solche Nachschlagewerke können dann zu Wörterbüchern gerechnet werden, nicht aber zu wortbezogenen Informationssystemen, wenn der Anteil an lexikografisch erarbeiteter Information überwiegt und zwischen automatisch und redaktionell erarbeiteten Informationen redaktionell erstellte Verknüpfungen bestehen. In diesem Beitrag werden sowohl lexikografisch bearbeitete elektronische Wörterbücher wie wortbezogene Informationssysteme betrachtet, weil Letztere erst im Zusammenhang mit dem fortschreitenden Ausbau umfangreicher elektronischer Korpora und der Entwicklung korpustechnologischer und informationstechnologischer Verfahren möglich wurden bzw. werden.

Ein Korpus ist eine Sammlung von Texten, die zum Zweck der linguistischen Analyse nach bestimmten Auswahlkriterien zusammengestellt wurde. Dabei wird der Anspruch erhoben, dass diese Sammlung natürliche, authentische Sprache repräsentiert (vgl. Tognini-Bonelli 2001, 2). Mit ‚Korpus' ist häufig implizit außerdem eine in elektronischer Form zur Verfügung stehende Textsammlung gemeint, die mithilfe von elaborierter Recherche- und Analysesoftware erschlossen werden kann. Im Folgenden wird unter ‚Korpus' im lexikografischen Kontext ein elektronisches Korpus verstanden, welches für ein bestimmtes Wörterbuch gezielt zusammengestellt wurde (vgl. Landau 2001, 323). Doch was heißt eigentlich ‚auf der Basis eines Korpus' bzw. genauer ‚korpusgestützt' oder ‚korpusgebunden'?

Ein korpusgebundenes Wörterbuch wird ausschließlich auf der Basis des Wörterbuchkorpus und ohne Hinzuziehung anderer primärer Quellen, aber auch ohne Hinzuziehung sekundärer und/oder tertiärer Quellen erarbeitet. Solche sekundären Quellen sind

> „alle Wörterbücher, die nach dem Instruktionsbuch entweder obligatorisch oder fakultativ konsultiert werden sollen, und zu den tertiären Quellen gehören alle sonstigen Sprachmaterialien, die benutzt werden sollen, wie z. B. linguistische Monographien und Grammatiken [...]." (Wiegand 1998, 140).

Ein korpusgebundenes Wörterbuch (z. B. ein Autorenwörterbuch) bildet genau die Sprachwirklichkeit ab, die das zugrunde gelegte Korpus repräsentiert. In der korpusgestützten Lexikografie, die sich ebenfalls ausschließlich auf ein Wörterbuchkorpus als primäre Quelle stützt, werden dagegen noch sekundäre und/oder tertiäre Quellen hinzugezogen. Ein korpusgestütztes Wörterbuch bietet also keine 1:1-Abbildung der Sprachwirklichkeit des zugrunde gelegten Korpus, sondern ergänzt, wo nötig, das Bild, das aus der Korpusanalyse gewonnen wurde. Innerhalb der korpusgestützten Lexikografie selbst können wiederum zwei unterschiedliche Ansätze verfolgt werden: der corpus-based-Ansatz und der corpus-driven-Ansatz: Unter ‚corpus-based' versteht man eine Methode, die von einem Korpus Gebrauch macht, um Theorien und Beschreibungen darzulegen, zu testen oder zu veranschaulichen. Man sucht im Korpus nach Beispielen, die bestimmte linguistische Argumente unterstützen oder theoretische Annahmen validieren sollen (vgl. Tognini-Bonelli 2001, 65). Im Gegensatz dazu verpflichtet sich ein Lexikograf beim corpus-driven-Ansatz den Korpusdaten als Ganzes, er wertet sie insgesamt aus ohne Vorannahmen und beschreibt die so gewonnenen Daten vollständig (vgl. Tognini-Bonelli 2001, 84). In der lexikografischen Praxis werden beide Ansätze auch kombiniert (vgl. zu einem Beispiel aus der Praxis in *elexiko* [Storjohann 2005]).

2.1. ‚Wortschatz Deutsch'

Seit etwa 10 Jahren wird an der Universität Leipzig das Projekt ‚Wortschatz Deutsch' erarbeitet mit dem Ziel, auf der Basis großer Textkorpora die Möglichkeiten der automatischen Sprachverarbeitung auszuschöpfen und die auf diese Weise ermittelten wortbezogenen Ergebnisse im Internet anzubieten. Die Quellen, also die Korpustexte, stammen ausschließlich aus dem Internet. Auf den Projektseiten werden zu über sechs Millionen deutschen Wörtern Frequenzangaben, grammatische Angaben, Angaben zu ihrer thematischen Einordnung, Kookkurrenzangaben und Korpusbeispiele gemacht. ‚Wortschatz Deutsch' ist vermutlich die größte Datensammlung zum Deutschen, die gegenwärtig konsultiert werden kann.

Gesetzt den Fall, ein Nutzer von ‚Wortschatz Deutsch' möchte in einer Situation der Textrezeption die Bedeutung des Wortes *Wörterbuch* nachschlagen, dann helfen ihm für das Verständnis dazu, was ein Wörterbuch ist, was man damit machen kann, wie es aussieht usw., im Online-Angebot von ‚Wortschatz Deutsch' am ehesten die Beispiele (vgl. Abb. 1). Unglücklicherweise erscheinen beim Suchwort *Wörterbuch* drei Textbelege, die relativ wenig Normales zum Konzept ‚Wörterbuch' transportieren. Als adjektivische Mitspieler zu *Wörterbuch* stehen *visuell, hässlich, groß* und *klein*, ein Belegkontext ist offensichtlich computersprachlich, in einem Beispiel wird *Wörterbuch* im Sinne von ‚Wortschatz' verwendet usw. Öffnet man weitere Textbelege, lässt sich insgesamt mehr über den Ausdruck *Wörterbuch* erfahren, doch muss man als Nutzer dann das tun, was einem sonst die Lexikografen abnehmen, nämlich aus möglichst vielen Textbelegen die Bedeutung eines Wortes zu ermitteln. Es wird außerdem deutlich, dass eine vollautomatische Belegauswahl nicht die gleiche Qualität haben kann wie eine lexikografische Auswahl.

Beispiel(e):
Die Keypatches werden mit einem visuellen **Wörterbuch** verglichen. (Quelle: *welt.de vom 13.01.2005*)
[...] um gleich zwei Vokabeln aus dem hässlichen **Wörterbuch** der Grünen zu zitieren. (Quelle: *spiegel.de vom 14.01.2005*)
[...]: kleines Wörterbuch für die Disko, großes **Wörterbuch** für die Schule. (Quelle: *archiv.tagesspiegel.de vom 15.01.2005*)
weitere Beispiele

Abb. 1: Beispiele für *Wörterbuch* in ‚Wortschatz Deutsch' (Ausschnitte)

Unter der Überschrift „Relationen zu anderen Wörtern" (vgl. Abb. 2) erfährt man zum Suchwort *Wörterbuch*, dass *Lexikon*, *Wörterverzeichnis* und *Zitatensammlung* Synonyme zu *Wörterbuch* sind, dass *Wörterbuch* mit den Bezeichnungen *Duden* und *Lexikon* zu vergleichen ist, dass *Wörterbuch* Synonym von *Enzyklopädie*, *Fibel*, *Lexikon* und *Nachschlagewerk* ist und referenziert wird von *Nachschlagewerk*. Gesetzt den Fall, ein Nutzer von ‚Wortschatz Deutsch' sucht in einer Situation der Textproduktion nach Synonymen für *Wörterbuch*, wird er überlegen müssen, warum *Wörterbuch* z. B. zwar als

Synonym von *Nachschlagewerk* bezeichnet wird, *Nachschlagewerk* aber nicht als Synonym zu *Wörterbuch* gebucht ist. Man erfährt über die paradigmatischen Angaben im Leipziger Wortschatzinformationssystem auch nicht wirklich, wie sich die Bezeichnungen *Wörterbuch, Enzyklopädie* oder *Nachschlagewerke* unterscheiden bzw. in welcher Relation sie zueinander stehen (*Wörterbuch* und *Enzyklopädie* sind Kohyponyme zum Oberbegriff *Nachschlagewerk*). Nutzer können aber erkennen, dass all diese Wörter auf eine bestimmte Art und Weise miteinander verwandt sind.

Relationen zu anderen Wörtern:

- Synonyme: Lexikon, Wörterverzeichnis, Zitatensammlung
- vergleiche: Duden, Lexikon
- ist Synonym von: Enzyklopädie, Fibel, Lexikon, Nachschlagewerk
- wird referenziert von: Nachschlagewerk

Abb. 2: Paradigmatische Relationen zu *Wörterbuch* in „Wortschatz Deutsch"

Ein kurzes Fazit zum Informationssystem ‚Wortschatz Deutsch' kann lauten: Ohne eine elektronische Korpusbasis und ohne die Möglichkeiten der automatischen Sprachverarbeitung könnte es solch ein System nicht geben. Als weiterer Vorteil liegt die riesige Menge an angebotenen Informationen auf der Hand. Diese gehen in ihrer Art zum Teil über das hinaus, was man üblicherweise aus Wörterbüchern gewohnt ist (z. B. mit der Auflistung signifikanter Kookkurrenzen oder ihrer grafischen Präsentation). Insofern sind für ein solches System vielleicht auch neue Nutzungssituationen anzunehmen, die nicht an Textproduktion oder -rezeption gebunden sind. ‚Wortschatz Deutsch' kann man z. B. als Lehrer des Deutschen als Fremdsprache nutzen, um Wortschatzübungen vorzubereiten. Problematisch ist ein rein automatisch erstelltes Wortschatzinformationssystem insofern, als die Angaben möglicherweise nicht korrekt sind. ‚Wortschatz Deutsch' bietet zwar eine große Quantität, aber die Qualität lässt manches Mal zu wünschen übrig. Abweichend von an Wörterbüchern klassischer Machart geschulten Nutzungsgewohnheiten muss ein Nutzer von ‚Wortschatz Deutsch' daher die gebotenen Informationen wesentlich intensiver und sorgfältiger interpretieren.

2.2. Das ‚Digitale Wörterbuch der deutschen Sprache' (DWDS)

Das Projekt ‚Digitales Wörterbuch der deutschen Sprache' entsteht an der Berlin-Brandenburgischen Akademie der Wissenschaften und setzt sich folgende Ziele:

> „[...] die Schaffung eines ‚Digitalen Lexikalischen Systems' — einer umfassenden, jedem Benutzer über das Internet zugänglichen Datenbank, die Auskunft über den deutschen Wortschatz in Vergangenheit und Gegenwart gibt. Dazu wird eine Benutzeroberfläche geschaffen, die zum einen als ‚lexikografischer Arbeitsplatz' für die wissenschaftliche Analyse des deutschen Wortschatzes

fungiert, zum anderen aber jedem Interessierten viele Suchmöglichkeiten eröffnet. [...] [Das Projekt] soll das verfügbare lexikalische Wissen, wie es in den bisherigen großen Wörterbüchern seinen Niederschlag gefunden hat, zusammenführen und auf den neuesten Stand bringen. Es soll ein Digitales Lexikalisches System entwickeln, das 1. Belege für die möglichen Verwendungen eines Wortes — aus gut erschlossenen Korpora — und eine wissenschaftlich verlässliche Beschreibung der verschiedenen Eigenschaften dieses Wortes miteinander verbindet, 2. sich jederzeit flexibel erweitern und korrigieren lässt, und 3. für viele — wissenschaftliche wie nichtwissenschaftliche — Zwecke nutzbar ist." (URL 1)

Das Suchergebnis präsentiert sich in der Standardansicht in einem viergeteilten Bildschirm, in dem sehr unterschiedliche Angaben verbunden werden, was ihre Erarbeitung (lexikografisch oder automatisch) und ihre Art (klassische Wörterbuchangaben, Korpusextrakte, grafische Präsentation) anbelangt. Das DWDS-Wörterbuch im ersten Teilfenster beruht auf dem retrodigitalisierten, sechsbändigen ‚Wörterbuch der deutschen Gegenwartssprache' (WDG) (vgl. Abb. 3). Diese elektronische Fassung bietet die Möglichkeit, nur bestimmte Angaben im Artikel anzeigen zu lassen (z. B. können Zusammensetzungen mit dem Stichwort zusätzlich geöffnet werden). Inhaltlich ist das WDG allerdings (noch) nicht überarbeitet worden.

Abb. 3

Im OpenThesaurus-Fenster „werden, nach Gruppen geordnet, die dem Stichwort nach der Bedeutung verwandten Wörter angezeigt" (URL 2). Das ‚Wörterbuch für Synonyme und Assoziationen', aus dem diese Angaben stammen, wird nicht innerhalb des DWDS-Projektes entwickelt und gepflegt, sondern es handelt sich hierbei um „ein freies deutsches Synonymwörterbuch, bei dem jeder mitmachen kann" (URL 3). Als Synonyme werden im hier gewählten Beispiel *Wörterbuch* (vgl. Abb. 4) die Wörter *Lexikon* und *Verzeichnis* genannt, als Oberbegriffe *Kompendium* und *Nachschlagewerk*. Warum beispielsweise *Verzeichnis* als Synonym, nicht aber als Oberbegriff (‚ein Wörterbuch ist eine Art von Verzeichnis') eingeordnet wurde, bleibt offen. Da unklar ist, wer diese Angaben auf welcher Grundlage (auf der Basis der eigenen Sprachkompetenz, durch Auswertung des DWDS-Korpus oder anderer Korpora, durch Abgleich mit dem WDG oder anderen Wörterbüchern?) erarbeitet hat, ist fraglich, ob sie den im Projekt selbst gestellten, oben zitierten Anspruch der „wissenschaftlichen Verlässlichkeit" erfüllen. Mit den Angaben im DWDS-Wörterbuch sind die Einträge im Thesaurus jedenfalls nicht verknüpft, weder durch automatische Verlinkung, noch durch redaktionellen Abgleich: Das Wort *Verzeichnis* in der Bedeutungserläuterung zu *Wörterbuch* kann man beispielsweise nicht anklicken, um dadurch den Thesaurus zu aktivieren. Und obwohl der Ausdruck *Verzeichnis* in dieser Paraphrase eindeutig als Oberbegriff verwendet wird, nennt der Thesaurus ihn als Synonym.

Abb. 4

Da Zitate in den Wortartikeln des WDG eher selten vorkommen (im Beispiel *Wörterbuch* etwa gar nicht, vgl. Abb. 3), sind die im DWDS-Kernkorpus-Fenster angebotenen KWIC-Zeilen mit den zugehörigen Belegen an sich eine willkommen Ergänzung. So bereichern z. B. die adjektivischen Mitspieler zu *Wörterbuch* in den Belegen (*klinisch, Thüringer*, vgl. Abb. 5) die Angaben im Wörterbuch. Und aus den Belegen erfährt man, dass ein *Wörterbuch* beispielsweise bei etwas *hilft*, *Einblick gibt* oder etwas *sagt* — Kontexte, die im Wörterbuchartikel fehlen. Ansonsten gilt das oben zur automatischen Belegauswahl im Projekt ‚Wortschatz Deutsch' Gesagte auch hier: Nicht immer sind die Belege optimal, um über die Bedeutung des Stichwortes zu informieren, und vor allem muss sie der Nutzer selbst interpretieren. Doch haben die DWDS-Belege insgesamt eine bessere Qualität, weil sie aus einem Korpus stammen, dessen Zusammensetzung sorgfältig geplant wurde. Allerdings war das DWDS-Korpus natürlich nicht die Primärquelle des WDG, das auf der Basis einer klassischen Belegsammlung erarbeitet wurde. Diese Belegsammlung bzw. die vollständigen Quellentexte sind wohl auch nicht in digitalisierter Form ins DWDS-Korpus eingegangen. Das DWDS-Wörterbuch ist im oben definierten Sinn kein korpusgestütztes Wörterbuch und es würde auch nicht durch eine automatische Verknüpfung mit dem DWDS-Korpus zu einem solchen.

Abb. 5

Schließlich gibt es das DWDS-Wortprofil. Dieses zeigt die Wörter oder Phrasen an,

> „die mit dem Stichwort in der ausgewählten Wortart typischerweise vorkommen. Je auffälliger der Zusammenhang mit dem Stichwort ist, desto größer wird dieses Wort bzw. diese Phrase dargestellt. Die Wörter und Phrasen sind als sog. Wörterwolkwolke [sic!] angeordnet." (URL 4)

Zu den „relevantesten syntaktischen Relationen" für ein *Wörterbuch* zählen: *Gegenwartssprache, einsprachig, Autorenporträt, Textauszug, Partnerverlag, grimmesch, digital* usw. (vgl. Abb. 6). Es ist nicht leicht, diese Angaben zu interpretieren: Inwiefern sind *Autorenporträt, Partnerverlag* und *Textauszug* diejenigen Wörter, die besonders typisch zusammen mit *Wörterbuch* vorkommen? In welchen Kontexten könnte dies so sein? Zeigt man mit der Maus auf eines dieser Wörter, erscheint in einem kleinen Fenster immerhin die Zusatzinformation „Wörterbuch hat Genitivattribut" (d. h. dieser Angabe liegt das Muster *Wörterbuch der Gegenwartssprache* zugrunde). Es verwundert, dass fast keines dieser Wörter, die mit dem Stichwort typischerweise vorkommen, im DWDS-Wörterbuchartikel erscheint. Umgekehrt kommt praktisch keines der Wörter, die im DWDS-Wörterbuchartikel zur Bedeutungserläuterung oder in den Beispielen verwendet werden, im automatisch erstellten Wortprofil vor. Die automatisch erstellten Angaben im DWDS-Wortprofil sind also redaktionell nicht mit den Angaben im DWDS-Wörterbuch abgestimmt; die Wörter, die im Wortprofil erscheinen, sind leider auch nicht als Hyperlinks zu den entsprechenden Wörterbuchartikeln realisiert. Dafür können aber immerhin durch Doppelklick auf eines der Wörter im Wortprofil passende Korpusbelege geöffnet werden. Aus diesen wird dann z. B. ersichtlich, dass der Partner *Sprache* immer nur

adjektivisch attribuiert als Genitivattribut zu *Wörterbuch* erscheint (z. B. in *Wörterbuch der philosophischen Sprache, Wörterbuch der deutschen und japanischen Sprache*).

Abb. 6

Auch für das ‚Digitale Wörterbuch der deutschen Sprache' gilt: Ohne eine elektronische Korpusbasis und ohne die Möglichkeiten der automatischen Sprachverarbeitung könnte es solch ein System nicht geben. Es ist nicht einfach, das DWDS als Wörterbuch oder als Wortschatzinformationssystem einzuordnen, weil im Grunde beides unter einer Oberfläche verbunden wird. Jedenfalls ist mit diesem Angebot an menschliche Nutzer gedacht, denen sich einerseits eine große Fülle an sprachlichen Informationen bietet, denen aber andererseits eine gehörige Portion an Sprachwissen und Wissen über die Benutzung digitaler Ressourcen abverlangt wird, damit sich ihnen diese Informationen erschließen.

2.3. ‚elexiko'

Das Online-Wörterbuch ‚elexiko' (vgl. Haß 2005) entsteht am Mannheimer Institut für Deutsche Sprache (IDS) und kann seit 2004 im Internet kostenlos benutzt werden. ‚elexiko' hat das Ziel, ein sehr umfangreiches einsprachiges Wörterbuch des Gegenwartsdeutschen zu erarbeiten: Die Wortliste umfasst 300.000 Einträge. In diesem Wörterbuch werden Bedeutung und Verwendung, Grammatik, Rechtschreibung und Wortbildung der einzelnen Wörter korpusgestützt beschrieben. Die Primärquelle, aus der alle Angaben gewonnen werden, ist das ‚elexiko'-Korpus mit über 2,8 Milliarden einzelner Wortformen aus Zeitungs- und Zeitschriftentexten. Es ist ein virtuelles Korpus aus dem ‚Deutschen Referenzkorpus' des IDS, das regelmäßig ergänzt und aktualisiert wird. Der Auswertung dient die am IDS entwickelte automatische Recherche- und Analysesoftware COSMAS II. Einen besonderen Stellenwert für ‚elexiko' hat daneben das statistische Verfahren der Kookkurrenzanalyse (Belica 1995). Auf dem Korpus basiert außerdem die Stichwortliste, die bis zu einem gewissen Grad dynamisch und offen ist.

Der für ‚elexiko' vorgesehene elektronische Publikationsweg ermöglicht es, die Beschreibung der ermittelten Lemmata nicht von A bis Z vorzunehmen. Zum einen werden daher z. T. automatisch erzeugte, auf Breite angelegte Informationen über die gesamte Stichwortstrecke hinzugefügt (z. B. Angaben zur Rechtschreibung). Zum anderen werden innerhalb eingegrenzter Wortschatzbereiche detaillierte, komplexe und in die Tiefe gehende Informationen hinzugefügt (z. B. die genaue Bedeutungs- und Verwendungsbeschreibung eines Wortes). Der Artikelbestand wird also modular ausgebaut, wobei Module unterschiedlich definiert sind: Wörter in einem Modul können nach ihrer Frequenz im ‚elexiko'-Korpus ausgewählt sein oder weil sie als Wortfamilie oder Wortfeld zusammengehören usw. Derzeit werden im sogenannten ‚Lexikon zum öffentlichen Sprachgebrauch' alle Stichwörter, die mindestens 10.000 mal im Korpus belegt sind, sehr ausführlich und mithilfe des *corpus-driven-* und des *corpus-based-*Ansatzes beschrieben. Ein weiteres Modul widmet sich den niedrig frequenten Stichwörtern in der Wortliste, d. h. solchen Wörtern, die weniger als 500 mal im ‚elexiko'-Korpus belegt sind. Dieses Modul

wird ausschließlich mit (teil-)automatischen Angaben gefüllt. Um eine hohe Verlässlichkeit der Informationen in den Wortartikeln zu erreichen, werden die automatisch gewonnenen Angaben, z. B. mithilfe eines Morphologisierungstools ermittelte Wortbildungsanalysen, redaktionell überprüft, bevor sie on-line erscheinen. Dies hat zur Folge, dass die Füllung der niedrig frequenten Stichwörter langsamer vorangeht, als dies bei einer ausschließlich automatischen Methode der Fall wäre.

Das Suchwort *Wörterbuch* ist in *elexiko* ausschließlich mit automatisch ermittelten Angaben versehen, nämlich Angaben zur Orthografie und Silbentrennung, die von drei automatisch ausgewählten Belegen ergänzt werden (vgl. Abb. 7). Hinzu kommen Informationen zur Korpusbelegung und Links zur Kookkurrenzdatenbank CCDB des IDS und zum Online-Angebot canoo.net, wo man grammatische Angaben (z. B. Angabe der Wortart, Flexionstabellen) zum Stichwort aufrufen kann. Die automatische Auswahl der Belege erfolgt in ‚elexiko' nicht völlig zufällig, sondern anhand bestimmter Kriterien, die die Qualität der angezeigten Belege zu verbessern hilft: Die Belege müssen z. B. aus drei verschiedenen Quellen und Jahrgängen stammen. Zusätzlich zu den Belegen wird angezeigt, zu welcher Frequenzschicht das Stichwort in ‚elexiko' gehört und aus wie vielen verschiedenen Quellen und Jahrgängen es stammt. Hieraus lässt sich erschließen, wie verbreitet das Wort ist. Ob die in ‚elexiko' gezeigten Belege tatsächlich eine bessere Qualität ausweisen, als die automatisch ausgewählten Belege in DWDS oder ‚Wortschatz Deutsch', lässt sich pauschal kaum beantworten. Jedenfalls ist es bei vielen Stichwörtern so, dass die drei in ‚elexiko' gezeigten Belege sehr viel über Bedeutung und Verwendung des jeweiligen Wortes vermitteln und häufig auch verschiedene Einzelbedeutungen des Wortes illustrieren. Die Möglichkeit, Kookkurrenzangaben oder grammatische Angaben zum Stichwort aufzurufen, ist in ‚elexiko' so gestaltet, dass der Nutzer die Seiten des Projektes verlässt. Es wurde keine integrative Darstellung gewählt, wie dies z. B. beim DWDS der Fall ist, weil im Projekt ‚elexiko' für die in der CCDB und bei canoo.net gezeigten Inhalte keine Verantwortung übernommen werden kann. Dass aber überhaupt dorthin verlinkt wird, zeigt, dass die dort erhältlichen Informationen als weitgehend zuverlässig eingestuft werden. Geplant ist darüber hinaus, das Stichwort aus ‚elexiko' heraus direkt im ‚elexiko'-Korpus nachschlagen zu können, z. B. um weitere Belege zu erhalten. Derzeit scheitert die Realisierung vor allem noch daran, dass nicht alle Texte im ‚elexiko'-Korpus frei zugänglich sind.

Abb. 7

Bei allen Wörtern, die in *elexiko* vollständig redaktionell bearbeitet werden, werden die automatisch generierten Angaben überprüft (z. B. die Silbentrennung) oder im Zuge der Artikelbearbeitung ersetzt (z. B. die automatisch ausgewählten Belege), die Links auf die Kookkurrenzdatenbank und canoo.net werden gelöscht, weil sie durch eigene Angaben zu lexikalischen Mitspielern des Stichwortes und grammatische Angaben ersetzt werden. Zukünftig sollen aber auch automatisch gewonnene Angaben zu Wortbildungsprodukten mit dem Stichwort als Basis über die gesamte Stichwortliste erstellt werden, die sowohl bei noch nicht bearbeiteten Wörtern wie ausgearbeiteten Stichwörtern angezeigt würden. Es wird dann zu überlegen sein, ob online der unterschiedliche Status dieser Angaben deut-

lich gemacht werden sollte und wie dies gegebenenfalls geschehen könnte. Jedenfalls werden auch diese Angaben ausschließlich auf der Auswertung des *elexiko*-Korpus beruhen.

Die konsequente Fundierung auf einem sehr umfangreichen Korpus spiegelt sich in *elexiko* in einer Fülle verschiedener, den aktuellen Sprachgebrauch beschreibende Angaben in den Wortartikeln wider. Diese zu erarbeiten, kostet natürlich Zeit. Bis wesentlich mehr Wörter redaktionell bearbeitet sind als derzeit, können automatisch ermittelte Angaben oder Links zu anderen sprachbezogenen Online-Nachschlagewerken zu möglichst vielen Wörtern hilfreich in Situationen der Textproduktion und -rezeption sein, wenn sie gewissen Qualitätsansprüchen genügen. Woran auch im Projekt *elexiko* noch gearbeitet werden muss, ist eine bessere Verknüpfung der Angaben untereinander. So ist es beispielsweise in *elexiko* ebenso wenig wie im DWDS möglich, durch Anklicken eines Wortes in einem Textbeleg den entsprechenden Wörterbucheintrag zu öffnen.

3. Schlussgedanken

Mit den gezeigten Beispielen sollten die Chancen und Probleme korpusgestützter Lexikografie besonders in Hinblick auf sprachbezogene Online-Nachschlagewerke zumindest angerissen werden. Dabei sollte auch deutlich geworden sein, dass sich die Nutzer solcher Wörterbücher oder Wortschatzinformationssysteme zwar über eine Fülle verschiedener, aus Korpora gewonnener Angaben freuen können, dafür aber auch bereit sein müssen, unter Umständen von alten Wörterbuchbenutzungsgewohnheiten abzuweichen. Die interpretatorische Eigenleistung der Nutzer wird stärker gefordert.

Auf der Basis sehr umfangreicher elektronischer Textkorpora zu arbeiten und dabei die Möglichkeiten intelligenter Korpustools wie der automatischen Sprachverarbeitung auszuschöpfen, wirkt sich außerdem sowohl auf die Zusammensetzung lexikografischer Arbeitsgruppen aus als auch auf den lexikografischen Prozess. Ohne informatische, computer- und korpuslinguistische Unterstützung könnte keines der vorgestellten Projekte funktionieren. Die Lexikografen selbst müssen zunehmend neben dem klassischen lexikografischen Rüstzeug Verständnis für die technischen Gegebenheiten entwickeln und bereit sein, eine gelungene Mischung zwischen Angabetypen unterschiedlicher Art herzustellen.

Literaturverzeichnis

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim. canoo.net: http://www.canoo.net/ [21.10.2009].
COSMAS II: http://www.ids-mannheim.de/cosmas2/ [21.10.2009].
Deutsches Referenzkorpus DEREKO.
http://www.ids-mannheim.de/kl/projekte/korpora/ [21.10.2009].
Digitales Wörterbuch der deutschen Sprache: http://beta.dwds.de/ [21.10.2009].

elexiko (2003ff.). In: OWID — Online Wortschatz-Informationssystem Deutsch. Hg. vom Institut für Deutsche Sprache, Mannheim. www.owid.de/elexiko_/index.html [21.10.2009].

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

Klosa, Annette (2007): Korpusgestützte Lexikografie: besser, schneller, umfangreicher? In: Kallmeyer, Werner/Zifonun, Gisela (Hg.): Sprachkorpora — Datenmengen und Erkenntnisfortschritt. Berlin/New York (=Jahrbuch des Instituts für Deutsche Sprache 2006), S. 105-122.

Kookkurrenzdatenbank CCDB des IDS: http://corpora.ids-mannheim.de/ccdb/ [21.10.2009].

Landau, Sidney I. (2001): Dictionaries: The art and craft of lexicografy. New York.

Müller-Spitzer, Carolin (2003): Ordnende Betrachtung zu elektronischen Wörterbüchern und lexikografischen Prozessen. In: Lexicografica 19, S. 140-168.

Statistische Kookkurrenzanalyse.
http://www.ids-mannheim.de/kl/projekte/methoden/ka.html [21.10.2009].

Storjohann, Petra (2005): Paradigmatische Relationen. In: Ulrike Haß (Hg.) (2005), S. 249-264.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam/Philadelphia.

URL 1: http://beta.dwds.de/project/ [21.10.2009].

URL 2: http://beta.dwds.de/help/panel/21/ [21.10.2009].

URL 3: http://www.openthesaurus.de/ [21.10.2009].

URL 4: http://beta.dwds.de/help/panel/30/ [[21.10.2009].

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

Wörterbuch für Synonyme und Assoziationen: http://www.openthesaurus.de/ [21.10.2009]

Wortschatz Deutsch: http://wortschatz.uni-leipzig.de/ [21.10.2009].

DWDS-Wörterbuch

Wörter|buch

Aussprache:

Grammatik: das

meist alphabetisch geordnetes Verzeichnis von Wörtern, die nach bestimmten Gesichtspunkten ausgewählt und erklärt sind

Klappe alles auf

Eintrag: **Wörter|buch**

Zusammensetzungen | Eintrag | Zitate | Beispiele

Abb. 3: Eintrag *Wörterbuch* im DWDS-Wörterbuch

OpenThesaurus

synonyme Wortgruppen für: **Wörterbuch**

Lexikon, Verzeichnis, Wörterbuch

Oberbegriff: Kompendium, Nachschlagewerk

Openthesaurus DB Version 2009-07-23

Openthesaurus Webseite

Abb. 4: Einträge im OpenThesaurus des DWDS zum Suchwort *Wörterbuch*

DWDS-Kernkorpus (eingeschränkte Version)

Treffer: 54 1 2 3 4 >>|

1	...t Jahren auch Des Teufels	Wörterbuch	, eine Sammlung von Misan...
2	... der sich laut klinischem	Wörterbuch	darin äußert, "daß man si...
3	...wohl. Der Welt größtes	Wörterbuch	als Brockhochhaus anzupre...
4	.... Eine Toolbox und ein	Wörterbuch	helfen beim Forschen, und...
5	...zu bemühen sie ein ganzes	Wörterbuch	modischer Unternehmensber...
6	...e Eingang finden wird ins	Wörterbuch	des Unmenschen: der "Einw...
7	...ht, und was immer uns das	Wörterbuch	dazu sagt, es bedeutet, d...
8	...gand: - Kleines Thüringer	Wörterbuch	- Reclam-Verlag, Leipzig;...
9	...paß. Nebenbei gibt das	Wörterbuch	Einblick in die vorwiegen...
10	...ment das Kleine Thüringer	Wörterbuch	. Davon abgesehen, daß...

DDC-Query | Darstellungsoptionen | Suchfilter

Abb. 5: KWICs zum Suchwort *Wörterbuch* im DWDS

DWDS-Wortprofil

Wortprofil für **Wörterbuch** als **NN** Frequenz: 1603

Ausgabe Autorenporträt
Gegenwartssprache Partnerverlag
Philosophie Soziologie Sprache Textauszug Unmensch
Version akademisch bestimmen bieten|über digital
einsprachig grimmesch kulturpolitisch philosophisch

Wortart: NN - Zeige Tags Tabellenansicht

Abb. 6: DWDS-Wortprofil zum Suchwort *Wörterbuch*

Orthografie

Normgerechte Schreibung: Wörterbuch
Worttrennung: Wör|ter|buch

Belege (automatisch ausgewählt)

Die fünfte Klasse von Helfrich-Rall hat es inzwischen fast geschafft, auf ein einheitliches Niveau zu kommen, auch in Deutsch. "Jetzt hat die Reform richtigen Wettkampfcharakter bekommen", erzählt die Lehrerin lächelnd. Den Kleinen bereite es eine diebische Freude, sie beim Falschschreiben an der Tafel zu erwischen. Sowohl Helfrich-Rall als auch Vater müssen durchaus noch das **Wörterbuch** bemühen. Denn selbst, "wenn wir uns bei ein paar Dingen wie dem Doppel-S problemlos umgestellt haben", meint Vater, "gibt es doch anderes, was der Gewöhnung bedarf." (M98/MAI.39667 Mannheimer Morgen, 12.05.1998, Ressort: Welt und Wissen; Kaum hat man alles kapiert, beginnt das Umlernen von neuem)

"ich fürchte, daß mir hier meine Ehre genommen werden soll" könne "verwerflich" für den Juristen etwas anderes sein als für den Laien? nein - "verwerflich" bedeute auch und gerade für den Juristen: ruchlos. Grimms **Wörterbuch** nennt als Beispiel den "verwerflichen Richter", der das Recht beugt: "und dieser Verwerflichkeit will man uns zeihen. (H85/FZ1.15914 Die Zeit, 25.01.1985, S. 06; Was heißt hier verwerflich?)

Er ist der Porsche unter den elektronischen Sprachencomputern: der neue Attaché von Hexaglot. Ausgestattet mit Sprachausgabe, SD-Card-Technologie, Lernsystem und Trainingsmodul führt die neueste Entwicklung der Langenscheidt-Tochter mit Sitz in der Sportallee 41 in zwei Sprachen (Deutsch, Englisch) sowie mit einem gastronomischen Spezialwortschatz in fünf Sprachen wortgewandt durch Reisen rund um den Globus. Insgesamt verfügt der Attaché über mehr als 5,1 Mio. Einträge. Mit Hilfe von SD-Cards kann das kleine Allround-Talent jederzeit um diverse **Wörterbücher** und Wortschätze ergänzt werden. Preis: 279,90 Euro. (HMP07/MAR.01453 Hamburger Morgenpost, 13.03.2007, Beilage S. 7; Premiere für das Sprachgenie)

Dieses Stichwort gehört im *elexiko*-Korpus der Frequenzschicht VII (1.001-5.000 mal belegt) an. Es ist in 15 verschiedenen Zeitungen oder Zeitschriften aus 21 Jahrgängen belegt.

Weitere Informationen:
Automatisch ermitteltes Kookkurrenzprofil von **Wörterbuch** in der CCDB
Grammatische Informationen (z.B. Angabe der Wortart, Flexionstabellen) unter canoo.net

Abb. 7: Automatisch generierte Angaben zum Suchwort *Wörterbuch* in *elexiko*