Erschienen in: Kallmeyer, Werner/Zifonun, Gisela (Hrsg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. - Berlin/New York: de Gruyter, 2007. (Jahrbuch des Instituts für Deutsche Sprache 2006), S. 105-122.

ANNETTE KLOSA

Korpusgestützte Lexikographie: besser, schneller, umfangreicher?

Abstract

In diesem Beitrag geht es einerseits um eine Definition dessen, was korpusgestützte Lexikographie ist, und andererseits um eine Bestandsaufnahme der gegenwärtigen Praxis korpusgestützter Lexikographie. Dabei wird ein Schwerpunkt gelegt auf allgemeinsprachige Wörterbücher der Gegenwartssprache, deren Inhalt die Beschreibung von Bedeutung und Verwendung von Lexemen ist. Außerdem liegt die Einschätzung zugrunde, dass die Auswertung elektronischer Korpora die Wörterbucharbeit weitgehend positiv beeinflusst und verändert, vorausgesetzt, dass zugrunde gelegte Korpus wurde für das geplante Wörterbuch so gut wie möglich in Umfang und Zusammensetzung eingerichtet.

John Sinclair, ein erfahrener Lexikograph und Korpuslinguist, hat vor kurzem bekannt, dass er zu Beginn seines COBUILD-Wörterbuch-Projektes im Jahr 1980 angenommen hatte, dass der Einsatz eines Korpus Genauigkeit und Reichhaltigkeit verbessern und den Prozess der Lexikographie [...] beschleunigen würde. Sinclair stellt weiter fest, dass sich manches davon bewahrheitete, er aber vor allem den Effekt der neuen Informationen, die das Korpus lieferte, grob unterschätzt habe:

> At the start of the Cobuild project in 1980 I assumed that the use of a corpus would improve accuracy and comprehensiveness, and it would speed up the process of lexicography because of the clarity of the descriptions and the organising power of the computer. Some of this proved to be correct, but I grossly underestimated the effect of the new information that the corpus supplied, and in particular the total lack of fit between the evidence coming from the corpus and the accepted categories of English lexicography. (Sinclair 2004, S. 9)

Vor diesem Hintergrund ist zu fragen: Ist korpusgestützte Lexikographie wirklich besser, schneller, umfangreicher, in dem Sinn, dass sie schneller zu besseren und umfangreicheren Ergebnissen kommt, wobei „besser" als wissenschaftlich angemessener zu verstehen ist? Bei der Beantwortung dieser Fragen gehe ich von meinen praktischen Erfahrungen als Lexikographin aus und zeige Beispiele aus der Wörterbucharbeit, um zu einer Bestandsaufnahme der gegenwärtigen Praxis korpusgestützter deutschsprachiger Lexikographie zu kommen. Dabei wird auch zu überlegen sein, ob es für die Lexikographie tatsächlich „heute zum Standard [gehört], mit Unterstützung durch große maschinenlesbare Korpora zu arbeiten – allerdings, bedingt durch die verfüg-

baren Korpora, praktisch exklusiv mit schriftlichen Texten", wie es im Ausschreibungstext für die Jahrestagung 2006 des IDS formuliert wurde. Bevor jedoch die Wertungsfragen beantwortet werden können, muss im ersten Teil des Beitrags differenzierter dargestellt und abgegrenzt werden, was hier unter korpusgestützter Lexikographie verstanden wird.

1. Was heißt korpusgestützte Lexikographie?

a) Lexikographie und (Sprach-)Wörterbücher

Mit Wiegand (1998) kann man ein Sprachwörterbuch (im Folgenden immer nur noch: Wörterbuch) als ein „Nachschlagewerk" verstehen, „dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann" (Wiegand 1998, S. 58). Dabei stehen hier Nachschlagewerke für menschliche Benutzer im Zentrum der Betrachtung. Offen bleibt noch die Frage, vor allem im Bereich der korpusgestützten Lexikographie, ob die Daten automatisch oder von Menschen erstellt sind.

Hier sollte man (mit Müller-Spitzer 2003) differenzieren zwischen automatisch erstellten Wortschatzinformationssystemen bzw. lexikographisch bearbeiteten Wörterbüchern. Unter lexikographischer 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, S. 150). Solche lexikographisch bearbeiteten, gedruckten oder elektronischen Wörterbücher stehen hier im Vordergrund. Automatisch erstellte Wortschatzinformationssysteme werden aber auch kurz erwähnt, weil sie erst im Zusammenhang mit dem fortschreitenden Ausbau umfangreicher elektronischer Textsammlungen und der Entwicklung korpustechnologischer und informationstechnologischer Verfahren möglich wurden bzw. werden.

b) Korpus

Von den verschiedenen möglichen und vorgeschlagenen Definitionen für „Korpus" soll hier Folgende gelten: Ein Korpus ist eine Sammlung von Texten, von denen man annimmt, sie sei repräsentativ[1] für eine bestimmte Sprache. Diese Sammlung wurde zum Zweck der linguistischen Analyse zusammengestellt. Daneben gelten außerdem die Annahmen, dass es sich bei den Texten um natürlichsprachliche Äußerungen (häufig in schriftlicher, aber auch in mündlicher) Form handelt und dass die Sammlung nach bestimmten Auswahlkriterien für einen bestimmten Zweck kompiliert wurde. Dabei wird der Anspruch erhoben, dass diese Sammlung authentische Sprache repräsentiert (vgl. Tognini Bonelli 2001, S. 2).

---

[1] „Repräsentativität" wird hier nicht im statistischen Sinne verwendet.

Wichtig ist außerdem, dass heute mit Korpus häufig implizit eine in elektronischer Form zur Verfügung stehende Textsammlung gemeint ist, die mithilfe von elaborierter Recherche- und Analysesoftware erschlossen werden kann. Zurzeit darf man für lexikographische Zwecke von einer solchen Software mindestens folgende Funktionalitäten erwarten (im Einzelnen vgl. hierzu Sinclair 1991, S. 21–37):

- Die Software kann zwischen Lemma und Wortformen unterscheiden; sie verfügt über ein Lemmatisierungsprogramm.
- Die Software erstellt Frequenzlisten zu Lemmata oder Wortformen der unterschiedlichsten Art (z. B. alphabetisch oder nach Häufigkeit sortiert).
- Die Software bietet Fundstellenkonkordanzen in Form von Key-Word-in-Context-Zeilen (= KWICs) an, die je nach Bedarf nach unterschiedlichen Kriterien definiert und sortiert werden können (z. B. alphabetisch nach dem folgenden oder dem vorausgehenden Wort).
- Die Software bietet eine Kookkurrenzanalyse an, die auf der Basis mathematisch-statistischer Verfahren das Aufdecken von signifikanten Regelmäßigkeiten bei der Verwendung von Wortkombinationen ermöglicht, wie sie beispielsweise am IDS entwickelt wurde (vgl. Belica 1995 und http://www.ids-mannheim.de/kt/misc/tutorial.html).

Im Folgenden wird unter Korpus – zumindest im lexikographischen Kontext – ein elektronisches Korpus verstanden. Im lexikographischen Zusammenhang ist außerdem wichtig, dass im Idealfall für ein bestimmtes Wörterbuch gezielt ein eigenes Korpus zusammengestellt wird (vgl. Landau 2001, S. 323).

c) Andere Wörterbuchquellen

Neben einem elektronischen Korpus können auch andere primäre Quellen im Sinne von Wiegand (1998, S. 140) einem Wörterbuch zugrunde liegen, und zwar traditionellerweise Belegzettel. Diese sind nach bestimmten Vorgaben angefertigte Exzerpte aus Texten; diese Texte als Materialbasis für ein Wörterbuch werden wiederum ebenfalls nach bestimmten Kriterien (z. B. Ausgewogenheit hinsichtlich der Textsorten) zusammengestellt. Alle Belegzettel sind in einem Belegzettelarchiv bzw. lexikographischen Zettelarchiv zusammengefasst. Beispiel hierfür ist etwa das Belegarchiv des „Deutschen Fremdwörterbuchs“ von Schulz/Basler, das am IDS neu bearbeitet wird:

> Das DFWB beruht, vor allem um der Darstellung seiner Wort- und Bedeutungsgeschichten ein tragfähiges Fundament zu geben, auf einer breiten Material- und Quellenbasis. Den entsprechenden Grundstock auch für die Neubearbeitung bildet das Schulz/Baslersche Quellen- und Belegmaterial: Die Sammlung, die ca. 2 Mio. Belegzettel umfasst, beruht in erster Linie auf einer systematischen Exzerption von mehr als 10000 gedruckten Quellen, die sich auf den ganzen neuhochdeutschen Zeitraum von etwa 1450 bis 1980/90 und auf ein breit gefächertes Textsortenspektrum erstrecken. (http://www.ids-mannheim.de/lexik/fremdwort/quellen.html)

Ein solches Belegzettelarchiv kann neben der klassischen Zettelform außerdem in digitaler Form fortgeführt werden, wie dies z. B. in der Dudenredaktion geschieht:

> Auf ganz traditionelle Weise ‚durchkämmen' wir Texte außerdem per Hand, besser gesagt mit dem Auge. Diese recht zeitintensive Tätigkeit, an der wir wegen der Qualität der Ergebnisse seit Jahrzehnten festhalten, übernehmen erfahrene Sprachbeobachter für uns. [...] Die auf diese Art vorselektierten Fundstellen – einzelne Sätze oder kürzere Textpassagen – werden in die Duden-Sprachkartei aufgenommen. Sie ist inzwischen auf über 3 Millionen Belege angewachsen und wird seit 1998 elektronisch geführt. (http://www.duden.de)

Außerdem können Zettelbelegarchive auch retrodigitalisiert werden, was aber z. B. weder für die Duden-Sprachkartei noch für andere Belegzettelarchive realisiert wurde[2]. Zwar erhöht eine digitale Form eines Belegarchivs den Nutzwert, indem beispielsweise Volltextsuchen über alle Belegtexte möglich werden oder einzelne Belege verschiedenen Lemmata zugewiesen werden können. Doch ist ein elektronisches Belegarchiv offensichtlich nicht das Gleiche wie ein Korpus im eben definierten Sinn. Belegtexte sind für die linguistische Analyse nur noch bedingt repräsentativ, auch wenn die zugrunde liegende Textauswahl dies war, weil menschliche Exzerptoren eher dazu neigen, das Ungewöhnliche als das Normale auf Belegzetteln festzuhalten.

Bei der Zusammenstellung der primären Quellen für ein Wörterbuch ist das Problem der geeigneten Auswahl sowohl für Belegarchive wie für Korpora zu berücksichtigen; beide sollen „ein zuverlässiges Abbild des Lexembestands und Lexemgebrauchs der Objektsprache bilden" (Schlaefer 2002, S. 104) – und das ist nicht immer leicht zu erreichen. In der deutschsprachigen Lexikographie werden zurzeit verschiedene Varianten der Korpuszusammenstellung speziell für lexikographische Zwecke neben der immer noch andauernden Pflege existierender Belegarchive praktiziert:

- Die Anlage eines eigenen lexikographischen Korpus für genau ein Wörterbuch, z. B. für das „Mittelhochdeutsche Wörterbuch", das in Trier und Göttingen neu erarbeitet wird:

  > Den Kern seiner Quellenbasis bildet ein Corpus von philologisch gesicherten Texten aller Textsorten der Periode. Auf der Grundlage dieses Quellenkorpus ist ein digitales Textarchiv angelegt und aus diesem durch computergestützte Lemmatisierung ein digitales Belegarchiv erstellt worden, das bei der Ausarbeitung des Wörterbuches zugrunde gelegt und durch Nachexzerption ständig ergänzt wird. (http://www.adwmainz.de/2005/index.php?sektion=vorh_gsk&ID=38)

---

[2] Auch nicht für das „Deutsche Rechtswörterbuch", wie Schlaefer (2002, S. 107) schreibt; hierzu liegt eine anders lautende E-Mail-Auskunft von Ingrid Lemberg vom 15. 02. 2005 vor.

- Die Anlage eines lexikographischen Korpus als primäre Quelle für verschiedene lexikographische Produkte aus einer Bearbeitungsstelle, z. B. für die Wörterbücher aus der Dudenredaktion:

  Wir bauen in der Dudenredaktion gerade ein eigenes elektronisches Korpus auf, das genau auf unsere Zwecke zugeschnitten sein wird: das ‚Duden-Korpus'. Es wird zunächst ca. 500 Millionen Wortformen umfassen und aus unterschiedlichsten Textsorten (Romanen, Zeitungsartikeln, Gebrauchsanleitungen etc.) zusammengesetzt sein. Darüber hinaus ist es ‚annotiert', was bedeutet, dass jede der 500 Millionen Wortformen mit besonderen sprachlichen Informationen angereichert ist. (http://www.duden.de)

- Die Zusammenstellung eines virtuellen lexikographischen Korpus aus einem größeren, allgemeinen Korpus, z. B. für das Projekt *elexiko* am IDS in Mannheim:

  Um für die Auswertung sprachlicher Daten eine gute empirische Basis zugrunde legen zu können, wurde nach formalen und inhaltlichen Kriterien aus den gesamten IDS-Korpora ein umfangreiches digitales Textkorpus zusammengestellt. Dieses Korpus ist ein dynamisches Korpus (ein sogenanntes Monitorkorpus), welches regelmäßig erweitert und aktualisiert wird, um die jeweils neuesten Entwicklungen verfolgen und damit aktuelle Beschreibungen liefern zu können. Derzeit (Dezember 2005) umfasst es ca. 1,3 Milliarden Textwörter. (http://www.elexiko.de/Korpusbasiertheit.html)

Neben Korpora und Belegarchiven als primäre Quellen greifen Lexikographen bei ihrer Arbeit auch auf elektronische Textsammlungen im weiteren Sinn zu. Dies können z. B. Gesamtausgaben bestimmter Autoren auf CD-ROM, digitale Zeitungsarchive auf CD-ROM und nicht zuletzt auch Texte im Internet sein, die mithilfe einer Suchmaschine erschlossen werden. Solch ein Verfahren wird etwa im Folgenden beschrieben:

> Zusätzlich zur Recherche im Duden-Korpus suchen wir punktuell auch in anderen elektronischen Quellen nach neuen oder bislang noch nicht verzeichneten Wörtern. Allen voran ist hier natürlich das Internet zu nennen, aber auch diverse Wirtschaftsdatenbanken oder Korpora anderer Institute. (http://www.duden.de)

Bei solchen Textsammlungen und insbesondere dem Internet handelt es sich aber nicht um Korpora, denn sie genügen wichtigen Kriterien nicht: Sie wurden nicht zum Zweck linguistischer Analyse nach bestimmten Kriterien zusammengestellt und sie sind nicht mithilfe von elaborierter Recherche- und Analysesoftware auswertbar. Wie Belica/Perkuhn (2006, S. 4) erklären, haben Internetsuchanfragen auch eine andere und „ganz spezielle Funktion": Sie dienen nämlich ausschließlich dem Aufspüren von Dokumenten im Internet. Allerdings kann man natürlich aus solchen elektronischen Textsammlungen, z. B. dem Internet, Korpora zusammenstellen (vgl. hierzu z. B. Thelwall 2005). Werden solche elektronischen Textsammlungen bei der lexikographischen Arbeit benutzt, um beispielsweise gezielt nach Belegen für neue Wörter zu suchen, dann stellen sie nur eine Form von primärer Quelle dar mit dem Nachteil, nicht unbedingt in die ursprüngliche Quellenkonzeption für das Wörterbuch eingebunden zu sein.

Überhaupt ist zumindest für die deutsche allgemeinsprachige Lexikographie der Gegenwart festzustellen, dass (noch) überwiegend eine Kombination verschiedener primärer Quellen als Wörterbuchbasis dient, wie etwa das Wörterbuch „Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen" zeigt: Primärquellen dieses Wörterbuches sind einerseits das virtuelle, so genannte neo-Korpus, das aus den IDS-Korpora der geschriebenen Sprache zusammengestellt wurde, und andererseits eine projekteigene Wortkartei mit durch Exzerption gewonnenen Belegen „aus den verschiedensten gedruckten Texten, [daneben] Hörbelege aus Fernseh- und Rundfunksendungen des Erfassungszeitraumes sowie Internetbelege" (Herberg/Kinne/Steffens 2004, S. XVI).

Das Projekt *elexiko,* das ebenfalls am IDS beheimatet ist, verfolgt dagegen den Ansatz, ein Wörterbuch auf der Basis eines eigenen Korpus vollständig neu zu erarbeiten – ein Ansatz, der in der englischsprachigen Lexikographie schon 1987 mit dem „Collins COBUILD English Language Dictionary" von John Sinclair und seinem Team realisiert wurde. Dieses Wörterbuch wird deshalb wohl zu Recht als „ein erstes beeindruckendes Ergebnis strikt korpusbasierter Lexikographie" (Engelberg/Lemnitzer 2001, S. 206) bezeichnet. Für die englischsprachige Lexikographie ist mit Sicherheit Hunston (2002, S. 96) zuzustimmen, dass Korpora das Schreiben von Wörterbüchern in solchem Ausmaß revolutioniert haben, dass es großen Verlagen heutzutage praktisch unmöglich ist, beispielsweise ein Lernerwörterbuch zu veröffentlichen, das nicht beanspruchen kann, auf einem Korpus zu basieren:

> Corpora have so revolutionised the writing of dictionaries and grammar books for language learners (or rather, for learners of English) that it is by now virtually unheard-of for a large publishing company to produce a learner's dictionary or grammar reference book that does not claim to be based on a corpus. As a result, this is probably the application of corpora that is most far-reaching and influential, in that even people who have never heard of a corpus are using the product of corpus investigation. (Hunston 2002, S. 96)

d) korpusgestützt – korpusgebunden

Was heißt korpusgebunden? Und was ist der Unterschied zu korpusgestützter Lexikographie? 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 nach Wiegand (1998, S. 140) „alle Wörterbücher, die nach dem Instruktionsbuch entweder obligatorisch oder fakultativ [als Hilfsmittel bei der lexikographischen Arbeit] konsultiert werden sollen, und zu den tertiären Quellen gehören alle sonstigen Sprachmaterialien, die benutzt werden sollen, wie z. B. linguistische Monographien und Grammatiken [...]." Ein korpusgebundenes Wörterbuch bildet genau die Sprachwirklichkeit ab, die das zugrunde gelegte Korpus repräsentiert – aber nicht mehr und auch nicht weniger. Dies wäre z. B. bei einem Autorenwörterbuch oder einem Wörterbuch zu einem bestimmten Text der Fall.

In der korpusgestützten Lexikographie, die sich auch ausschließlich auf ein Wörterbuchkorpus als primäre Quelle stützt, ist es dagegen möglich, neben der Auswertung des Wörterbuchkorpus noch sekundäre und/oder tertiäre Quellen hinzuzuziehen. Dieses Vorgehen wird beispielsweise an der Beschreibung der Praxis der Lesartendisambiguierung im „Cambridge International Dictionary of English" (kurz: CIDE) deutlich:

> Lexicographers considered how to split up a polysemic word among a number of guidewords, using their own understanding and a study of other dictionaries, and testing this understanding against the picture emerging from the data. Through using the corpus, CIDE [= Cambridge International Dictionary of English] lexicographers often found that previous dictionaries defined quite rare senses of words but missed important, common ones. (Baugh/Harley/Jellis 1996, S. 41)

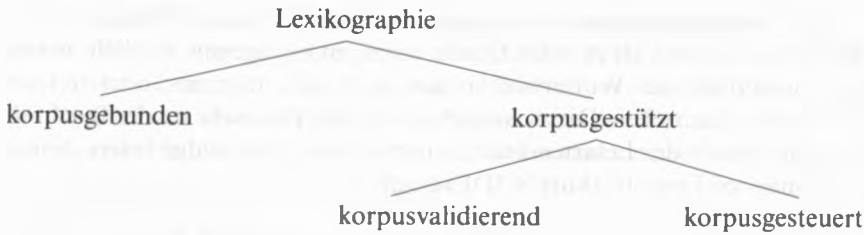
Ein korpusgestütztes Wörterbuch bietet also keine 1 : 1-Abbildung der Sprachwirklichkeit des zugrunde gelegten Korpus, sondern ergänzt, wo nötig, das Bild, das aus der Korpusanalyse gewonnen wurde. Mit korpusgestützter Lexikographie soll aber noch mehr gemeint sein: Hier genügt es nicht, dass ein Korpus (im oben definierten Sinne einer speziell zusammengestellten Sammlung elektronisch aufbereiteter Texte) als primäre Quelle zugrunde liegt, sondern dieses muss mithilfe korpuslinguistischer Verfahren erschlossen werden. Insofern ist etwa das oben erwähnte „Mittelhochdeutsche Wörterbuch" kein Vertreter korpusgestützter Lexikographie, denn dort wird aus dem Korpus unter Einsatz eines Lemmatisierungstools ein elektronisches Belegarchiv erstellt, das wiederum die Grundlage für die Artikelbearbeitung ist. Andere korpuslinguistische Verfahren wie z. B. Kontextanalysen kommen aber (noch) nicht zum Einsatz. Der Unterschied zwischen korpusgestützt erarbeiteten Wörterbüchern und nicht korpusgestützt erarbeiteten Wörterbüchern ist also hauptsächlich ein methodischer.

Zurzeit wird in der mit Korpora arbeitenden Lexikographie in verschiedenen Ländern allgemein dem korpusgestützten Vorgehen gegenüber dem korpusgebundenen Vorgehen der Vorzug gegeben. Als Beispiel hierfür können folgende Zitate, die sich auf das „Longman Dictionary of Contemporary English" in der dritten Auflage von 1995 beziehen, gelten:

> The corpus is a massively powerful ressource to aid the lexicographer, which must be used judiciously. Our aim at Longman is to be corpus-based, rather than corpus-bound. (Summers 1996, S. 262)
>
> [...] in dictionary making editorial judgment is of paramount importance, because blindly following the corpus, no matter how carefully it may be constructed to represent the target language type accurately, can lead to oddities. (Summers 1996, S. 266)

e) korpusvalidierend – korpusgesteuert

Innerhalb der korpusgestützten Lexikographie selbst können wiederum zwei unterschiedliche Ansätze verfolgt werden: der korpusvalidierende (engl. *corpus based*) Ansatz und der korpusgesteuerte (engl. *corpus driven*) Ansatz.



Unter *korpusvalidierend* versteht man eine Methode, die von einem Korpus Gebrauch macht, um Theorien und Beschreibungen darzulegen, zu testen oder zu veranschaulichen:

> [...] the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study. (Tognini Bonelli 2001, S. 65)

Im Gegensatz dazu wird beim korpusgesteuerten Ansatz ein Korpus ganz anders als für die Suche nach Beispielen, die bestimmte linguistische Argumente unterstützen oder theoretische Annahmen validieren sollen, genutzt[3]:

> In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back preexisting theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. (Tognini Bonelli 2001, S. 84)

An Beispielen aus der Lexikographie sollen die unterschiedlichen Ansätze im Folgenden deutlich werden. Als Vertreter eines korpusvalidierenden Ansatzes kann das schon zitierte „Cambridge International Dictionary of English" gelten. Im oben genannten Zitat von Baugh/Harley/Jellis (1996, S. 41) wird das Vorgehen deutlich: Etwas überspitzt formuliert geht der Erarbeitungsweg von Introspektion über Konsultation von Sekundärquellen schließlich hin zum Vergleich mit der Korpusauswertung. Oder, in einer Formulierung zu dem Wahrig-Wörterbuch „Fehlerfreies und gutes Deutsch" von 2003:

> Empirisch abgesichert sind alle Schreibungen und Wortverwendungen durch einen Abgleich [Unterstreichung hinzugefügt] mit dem WAHRIG Textkorpusdigital, einer rd. 500 Mio. Wörter umfassenden, digitalen Dokumentation der deutschen Gegenwartssprache, die den aktuellen Sprachgebrauch authentisch widerspiegelt. (Wahrig 2003, S. VII)

Als Beispiel für den überwiegenden Einsatz des korpusgesteuerten Verfahrens in der Lexikographie kann das ebenfalls schon genannte „Collins COBUILD

---

[3] In Steyer (2004, S. 93) werden für die beiden genannten empirischen Prinzipien die Termini Konsultationsparadigma (für den korpusvalidierenden Ansatz) und Analyseparadigma (für den korpusgesteuerten Ansatz) vorgeschlagen.

English Language Dictionary" gelten. Das Lexikographenteam schätzt seine Arbeit als Antwort auf die Frage „Was zeichnet wahre Korpuslexikographie aus?" folgendermaßen ein:

> What characterizes true corpus lexicography? Each entry in a dictionary represents a detailed examination of the corpus evidence. [...] The entry reflects what is found – and what our users are likely to find in the real text world – rather than what was believed to be the case. (Clear u. a. 1996, S. 308f.)

An einem Beispiel aus der lexikographischen Arbeit im Projekt *elexiko* soll deutlich werden, wie korpusvalidierender und korpusgesteuerter Ansatz in korpusgestützter Lexikographie sinnvoll verbunden werden können. Durch den Einsatz der statistischen Kookkurrenzanalyse[4] werden im Projekt korpusgesteuert erste Kandidatenwörter für paradigmatische Relationen gewonnen, ohne dass sich die Lexikographen auf die eigene Intuition stützen. Wichtig ist, dass dies nicht bedeutet, „dass man auf eine lexikografische Auswertung der dort extrahierten Ergebnisse verzichten darf, denn Sinnrelationen müssen identifiziert, bestätigt, zugeordnet und illustriert werden" (Storjohann 2005a, S. 254).

Mithilfe des korpusgesteuerten Ansatzes kommt es nämlich nicht immer zu einer umfassenden Beschreibung sinnverwandter Wörter, sodass in einem zweiten Schritt sekundäre Quellen, z. B. Synonymwörterbücher, konsultiert werden, in denen weitere Kandidatenwörter zu finden sind. Diese werden dann gezielt im Korpus gesucht, das damit nicht Ausgangspunkt der lexikographischen Beschreibung ist, sondern der Rückprüfung dient. Die mithilfe dieses korpusvalidierenden Ansatzes gefundenen paradigmatischen Partnerwörter müssen allerdings den gleichen Aufnahmekriterien für den Wortartikel genügen (Frequenz, Streuung über mehrere Quellen und mehrere Jahrgänge) wie die mithilfe des korpusgesteuerten Ansatzes gewonnenen Partner.

Kurz zusammengefasst kann man sagen: Korpusgestützte Lexikographie erarbeitet Wörterbücher auf der Grundlage elektronischer, gezielt zusammengestellter Textsammlungen, die (bezogen auf die Gegenwartssprache) authentische Sprache repräsentieren sollen und die mithilfe geeigneter Recherche- und Analysesoftware erschlossen und ausgewertet werden. Alle Daten werden redaktionell auch unter Einbezug sekundärer und/oder tertiärer Quellen geprüft und bewertet, wobei verschiedene Wörterbücher entweder vom Korpus selbst ausgehen (korpusgesteuert) oder das Korpus zur Rückprüfung benutzen (korpusvalidierend) oder beide Methoden kombinieren. Die meisten der bisher gezeigten Beispiele waren gegenwartssprachliche Wörterbücher, deren Gegenstand zumeist Bedeutungs- und Verwendungsbeschreibung ist. Korpusgestützte Lexikographie hat sich genau in diesem Bereich, besonders aber auch für Lernerwörterbücher entwickelt. Eine Aus-

---

[4] Am IDS durch Cyril Belica entwickelt und seit 1995 als Teil der Korpusrecherche- und Analysewerkzeuge der COSMAS-Plattform angeboten.

weitung der Methoden z. B. auf fachsprachliche oder sondersprachliche Wörterbücher, auf dialektale oder regionale Wörterbücher, oder auf Antonymwörterbücher usw. steht weitgehend noch aus.

2. Wie wirkt sich Korpusgestütztheit auf Wörterbücher aus? – Zwei Beispiele

In ihrer Darstellung zum „Cambridge International Dictionary of English" beschreiben Baugh/Harley/Jellis (1996) detailliert, wie sie Korpusdaten sowie Korpusrecherche- und -analysemethoden eingesetzt haben, um dieses Wörterbuch zu erarbeiten. Sie gehen dabei auf die Entscheidung über die Aufnahme von Stichwörtern, die Lesartendisambiguierung, die Arbeit mit Belegen, die Angabe von Kollokationen, die Erarbeitung von grammatischen und morphologischen Angaben und weiteres mehr ein. Von diesen Punkten wird hier nur die besonders relevante Frage des Stichwortansatzes und der Stichwortliste sowie die Angabe von Wortverbindungen bzw. der Einbezug des Kontextes berücksichtigt. Dabei soll auch überlegt werden, ob die Benutzung von Korpora diese Arbeitsvorgänge tatsächlich schneller und die Ergebnisse besser und umfangreicher macht.

a) Die Frage des Stichwortansatzes und der Stichwortliste

Korpusauswertung spielt bei der Frage des Stichwort*ansatzes* eine Rolle und ist außerdem bei der Frage der Lemma*liste* in zweierlei Hinsicht relevant: Für schon bestehende Wörterbücher kann für Überarbeitungen und Neuauflagen eine systematische Korpusauswertung dabei helfen, Kandidaten für neu aufzunehmende Stichwörter zu ermitteln und so genannte Wörterbuchleichen aufzuspüren. Für neu zu erstellende Wörterbücher können vollständig neue Lemmalisten durch systematische Korpusauswertungen erarbeitet werden.

Bevor überhaupt eine Stichwortliste erarbeitet werden kann, ist aber zu klären, in welcher Form die im Wörterbuch zu beschreibenden Wörter in diese Liste aufgenommen werden sollen. Traditionellerweise geschieht dies in der so genannten Nennform, Grundform oder Zitierform. In Textkorpora kommen solche Grundformen allerdings in den seltensten Fällen vor, sondern in Texten liegen flektierbare Wörter natürlich in verschiedenen Wortformen vor, die entweder intellektuell oder unter Einsatz eines Lemmatisierungsprogrammes auf eine Wortform zusammengeführt werden können. John Sinclair nennt hierzu die Schwierigkeiten:

> Lemmatization looks fairly straightforward, but is actually a matter of subjective judgement by the researcher. There are thousands of decisions to be taken. Also, it is not yet understood how meanings are distributed among forms of a lemma, and a new branch of study is looming – the interrelationships of a lemma and its forms. (Sinclair 1991, S. 41)

Sinclair überlegt weiter, ob es tatsächlich so sein muss, dass die Grundform eines Wortes als Stichwort im Wörterbuch dient, selbst wenn diese Grund-

form in einem Korpus fast nicht belegt ist. Könnte nicht stattdessen die am meisten belegte Form das Lemma werden?:

> Another interesting question in this area is how to decide what the physical form of a lemma should be. Traditionally, the 'base', or uninflected, form is used even when that form is hardly ever found on its own, or hardly ever found at all. But a case could be made for any of a number of alternatives, for example, that the most frequently-encountered form should be used for the lemma; and the first-stage evidence from the computer can provide a good basis for planning new methods of access to the word-forms for the language. (Sinclair 1991, S. 42)

Aber selbst das COBUILD-Wörterbuch, oben schon als das erste tatsächlich korpusgestützte Wörterbuch bezeichnet, geht diesen Weg nicht, sondern bleibt bei der traditionellen Form des Stichwortansatzes nach der Grundform. Dies hängt zum einen sicherlich mit der Frage der Benutzerfreundlichkeit zusammen, aber auch mit den allgemein üblichen Schritten, wie Stichwortlisten erstellt werden.

Bei der vollständig neuen Erarbeitung einer Stichwortliste für ein Wörterbuchprojekt kann man entweder „bereits vorhandene Stichwortlisten auswählen, auswerten und schließlich mehr oder minder modifiziert übernehmen" oder „man kann sie auf der Basis von Korpora komplett neu erstellen" (Schnörch 2005, S. 73). Für das Projekt *elexiko,* das hier als Beispiel für korpusgestützte Lexikographie in diesem Bereich dienen soll, wurde der zweite Weg gewählt, dabei aber zusätzlich der Abgleich mit anderen Stichwortlisten als Ergänzung und Korrektiv eingesetzt. So ist zunächst und in kurzer Zeit eine Liste an Stichwortkandidaten entstanden, die mit ca. 320.000 Einträgen sehr umfangreich war[5]. Die redaktionelle Bearbeitung dieser Stichwortkandidatenliste hin zu einer wirklichen Stichwortliste dauerte dagegen viel länger; in Summe waren damit drei Lexikographen jeweils mehrere Monate beschäftigt, und noch immer überprüfen studentische Hilfskräfte einzelne unklare Stichwortkandidaten.

Warum ist aber eine redaktionelle Überprüfung der Stichwortkandidatenliste überhaupt nötig? Zunächst einmal entspricht solch ein redaktioneller Überarbeitungsgang grundsätzlich dem korpusgesteuerten Verfahren, bei dem ja der Lexikograph das Bild, das sich ihm aus dem Korpus darstellt, betrachtet, analysiert, bewertet, sortiert und schließlich beschreibt. So wurde in *elexiko* also jeder mithilfe der Korpustechnologie vorgeschlagene Stichwortkandidat angesehen, nach bestimmten Kriterien (wie „richtig", „unklar", „falsch") bewertet, in verschiedene Teilmengen einsortiert und dann in Form der öffentlich zugänglichen Stichwortliste[6] von *elexiko* dokumentiert. Doch

---

[5] Natürlich muss die korpusgestützte Erstellung einer Stichwortliste nicht zu solch einem Umfang führen, denn auch aus einem viele Millionen Wortformen enthaltenden Wörterbuchkorpus kann prinzipiell eine deutlich kleinere Stichwortliste ermittelt werden, wenn man etwa andere Frequenzen als Kriterien vorgibt.

[6] Im Internet zu erreichen unter http://www.elexiko.de.

ist dies nicht nur wegen der angewandten Methode nötig gewesen, sondern auch weil die Zusammenführung verschiedener Wortformen auf eine Grund-, Nenn- bzw. Zitierform mithilfe eines automatischen Lemmatisierers natürlich nicht immer fehlerfrei funktioniert und teilweise nach lexikographischen Gesichtspunkten revidiert werden muss[7].

Als Fazit ist festzuhalten: Es geht korpusgestützt schnell, die Stichwortkandidaten ermitteln zu lassen, und man kann sehr umfangreiche Kandidatenlisten erzeugen, ob aber die lexikographische Überprüfung dieser Kandidaten und die Erarbeitung der wirklichen Stichwortliste schneller geht als etwa das Kompilieren aus schon existierenden Wörterbüchern, ist schwer zu beurteilen. Außerdem ist zu berücksichtigen, dass zumindest dann, wenn einem Wörterbuch ein dynamisches Korpus zugrunde liegt, die Stichwortliste in regelmäßigen Abständen aktualisiert werden muss. Ob Stichwortlisten besser werden, wenn sie korpusgestützt erarbeitet sind, ist ebenfalls differenziert zu beantworten. Sie bilden zunächst einmal direkt das Korpus ab, und wenn dieses für den Wörterbuchzweck gut zusammengestellt wurde, dann ist auch die Stichwortliste gut. Wenn aber das zugrunde gelegte Korpus nicht wirklich für den Wörterbuchzweck geeignet ist, muss die Stichwortliste u. U. redaktionell ergänzt werden, damit sie gut wird. Werden Stichwortlisten aus anderen Listen kompiliert, stellt sich natürlich die Frage der Auswahlkriterien und der Vollständigkeit – eine gute und aktuelle Stichwortliste zu erhalten, ist mit dieser Methode keineswegs leichter.

Die Ergänzung und Überprüfung einer schon erarbeiteten Stichwortliste ist der dritte Punkt, der hier betrachtet werden soll. Die Unterstützung für den Lexikographen durch Korpusauswertung ist hierbei wirklich enorm:

> Sprache bleibt bekanntlich nicht stehen; neue Wörter kommen hinzu, alte erhalten neue Bedeutungen. Was früher in langer, ebenso akribischer wie mühsamer Handarbeit einzeln zusammengesucht werden musste, kann nun schnell und systematisch auf der Grundlage eines großen Korpus der deutschen Gegenwartssprache ermittelt werden. (http//www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html)

Tatsächlich bedeutet der Einsatz von Korpora in diesem Bereich eine Verbesserung hinsichtlich der Aktualität. Dann kann gelten, was für das COBUILD-Projekt schon 1987 formuliert wurde: „The main criterion for inclusion or exclusion of headwords and senses was the strength of the corpus evidence“ (Sinclair 1987, S. 65).

Auf der anderen Seite bedeutet es redaktionellen Aufwand, die aus dem Korpus generierten Vorschläge für Neuaufnahme- oder Streichkandidaten zu überprüfen. Aus den etwa 5.000 „interessanten neuen Wörtern“, die aus dem zugrunde gelegten Korpus für das Wahrig-Wörterbuch der „Deutschen Rechtschreibung“ als Kandidaten ermittelt wurden, sind beispielsweise dann doch nur etwa 1.100 tatsächlich in die Neuauflage des genannten Wörter-

---

[7] Zu einzelnen Beispielen für Fehllemmatisierungen vgl. Schnörch (2005, S. 77f.).

buches eingegangen[8] – und solch ein Arbeitsgang kostet (wenn auch lohnende) Zeit. Denn nicht immer sind Wörter, die als sehr frequent auffallen, auch wirklich für ein Wörterbuch interessant, und nicht immer muss ein niedrig frequentes Wort tatsächlich aus dem Wörterbuch gestrichen werden. Im Einzelfall hängt dies sehr vom einzelnen Wörterbuchgegenstand und -ziel ab. Kriterien für den Ausschluss eines Stichwortes trotz hoher Frequenz können z. B. die fehlende Streuung über verschiedene Quellen oder über längere Zeiträume sein, aber auch das jeweilige Register, die Varietät oder die Periode, denen das Wort zuzuordnen ist (vgl. hierzu Baugh/Harley/Jellis 1996, S. 49).

Als Fazit für die Überarbeitung bestehender Stichwortlisten ist zu ziehen: Systematische Korpusauswertung vereinfacht und beschleunigt die Ermittlung von Aufnahme- oder Streichkandidaten. Neben der Frequenz können weitere Korpusdaten (z. B. Zahl der Quellen oder Jahrgänge, in denen das Stichwort belegt ist) dem Lexikographen bei der Entscheidung über Aufnahme oder Streichung helfen – ihm die Entscheidung abnehmen können sie aber nicht. Trotzdem ist dieses korpusgesteuerte Verfahren besser als eine rein redaktionelle Belegsuche nach zufällig aus anderen Quellen, z. B. bei der eigenen Lektüre einer Tageszeitung, entdeckten Aufnahmekandidaten. Durch den Einsatz des korpusgesteuerten Verfahrens nimmt der Umfang an korrigierenden Eingriffen in eine Stichwortliste eher zu, und zumindest in der Verlagslexikographie wird mit jeder Neuaufnahme auch geworben:

> Mit rund 5000 neu aufgenommenen Wörtern, wie beispielsweise Billigflieger, Dosenpfand, Fotohandy, Genmais, Ich-AG, LAN-Party, Minijob und Sars, ist das Wörterverzeichnis auf den aktuellen Stand gebracht. (Duden 2004, S. 5)

Offensichtlich ist die Zeit gekommen, in der für viele Wörterbücher dem korpusgesteuerten Verfahren bei der Überarbeitung und Aktualisierung der Stichwortliste der Vorzug gegeben wird.

b) Wortverbindungen und Kontext

Korpusgestützte Lexikographie ist in besonderer Weise durch den Einbezug des Kontextes und die Berücksichtigung von Wortverbindungen gekennzeichnet, wie bereits Teubert (1999, S. 312) beobachtet:

> Längst hat mit der Verfügbarkeit größerer Korpora und fortgeschrittener Korpussoftware die Korpuslinguistik auch hierzulande Eingang in die allgemeinsprachige Lexikographie gefunden. Man wendet sich größeren Bedeutungseinheiten zu und bezieht vermehrt den Kontext in Bedeutungsbeschreibungen ein. Mittlerweile setzt sich die Überzeugung durch, dass die nächste Wörterbuchgeneration, die einsprachige ebenso wie die zweisprachige, zumindest korpusvalidiert, wenn nicht korpusbasiert sein muss. (Teubert 1999, S. 312)

In der dritten Auflage des „Longman Dictionary of Contemporary English" von 1995 tauchen erstmals so genannte „lexical units" mit eigenen Bedeu-

---

8 Vgl. http://www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html.

tungspunkten unter einem Einwort-Stichwort auf, wofür Summers (1996, S. 263) den Wörterbucheintrag *to be liable* als Beispiel bietet:

> **1 be liable to do something** to be likely to do or say something or to behave in a particular way, especially because of a fault or natural tendency: *The car is liable to overheat on long trips.* **2** [not before the noun] legally responsible for the cost of something: [+**for**] *Tenants have legal liability for any damage they cause.* **3** likely to be affected by a particular kind of problem, illness etc: [+**to**] *You're more liable to injury when you don't get regular exercise.* [...] (Summers 1996, S. 263)

Summers erklärt, dass „Phraseologie" ein wichtiges Merkmal für die Neuauflage des Wörterbuches werden sollte, weil fortgeschrittene Lerner des Englischen als Zielgruppe galten. Allerdings hatte die Redaktion nicht damit gerechnet, wie weitreichend diese Entscheidung sich erweisen würde – für das Aussehen des Wörterbuches, für den Aufbau der Einträge und für die Form der Bedeutungserläuterungen selbst (vgl. Summers 1996, S. 262). Zumindest wird diese Entscheidung der Tatsache gerecht, dass Wörter in Texten eben nicht isoliert, sondern mit anderen Wörtern zusammen auftreten.

Häufig treten aus Korpora gerade mehr oder weniger feste Wortverbindungen hervor. Für Lexikographen stellt sich natürlich die Frage, wie mit ihnen umzugehen ist. Wie im Falle des „Longman Dictionary" können sie in Einzelwortartikel integriert werden. Dieses Vorgehen entspricht dabei durchaus dem korpusgesteuerten Verfahren, weil ja durch die Korpusanalyse die Wortverbindung zunächst als solche hervortritt und erst dann der Lexikograph die Entscheidung trifft, ob und wie sie im Wörterbuch zu behandeln ist. Dieses Vorgehen bleibt aber natürlich ein lexikographisches. Anders ist der korpuslinguistische Ansatz, der nämlich solche Wortverbindungen gar nicht mehr im Rahmen eines am einzelnen Wort interessierten Wörterbuches beschreiben will, sondern ganz herauslöst aus diesem Kontext, wie dies am IDS etwa im Projekt „Usuelle Wortverbindungen" geschieht:

> Im Projekt UWV werden rekurrente Muster des Sprachgebrauchs, so wie sie sich in Kookkurrenzprofilen vor allem im hochfrequenten Bereich manifestieren, aus den IDS-Korpora herausgefiltert und nach linguistischen Kriterien systematisiert und beschrieben. Dazu werden die am IDS entwickelten mathematisch-statistischen Analyse- und Clusteringverfahren, speziell die statistische Kookkurrenzanalyse, systematisch angewendet und einer ständigen linguistischen Reflexion unterzogen. Langfristiges Ziel ist die Visualisierung solcher Gebrauchsmuster in online abrufbaren Wortverbindungsnetzen für unterschiedliche Nutzungssituationen. (http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen/)

Wolfgang Teubert stellte fest, dass der Kontext in korpusgestützter Lexikographie bei den Bedeutungsbeschreibungen eine größere Rolle als zuvor spielt. Dies bestätigen die Autoren des COBUILD-Wörterbuches:

> The corpus also, more indirectly, influenced the definition style which we developed and which has been widely imitated. We were continually confronted by the phraseology and lexicogrammatical patterns associated with individual words and meanings, and we realized that the clearest way to communicate these was to build them into the definition

text, and to define an item **in its context**, rather than as disembodied entities. (Clear et al. 1996, S. 309)

Im Projekt *elexiko* werden Definitionen auch in ganzen Sätzen formuliert, unter anderem deshalb, weil dies erlaubt, „die am Gebrauch des Wortes beteiligten typischen Beziehungspartner [... in die semantische Paraphrase] aufzunehmen" (Storjohann 2005b, S. 190). Ergänzend dazu sind zwei weitere Angabetypen in der *elexiko*-Artikelstruktur zu sehen: die typischen Verwendungsmuster und die so genannte Mitspielerangabe. Beide Angaben sind innerhalb der Artikelstruktur im Bereich „Bedeutung und Verwendung" verankert, und gemeinsam mit der Beschreibung von Verwendungsspezifika bilden sie zusammen mit der Bedeutungsparaphrase den Korpusbefund zu diesem Bereich ab:

> Die Angaben zu ‚Semantischer Umgebung und Mitspielern' [...] sollen einen bestimmten Typ kognitiver Assoziationen zwischen dem Stichwort in einer Lesart und Wörtern in dessen Textumgebung darstellen. Dabei ist der empirische Zugang zur Textumgebung qua Korpus entscheidend. Es hat sich gezeigt, dass die Listen der signifikanten Kookkurrenzpartner [...] Lexeme enthalten, die sich als Realisierung semantischer Rollen auffassen lassen, und zwar unabhängig von der Wortart. (Haß 2005, S. 228)

Mit diesem Angabetyp liegt also ein Resultat korpusgestützter Lexikographie vor: Mithilfe intelligenter Korpusrecherche- und -analysesoftware ermittelte Informationen werden nicht etwa verworfen oder möglicherweise zwanghaft in übliche Wörterbuchstrukturen integriert, sondern bekommen einen neuen Platz im Wortartikel zugewiesen. Natürlich unterliegt das Projekt *elexiko* wegen der Publikation im Internet dabei keinen Platzbeschränkungen und kann vor diesem Hintergrund leicht einen neuen, Raum beanspruchenden Angabetyp anbieten.

Ist korpusgestützte Lexikographie, was diesen Bereich betrifft, besser, schneller, umfangreicher? Gerade hier führt das Prinzip der Korpusgestütztheit zu einem deutlichen Qualitätsgewinn von allgemeinsprachigen Bedeutungswörterbüchern, besonders aber auch Lernerwörterbüchern. Bedeutungsparaphrasen sind so enger und aktueller am tatsächlichen Sprachgebrauch orientiert, sie können mit frequenten, aktuellen Wortverbindungen ergänzt werden; Wortverbindungen können in sinnvoller Weise in Wortartikel integriert werden, wenn sie nicht alleiniger Gegenstand von Untersuchungen und Darstellungen werden. Dabei sind die Lexikographen nicht mehr auf Introspektion angewiesen, sondern Usuelles, Zentrales wird aus dem Wörterbuchkorpus ermittelt und beschrieben. Dass eine Ausweitung von Angabetypen (wie am Beispiel von *elexiko* gezeigt) natürlich gleichzeitig eine längere Bearbeitungszeit bedeutet, ist zu berücksichtigen. Auf der anderen Seite ermöglicht intelligente Korpussoftware, dass Wörterbuchbeschreibungen den Sprachgebrauch in einer bislang nicht gekannten Nähe abbilden können, weil sie bei der Wahl zwischen Signifikantem und Nicht-Signifikantem hilft.

Auch im Vergleich zu automatisch erstellten wortbezogenen Informationssystemen ist der Qualitätsaspekt wichtig. Diese können zwar in relativ kurzer

Zeit eine große Menge an lexikalischen Informationen zusammentragen, doch haben diese Informationen, weil sie nicht durch Lexikographen aufbereitet sind, nicht die gleiche Qualität wie die Angaben in Wörterbüchern.

3. Ausblick

Die Arbeit mit elektronischen Korpora wirkt sich in vielerlei Hinsicht auf die gängige lexikographische Praxis aus. Dies gilt bereits im Rahmen der Fortführung traditioneller Wörterbucharbeit, die von Einzelwörtern in ihrer Grundform ausgeht. Es bleibt dabei, zu diesen Wörtern Lesarten anzugeben, doch wird inzwischen auch teilweise dazu übergegangen, diese nach ihrer Frequenz im Korpus anzuordnen und nicht beispielsweise nach ihrem etymologischen Zusammenhang. Es bleibt dabei, diese Lesarten zu paraphrasieren und neben grammatischen Angaben z. B. typische Verwendungen zu nennen, sowie auch illustrierende Textbelege zu bringen. Geschieht dies alles korpusgestützt und unter bevorzugtem Einsatz des korpusgesteuerten Verfahrens, dann ist dies an sich ein deutlicher Qualitätsgewinn gegenüber nicht-korpusgestützter Lexikographie, da die Beschreibungen dem durch das Korpus repräsentierten Sprachgebrauch näher sind.

In gewisser Weise bleibt solch ein Vorgehen aber trotzdem traditionell, indem es das Einzelwort betrachtet und beschreibt. Wolfgang Teubert formuliert im Kontrast dazu eine andere Perspektive:

> Dass die Bedeutung der Wörter Textbelegen zu entnehmen ist, war seit Adelung und Campe das Grundprinzip der klassischen Lexikographie, das von der Korpuslinguistik aufgegriffen wird. Indessen unterscheidet sie sich von der klassischen Lexikographie in bestimmten Einzelheiten. Einmal wertet die Korpuslinguistik das Korpus systematisch und nicht nur exemplarisch aus. Zum anderen verzichtet sie [...] darauf, die Bedeutung eines Wortes [...] in Isolation zu beschreiben. Drittens bemüht sie sich, auf unterschiedlichen Kontexten beruhende Gebrauchsunterschiede zu verzeichnen und nicht, wie die klassische Lexikographie, ein Wort [...] in Bedeutungen zu zerlegen. Viertens geht es der Korpuslinguistik weniger um das Einzelwort als vielmehr um die Wechselwirkung zwischen Textelement und Kontext. (Teubert 1999, S. 302)

So ist zu erwarten, dass sich Lexikographie noch stärker verändern wird, wenn die Arbeit mit Korpora weiteren Einfluss nimmt[9]. John Sinclair beschreibt im Rückblick auf die COBUILD-Arbeit (2004, S. 9), dass eben nicht nur Details, sondern Beschreibungskategorien und schließlich ganze theoretische Annahmen revidiert werden mussten. Und von der Konzentration auf das Wort als Träger der lexikalischen Bedeutung bewegte sich Sinclair hin zur Vorstellung des „lexical item", das mehrere Wörter lang sein kann und der Hauptträger der lexikalischen Bedeutung ist. Wie wird sich das erst auf die Lexikographie auswirken? Und wie auf die Benutzer? Dies bleibt abzuwarten.

---

[9] An Beispiel des Projektes *elexiko* sind diese Veränderungen zumindest teilweise schon zu beobachten.

Offen ist auch, ob es dabei bleiben kann, unter Lexikographie das redaktionelle Schreiben von Wörterbüchern zu verstehen. Es entstehen neue Arten von automatisch erstellten wortbezogenen Informationssystemen, die vorzugsweise im Internet publiziert werden. Ein Beispiel hierfür ist das Projekt „Digitales Wörterbuch der deutschen Sprache", das an der Berlin-Brandenburgischen Akademie der Wissenschaften entsteht und in dem Korpuslinguisten die retrodigitalisierte Fassung eines traditionellen Wörterbuches um automatisch aus dem Korpus generierte Wortinformationen (z. B. automatisch ermittelte Belege) anreichern und erweitern. Wortbezogene Informationssysteme können aber auch ohne Lexikographen oder Korpuslinguisten erarbeitet werden, wie das Projekt „Wortschatz-Lexikon" zeigt, das in der Hand von Informatikern an der Universität Leipzig liegt und ausschließlich automatisch ermittelte Informationen (z. B. Belege, Angaben zu Synonymen) zu einzelnen Wörtern anbietet. Auch hierbei bleibt abzuwarten, was von den Benutzern in welchen Benutzungssituationen angenommen werden wird und was sich damit durchsetzen wird.

Insofern befindet sich die Lexikographie in einer hoch spannenden Phase; sie sollte sich auch weiterhin mit den Methoden und Möglichkeiten der Korpuslinguistik auseinander setzen, wie auch die Korpuslinguistik von der Auseinandersetzung mit den Methoden und Möglichkeiten der Lexikographie profitieren kann.

4. Literatur

Baugh, Simon/Harley, Andrew/Jellis, Susan (1996): The Role of Corpora in Compiling the Cambridge International Dictionary of English. In: International Journal of Corpus Linguistics 1/1 1996, S. 39–59.

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim.

Belica, Cyril/Perkuhn, Rainer (2006): Korpuslinguistik – Das unbekannte Wesen. Oder: Mythen über Korpora und Korpuslinguistik. In: Sprachreport 1/2006, S. 2–8.

Clear, Jeremy/Fox, Gwyneth/Francis, Gill/Krishnamurthy, Ramesh/Moon, Rosamund (1996): COBUILD: The State of the Art. In: International Journal of Corpus Linguistics 1/2 1996, S. 303–314.

Duden (2004): Die deutsche Rechtschreibung. 23., völlig neu bearbeitete und erweiterte Auflage. Herausgegeben von der Dudenredaktion. Auf der Grundlage der neuen amtlichen Rechtschreibregeln. Mannheim/Leipzig/Wien/Zürich.

Engelberg, Stefan/Lemnitzer, Lothar (2001): Lexikographie und Wörterbuchbenutzung. Tübingen.

Haß, Ulrike (2005): Semantische Umgebung und Mitspieler. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. S. 227–234.

Herberg, Dieter/Kinne, Michael/Steffens, Doris (2004): Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Unter Mitarbeit von Elke Tellenbach und Doris al-Wadi. Berlin/New York. (Schriften des Instituts für Deutsche Sprache Band 11).

Hunston, Susan (2002): Corpora in Applied Linguistics. Cambridge.

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

Müller-Spitzer, Carolin (2003): Ordnende Betrachtung zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: Lexicographica 19/2003, S. 140–168.

Schlaefer, Michael (2002): Lexikologie und Lexikographie. Eine Einführung am Beispiel deutscher Wörterbücher. Berlin.

Schnörch, Ulrich (2005): Die *elexiko*-Stichwortliste. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. S. 71–90.

Sinclair, John (1987): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London.

Sinclair, John (1991): Corpus Concordance Collocation. Oxford.

Sinclair, John (Hg.) (2004): How to Use Corpora in Language Teaching. Amsterdam/Philadelphia.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethode, linguistisches Modell, lexikografische Perspektiven. In: Kathrin Steyer (Hg.) 2004: Wortverbindungen – mehr oder weniger fest. S. 87–116.

Storjohann, Petra (2005a): Paradigmatische Relationen. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. S. 249–264.

Storjohann, Petra (2005b): Semantische Paraphrasen und Kurzetikettierungen. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. S. 182–203.

Summers, Della (1996): Computer lexicography: the importance of representativeness in relation to frequency. In: Jenny Thomas/Mick Short (Hg.) 1996: Using Corpora for Language Research. Studies in the Honour of Geoffrey Leech. S. 260–266.

Teubert, Wolfgang (1999): Korpuslinguistik und Lexikographie. In: Deutsche Sprache 4/1999, S. 292–313.

Thelwall, Mike (2005): Creating and using Web corpora. In: International Journal of Corpus Linguistics 10/4 2005, S. 517–541.

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

Wahrig (2003): Fehlerfreies und gutes Deutsch. Das zuverlässige Nachschlagewerk zur Klärung sprachlicher Zweifelsfälle. Gütersloh/München.

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

5. Internetquellen (zuletzt eingesehen März 2006)

http://www.adwmainz.de/2005/index.php?sektion=vorh_gsk&ID=38
http://www.duden.de/
http://www.dwds.de
http://www.elexiko.de/Korpusbasiertheit.html
http://www.ids-mannheim.de/kt/misc/tutorial.html
http://www.ids-mannheim.de/lexik/fremdwort/quellen.html
http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen/
http://www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html
http://www.wortschatz.uni-leipzig.de/