Stefaniya Ptashnyk

Neue Methoden und Publikationsformen in der Lexikologie/Lexikografie

Neue Methoden und Publikationsformen in der Lexikologie und Lexikografie waren das Thema eines wissenschaftlichen Kolloquiums, das im Anschluss an die 39. Jahrestagung des IDS in Mannheim stattfand und das Ziel hatte, online- sowie korpusbezogenen Projekten ein Forum zu bieten. Die Aktualität dieses Kolloquiums lag auf der Hand: Die Tagung des IDS zeigte deutlich, dass der phraseologische Bereich zunehmend eine Tendenz zur Ausweitung in Richtung Kollokationsforschung aufweist und dadurch von der Korpusproblematik nicht zu trennen ist. Kollokationsanalyse und Korpusauswertung gewinnen immer mehr an Bedeutung für die praktische Lexiko- und Phraseografie als herausragendes Instrumentarium zur Beschreibung des Wortschatzes einer Sprache.

Das Kolloquium stieß beim Fachpublikum auf großes Interesse: Es versammelten sich etwa 100 Interessenten aus verschiedenen Ländern im Vortragssaal des IDS. Nach der Begrüßung durch den Institutsdirektor **Ludwig Eichinger** und die Initiatorin des Forums, **Kathrin Steyer,** wurden ausgewählte Projekte präsentiert, die sich weltweit korpusbasiert mit der deutschen Sprache befassen.

Das Podium eröffnete **Wolfgang Teubert** (Birmingham), der über die korpusbasierte Lexikografie am Birmingham Centre for Corpus Linguistics (http://www.corpus.bham.ac.uk) berichtete. Der Birminghamer Ansatz der Bedeutungsbeschreibung beruht auf der Annahme, dass das, was sich über die Bedeutung eines Wortes oder einer Phrase sagen lässt, bereits im Diskurs enthalten sein muss. Was Wörter und Phrasen bedeuten, wurde und wird im Diskurs von den Mitgliedern der Diskursgemeinschaft ausgehandelt. Die Ergebnisse findet man in allen (mündlichen und schriftlichen) Texten des Diskurses einschließlich zurückliegender Texte in der Form von Paraphrasen, Erklärungen, Zuordnungen und mehr oder weniger vollständigen Definitionsansätzen. Um diese Paraphrasen für die lexikografische Bedeutungsbeschreibung zugänglich zu machen, gilt es, ein geeignetes Korpus zusammenzustellen. In diesem Korpus lassen sich dann unter Verwendung spezieller Suchalgorithmen die Paraphrasen der gesuchten Ausdrücke, die in einer Vielzahl von Formeln und Mustern realisiert werden, identifizieren. Ein ers-

ter Anfang ist die Suche nach Wortformen von *bedeuten, heißen, nennen* oder *meinen* im näheren Kontext der jeweiligen Bedeutungseinheit, oder, bei Substantiven, beispielsweise Satzanfänge wie *Globalisierung ist ...*

Teubert demonstrierte am Beispiel des Neologismus *nachhaltig*, wie dieser Ansatz operationalisierbar und für die Lexikografie nutzbar gemacht werden kann. Polyseme Ausdrücke müssen allerdings erst disambiguiert werden. Dies geschieht durch die Ermittlung ihrer ‚Kollokationsprofile' (Sets signifikant häufiger Kontextwörter), die sich je nach Bedeutung unterscheiden. So wurden zunächst die Belege für *nachhaltig* im Sinne der Ökologie identifiziert. Anschließend wurden dann für diese Untermenge die Belege ermittelt, in denen *nachhaltig* paraphrasiert wurde. Ein Nachteil dieses Verfahrens ist, dass Paraphrasen am häufigsten da anzufinden sind, wo die Diskursgemeinschaft die Bedeutung bestimmter brisanter Ausdrücke verhandelt, oder wo neue Ausdrücke in den Diskurs eingeführt werden. Um für alle Lemmata Paraphrasen zu finden, werden Korpora benötigt, die in den mehrstelligen Milliardenbereich gehen. Alternativ muss daher verstärkt das Internet als virtuelles Korpus herangezogen werden.

Für die Zwecke der praktischen Lexikografie wurde das „Wahrig Textkorpus digital" (WTd) geschaffen, das als Resultat einer inzwischen mehr als zweijährigen Zusammenarbeit zwischen der Firma CLT Sprachtechnologie und dem Wissen Media Verlag (vormals Bertelsmann Lexikon Verlag) entstanden ist. Dieses deutsche Korpus mit ca. 575 Mio. Textwörtern wurde von **Andrea Kowalski** (Saarbrücken) vorgestellt. Es enthält mehrere Ebenen der linguistischen Annotation (wie POS, Lemmatisierung, partielle Konstituentenstrukturen), verschiedene Erweiterungen sind in Planung. Von den aufgenommenen Textwörtern können durchschnittlich 11% nicht lemmatisiert werden. Daraus werden Eigennamen, Abkürzungen, „Nichtwörter", Straßennamen, Ortsangaben u. ä. herausgefiltert. Die verbleibenden Wortformen werden auf das entsprechende Lemma zurückgeführt.

Die von CLT vorgenommenen Auswertungen des WTd gehen in die Neuauflagen der verschiedenen Wörterbücher aus der Wahrig-Reihe ein. Bisher stand dabei die Erzeugung von wörterbuchspezifischen Lücken- und Neologismenlisten im Mittelpunkt. Von Nutzen ist auch die automatische Erkennung von rechtschreibschwierigen Wörtern und häufigen Schreibfehlern (z. B. *aggresiv mit einfachem s, *standart- mit t in den Komposita, *Entgeld mit d usw.), Anglizismen (mit heterogenen Flexionseigenschaften), Kollokationen etc., die auf der Basis des WTd vorgenommen werden. Diese Auswertungen gehen als Empfehlungen an die Lexikografen, die letztlich über Neuaufnahmen in den Wörterbuchbestand sowie Aktualisierungen oder Streichungen von veralteten Einträgen entscheiden.

Anschließend präsentierten **Katharina Falkson** und **Ingrid Lemberg** (Heidelberg) die Online-Fassung des Deutschen Rechtswörterbuches (DRW), die seit einigen Jahren parallel zur Printversion existiert (zu finden unter: http://www.rzuser.uni-heidelberg.de/~cd2/drw/index.htm). Das DRW

ist ein historisches Bedeutungswörterbuch zur westgermanisch-deutschen Rechtssprache, in dem sowohl fachsprachliche Termini als auch der Alltagswortschatz in seinen rechtlichen Bezügen erschlossen werden. Es entsteht an der Heidelberger Akademie der Wissenschaften auf der Basis von rund 8.000 Quellen. Bisher sind zehn der geplanten sechzehn Bände im Druck erschienen. Die Online-Fassung des DRW weist ein gut durchdachtes Hypertextualisierungskonzept auf und verfügt über eine reichhaltig ausgestattete Benutzeroberfläche. Mehrfache Zugriffsstrukturen, Verknüpfungen mit Textarchiven und Faksimiles sind das Ergebnis einer etwa achtjährigen Arbeit. Verbesserungsvorschläge rief allerdings die Aufteilung der einzelnen Frames hervor: Eine bessere Übersichtlichkeit wäre hier wünschenswert.

Hans Bickel (Basel) stellte ein grenzüberschreitendes Projekt vor: das Wörterbuch der nationalen und regionalen Varianten der deutschen Standardsprache, das sich zum Ziel setzt, den regionalen Wortschatz aus Deutschland, der Schweiz und Österreich gleichberechtigt nebeneinander vorzustellen. An dem Projekt arbeiten seit fünf Jahren lexikografische Teams in Duisburg, Innsbruck und Basel. Dabei stützen sich die Lexikografen auf ein Korpus moderner Texte, das verschiedene Textsorten umfasst: Tageszeitungen, Illustrierte, populäre Fachzeitschriften und Fachbücher, amtliche Texte (Formulare), Werbeprospekte und belletristische Prosa. Für die Quellenexzerption wurde ein Rundlaufverfahren zwischen den einzelnen Arbeitszentren entwickelt, das Internet dient systematisch als empirische Basis zur Aufspürung, Überprüfung und Bewertung von Varianten. Ins Wörterbuch werden Lexeme mit einer spezifisch nationalen oder regionalen Verwendung, Verwendungsfrequenz oder regional differenzierter Bedeutung aufgenommen (z. B. *Marille, Estrich, Velo* u. a.) sowie Wörter mit unterschiedlicher Aussprache und Betonung, jedoch keine mundartlichen Ausdrücke.

Mit einem derartigen Wörterbuch betritt das Projekt Neuland. Es zeigt das Deutsche als eine plurizentrische Sprache und eröffnet dadurch – im Vergleich zu bisherigen Wörterbüchern – einen neuen Blick auf die Standardsprache. Allerdings stößt dieses Projekt auch auf Schwierigkeiten. So ist es zum Beispiel fraglich, ob Lexeme wie *Brötchen, Laibchen, Weckerl, Brötli* und *Kanapee*, die sich zwar auf denselben Wirklichkeitsbereich beziehen, lexikografisch auf der selben Ebene anzusetzen sind. Eine gedruckte Ausführung des Wörterbuches wird im Jahr 2004 erwartet.

Die praktische Arbeit an korpusbasierter Lexikografie in Tschechien schilderte **Marie Vachkova** (Prag) in ihrem Vortrag über das Große Deutsch-Tschechische Akademische Wörterbuch. Dieses Werk ist als ein allgemeines Übersetzungswörterbuch mit etwa 130000 Einträgen geplant.

Das Projekt, das seit dem Jahr 2000 am Institut für germanische Studien der Philosophischen Fakultät der Karls-Universität läuft und von der „Grant Agency of the Czech Republic" finanziert wird, kam hauptsächlich dank des persönlichen Engagements von Marie Vachkova zustande. Das Wörterbuch wird korpusbasiert und wortartenorientiert verfasst, wobei den Fachwort-

schätzen besondere Aufmerksamkeit zukommt. Weit gehend werden bei der lexikografischen Erfassung die gesprochene Sprache, pragmatische Besonderheiten und die Varietäten des Deutschen berücksichtigt.

Da dieses Projekt mit sehr geringen finanziellen Mitteln ausgestattet ist, werden zahlreiche Germanistikstudenten und Doktoranden in die lexikografische Arbeit involviert. So entstehen innerhalb des Projekts „Nebenprodukte" wie Dissertationen, kleinere terminologische Nachschlagewerke, Seminar- und Diplomarbeiten. Demnächst ist auch ein Workshop zur Kollokationsanalyse in Zusammenarbeit mit dem IDS in Prag geplant.

Annette Klosa (IDS) präsentierte das institutsinterne Projekt „elexiko – Wissen über Wörter" – das (derzeit noch im Aufbau befindliche) lexikalisch-lexikologische korpusbasierte Informationssystem des Instituts (http://www.ids-mannheim.de). Zur Erarbeitung dieses hypertextuellen Informationssystems wird bewährte lexikografische Praxis mit neuen empirischen Methoden der korpusbasierten Sprachuntersuchung verbunden. Die Form der Internetpublikation eröffnet komplexe und auf unterschiedliche Bedürfnisse abgestimmte Nutzungsmöglichkeiten. Bei der Beschreibung einzelner Lemmata werden u. a. solche Bereiche wie Bedeutung und Verwendung, Grammatik, Pragmatik sowie Etymologie abgedeckt. Die Stichwortliste wird mit Hilfe eines Lemmatisierungsprogramms auf Basis der Vorkommenshäufigkeit der Wörter erstellt. Geplant sind ca. 350 000 Lemmata.

Für das Projekt wird das IDS-Textkorpus COSMAS mit über 1,8 Mrd. Textwörtern genutzt. Durch die Auswertung der Korpora wird eine neue Beschreibungs- und Darstellungsqualität erreicht. Es handelt sich nicht nur um die Bedeutungserschließung, sondern auch um die Präsentation der Verwendungsspezifika. So zeigte Annette Klosa, dass das Wort *vorstellig* attributiv verwendet werden kann, obwohl es im Duden nur als Adverb markiert ist. Das Wort *morgendlich* sollte laut Angaben im Duden attributiv gebraucht werden, die Korpustexte belegen aber auch den prädikativen Gebrauch im Sinne von „jeden Morgen". Die Korpusauswertung erlaubt eine vernünftige Lösung der sprachlichen Zweifelsfälle. Auf die Frage, ob *Sims* z. B. ein Maskulinum oder Neutrum ist, liefert COSMAS folgende Antwort: 28 Belege für *der Sims* und 11 Belege für *das Sims*. Die elexiko-Datenbank ermöglicht dem Lexikografen zudem, die Angaben zu kommentieren.

Die Bearbeitung der Stichwortliste soll modular verlaufen, wobei zwei Verfahrensweisen vorgesehen sind: (1) allgemeine Informationen zu allen Lemmata und (2) tief gehende Informationen zu ausgewählten Lemmata. Viele geplante Anwendungen, wie etwa gezielter Zugriff für unterschiedliche Nutzerkreise, zahlreiche Vernetzungen u. a., sind derzeit im Versuchsstadium. Ab Herbst 2003 wird die Lemmaliste im Internet abrufbar sein.

Zum Abschluss des Kolloquiums stellte **Stefan Bordag** (Leipzig) das inzwischen weiten Nutzerkreisen bekannte Projekt „Deutscher Wortschatz" (http://wortschatz.informatik.uni-leipzig.de) vor. Zunächst als eine bloße Sammlung deutscher Wörter geplant, verfügt der „Deutsche Wortschatz"

heute über ein umfassendes Korpus deutscher Sprache mit 6,5 Mio. Wortformen und 23 Mio. Textwörtern, hauptsächlich aus Zeitungs- und populärwissenschaftlichen Texten. Im Rahmen des Projektes werden zur Zeit statistisch basierte Algorithmen sowie Verfahrensweisen zur voll- oder semiautomatischen Sammlung und Verarbeitung von natürlichsprachlichem Material entwickelt. Zur Verarbeitung zählen dabei sowohl Verfahren zur Qualitätssicherung (Rechtschreibfehler, widersprüchliche Informationen, fehlende Angaben) als auch Verfahren zur automatischen Generierung von Angaben. Auch das Kollokationsverfahren findet Anwendung, wodurch Satzkollokationen und Nachbarkollokationen extrahiert und anhand eines mehrdimensionalen Netzes visualisiert werden. Neue Entwicklungen im Projekt sind z. B. „Wörter des Tages" mit einem Häufigkeitsvergleich.

Als eindeutig positiv wurde in der Diskussion die ständige Online-Nutzbarkeit der Projekt-Ergebnisse hervorgehoben. Die linguistische Grundlage des „Deutschen Wortschatzes" rief aber kritische Stimmen hervor, denn die Präsentation vermittelte den Eindruck eines in erster Linie auf Informatikerbedürfnisse ausgerichteten Projekts.

Wie ein roter Faden wurde immer wieder die Frage nach der Kooperation einzelner elektronisch-lexikografischer und korpusbezogener Projekte angesprochen. Korpora entstehen an verschiedenen Stellen mit erheblichem finanziellen und personellen Aufwand, deren Autoren sich um eine große Zahl von Textwörtern und um die Verbesserung der Zugriffsmöglichkeiten für die Nutzer bemühen. Eine Koordination dieser Bemühungen wäre wünschenswert. Bisher scheitern Kooperationsverhandlungen vor allem an rechtlichen Fragen wie der Freigabe der Texte, Urheberrechte usw. Hier besteht noch ein sehr großer Bedarf an logistischen Überlegungen.

Ein weiteres Problem, das vom COSMAS-Autor **Cyril Belica** (IDS) thematisiert wurde, ist die Transparenz der Korpusmethode. Dem Nutzer sollen vielfältige Zugriffsmöglichkeiten geboten werden, vor allem die Option, je nach Zielsetzung oder Fragestellung ein eigenes Korpus auf der Basis der aufbereiteten und annotierten Texte zusammenzustellen. Auf diese Weise lässt sich ein größerer qualitativer Nutzen der Korpuslinguistik für die praktische Lexikografie erreichen.