Sprachkorpora – Datenmengen und Erkenntnisfortschritt

Bericht von der 42. Jahrestagung des Instituts für Deutsche Sprache

von Jens Gerdes

Die mittlerweile 42. Jahrestagung des IDS fand vom 14.-16. März 2006 statt. Das diesjährige Thema lautete „Sprachkorpora – Datenmengen und Erkenntnisfortschritt“. Wenn man also untersucht, welche Folgen die relativ junge Möglichkeit der Nutzung großer elektronischer Korpora für die linguistische Forschung hat, taucht natürlich gleichzeitig die Frage auf, wo die Grenzen korpusbezogener Arbeit liegen. Oder, anders gewendet: Welchen Status haben denn damit überhaupt noch die anderen traditionellen Arten des Datenbezugs und der Datengewinnung für den Linguisten? Alternativen sind ja klassischerweise die Introspektion und das gesteuerte Experiment. Eine solche Diskussion, so steht zu erwarten, liefe darauf hinaus, in welcher Form sich diese Vorgehensweisen beim Gewinn sprachwissenschaftlicher Erkenntnis ergänzen. Noch genereller sind dann Überlegungen, welche Konsequenzen diese verschiedenen Arten des Bezugs auf sprachliches Material für den Status der Linguistik als empirische Wissenschaft hätten. Solcherlei grundsätzliche Betrachtungen stellte denn auch der erste Referent der Tagung, **Christian Lehmann**, in seinem Vortrag über „Daten, Dokumentation, Korpora“ an. Als Folge seiner gründlichen Untersuchung der Kriterien, nach denen etwas in der Linguistik als wissenschaftliches Datum gelten kann, kam er allerdings bezüglich des gegenwärtigen Status der Sprachwissenschaft als empirische Wissenschaft zu einer eher skeptischen Einschätzung. Der Beitrag von **Anke Lüdeling** behandelte die Schwierigkeiten, Korpusdaten möglichst unbeeinflusst von Vorannahmen zu erheben. Wenn etwa die Entscheidung für bestimmte Annotationstypen auch Folgen für die quantitative Auswertung hat, dann leuchtet unmittelbar ein, dass es nützlich ist, die Annotationsebenen technisch von den Texten getrennt zu speichern. Anhand eines Lernerkorpus diskutierte Lüdeling die besondere methodische Herausforderung der Tatsache, dass man bei unterschiedlicher Annotation zu konfligierenden Ergebnissen kommen kann. Für die historische Sprachwissenschaft, die ja notwendigerweise immer schon eine Korpus-Wissenschaft war, ist dagegen der Vorteil, Texte und ihre Annotierungen elektronisch und in repräsentativer Weise zur Verfügung zu haben, offenkundig. Das stellte **Karin Donhauser** in ihrem Vortrag „Historische Korpora und Sprachgeschichtsforschung“ fest und berichtete darüber hinaus über verschiedene Aktivitäten, zu einem gemeinsam nutzbaren Korpus für die historische Erforschung des Deutschen zu kommen.

Prof. Dr. Ludwig M. Eichinger begrüßt die Tagungsteilnehmer

Nach diesen einführenden Vorträgen des ersten Vormittags ging es im Weiteren um theoretische und methodische Aspekte der Arbeit mit Korpora in Grammatik und Lexik. **Sam Featherston** betonte den Wert experimenteller Verfahren und gesteuerter Introspektion, besonders hinsichtlich der Kontrollierbarkeit von Verfahren und Ergebnissen. Damit plädierte er für eine Methodeninteraktion. **Stefan Müller** zeigte, dass sehr große Korpora zum Nach-

weis marginaler Phänomene etwas beitragen können, was durch andere Methoden nicht erreicht wird – die linguistische Bewertung seltener Belege dürfe aber dennoch nicht ausbleiben. Eine andere Seite der Korpusarbeit präsentierte **Ulrike Demske**, die von ihrem Projekt der syntaktischen Annotation historischer Texte berichtete, das den Beginn der Arbeit an einer Baumbank für das Frühneuhochdeutsche bedeutet. **Annette Klosa** schilderte die Arbeit an *elexiko*, dem korpusbasierten Wörterbuch des IDS, und diskutierte die Frage, ob und wie eine korpusgestützte Lexikographie zu wissenschaftlich angemesseneren Resultaten kommen kann, als sie das klassische Wörterbuch bietet. Ebenfalls aus dem Theorie-Praxis-Umfeld der Lexikographie kam der Beitrag von **Jörg Asmussen**, der über Verfahren und Probleme der Umsetzung von „Den Danske Ordbog“ in eine adäquate elektronisch nutzbare Form referierte. Insbesondere Fragen der Korpus-Recherche unter semantischen Aspekten standen hier im Vordergrund.

Der nächste Vormittag widmete sich dann den Bedingungen der Arbeit mit Korpora in Pragmatik und Soziolinguistik. **Gunter Senft** betonte die Notwendigkeit einer zusätzlichen Annotationsebene mit ethnolinguistischer Information, wenn man Texte bei kultureller Differenz adäquat verstehen will. In ähnlicher Weise, allerdings eher auf die strukturelle Seite der Beschreibung wenig erforschter Sprachen bezogen, wies **Dafydd Gibbon** in seinem Vortrag auf die Notwendigkeit hin, passende qualitative Kategorien zu entwickeln.

Die Untersuchung regionalsprachlicher Variation in gesprochensprachlichen Korpora wurde im Beitrag von **Alexandra Lenz** dargestellt. Eine Beschreibung der regional unterschiedlich verteilten relativen Häufigkeit des Rezipientenpassivs diente der Referentin dazu, die Einsatzmöglichkeiten der „Datenbank Gesprochenes Deutsch“ (DGD) im IDS für solche Zwecke zu veranschaulichen. **Werner Kallmeyer** sprach im Anschluss über die speziellen Schwierigkeiten der automatischen Recherche in Gesprächskorpora. In der Hauptsache sind dies naturgemäß die Koordinierung von Bilddaten, Tondaten und Transkripten sowie allgemein die Bearbeitungsmöglichkeiten der komplexen gesprächsanalytischen Transkriptionstypen. Den dritten Teil der Tagung, der sich spezifischer mit informatischen Fragen beschäftigte, eröffnete **Hans Uszkoreit**. Sein Vortrag behandelte nicht nur den Wert verschiedener Datentypen, sondern legte auch die Probleme einer adäquaten qualitativen Bearbeitung dar. Einen methodisch vollkommen anderen Zugang beschrieb der Beitrag von **Thorsten Brants**, der zeigte, wie bei der Internet-Suchmaschine „Google“ versucht wird, mit riesigen Datenmengen und dem Vergleich von „Vorkommensclustern“ maschinelle Übersetzungen zu ermöglichen. Abschließend berichtete **Michael Strube** von einem praktisch orientierten Projekt mit dem Ziel, zu einer automatischen Zusammenfassung des Inhalts von Dialogen zu kommen. Er ging vor allem auf die dialogspezifischen Schwierigkeiten bei diesem Vorhaben ein.

Die abschließende Podiumsdiskussion unter der Leitung von **Norbert Richard Wolf** beschäftigte sich unter verschiedenen Aspekten mit der Frage, wie in der Sprache vorfindliche Variation angemessen in Korpora zu repräsentieren sei.

Großen Anklang fand auch ein zusätzlicher Programmpunkt: Am Nachmittag des zweiten Tagungstages wurden in den Räumen des IDS hauseigene und fremde Systeme für Transkription, Text-Ton-Alignment, Datenbankrecherche und die Handhabung von Multimedia-Korpora präsentiert.

Zudem wurde das IDS aus Anlass dieser Tagung im Rahmen einer Initiative der Bundesregierung und des BDI (Bundesverband der Deutschen Industrie) als „Ort der Ideen“ ausgezeichnet. Eingeleitet durch ein Grußwort von Peter Kurz, dem Kulturbürgermeister der Stadt Mannheim, überreichte am Mittwochmorgen der Vertreter der Deutschen Bank, Josef Zimmermann, dem Institutsdirektor Ludwig Eichinger eine Urkunde. Direktor **Ludwig Eichinger** hielt aus diesem Grund am Nachmittag einen auch für sprachwissenschaftliche Laien verständlichen Vortrag unter dem Titel „Was ist los mit der deutschen Sprache? Aktuelle Tendenzen der Sprachentwicklung und Sprachforschung.“ Unter anderem stellte er darin natürlich auch die Arbeit des IDS vor.

Die nächste IDS-Jahrestagung findet unter dem Titel „Sprache – Kognition – Kultur“ vom 6. bis 8. März 2007 im Stadthaus in Mannheim statt.

Der Autor ist geprüfte wissenschaftliche Hilfskraft am Institut für Deutsche Sprache in Mannheim.

Foto: A. Trabold