In: Kallmeyer, Werner / Zifonun, Gisela (Hg.) (2007): Sprachkorpora – Datenmengen und Erkenntnisfortschritt (= Institut für Deutsche Sprache Jahrbuch 2006). Berlin / New York: de Gruyter, S. 1-8.

Ludwig M. Eichinger

Linguisten brauchen Korpora und Korpora Linguisten

Wege zu wohl dokumentierten und verlässlichen Aussagen über Sprache

1. Ein Anfang

Zur 42. Jahrestagung des Instituts für Deutsche Sprache, der des Jahres 2006, konnte der Direktor nicht nur wieder an die vierhundertsiebzig Teilnehmer begrüßen, sondern darunter auch und ganz besonders den 26. Preisträger des Konrad-Duden-Preises der Stadt Mannheim. Mit Heinrich Löffler fiel die Wahl des Preisgerichts auf einen Kollegen, der nicht nur ein regelmäßiger Gast auf den Jahrestagungen des IDS war, sondern der dem Institut auch sonst seit vielen Jahren und besonders eng verbunden ist. Nicht zuletzt war er von der Gründung dieses Gremiums im Jahre 1998 bis 2002 der Vorsitzende des Wissenschaftlichen Beirats des Instituts. Er hat in dieser Funktion das Institut und seinen Beirat mit Bedacht aber auch Entschlossenheit an die noch neue Welt der Evaluationen und Leistungskontrollen herangeführt. Das IDS freut sich, dass dieser Preis einem unserem Institut solcherart nahe stehenden Kollegen zuerkannt wird, einem Kollegen zudem, der sich in vielerlei Hinsicht in die germanistische Pflicht nehmen ließ. Für viele von uns, die wir einer etwas jüngeren Wissenschaftlergeneration angehören, repräsentiert Heinrich Löffler in seiner wissenschaftlichen Arbeit in bemerkenswerter Weise die Generation und den Typ von germanistischem Linguisten, der sich ausgehend von einer zeitbedingt traditionellen germanistischen Basis auf einen eigenständigen Weg in die neue Welt der Linguistik gemacht hat.

Wir freuen uns auf die Verleihung dieses Preises an Heinrich Löffler und mit ihm über diesen Preis; diese Gelegenheit bietet aber auch den willkommenen Anlass, um den verleihenden und geldgebenden Institutionen, der Stadt Mannheim und dem DUDEN-Verlag, dafür zu danken, dass sie nunmehr zum sechsundzwanzigsten Mal die erfolgreichen Mühen eines der im Steinbruch der deutschen Sprachwissenschaft Tätigen mit diesem Preis würdigen.

2. Drei wichtige Punkte

Unsere Tagung steht dieses Jahr unter dem Thema „Sprachkorpora – Datenmengen und Erkenntnisfortschritt" – wer will, kann sich ein Fragezeichen dazu denken. Es ist das aus drei Gründen ein wichtiges und aktuelles Thema.

- Zum einen haben sich in den letzten Jahren die technischen Möglichkeiten zur Erstellung sprachlicher Korpora in einem Ausmaß verbessert, dass die Lage in dieser Hinsicht nicht mehr mit der noch vor fünf, sechs Jahren zu vergleichen ist. An den Ergebnissen dieser Entwicklung kann man nicht mehr einfach vorübergehen.
- Zum anderen haben sich dadurch die Verhältnisse auch in theoretischer Hinsicht geändert: selbst die linguistischen Theorien, denen im Prinzip die Introspektion als methodisches Instrumentarium genügt hatte, denen reale Daten eher als verschmutzte Reflexe abstrakter Prinzipien galten, sehen nun in der Beobachtung großer Korpora eine weitere Möglichkeit, auch das theoretisch relevante Wissen über Sprachliches zu erweitern. Und das gilt natürlich umso problemloser für jene Herangehensweisen, die den Weg zu empirischer linguistischer Erkenntnis in einer hypothesengeleiteten Belegsammlung suchten.
- Zum dritten haben diese drei Weisen, sich dem sprachlichen Datum anzunähern, erst durch die bemerkenswerten Entwicklungen der Korpustechnologie einen Status erreicht, der sie nun auf gleicher Höhe einander gegenüberzustellen erlaubt. Tatsächlich hat es die anscheinend und scheinbar grenzlose Verfügbarkeit von Belegen in elektronischen Korpora nicht nur möglich gemacht, sondern auch erzwungen, in neuer Weise klarzulegen, was in der Linguistik als ein empirisches Datum verstanden werden soll. Sofern in vielleicht naiver Weise mit allen drei methodischen Herangehensweisen der Anspruch verbunden wird, sprachliche Empirie zu dokumentieren, ist jedenfalls damit Unterschiedliches gemeint. Die hypothesengeleitete Datenzusammenstellung und mehr noch die auf das eigene Sprachkönnen und seine Verlässlichkeit rekurrierende Introspektion berufen sich auf die nur relative Unabhängigkeit der sprachlichen Erzeugung vom erzeugenden (und perzipierenden) Subjekt, das daher mehr Verfügungsmacht über die entsprechenden Erzeugnisse habe als bei naturgegebenen Objekten oder Artefakten. Dagegen sind Korpora im Prinzip der Versuch, ohne Rekurs auf solche Tatbestände der sprachlichen Wirklichkeit über statistische Untersuchungen und mathematische Modellierungen nahe zu kommen.

In gewisser Weise stoßen wir hier – zumindest in den Extremen – auf die wichtige Frage, wie sich eine solche in der Tradition eines klassischen Empirizismus stehende Vorgehensweise, die sich erst jetzt so recht auf Sprache anwenden lässt, zu einem holistischen Inferentialismus verhält, der davon ausgeht, dass das verstehende Subjekt, und nur es, in der Lage ist, die Einbettung des sprachlichen Datums in die Welt unserer interaktionalen Gesetzmäßigkeit aufzuweisen.[1]

[1] Vgl. dazu die Ausführungen in Brandom 2000, S. 147 ff. bes. 152.

Die umgekehrte Frage stellt sich allerdings erst mit der Verfügbarkeit großer elektronischer Korpora: ohne jeden Zweifel ist es für das erkennende Subjekt außerordentlich schwierig, ohne solch eine Datenübersicht all diese Phänomene und ihre Größenordnung ins Bewusstsein zu rufen, die unter bestimmte Klassifikationen fallen würden.

3. Wachsende Korpora – wachsende Bedeutung?

Damit gewinnen Korpora auch über den Bereich hinaus an Bedeutung, in denen ihr statistischer Wert augenfällig ist.

- Augenfällig ist er bei der Untersuchung lexikalischer Kookkurrenzen, insofern lexikalische Information einer unmittelbaren Suche ohne weitere Einschränkungen und Annahmen unmittelbar zugänglich ist. Es bedarf dazu nicht viel mehr als der Grundannahme, dass die Rekurrenz von lexikalischen Einheiten, als „Kernen" von Wörtern, ein relevantes Merkmal einer Sprache darstelle. Diese Behauptung dürfte wohl inzwischen als unstrittig gelten, auch wenn sie in der Geschichte der sprachwissenschaftlichen Erkenntnis längere Zeit von dem Tatbestand verdunkelt wurde, dass Sprachtypen im Mittelpunkt der Beschreibung standen, bei denen die relationale Kraft grammatischer Mittel, vor allem auch der Flexive, die Bedeutung der lexikalischen Kookkurrenz verdunkelten. Das objektive Problem, das hinter dieser Differenz steht, schlägt sich in unserem thematischen Zusammenhang zum Beispiel auf der Ebene des Wortarten-Tagging nieder. Wenn hier häufig an den gängigen Wortarten des europäischen Sprachtyps entlanggegangen wird,[2] projiziert das Annahmen über den prototypischen Charakter bestimmter Wortarten auf die zu interpretierenden Wortfolgen. Damit ergibt sich eine gewisse, aber je nach Sprache unterschiedlich direkte Korrelation zu den einem Lemma zugehörigen lexikalisch-morphologischen Varianten. Das heißt, dass hier die Kookkurrenzen bereits nach einem – grob gesagt am indoeuropäischen Muster orientierten – Inventar klassifiziert sind, bevor ihre Kontexte vorurteilslos betrachtet worden sind.[3] Die erfolgreiche Arbeit der Korpuslinguistik im Bereich der Kollokationen weist auf die Bedeutung der Ebene der Kookkurrenz lexikalischer Kerne, ihre Verschränkung mit strukturellen Erscheinungen stellt eine jener Herausforderungen für eine linguistische Beschreibung dar, als deren mögliche Antwort sich zum Beispiel Ansätze der *construction grammar* verstehen.[4]

[2] Vgl. z. B. die Vorgaben des STTS, s. Schiller et al. 1999.

[3] Das soll nicht verstanden werden als Aufforderung, gänzlich von bereits gewonnenem Wissen abzusehen; vgl. aber, was z. B. in Zifonun et al. (1997, S. 21) zu diesem Punkt gesagt wird.

[4] Vgl. aber mit pragmatischerer Zielsetzung z. B. auch, was Feilke (1996, S. 181 ff.; bes. S. 186) zu „Lexikalisierung" schreibt.

- Logischerweise weniger augenfällig ist daher die Nutzung korpuslinguistischer Methoden für syntaktische Analysen, wenn sie etwas anderes sein sollen und wollen, als ein Beleg- und Operationsinventar für vorgängig entwickelte Beschreibungsweisen und theoretische Ansätze. Aus verständlichen Gründen überwiegen daher derzeit solche Ansätze, die sich eher als Datenbanken für syntaktische Muster verstehen lassen.[5] Aber eigentlich ist das nur die eine Seite dessen, was man auf dieser Ebene von korpuslinguistischen Analysen erwarten würde. Korpora können zwar auf jeden Fall als ein Inventar grammatischer Strukturmuster verstanden werden. Die Frage ist dabei nur, wie man auf dieser Ebene ebenfalls die Vorteile eines Korpus nützen könnte, die darin liegen, Regularitäten ohne allzu viel vorgängige Interpretation aus ihnen zu ermitteln. Das betrifft vor allem den Punkt, welche Arten von Annotationen gewählt werden, und inwieweit sie ein Präjudiz im Hinblick auf die gewünschten Ergebnisse darstellen. Ohne Zweifel ist es schon ein wünschenswertes Ergebnis, festzustellen, von welchen syntaktischen Rekurrenzen ein Korpus geprägt ist, wenn man ein bestimmtes Analyseschema vorgibt, das sich in der grammatischen Tradition in der einen oder anderen Weise bewährt hat. Aber noch schöner wäre es natürlich aus Korpussicht, wenn sich die Emergenz von Strukturen auf einem rein statistischen Weg ergäbe. Und eigentlich nicht nur aus Korpussicht, in Anbetracht der Vielfältigkeit der sprachlichen Praxis wäre zu hoffen, dass sich die emergenten Strukturzüge vom verstehenden linguistischen Interpreten in ein inferentielles Bezugsfeld von Behauptungen eingebunden werden könnte. Man sollte an dieser Stelle nicht überrascht sein, gelegentlich mit Konstellationen konfrontiert zu werden, mit denen man so nicht gerechnet hatte.[6]

4. Generelles zu Korpora

4.1 Die Art von Korpora

Um an die Emergenz von Phänomenen aus korpusbasierten Untersuchungen zu glauben, muss man natürlich einigermaßen sichern, dass man sich auf die Korpora verlassen kann. Im Prinzip gibt es dazu zwei Wege.

- Zum einen kann man sich ein Korpus vor- und zusammenstellen, das im Hinblick auf die vermutete Repräsentativität für bestimmte Phänomene oder für bestimmte Varietäten zusammengestellt ist.
- Zum anderen kann man sich vornehmen, über die schiere Größe und Vielfalt von Korpora die Reliabilität von Korpora zu erhöhen.[7]

[5] Etwa das TIGER-Korpus; s. König/Lezius 2003.

[6] Ohne dass das die Ergänzung durch andere Methoden überflüssig machen würde; manchmal vermeiden der Rekurs auf Verstehen und das dadurch geleitete kontrollierte Experiment ausführliche Umwege im Korpus ebenso wie Fehlschlüsse.

[7] Vgl. dazu die Beschreibung des Programmbereichs Texttechnologie auf den Internet-

Wobei der Vorteil der zweiten Methode unter anderem darin zu liegen scheint, dass sich zielorientierte Ausschnitte aus den Großkorpora gewinnen lassen, so dass in gewisser Weise die erste Möglichkeit – der klug gewählte Ausschnitt – in der Menge der gesamten gesammelten Daten enthalten ist.

4.2 Methoden und Methode

Offenbar ist, dass es jedenfalls klug gewählter Methoden bedarf, um zu eruieren, was wir aus den Korpora lernen können. Und je größer die Korpora, desto höher sollte zwar einerseits der mögliche Erkenntnisgewinn sein, es ist aber auch klar, dass die intuitive Zugänglichkeit und Nachvollziehbarkeit im gleichen Ausmaße sinkt, in der die Größe des Korpus wächst. Das beginnt schon ganz banal damit, dass uns eine zu lange Reihe an Belegen jeglicher Übersichtlichkeit beraubt.

Aber auch abgesehen von diesem Problem bleibt die Tatsache bestehen, dass Korpusbefunde in der einen oder anderen Weise linguistisch interpretiert werden müssen, dass eine qualitative Einschätzung dessen vonnöten ist, was an Phänomenen aus dem Korpus auftaucht. Nicht jede Verwendung ist gleich gut: Texte enthalten in verschiedener Weise markierte Verwendungen, die zum Beispiel die Frage der Grammatikalitätsurteile als nicht unkompliziert erscheinen lassen. Welche Bedeutung haben, wenn das so ist, kontrollierte Experimente und Manipulationen des Materials – die immerhin auch von der Nutzung real vorkommender Variation profitieren kann – und welche Bedeutung kommt dann der introspektiven Abwägung der verschiedenen belegten Erscheinungen zu?

5. Besonderheiten

5.1 Sprachstufenkorpora

In mancher Hinsicht ist das Arbeiten mit Korpora nichts völlig Neues: vor allem wer mit historischen Sprachstufen arbeitet, steht ganz offenkundig vor dem Problem, dass er eigentlich außer Korpora nicht viel hat.

Wenn er Glück hat, kann er sich über die Betrachtung großer Textmengen eine Art Ersatzkompetenz aneignen. Aber auch in diesen Fällen kommt man allmählich an die Grenzen des Problems, das die Gegenwartssprache und die Beschäftigung mit ihr so ganz entscheidend prägt, nämlich die prinzipielle Unbegrenztheit des Materials, mit dem wir es zu tun haben. Um zu verlässlichen Aussagen zu kommen, ist es nötig, die Texte aus signifikanten Kommunikationssituationen zu nehmen und gleichzeitig ihren Verständnishintergrund mit in Betracht zu ziehen – was übrigens nicht nur für historische, sondern für alle kulturell differenten Datenmengen gilt.

Seiten des IDS: „Forschungsgegenstand des Programmbereichs ist die explorative Analyse von sehr großen Sammlungen natürlichsprachlicher Texte („very large corpora")."; vgl. auch Perkuhn/Belica (2005).

Für die historischen Korpora gilt zudem in Sonderheit, dass auch die Frage der adäquaten Annotationskategorien eigens zu beantworten ist.

5.2 Mediales

Was sich beim historischen Blick schon andeutet, findet seine Bestätigung und volle Ausprägung bei der Betrachtung diamedialer Differenzierung. Die Bedingungen moderner europäischer Schriftlichkeit, wie sie für das Deutsche gelten, filtern schon eine ganze Menge von Variation und Begleitbedingungen weg, wie sie für gesprochene Sprache typisch sind. Aufgrund der vielfältigen und multimodalen Einbettung des Sprechens stellen bereits die angemessene Aufzeichnung und dann die elektronische Verarbeitung der Daten gesprochener Sprache weitaus komplexere Aufgaben dar, als das bei geschriebener Sprache der Fall ist.[8] So ist es denn kein Zufall, dass Korpora gesprochener Sprache bei weitem nicht den Bearbeitungsstatus und auch den Umfang erreicht haben, der bei den Korpora geschriebener Sprache nun gängiger Standard ist. Zwar haben sich die Möglichkeiten der Aufnahme gesprochener Sprache in den letzten Jahren dramatisch verbessert, die elektronische Verarbeitung und Analyse kennt aber ihre ganz eigenen Schwierigkeiten. Und das gilt insbesondere, wenn die Untersuchung gesprochener Sprache nicht nur an eigentlich strukturellen Merkmalen der sprachlichen Form interessiert ist, sondern darüber hinaus bemüht ist, das damit verbundene Gefüge der Interaktion zu beschreiben.

6. Der Tagungsablauf

Eigentlich mit all den Punkten, die jetzt so beiläufig angesprochen worden sind, beschäftigen sich die Vorträge, die auf der hier dokumentierten Tagung gehalten worden sind. Das beginnt mit zwei grundsätzlichen Referaten zu Methodologie und Methoden einer Linguistik, die sich auf Korpora einlässt und einem zu den Folgen von Korpusbezug für Deutungen zur Sprachgeschichte. Anschließend geht es zunächst aus verschiedenen – grammatiktheoretischen – Blickwinkeln um Möglichkeiten und Grenzen quantitativer und qualitativer Methoden bei syntaktischen Untersuchungen, danach um den Nutzen von Korpusbasiertheit für die moderne Lexikographie. Dass Korpora nur scheinbar für sich stehen, dass sie einer deutlichen Wissenseinbettung bedürfen, davon sprechen die Vorträge des nächsten Abschnitts, daneben von der Handhabung von Variation und komplexen interaktionellen Verhältnissen im Fall von gesprochener Sprache. In der nächsten Einheit zeigen die Computerlinguisten, was man mit Korpora und passender Technik machen kann – und was es darüber hinaus braucht. Am Schluss sollte eine Podiumsdiskussion Klarheit über das bis dahin Besprochene bringen.

[8] Vgl. die inzwischen technisch historischen aber grundsätzlich bedenkenswerten Überlegungen in Thun (1993).

Der Besichtigung der korpuslinguistischen Praxis diente ein Block mit Präsentationen, mit denen ein Nachmittag gefüllt war, und die im Gebäude des IDS stattfanden.

Was sonst noch auf dem Programm zu finden war, wird die Teilnehmer überrascht haben: unsere Jahrestagung ist als ein Ort der Ideen ausgewählt worden, eine Ehre, die wir unseren Gästen verdanken, ist es doch die regelmäßige Zusammenkunft von Germanisten aus aller Herren Länder, der wir diese Ehrung nicht zuletzt verdanken.[9] In diesem Kontext fand neben dem normalen Tagungsprogramm ein Vortrag statt, in dem einem allgemeinen Publikum dargestellt wurde, inwieweit linguistische Kenntnis beim Verständnis alltäglich auffälliger Phänomene in der deutschen Sprache hilfreich sein kann.

7. Zum Schluss

7.1 Ein Wunsch

Korpusarbeit ist ein nicht unaufwendiges Geschäft und alle, die sich länger damit beschäftigen, wissen das. So gesehen ist diese Tagung auch eine Art Aufruf, einen Weg zu einer effizienten Zusammenarbeit der Interessierten, zu einer angemessenen Konzentration der Korpusarbeit zu finden. Wir sollten gemeinsam über die Phase der Einzelkämpferkorpora hinauskommen.

7.2 Ein Dank und noch ein Wunsch

Dass alle Teilnehmenden eine interessante und anregende Tagung und einen angenehmen Aufenthalt in Mannheim gehabt haben, hat sicherlich verschiedene Gründe.

Einer ist aber sicher, dass das Programm der Tagung erfüllt hat, was man sich von einer Jahrestagung mit diesem Thema erwartet. Die Anerkennung und der Dank dafür gebührt den Mitgliedern des Vorbereitungsausschusses, dem neben Frau Zifonun und Herrn Kallmeyer aus dem IDS Hans Uszkoreit als Mitglied unseres wissenschaftlichen Beirats und Gunter Senft vom MPI Nijmwegen angehörten. Ihnen allen sei ebenso herzlich für ihre Tätigkeit gedankt wie all den Mitarbeiterinnen und Mitarbeitern meines Hauses, die mit gewohnter Professionalität an der Vorbereitung und Durchführung dieser Tagung mitgewirkt haben. Und zuvorderst und zuletzt gilt mein Dank den Referentinnen und Referenten, die ja mit ihrer Arbeit die Tagung eigentlich ausgemacht haben.

[9] Die Wahl eines „Orts der Ideen" für jeden Tag des Jahres ist Teil der mit dem Jahr der Fußballweltmeisterschaft verbundenen Imagekampagne „Land der Ideen", deren Schirmherr der Bundespräsident ist (vgl. www.land-der-ideen.de).

8. Literatur

Brandom, Robert B. (2000): Expressive Vernunft. Begründung, Repräsentation und diskursive Festlegung. Frankfurt am Main: Suhrkamp.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt am Main: Suhrkamp.

König, Esther/Lezius, Wolfgang (2003): The TIGER language – A Description Language for Syntax Graphs, Formal Definition. Technical report. Institut für Maschinelle Sprachverarbeitung. University of Stuttgart

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

Schiller, Anne/Teufel, Simone/Stöckert Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Universität Stuttgart. Institut für maschinelle Sprachverarbeitung/Universität Tübingen Seminar für Sprachwissenschaft (zu finden unter http://www.ifi.unizh.ch/CL/siclemat/man/SchillerTeufel99STTS.pdf).

Steyer, Katrin (Hg.) (2004): Wortverbindungen – mehr oder weniger fest (=IDS Jahrbuch 2003). Berlin/New York: de Gruyter.

Thun, Harald (1993): Was hat sich in der Dialektologie durch die Benutzung von Tongeräten geändert? In: Schlieben-Lange, Brigitte (Hg.): Materiale Bedingungen der Sprachwissenschaft. Zeitschrift für Literaturwissenschaft und Linguistik 90/91, S. 139–156.

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno u. a. (1997): Grammatik der deutschen Sprache. (= SIDS 7.1–3). Berlin/New York: de Gruyter.