Anton Näf/Rolf Duffner (Neuchâtel)

Korpuslinguistik im Zeitalter der Textdatenbanken

Vorwort der Herausgeber zu *Linguistik online* 28, 3/06

First throw away your evidence! (John Sinclair)
Theoriae volant, data manent.

Kein Zweifel: Nicht bloss die Art und Weise wie wir uns kleiden, sondern auch so seriöse Dinge wie die bevorzugten Themen und dominierenden Methoden einer wissenschaftlichen Disziplin sind im Laufe der Zeit gewissen Schwankungen, ja eigentlichen Moden unterworfen. Dies gilt auch für die Sprachwissenschaft. Während vieler Jahrzehnte, bis in die Zeit nach dem zweiten Weltkrieg hinein, war diese stark diachronisch ausgerichtet. In der zweiten Hälfte des 20. Jahrhunderts änderte sich dies dann aber grundlegend. Im Zusammenhang mit der – auch institutionellen – Etablierung einer synchronisch und theoretisch orientierten, nun meist Linguistik genannten Wissenschaft von der Sprache lösten einander in schnellem Rhythmus verschiedene Strömungen und Schulen ab. Im Zentrum des wissenschaftlichen Arbeitens standen dabei meist nicht die Beobachtung von Sprachproduktionen sowie die Sammlung und Klassifizierung von einschlägigen Belegen, sondern – insbesondere bei der lange Zeit dominierenden generativen Grammatik – die Diskussion von selbstkonstruierten Einzelsätzen. Durch diese theoretisch-spekulative Arbeitsweise wurde methodologisch gesehen die Introspektion als *via regia* der Forschung etabliert, wobei man glaubte, sich den "Umweg" über die – naturgemäss mühsame und zeitaufwendige – Performanz ersparen zu können. Wer eine Bilanz der Jahrzehnte des Linguistikbooms ziehen wollte, käme wohl nicht um die Feststellung herum, dass der Ertrag an empirisch abgesicherten, "zeitresistenten" Forschungsergebnissen insgesamt eher bescheiden ausfällt.

Seit einem guten Jahrzehnt hat sich nun aber die Ausgangslage für die sprachwissenschaftliche Forschung grundlegend verändert, dies durch die Verfügbarkeit und die Abfragemöglichkeiten von digitalen Grosskorpora. Von der zeitaufwendigen Arbeit der Belegsuche befreit, können sich die Forscher nun umso intensiver der Auswertung und Interpretation von repräsentativen Stichproben widmen. Was wir gegenwärtig mit dem Aufkommen der Korpuslinguistik und der Hinwendung – zum Teil ist es eine Rückkehr unter veränderten Bedingungen – zur empirischer Forschung beobachten können, wird von vielen als eine eigentliche *revanche de l'empirisme* (B. Habert) wahrgenommen. Jedenfalls hat das Pendel seit einiger Zeit wieder ganz deutlich auf die Seite der empirischen Analyse von Korpora (gemäss dem klassischen Dreischritt beobachten – beschreiben – erklären) ausgeschlagen.

Auch die linguistische Forschung wird in Zukunft immer stärker anhand von Massstäben beurteilt werden, wie sie in der empirischen Sozialforschung schon seit längerem Standard sind. Gleichzeitig werden replizierbare Forschungsresultate zu einer Entideologisierung der Debatten beitragen. Eine solche wohltuende Versachlichung der Forschung, in deren Zentrum die *faits de langue* (de Saussure) und das Funktionieren von Sprache stehen, bietet die Chance eines Neubeginns. Auch wenn wir hier nicht – einmal mehr – von einem Paradigmenwechsel sprechen sollten, scheint es doch so zu sein, dass die moderne Korpuslinguistik, die einen eigentlichen qualitativen Sprung bedeutet, in absehbarer Zukunft zu einem viel verlässlicheren Abbild der Sprachwirklichkeit führen wird.

Ein Beispiel: Während noch vor kurzem ein ganzes Forscherleben kaum genügt hätte, um die schlichte Frage zuverlässig zu beantworten, mit welchen Adjektiven das Nomen *Miene* am häufigsten kombi-

ISSN 1615-3014

niert auftritt (*versteinert, ernst, unbewegt, finster* usw.), lassen sich solche Rohdaten heute sozusagen "auf Knopfdruck" aus riesigen Textmengen herausfiltern. Entsprechendes gilt für die Verben, bei denen *Miene* typischerweise als Subjekt (*sich verfinstern, sich aufhellen* usw.) bzw. als Objekt (*verziehen, aufsetzen* usw.) auftritt. Derartige Erkenntnisse sind nicht zuletzt auch von unmittelbar praktischem Nutzen und werden zu der dringend nötigen Verbesserung der einsprachigen – und in einem zweiten Schritt auch der zweisprachigen – Wörterbücher beitragen.

Das vorliegenden Heft 28/2006 von *Linguistik online* möchte zu einer empirisch fundierten Neuorientierung der Linguistik beitragen. Hauptgeschäft einer künftigen Sprachwissenschaft muss wieder das Verifizieren bzw. Falsifizieren von Hypothesen werden, nicht zuletzt auch von solchen, die noch zu "vordigitalen" Zeiten aufgestellt wurden. Die empirische Erforschung von sprachlichen Fakten muss wieder ins Zentrum des wissenschaftlichen Arbeitens und Argumentierens gestellt werden. Diese Auffassung von Sprachwissenschaft liegt – bei allen Unterschieden im einzelnen – allen in diesem Heft vereinigten Beiträgen zur Korpuslinguistik zugrunde.

Die Beiträge spannen einen weiten Bogen ausgehend von vergleichenden Überblicksartikeln über Grundsatzdebatten bis hin zur Anwendung korpuslinguistischer Verfahren auf konkrete linguistische Fragestellungen. Die beiden Orientierungsartikel zu Beginn sind den fernabfragbaren linguistischen Textkorpora (R. Duffner/A. Näf) und den Einführungen in die Korpuslinguistik (G. Kolde) gewidmet. Darauf folgt ein programmatisches Plädoyer für eine hermeneutische Korpuslinguistik (W. Teubert) sowie eine Replik darauf (R. Berthele). Auf einen Aufsatz zum Internet als linguistischem Korpus (H. Bickel) folgen dann drei Anwendungen auf konkrete sprachwissenschaftliche Themen, nämlich die Satzarten (A. Näf), die Funktionsverbgefüge (A. Kamber) und die Satzadverbien (R. Duffner). Den Band beschliesst ein kontrastiver Beitrag zu den Kollokationen (A. Reder). Die meisten Aufsätze sind überarbeitete Versionen von Vorträgen, welche im Jahre 2004 am Doktoranden-Seminar (*Troisième cycle*) der Westschweizer Universitäten in Neuchâtel gehalten worden sind (20. –23. Mai 2004).

Der Eröffnungsbeitrag von **Rolf Duffner** und **Anton Näf** bietet einen umfassenden Vergleich der Möglichkeiten und Grenzen von fünf wichtigen, öffentlich zugänglichen Textdatenbanken zum Gegenwartsdeutsch (Leipzig, DWDS, COSMAS II, COSMAS tagged, TIGER). Schritt für Schritt werden deren Abfragemöglichkeiten im Bereich von Lexikon, Morphologie und Syntax vorgestellt. Die in diesem Beitrag enthaltenen Informationen werden darüber hinaus in Form von vier – auch ohne Rückgriff auf den Fliesstext verständlichen – synoptischen Tabellen kondensiert.

Wenn es noch eines Beweises bedürfte, das die Korpuslinguistik – verstanden als linguistische Teildisziplin, die auf der Grundlage von digital gespeicherten Grosskorpora arbeitet – in voller Entwicklung begriffen ist, dann liefert diesen der Beitrag von **Gottfried Kolde**. Auf der Suche nach einer Einführung in die Korpuslinguistik wurde er nur im ausserdeutschen Sprachraum fündig. So entschloss er sich denn, seiner vergleichenden Analyse die englischsprachige Einführung von D. Biber et al. (1998) und die französischsprachige von B. Habert et al. (1997) zugrunde zu legen. Von diesen beiden erweist sich bei näherem Hinsehen allerdings bloss die erstgenannte als echtes, auch für Anfänger geeignetes Lehrmittel. Kurz vor Redaktionsschluss dieser Nummer von *Linguistik online* erschien dann aber die – von Kolde noch in Form eines Nachtrags berücksichtigte – deutschsprachige Einführung von L. Lemnitzer/H. Zinsmeister (2006), welche die Ansprüche, die an eine Einführung in diese Teildisziplin zu stellen sind, nun in hohem Masse erfüllt.

Der Beitrag von **Wolfgang Teubert**, Inhaber des Lehrstuhls für Korpuslinguistik an der Universität Birmingham, ist ein engagiertes Plädoyer für die Einbettung der Korpuslinguistik in den übergeordneten Rahmen der Hermeneutik. Zwar bleiben auch für Teubert Verfahren wie etwa die Kookkurrenzanalyse weiterhin unerlässliche Hilfsmittel; diese genügten aber für sich allein noch nicht für die

Bestimmung der Bedeutung von "Diskursobjekten". Was Begriffe wie *arrangierte Ehe* oder *Zwangsehe* bedeuten, lässt sich erst durch eine Beobachtung der Diskurse bestimmen, in welchen die Bedeutung dieser begrifflichen Konstrukte – in einer gegebenen Diskursgemeinschaft – ausgehandelt wird. Dies lässt sich mit exemplarischer Deutlichkeit für gesellschaftlich brisante Diskursobjekte nachzeichnen, deren Geltung in einem – grundsätzlich nicht abschliessbaren – polyphonen Strom von Definitionsversuchen, Paraphrasen und Stellungnahmen zur Debatte steht. Bedeutung gibt es somit nicht abstrakt und für sich abgehoben, sondern bloss als ein von den Sprachteilhabern ausgehandeltes soziales Konstrukt, wobei frühere Texte – in einem weit verzweigten Netz von intertextuellen Bezügen – ihre Spuren in späteren hinterlassen.

Raphael Bertheles Beitrag ist eine engagierte Replik auf den programmatischen Artikel von Teubert, in der bestimmte Prämissen und Konsequenzen von dessen Konzeption einer "hermeneutischen Korpuslinguistik" kritisch hinterfragt werden. Für Berthele gibt es keinen Grund, bei der Bedeutungsbestimmung auf die bewährten linguistischen Verfahren und Kategorien zu verzichten. Vielmehr plädiert er dafür, am Konzept einer – vom Kontext relativ unabhängigen – Bedeutung (Denotat) festzuhalten, da nicht alles, was in Diskursen ausgehandelt werde, zur Bedeutung gehöre. Nicht einverstanden ist Berthele ferner damit, dass bei der hermeneutischen Korpuslinguistik das Sprachsystem zu sehr in den Hintergrund gerückt wird, indem etwa der Begriff Wort durch die sozial konstruierte Grösse Diskursobjekt ersetzt wird. Überdies greife ein Verständnis von Texten als Reformulierung und Neuinterpretation von bereits Gesagtem und Geschriebenem zu kurz. Für Berthele ist die Korpuslinguistik im übrigen keine selbstständige Theorie, sondern bloss eine – allerdings sehr potente – neuartige Forschungsmethode, dank der in Zukunft Aussagen zu zentralen Fragen des Lexikons und der Grammatik (etwa zu den Kollokationen) auf eine völlig neue, empirisch abgesicherte Basis gestellt werden können.

Als Ergebnis von **Hans Bickels** methodisch innovativem Beitrag mit dem Titel *Das Internet als linguistisches Korpus* kann man festhalten, dass sich mit Hilfe von Internet-Suchmaschinen wie Altavista oder Google in der lexikologischen Forschung konsistente und reproduzierbare Resultate erzielen lassen. Trotz ständiger Variabilität verfügen diese Korpora, etwa was die Frequenzen betrifft, über eine erstaunliche lexikalische Stabilität, was mit ihrer textlichen Vielfalt und immensen Grösse zusammenhängt. Ausgehend von seinen Erfahrungen als Mitautor des *Variantenwörterbuchs des Deutschen* (2004) kann Bickel – mit Hilfe der sog. Internetdomains .de/.at/.ch – einen doppelten Nachweis erbringen: Zum einen treten regional nicht markierte Wörter in den drei deutschsprachigen Ländern ziemlich genau proportional zur Gesamtzahl der jeweiligen Internetseiten auf (grob gesagt: .de: 80%, .at: 10% und .ch: 10%). Zum andern lassen sich national markierte Varianten deutlich als solche verifizieren (z. B. .de: *Abiturient*, .at: *Maturant*, .ch: *Maturand*) bzw. falsifizieren. Von grossem Interesse ist dabei, dass mit diesem Instrumentarium erstmals auf verlässliche Weise auch der – negative – Nachweis erbracht werden kann, dass ein bestimmtes Wort in den andern Sprachgebieten nicht verwendet wird, dass z. B. in der Schweiz *Bostitch* gesagt wird, aber nicht *Tacker* oder *Hefter* (.de) oder *Klammerlmaschine* (.at).

Anton Näf zeigt in seinem Beitrag *Satzarten unterscheiden – Kann das der Computer?*, dass auch in einem Korpus wie COSMAS II, das ja weder mit einem Tagger noch mit einem Parser vorbehandelt wurde, syntaktische Fragestellungen erforscht werden können. Anhand des Beispiels der Satzarten wird das Verfahren der *Anfragezuspitzung* vorgestellt, mit dessen Hilfe sich syntaktische Phänomene durch eine gezielte Kombination von Suchbefehlen zwar zum Teil nicht exhaustiv erfassen, aber immerhin in ihren Konturen und Grössenordnungen einkreisen lassen. "Angriffshebel" sind dabei zum einen graphische Zeichen (Satzschlusszeichen, Grossbuchstaben usw.) und grammatische Funktions-

ISSN 1615-3014

wörter (z. B. *wie*), zum andern aber auch – etwa im Fall des für das Abfragewerkzeug von COSMAS II unverständlichen Begriffs Verb-Erststellung – "von Hand" erhobenes Vorwissen über die typische lexikalische Füllung von Strukturstellen. Die Ergiebigkeit eines solchen Zugriffs wird anhand von intonatorischen Minimalpaaren mit Verb-Erststellung (Interrogativsatz vs. Exklamativsatz: *War das eine gute Idee?* vs. *War das eine gute Idee!*) aufgezeigt. Bei den entsprechenden *wie*-Sätzen mit Zweitstellung zeigt sich bei zahlreichen Adjektiven eine deutliche Distributionspräferenz für den Exklamativsatz (*wie reizend, wie widerlich* usw.).

Zu den Funktionsverbgefügen gibt es seit gut vierzig Jahren eine umfangreiche Forschung. Diese hat sich aber überwiegend mit Abgrenzungs- und Definitionsfragen befasst und darüber schlicht die empirische Erhebung der sprachlichen Fakten vernachlässigt. Im Rahmen seiner 2006 an der Universität Neuchâtel verteidigten Dissertation hat nun **Alain Kamber** erstmals die häufigsten Funktionsverbgefüge auf einer breiten empirischen Grundlage (Analysekorpus von 5 Millionen und Kontrollkorpus von 60 Millionen Textwörtern) untersucht und nach ihren syntaktischen (Erweiterbarkeit, Komplementierbarkeit usw.) und semantischen Verwendungsweisen analysiert. Anhand des hochfrequenten Funktionsverbs *kommen* stellt Kamber in diesem Beitrag sein Vorgehen und seine Ergebnisse vor. Diese werden sodann mit den Einträgen in den gängigen Wörterbüchern und Grammatiken verglichen. Kambers Resultate sind auch von unmittelbar praktischem Nutzen, zum einen bei der Überarbeitung der ein- und zweisprachigen Wörterbücher, zum andern für die Praxis des DaF-Unterrichts (z. B. die nach Archilexemen geordneten Zusammenstellungen von (quasi)synonymen Funktionsverbgefügen).

Der Beitrag von **Rolf Duffner** zu den Satzadverbien im Gegenwartsdeutsch ist ein Werkstattbericht über eine breit angelegte empirische Studie zu diesem Phänomen. Anhand von zwei Beispielen, dem sehr frequenten *glücklicherweise* und dem selteneren *paradoxerweise* (immerhin noch 1296 Belege in COSMAS II) stellt Duffner das von ihm angewandte Verfahren zur Erstellung von Kookkurrenzprofilen vor. Diese sind das Gesamtresultat von mehrfachen Kookkurrenzanalysen zu einem Item bei jeweils veränderter Einstellung der Suchparameter. Ziel ist dabei die Eruierung typischer Verwendungskontexte (Kookkurrenzpartner, syntaktische Verkettungen usw.) dieser sprachlichen Einheiten (z. B. *Glücklicherweise wurde niemand verletzt*). Durch eine sog. Reziprokanalyse wird dann geprüft, ob auch bei umgekehrter Suchrichtung eine hohe Affinität besteht oder nicht, was etwa bei *verletzt* für *glücklicherweise* (sowie für das synonyme *zum Glück*) zutrifft. Während für *glücklicherweise* vor allem Autosematika als präferierte Kookurrenzpartner auftreten, sind es bei *paradoxerweise* in erster Linie Fokuspartikeln (*gerade, ausgerechnet* usw.). Die neu gewonnenen Erkenntnisse werden anschliessend mit den Angaben zu den Satzadverbien in Wörterbüchern konfrontiert.

Am Beispiel des Sprachenpaars Ungarisch-Deutsch bietet **Anna Reders** Beitrag zunächst einen Forschungsüberblick zum Begriff Kollokation sowie im Anschluss daran eine wohl durchdachte kontrastive Typologie zur Klassifizierung von Kollokationen. Zum andern berichtet die Autorin über eine empirische Studie zum Transferverhalten ungarischer Deutschlerner (mit insgesamt 3 x 200 Versuchspersonen aus drei Lernniveaus). Bei der Auswertung der Resultate erweist sich neben der Lernstufe auch die Art der Kollokation als relevante unabhängige Variable: Die Probanden der beiden höheren Lernstufen produzieren bei metaphorischen Kollokatoren deutlich seltener durch negativen L1-Transfer bedingte Fehler als bei den nicht metaphorischen, während sich in der ersten Lernstufe kein solcher Unterschied manifestiert. Zum Schluss macht Reder Vorschläge für eine verbesserte Kollokationsdidaktik, welche dazu beitragen soll, dass ungarische Deutschlerner Ketten wie *jemandem ins Wort <schneiden>* (statt *fallen*) vermeiden lernen.

Neuchâtel, 23. Oktober 2006 Anton Näf / Rolf Duffner

ISSN 1615-3014