Kapitel 7 aus der 1. und 2. Auflage von Lemnitzer & Zinsmeister (2006/2010): Korpuslinguistik. Eine Einführung. Tübingen: Narr. In der 3., erweiterten Auflage des Buches (2015) nicht mehr enthalten.

Man muss sich mit einem Korpus anfreunden – Erfahrungen von Linguisten mit Korpora

Im Folgenden präsentieren wir die Antworten einiger Linguisten auf vier Fragen, die wir im Interview stellten. Die Fragen beziehen sich auf die Erfahrungen, die die Befragten bei ihrer linguistischen Forschung mit Korpora gemacht haben, und auf Empfehlungen, die sie Neulingen der Korpuslinguistik geben. Einige der Befragten haben per Mail geantwortet, einige haben sich zu einem mündlichen Interview bereit erklärt. Dies ist einer der Gründe, warum die Antworten unterschiedlich lang ausgefallen sind. Wir haben in den meisten Fällen die Antworten gekürzt, da sie sonst den Rahmen dieses Kapitels gesprengt hätten.

Wir haben versucht, Interviewpartner zu gewinnen, die Experten auf mindestens einem der mit der Korpuslinguistik verbundenen Gebiete sind, nämlich theoretische Linguistik, Lexikographie, Computerlinguistik und Sprachlehre. Dass die Biographien der meisten Interviewten in irgendeiner Weise mit Tübingen verbunden sind, hängt sicher auch mit der wissenschaftlichen Biographie der Autoren dieses Buches zusammen.

Wir stellen zunächst die befragten Experten vor und fassen im Anschluss daran alle Antworten zu jeweils einer Frage in einem Abschnitt zusammen. Wir haben die Beiträge dabei alphabetisch geordnet.

1 Porträts

Julia Berman Studium Deutsch als Fremdsprache, der Germanistischen Linguistik und der Romanistik an der Universität München, anschließend Wissenschaftliche Angestellte am Institut für Maschinelle Sprachverarbeitung und Mitglied im Graduiertenkolleg *Linguistische Grundlagen der Sprachverarbeitung* in Stuttgart. Promotion im Bereich der Syntax des Deutschen im Jahr 2000. Zur Zeit Studienrätin im Hochschuldienst am Germanistischen Institut der Ruhr-Universität Bochum, Lehrstuhl für Germanistische Linguistik. Mitverfasserin eines Arbeitsbuchs zur deutschen Syntax, erschienen im Gunter Narr Verlag.

Sabine Braun Senior Lecturer for Translation Studies an der Universität Surrey, UK. Zuvor wissenschaftliche Angestellte am Seminar für Englische Philologie der Universität Tübingen. Promotion in englischer Linguistik. Forschung u.a. im Bereich der Verwendung von Korpora für die Sprachlehre (*ELISA: English-Language Interview Corpus as a Second-Language Application*).

Veronika Ehrich promovierte an der Universität Bielefeld; Habilitation an der Universität zu Köln im Bereich der germanistischen Linguistik. Aufenthalte als Gastwissenschaftlerin am Max-Planck-Institut in Nijmegen und an der Universität Berkeley. Seit 1993 Professorin für germanistische Linguistik an der Universität Tübingen. Projektleiterin im Sonderforschungsbereich *Linguistische Datenstrukturen*.

Christiane Fellbaum promovierte an der Princeton University, wo sie jetzt als Linguistin und kognitive Wissenschaftlerin als *Senior Research Psychologist* forscht. Leitende Entwicklerin des *Princeton WordNet* einer lexikalisch-semantischen Ressource des Englischen. Sie erhielt 2001 den Wolfgang-Paul-Preis der Humboldt-Stiftung.

Erhard Hinrichs Promotion (1985) auf dem Gebiet der Formalen Semantik an der Ohio State University. Research Scientist bei Bolt Beranek and Newman Laboratories, Cambridge, Mass. (1985-1987). Assistant Professor am Department of Linguistics, University of Illinois at Urbana-Champaign (1987-1990). Professor für Allgemeine Sprachwissenschaft und Computerlinguistik an der Universität Tübingen (seit 1991). Gastprofessuren an der Universität des Saarlandes, Saarbrücken (1988/89), bei NTT Laboratories, Yokosuka (1993/94) und an der Ohio State University (seit 1997). Seit 1998 ist er Projektleiter für große Annotationsprojekte zum Gegenwartsdeutschen, u.a. der Tübinger Baumbanken für gesprochene Sprache und für Schriftsprache.

Annette Klosa Promotion an der Universität Bamberg im Bereich der germanistischen Linguistik, verschiedene Tätigkeiten an den Universtitäten Bamberg und Sewannee in Tennessee. 1996-2001 Redakteurin in der Dudenredaktion. Zur Zeit wissenschaftliche Mitarbeiterin in der Abteilung Lexik des Instituts für deutsche Sprache in Mannheim, Leiterin des Projekts *elexiko*.

Anke Lüdeling promovierte 1998 in Tübingen über deutsche Partikelverben; danach arbeitete sie als wissenschaftliche Mitarbeiterin am Institut für maschinelle Sprachverarbeitung an der Universität Stuttgart und am Institut für Kognitionswissenschaft an der Universität Osnabrück. Seit 2002 Juniorprofessorin und seit 2009 Professorin für Korpuslinguistik an der Humboldt Universität zu Berlin. Sie baut dort u.a. ein fehlerannotiertes Korpus mit Beiträgen von Lernern des Deutschen als Fremdsprache auf.

Karin Pittner Studium der Anglistik und Germanistik in München. Wissenschaftliche Mitarbeiterin am Institut für Deutsche Philologie der Ludwig-Maximilians-Universität München. Habilitation 1997. Seit 1999 Professorin für Germanistische Linguistik an der Ruhr-Universität Bochum. Mitverfasserin eines Arbeitsbuchs zur deutschen Syntax, erschienen im Gunter Narr Verlag.

Marga Reis Promotion und Habilitation in der germanistischen Linguistik an der Universität München. Seit 1985 Professorin für germanistische Linguistik an der Universität Tübingen. Emeritiert seit 2008.

Angelika Storrer Dissertation zur Verbvalenz an der Universität Heidelberg. Forscherin am Institut für wissensbasierte Systeme der IBM, am Seminar für Srachwissenschaft der Universität Tübingen und am Institut für deutsche Sprache in Mannheim, in den Bereichen Grammatik und Lexik des Deutschen. Seit 2002 Professorin für Linguistik der deutschen Sprache und Sprachdidaktik an der Universität Dortmund.

2 Fragen und Antworten

Bei welchen linguistischen Untersuchungen haben Sie Korpora verwendet?

Berman: In meiner Dissertation zur Frage der Stellung der Komplementsätze und in einer Arbeit zu Prädikativen.

Braun: Vor einigen Jahren haben wir angefangen, ein kleines Korpus für Sprachlernzwecke zu entwickeln. Das *ELISA* Korpus ist eine Sammlung von Interviews mit englischen Muttersprachlern, die auf Video aufgezeichnet und transkribiert wurden. Die Sprecher berichten alle über ihr Berufsleben und über ihre berufliche Entwicklung und geben Beispiele aus ihrer Projektarbeit. Sie reden also alle etwa über das gleiche Thema. Wir haben allerdings in den Interviews die Fragen nur angerissen und haben die Leute dann reden lassen, um einen möglichst natürlichen Redefluss zu bekommen. Das Interessante an diesem Korpus für Sprachenlerner ist, dass es thematisch homogen ist, aber trotzdem lebendig und z.T. auch witzig. Die Interviews werden auch dadurch attraktiv, dass einige ungewöhnliche Berufe dabei sind, z.B. der Direktor eines tropischen Zoos. Dennoch geht es in allen Texten letztlich um Professional English, das für das Sprachenlernen heute in vielen Bereichen relevant ist. Außerdem sind viele Varietäten des Englischen vertreten (Australier, Engländer, Amerikaner, Iren, Schotten). Für einen Lerner ist es interessant, sich die Interviews im Ganzen anzuschauen, um überhaupt zu wissen, worum es da geht, oder sich dann mit einzelnen Teilen aus verschiedenen Interviews zu beschäftigen. Durch die Ähnlichkeit der Interviews können die Lerner z.B. vergleichen, wie verschiedene Leute über ihre Projektarbeit oder

über ihre Probleme bei der Arbeit reden. Dieser Materialreichtum ist interessant und lehrreich. Das kann dann ergänzt werden durch die Arbeit mit Wortlisten, Konkordanzen und weiterführenden Übungsmaterialien, so dass man lernt, welche Wörter und Phrasen wirklich häufig verwendet werden und wie und in welchem Kontext.

Ehrich: Ich habe Korpora vor allen Dingen bei Spracherwerbsuntersuchungen verwendet. Da kann man nur zwei Methoden anwenden: entweder Korpora oder Experimente. Ich habe bei meiner eigenen Arbeit Korpusuntersuchungen gemacht. Ich habe außerdem Korpora bei Untersuchungen zu Modalverben verwendet, und zwar die COSMAS Korpora.

Fellbaum: Das Projekt *Kollokationen im Deutschen des 20. Jahrhunderts* basiert ausschließlich auf Korpusdaten. Ziel des Projekts ist die empirische Erforschung der linguistischen Eigenschaften deutscher Idiome auf breiter Basis.

Hinrichs: Ich habe Korpora sowohl bei eher linguistischen Arbeiten als auch bei computerlinguistischen Untersuchungen genutzt. 1980 habe ich zum ersten Mal ein Korpus für eine linguistische Untersuchung verwendet, also lange bevor Korpuslinguistik wieder salonfähig wurde. Ich habe damals eine Arbeit geschrieben zur Semantik temporaler Anaphora. Es ging konkret darum, wie sich Zeitstrukturen in narrativen Texten abbilden. Ich habe relativ schnell eingesehen, dass man mit konstruierten Beispielen nur eine sehr begrenzte Datenmenge bekommt und habe mir deswegen damals sehr viele Krimis angeschafft. Bei Krimis spielen eben die Reihenfolge von Ereignissen und deren Auswirkungen eine viel größere Rolle als bei anderen belletristischen Texten. Es ist mir dann gelungen, interessante Abschnitte zu finden in Bezug auf temporale Anaphora. Diese Arbeit ist damals relativ viel zitiert worden, u.a. sicher darum, weil dort eine Korpusuntersuchung gemacht worden ist, auch wenn das eine Art von Korpus war, die man heute eher belächeln würde. Seitdem bin ich ein großer Fan von korpuslinguistischen Untersuchungen vor allem auch für theoretische Fragestellungen.

In jüngster Zeit habe ich Korpora unter zwei Forschungsperspektiven untersucht, und zwar einmal, wie sich große Datenmengen z.B. nach Wortklassen automatisch annotieren lassen – das sind Untersuchungen, die ich im SFB 441 gemacht habe – und wie sich auf Grund von getaggten Texten dann auch syntaktische Informationen automatisch generieren lassen. Für die letztgenannte Fragestellung verwenden wir bereits linguistisch annotierte Korpora. Ich habe vor allem ein Interesse an maschinellen, überwachten Lernverfahren, bei denen es notwendig ist, dass man linguistisch annotierte Daten in großer Menge vorrätig hat. Das jüngste Beispiel einer solchen Untersuchung ist die Verwendung von memory-basierten Lernverfahren und hybriden, statistischen und regelbasierten, Verfahren, um Anaphern aufzulösen. Diese Untersuchung, die ich zusammen mit Ekaterina Filipowa und Holger

Wunsch mache, dauert noch an. In der letzten Phase des SFB habe ich mich mit morphologischer Disambiguierung beschäftigt, d.h. wie sich Kasus- und Numerusinformationen automatisch kontextuell auflösen lassen. Diese Untersuchungen habe ich zusammen mit Julia Trushkina durchgeführt. Da ging es um die Verbindung von probabilistischen kontextfreien Grammatiken und regelbasierten Constraint-Grammatiken.

Lüdeling: Ich habe Korpora sowohl für qualitative als auch für quantitative morphologische Untersuchungen verwendet. Ich habe angefangen, mit Korpora zu arbeiten, als ich im Verbmobil-Projekt tätig war – damals habe ich anhand von englischen und deutschen Korpora Übersetzungsregularitäten gesucht und Transferregeln geschrieben. In einem späteren Projekt (*DeKo*) habe ich vor allem mit qualitativen Mitteln versucht, Wortbildungsregularitäten und bestimmte Wortbildungsmuster zu finden. Quantitative Untersuchungen bezogen sich auf Produktivitätsuntersuchungen (*Wie produktiv ist ein Wortbildungsmuster?*). Produktivität und Wortbildung interessieren mich auch heute noch sehr. Weiter verwende ich historische Korpora bei der Untersuchung von Ähnlichkeiten zwischen den historischen Stufen des Deutschen mit bioinformatischen Methoden. Das Projekt ist noch ziemlich neu und wir stehen noch am Anfang. Dann sind wir gerade dabei, ein Lernerkorpus aufzubauen, in dem man Daten von fortgeschrittenen Lernern des Deutschen als Fremdsprache findet. Auch hier arbeiten wir qualitativ (wir entwickeln Tagsets für ‚Fehler' oder Abweichungen auf allen linguistischen Ebenen und quantitativ (wir vergleichen unsere Lernerdaten statistisch mit entsprechend erhobenen Muttersprachlerdaten).

Klosa: Elektronische Korpora des geschriebenen Deutsch benutze ich seit einigen Jahren zur Beantwortung von wortbezogenen Sprachanfragen durch Laien, zur Klärung von Fragen der Wortbildung und hauptsächlich als Grundlage der Erarbeitung von Wortartikeln für ein hypertextuelles Informationssystem zur deutschen Gegenwartssprache (dies genauer zu erläutern, würde hier allerdings den Rahmen sprengen, weswegen auf `www.elexiko.de` hingewiesen sei). Der „Griff" zu Korpora ist immer dadurch motiviert, dass ich aus der Korpusrecherche und -analyse Ergebnisse erwarten kann, die den tatsächlichen Sprachgebrauch (wie er im jeweiligen Korpus festgehalten ist) dokumentieren und die zumindest zum Teil über das hinausgehen, was schon in Wörterbüchern, Grammatiken, Sprachratgebern, Wortbildungslehren etc. festgehalten ist. Ein Beispiel: Eine typische Sprachberatungsanfrage zielte darauf zu erfahren, ob in einem bestimmten Kontext das Nomen *Brache* korrekt verwendet wurde. Hierüber war es zu einem Streit zwischen einem Landwirt und dem zuständigen Amt für Landwirtschaft gekommen. Eine Überprüfung der Wörterbucheinträge zu Brache ergab, dass hiermit ein *brachliegendes Feld, Land* bezeichnet wird, wie die *Zeit, während der der Acker brachliegt*. Eine Recherche in den Korpora geschriebener Sprache des In-

stituts für Deutsche Sprache (IDS, Mannheim) ergab darüber hinaus, dass brachliegende Stücke Land nicht nur im landwirtschaftlichen Kontext thematisiert werden, sondern das Wort Brache auch für unbebaute, nur noch mit Schutt, Ruinen und wild wachsenden Pflanzen bedeckte innerstädtische Stücke von Land verwendet wird. In diesem Sinne konnte ich dann auch dem Landwirt bei seinem Streit helfen. Die systematische Auswertung der Suchergebnisse aus einer Korpusanalyse kann beispielweise auch Aufschluss über neue Wortbildungsmuster, die gerade erst entstehen bzw. ausgebaut werden, geben. Verben mit *zwischen-* sind hierfür ein gutes Beispiel (z. B. *zwischenlagern, zwischenlanden, zwischenschalten, zwischenfinanzieren*). Diese sind zwar z.T. schon in gegenwartssprachlichen Wörterbüchern erfasst, das Wortbildungsmuster ist als solches von der Wortbildungsforschung aber noch nicht beschrieben worden. In den IDS-Korpora der geschriebenen Sprache finden sich über 40 *zwischen*-Verben, die allerdings, wie eine chronologische Sortierung der Belegstellen zeigt, erst seit den 1980er Jahren vorkommen. Es handelt sich bei diesem Muster offensichtlich um ein noch junges, was auch die fehlende Beschreibung in Darstellungen zur verbalen Wortbildung erklärt.

Pittner: Ich habe Korpora zur Untersuchung von freien Relativsätzen verwendet. Mein Ziel war herauszubekommen, wie häufig der Kasus des Relativums von dem vom Matrixverb geforderten Kasus abweicht und ob diese Abweichungen einer bestimmten Regularität folgen. Zum anderen habe ich neulich in einer kleinen Studie untersucht, in welchen Funktionen *wo* in Relativsätzen auftritt. Hierzu verwendete ich vom IDS bereitgestellte Korpora, und zwar ein Zeitungskorpus und das Pfeffer-Korpus, da es sich hierbei um ein Korpus gesprochener Sprache mit Sprechern aus verschiedenen Regionen handelt, so dass sich regionale Unterschiede und Unterschiede zwischen gesprochener und geschriebener Sprache aufzeigen ließen. Gerade im Entstehen ist eine Studie zu adjektivischen Eigenschaften prädikativer Nomina, für die ich das World Wide Web als Korpus heranziehe.

Reis: Ich habe extensiv bei zwei Untersuchungen Korpora verwendet, und zwar einmal bei einer kritischen Nachuntersuchung zur Verteilung von Korrelat-*es*, da wollte ich nachprüfen, was Bengt Sandberg (Sandberg, 1998) gemacht hat, ein weiteres Mal bei Untersuchungen zum Gebrauch von *drohen* und *versprechen*. Es ist bei beiden kohärenten Verben notwendig, herauszufinden, inwieweit die dritte Konstruktion verbreitet ist. Die Daten habe ich nicht selbst erhoben, sondern von Mitarbeitern aus den IDS-Korpora extrahieren lassen. Ich habe sie aber selbst ausgewertet.

Storrer: Ich führe corpusgestützte Untersuchungen zu deutschen Nominalisierungsverbgefügen (NVG) durch, also zu Konstruktionen wie *in Verbindung bringen*, *Unterricht erteilen*. Dabei interessiert mich einerseits das Verhältnis zwischen NVGs und semantisch äquivalenten Basisverbkonstruktio-

nen, also z.B. zwischen *Unterricht erteilen* vs. *unterrichten*. In einer Vergleichsstudie habe ich untersucht, in welchen Fällen die beiden Konstruktionstypen tatsächlich wechselseitig paraphrasierbar sind, und welche Leistungen sie jeweils für die Textkohärenz und Informationsstrukturierung spielen. Andererseits habe ich bestimmte Annahmen über die Systematik der NVG evaluiert, u.a. die morphosyntaktische Variabilität der nominalen Bestandteile (*Unterricht*, *in Verbindung*), deren Pronominalisierbarkeit und deren Modifizierbarkeit durch Adjektivattribute oder Relativsätze.

Weiterhin nutzen wir Korpora zur empirischen Erforschung von sprachlichen Besonderheiten in der internetbasierten Kommunikation (E-Mail, Foren, Chat, Weblogs etc.). Der Schwerpunkt in meinem Dortmunder Arbeitsbereich liegt auf der Erforschung der Chat-Technologie in verschiedenen Anwendungskontexten (Freizeitbereich, Medien, Bildung + Weiterbildung). Meine Arbeitsgruppe hat dazu ein Korpus aufgebaut, das Chat-Logfiles aus verschiedenen Bereichen umfasst: universitäre Online-Seminare; Interviews per Chat; eventbegleitende Chats (z.B. *InterNetzer*), Beratungs-Chats (psychosoziale Beratung, Bibliotheks-Hotline, Beratung zur eBay-Nutzung) und themenunspezifische „Plauderchats“ . Die Logfiles sind in ihrer Struktur XML-annotiert und mit Metadaten (Chat-Typ, verwendetes Werkzeug, Informationen über das Geschlecht der Chatter etc.) versehen. Zu seiner Auswertung wurde ein darauf spezialisiertes Such- und Auswertungswerkzeug entwickelt. Auf dieser Basis untersuchen wir einerseits, wie häufig die als typisch für Chats geltenden Sprachmerkmale, also z.B. Emotikons, Abkürzungen wie **lol**, Asterisk-Ausdrücke wie **heul**, konzeptionell mündliche Elemente wie *haste*, *biste*, *net*, tatsächlich in den Logfiles der verschiedenen Anwendungsbereiche vorkommen. Andererseits untersuchen wir Probleme der Kohärenzbildung und Sprachhandlungskoordination, die auf die spezielle Situation im Chat (fehlende Wahrnehmbarkeit von Produktionshandlungen) zurückgehen. Wir typisieren Typen von Problembereichen (Referenz, Deixis, Überkreuzung von Adjazenzmustern) und untersuchen, wie die Chatter damit umgehen bzw. mit welchen Strategien sich solche Probleme ggf. vermeiden lassen.

Was hat Sie bei ihrer linguistischen Arbeit zur Verwendung von Korpora motiviert?

Berman: Die Suche nach Belegen, um theoretische Annahmen widerlegen zu können. So lautete z.B. eine Annahme, dass Gradpartikel nicht vor linksversetzten Ausdrücken stehen können. Dies hätte meiner Analyse widersprochen. Daher brauchte ich dringend authentische Gegenbeispiele, die ich auch problemlos gefunden habe. Allgemein: Der Wunsch, Belege zu finden, wenn die Daten sehr umstritten sind, und die Erfahrung, dass ich durch die

Korpora auf Phänomene gestoßen bin, auf die ich selbst gar nicht gekommen wäre.

Braun: Unser Interesse an Korpora ist eingebettet in die empirische Sprachlernforschung zum Zweitsprachenerwerb. Die Sprachlernforschung orientiert sich stark an konstruktivistischen, kognitiven und kommunikativen Lerntheorien. Diese Ansätze fordern, dass man im Sprachunterricht mit authentischen Materialien arbeitet und dass man Lerner möglichst zum autonomen Lernen befähigt. Für diese beiden Ziele eignet sich der Einsatz von Korpora sehr gut. Korpora sind eine wertvolle Quelle authentischer Sprachdaten und bieten für Sprachenlerner den Vorteil, dass man Sprachverwendungsmuster viel deutlicher sieht als in einzelnen Texten. Korpuslinguistische Methoden fördern induktives und autonomes Lernen. Außerdem können sie die Texterschließung besser unterstützen. Es ist ganz was anderes, ob ich einen Text einfach nur so lese oder ob ich ihn mir über Frequenzlisten und Konkordanzen weiter erschließe.

Ehrich: Bei den Kohärenzeigenschaften der Modalverben ist es so, dass diese skopusambig sind. Und ich wollte gerne wissen, ob diese Skopusambiguität, die in der Grammatik einfach unzweifelhaft besteht, sich in realen Daten wiederauffinden lässt. Ich muss leider sagen, dass, wie die realen Daten zeigen, es eine solch klare Präferenz für den weiten Skopus der Negation gibt, dass letzten Endes mit der Skopusambiguität der Modalverben nicht gut argumentieren ist, wenn man erstmal reale Daten anschaut. Allerdings habe ich hier nur geschriebene Daten verwendet. Ich bin immer noch davon überzeugt, dass der Skopus ambig ist, dass man dann aber die Intonation in die Untersuchung mit einbeziehen müsste. Dazu bräuchte man aber bessere Korpusdaten, wenigstens mit einer Audiospur.

Fellbaum: Ich glaube nicht, dass man heute noch linguistische Theorien bauen und vertreten kann, ohne die relevanten Phänomene in Korpora untersucht zu haben (das heißt natürlich nicht, dass *alle* Korpusdaten gleichwertig sind). Zweitens ging es bei meinem Thema – Kollokationen – darum, die unzulänglichen Darstellungen in Papierlexika durch eine digitale lexikalische Ressource, in der jeder Eintrag mit Korpusdaten verlinkt ist, zu ergänzen oder zu ersetzen.

Hinrichs: Es gibt zwei Motivationen hierfür. Bei rein linguistischen Arbeiten ist das die Einsicht, dass als Datentyp intuitive Sprecherurteile häufig nicht ausreichen, vor allem auch für semantische Fragestellungen, wo es weniger darum geht, ob ein Satz etwa wohlgeformt ist, sondern eher darum, wie ein Satz im Kontext zu interpretieren ist. Deswegen muss man ohnehin über die klassische Satzgrenze auf die Diskursebene übergehen. Wenn man dann sowieso bei Texten ist, dann liegt eine korpuslinguistische Untersuchung nahe, weil man noch stärker als bei syntaktischen Untersuchungen bei der Generierung von Daten nach der eigenen linguistischen Intuition an

die Grenze stößt. Man neigt ja dazu, nur solche Beispiele als Artefakte zu generieren, die zur eigenen Theorie passen. Um aus diesem Teufelskreis herauszukommen, bietet sich die Verwendung von Korpora einfach an.

Bei computerlinguistischen Untersuchungen ist es so, dass man ohnehin auf annotierte Korpora angewiesen ist, wenn man datenintensive Methoden verwenden will, also z.B. stochastische Verfahren und maschinelle Lernverfahren. Das gehört zur Methode von überwachten Lernverfahren und ist insofern nicht nur eine Frage der Motivation, sondern auch der Notwendigkeit, die sich aus der gewählten Methode ergibt.

Lüdeling: Die Fragestellungen sind bei mir zunächst theoretischer Natur, also zum Beispiel, wie Wortbildungsprozesse ablaufen. Da habe ich Korpora zunächst mal als Beispielsfundus verwendet, um darin Dinge zu finden, auf die ich selber nicht gekommen wäre. Ich habe die Daten überwiegend als Inspirationsquelle genutzt, ansonsten aber auf meine Intuition oder Grammatikalitätsurteile vertraut. Bei solchen Untersuchungen sortiere ich Daten aus, wenn ich denke, dass sie falsch oder unbrauchbar sind. Dann habe ich an der auch aus der theoretischen Linguistik kommenden Frage der Produktivität von Wortbildungseinheiten geforscht. Dafür gibt es außer Korpora kaum Quellen, denn ich brauche hier quantitative Daten und darüber hat man keine Intuition. Hier arbeite ich viel strikter oder im eigentlichen Sinne korpuslinguistisch und nehme alle Daten so an, wie ich sie finde (wie die Daten interpretiert werden, ist dann eine andere Sache). Für Untersuchungen, die sich auf historische Korpora und Lernerkorpora stützen, habe ich natürlich auch keine intuitiven Daten. Historische Sprachformen sind nicht anders zugänglich als in Korpusform und Lernerdaten kann ich auch nur bekommen, wenn ich Lerner befrage, d.h. in den beiden Fällen steht mir die Intuition als Datenquelle nicht zur Verfügung. Bei Lernern könnte man sicher noch mit psycholinguistischen Experimenten arbeiten.

Pittner: Die interessantesten Ergebnisse ergab bisher meine Untersuchung zu den freien Relativsätzen. Ich hatte angefangen, einzelne Belege zu freien Relativsätzen (und zu einer ganzen Reihe von anderen sprachlichen Phänomenen) zu sammeln. Im Fall der freien Relativsätze zeigte sich dann schnell, dass die Belege nicht den in den deutschen Grammatiken postulierten Regeln folgten. Dort wurde behauptet, dass der Kasus des Relativpronomens bei freien Relativsätzen stets der Kasusform entsprechen muss, die der Matrixsatz für die vom freien Relativsatz gefüllte Stelle verlangt. Die Belege für Abweichungen von diesen Regeln häuften sich. Ich befragte dazu auch Sprecher/innen, wie sie diese Belege beurteilten und sie wurden zum Teil nicht schlechter beurteilt als „regelkonforme“ freie Relativsätze. Es zeigte sich jedoch auch, dass bestimmte Abweichungen als deutlich besser beurteilt wurden als andere. In der Literatur zu freien Relativsätzen in anderen Sprachen stieß ich dann auf die Aussage, dass die Abweichungen einer Kasushier-

archie folgen. Es war sehr spannend zu sehen, dass die Abweichungen, die einer Kasushierarchie folgen, auch von den Sprecher/innen des Deutschen als wesentlich besser beurteilt wurden, als solche, die dies nicht tun. Außerdem fand ich keine Belege für Abweichungen, die nicht der Kasushierarchie genügen. Das Deutsche schien hier also genau den Regeln der Kasushierarchie zu folgen, die auch in einigen anderen Sprachen wirksam ist. Es stellte sich aber nun die Frage, ob es denn gar keine Ausnahmen zu dieser Hierarchie gibt, und zudem auch die grundlegende Frage, wie oft die Abweichungen von den in den Grammatiken formulierten Regeln auftreten. Handelt es sich dabei eher um ein marginales Phänomen, das man vernachlässigen kann, oder sind sie doch ziemlich häufig? Dies kann man anhand einer Belegsammlung nicht entscheiden. Es war also notwendig, systematisch die Häufigkeit von abweichenden und nicht abweichenden freien Relativsätzen zu untersuchen. Als Korpora wählte ich ein vom Institut für Deutsche Sprache in Mannheim zur Verfügung gestelltes Zeitungs- und Belletristik-Korpus sowie ein Korpus philosophischer Texte, das im Rahmen der digitalen Bibliothek erschienen ist. Diese CD ist zwar nicht als Korpus für linguistische Untersuchungen zusammengestellt, die bereitgestellten Suchwerkzeuge erwiesen sich jedoch für meine Fragestellung als ausreichend. Es zeigte sich, dass Abweichungen in manchen Korpora häufiger auftraten als in anderen. Die philosophischen Texte wiesen kaum Abweichungen auf. Insgesamt zeigte sich, dass es sich dabei um ein sehr häufiges Phänomen handelt, das man nicht einfach als Fehler wegerklären kann. Beispiele für Abweichungen, die nicht im Einklang mit der Kasushierarchie waren, traten praktisch nicht auf. Inzwischen werden die Abweichungen sowie die dabei wirksame Kasushierarchie auch von einigen neueren Grammatiken anerkannt. Das war für mich das Spannendste: zu sehen, dass Grammatiken bei einigen Phänomenen eher von normativen oder tradierten Auffassungen ausgehen, die wahrscheinlich noch nie auf ihre empirische Fundierung hin untersucht worden sind. Auch bei gut dokumentierten Sprachen wie dem Deutschen gibt es also für Korpuslinguist/inn/en Einiges zu entdecken. Die Überprüfung von Theorien und Hypothesen am konkreten sprachlichen Material halte ich für eine der spannendsten Aufgaben von Linguist/inn/en.

Reis: Bei der Untersuchung zu *drohen* und *versprechen* hat mich motiviert, dass man ohne authentische Belege niemanden überzeugen kann. Ich hatte zwei Thesen – erstens die, dass es sich eigentlich nicht um modale Varianten handelt. Dazu musste ich nachweisen können, dass es infinite Belege gibt. Zweitens wollte ich etwas Historisches herausfinden. Ich wollte eigentlich zeigen, dass die sogenannte dritte Konstruktion erst spät entstanden ist. Leider haben dann die Vergleichskorpora aus dem 19. Jahrhundert gefehlt, also konnte ich das nicht zeigen. Das ist eines der größten Probleme, dass es

keine historischen Vergleichskorpora, also z.B. Korpora von Zeitungstexten aus dem 19. Jahrhundert, gibt[1].

Bei der *es*-Arbeit ist der zentrale Punkt, dass immer unterschätzt wird, dass Korpusdaten Produktionsdaten sind und Grammatikalitätsurteile Rezeptionsdaten. Das sind beides komplexe kognitive Operationen ganz unterschiedlicher Art. Ich wollte die *es*-Arbeit von Sandberg, die korpusbasiert ist, nachprüfen und vergleichen mit Studien über Grammatikalitätsurteile.

Storrer: Zunächst der Wunsch nach größerer Beschreibungsadäquatheit. In den 1980er Jahren war ich am damaligen Forschungszentrum der IBM in Heidelberg am Aufbau einer Datenbank zur Valenz deutscher Verben beteiligt. Grundlage waren die Testsets aus dem Valenzwörterbuch von Helbig und Schenkel. Immer wieder ist es mir passiert, dass ich bei der abendlichen Zeitungslektüre auf einen Satz gestoßen bin, den ich am Nachmittag durch meine Valenzbestimmung als ungrammatisch ausgeschlossen hätte, der im gegebenen Kontext grammatisch aber völlig unauffällig war. Ich habe in diesem Zuge insgesamt eine gewisse Skepsis gegen den alleinigen Einsatz von linguistischen Tests entwickelt, allerdings waren damals noch kaum digitale oder gar linguistisch aufbereitete Korpora für die deutsche Sprache verfügbar. Aus meiner Sicht ist es eine große Bereicherung für die linguistische Forschung, dass inzwischen mehrere digitale Korpora geschriebener und gesprochener deutscher Sprache mit entsprechenden Werkzeugen kostenfrei verfügbar sind und auch die Menge linguistisch annotierter Daten wächst. In vielen Bereichen lassen sich dadurch sprachliche Phänomene besser beschreiben und systematisieren; besser in dem Sinne, dass die Beschreibungen das Verhalten der entsprechenden Einheiten in authentischen Texten adäquater wiedergeben, dass auch quantitative Aspekte berücksichtigt werden können, sodass typische/gebräuchliche von weniger typischen/ungewöhnlichen Strukturen unterschieden werden können.

Viele satzübergreifende Phänomene (Textkohäsion und -kohärenz, Informationsstrukturierung, thematische und rhetorische Strukturierung) lassen sich überhaupt erst sinnvoll an Korpusdaten untersuchen. Z.B. lässt sich die Frage, worin sich die Verwendungsmöglichkeiten von Nominalisierungsverbgefügen (*Unterricht erteilen, Hilfe leisten*) von denen scheinbar synonymer Basisverbkonstruktionen (*unterrichten, helfen*) unterscheiden, am besten an Verwendungen beider Konstruktionstypen in authentischen Textzusammenhängen beantworten.

[1] Abhilfe könnte das Projekt Deutsches Textarchiv (`www.deutsches-textarchiv.de`) schaffen. Dieses Archiv soll bedeutende und einflussreiche Texte aus der Zeit von 1650 bis 1900 enthalten. Dieses von der DFG geförderte Projekt hat gerade seine erste Phase abgeschlossen.

Gab es positive oder negative Überraschungen bei Ihrer Arbeit mit Korpora?

Berman: Positiv: Ich habe oft überraschende Daten gefunden. Negativ: Die Anfragesprache ist bei manchen Korpora zu kompliziert bzw. zu wenig dokumentiert. Man braucht zuviel Zeit und Geduld, die Anfragesprache zu verstehen und anwenden zu können.

Braun: Bei einigen Interviews haben wir Webseiten der Institution gefunden, aus der die Sprecher stammen. Die Sprecher erzählen oft das über ihre Firma oder ihr Unternehmen, was bekannt und veröffentlicht ist, und solche Informationen finden sich dann ja auch auf den Webseiten dieser Firmen wieder. So kann man sehr gut vergleichen zwischen der gesprochenen Variante und den Texten auf der Webseite. Hier war ich dann selbst überrascht, obwohl mir die Unterschiede natürlich bewusst sind, wie stark diese Unterschiede sind – und das selbst bei Sprechern, von denen man beim Zuhören den Eindruck hat, dass sie sozusagen „wie gedruckt" reden. Trotzdem verwenden sie viele Intensifier und Softener wie *really* oder *just* und auch jede Menge vager Ausdrücke von *lots of* bis *sort of*. Ich fand es sehr bemerkenswert, dass man das bereits in einem kleinen Korpus in dieser Deutlichkeit sieht. Solche einfachen Vergleiche kann man auch mit Schülern gut machen.

Ehrich: Das Technische an der Korpusarbeit kann ich nicht so gut kommentieren, weil ich zwar mit Korpusdaten gearbeitet habe, aber mir die Rohdaten von meinen Studierenden, von meinen Hilfskräften habe rausfiltern lassen. Ich habe nur die *CHILDES*-Daten technisch selbst gehandelt und das *CHILDES-Korpus* ist sehr benutzerfreundlich. Die inhaltlichen Überraschungen sind immer enorm, wenn man mit Korpora arbeitet, ich hab schon drauf hingewiesen im Zusammenhang mit den Modalverben. Wir haben auch Korpusuntersuchungen gemacht bei einem Projekt über *-ung*-Nominalisierungen, da sind auch die Überraschungen groß, weil man sieht, dass das, was sozusagen von der Grammatik vorgegeben ist, im Deutschen gar nicht vorkommt. Wir haben in der Regel Strukturen mit definitem Artikel und Nominalisierung und nicht viel mehr, d.h. komplexe Strukturen wie im Englischen kommen kaum vor, das fand ich überraschend.

Fellbaum: Natürlich! Introspektion („kann man XYZ sagen?") führt nur zu begrenzten, oft falschen, Einsichten, da man bei der Selbstbefragung schon theoretische Vorurteile oder Scheuklappen hat. Zweitens wollen wir die Sprache einer Gemeinschaft, nicht nur die einiger Linguisten, untersuchen. Die Reichhaltigkeit der Sprache überrascht uns sehr oft. Positiv natürlich, da neue Daten immer neue Herausforderungen stellen und auch zu neuen Einsichten führen können.

Hinrichs: Ich hatte schon von meiner Magisterarbeit erzählt, in der es um temporale Anaphora ging. Da habe ich die Verwendung authentischer Daten

als sehr positiv empfunden, weil sie einfach einen viel breiteren und unverstellten Blick ermöglicht und ich auf wesentliche neue Einsichten durch die empirischen Untersuchungen mit in dem Fall narrativen Texten gestoßen bin, was mit Sicherheit bei rein künstlich konstruierten Texten nicht passiert wäre. Man sollte aber auch einen nüchternen Blick für den Nutzen von Korpora und auch deren Grenzen haben. Der Nutzen liegt auf der Hand, man bekommt überraschende und häufig auch sehr interessante Daten. Man sollte jetzt aber auch nicht eine Art Ausschließlichkeit von Korpora als Datentyp postulieren, als einen Datentyp, der dann in toto die intuitiven Sprecherurteile ersetzen kann. Es ist so, dass man natürlich bei positiven Belegen Rückschlüsse aus Korpora ziehen kann. Aber aus dem reinen Nichtvorhandensein von Phänomenen in Korpora lassen sich zwar gewisse Rückschlüsse auf die Häufigkeit dieser Phänomene ziehen, es lässt sich aber kein grundsätzliches Urteil hinsichtlich etwa von Grammatikalität ableiten. Insofern ist wohl nur derjenige negativ überrascht, der sich von Korpusuntersuchungen ein Allheilmittel erhofft.

Lüdeling: Wenn die Fragestellung nicht von einem Korpus ausgeht, sondern theoretisch motiviert ist, dann stehen oft die Daten nicht zur Verfügung, die ich eigentlich bräuchte. Das passiert mir z.B. bei Wortbildungsuntersuchungen oft. Ich spreche von Arbeiten, die ich mit Stefan Evert zusammen gemacht habe. Wir haben die Produktivität von seltenen Wortbildungsmustern untersucht, zum Beispiel das nicht-medizinische *-itis*. Dafür haben wir das größte Korpus genommen, das es zur Zeit für das Deutsche gibt, nämlich das Akademiekorpus mit einer Milliarde Wortformen. Wir haben selbst dort sehr wenig Daten gefunden. Es ist also sogar bei großen Korpora schwierig, wenn man seltene Phänomene untersucht. Man muss sein Korpus sehr gut kennen, um zu wissen, welche Dinge man daran untersuchen kann, und man muss die Grenzen von Korpora einfach anerkennen. Außerdem ist es oft so, dass Korpora nicht gut aufbereitet sind. Nicht die Primärdaten sind das Problem, sondern die durch Taggen oder Lemmatisierung oder irgend eine andere Annotation hinzugefügten Daten, mit denen man manchmal nicht so viel anfangen kann, besonders bei den historischen Korpora. Das heißt, dass man einiges selber machen bzw. mit Heuristiken arbeiten muss. Das ist oft mit Datenverlusten verbunden und je nach Phänomen kann das schlimm sein. Bei manchen Untersuchungen ist man über jeden Beleg froh, weil man sowieso so wenige hat. Bei den Untersuchungen zur morphologischen Produktivität habe ich öfter überraschend festgestellt, dass irgendein Wortbildungsmuster produktiv ist, das ich gar nicht auf der Liste hatte. Ein Beispiel ist die nicht-medizinische Verwendung von *-itis*, die wir uns erst gar nicht angucken wollten und wo wir überrascht waren, dass es produktiv verwendet wird. Das ist auch qualitativ spannend, da es mit nativen Stämmen vorkommt.

Klosa: Natürlich gibt es immer wieder Überraschungen, z.B. wenn man Fragen der Wortbildungsproduktivität in umfangreichen elektronischen Textkorpora nachgeht. Warum etwa gibt es über 1.200 Komposita mit *Berlin* als Bestimmungswort in den IDS-Korpora der geschriebenen Sprache, aber nur rund 170 mit *Bonn* als Bestimmungswort? Beides (zumindest zeitweilig) deutsche Hauptstädte, werden sie doch offensichtlich sehr unterschiedlich thematisiert und in sehr unterschiedlichem Umfang. Wollte man hierauf eine genaue Antwort finden, müsste man aber über die reine Analyse der Vorkommen natürlich hinaus- und in die Textbelege hineingehen.
Reis: Die immer wieder unglaublichen Übergänge in Daten aus freier Wildbahn sind immer wieder überraschend. Da gibt es Konstruktionen, denen man als Grammatiker gar nicht begegnen möchte, die aber belegt sind. Man tut gut daran, wenn man sich irgend einen Reim drauf machen kann, wie solche Konstruktionen überhaupt entstehen können. Ich habe Beispiele gefunden zu *drohen* und *versprechen*, wo man sagen konnte, dass der Gegensatz, den man zwischen Kontroll- und Anhebungskonstruktionen sieht, nicht passt. Es gibt Konstruktionen, von denen ich nicht glaube, dass man die überhaupt analysieren können muss, aber dann muss man zumindest eine vernünftige Theorie haben, um das zu erklären. Da ich seit Jahrzehnten Wortbildungs- und Syntaxklausuren nur anhand von Realtexten stelle, weiß ich, dass man das überhaupt nicht erfinden kann, was da auf engem Raum vorkommt.
Storrer: Beim korpusgestützten Arbeiten erlebt man immer wieder Überraschungen in dem Sinne, dass man an die Daten mit Erwartungen herangegangen ist, die dann nicht erfüllt wurden. Ob man diese als positiv oder als negativ empfindet, hängt von der individuellen Situation und dem persönlichen Temperament ab. Sicherlich ist es schmerzlich, wenn die Daten eine liebgewonnene theoretische Hypothese nicht stützen oder gar massiv untergraben. Andererseits kann man, wenn man sich nicht gerade in der Endphase einer Arbeit befindet, die ganz zentral auf dieser Hypothese aufbaut, deren Falsifizierung positiv als Erkenntnisgewinn deuten, der zu neuem Nachdenken über den entsprechenden Bereich anregt.

In unseren Untersuchungen zu Sprachmerkmalen im Chat hat sich beispielsweise gezeigt, dass es DIE „Sprache des Chats“ nicht gibt. Viele der angeblich chattypischen Merkmale, die in den Medien als Anzeichen für den Verfall unserer Schreibkultur beargwöhnt werden, sind zwar typisch für Freizeit- und Plauderchats, treten aber in anderen Anwendungsbereichen der Chattechnologie (Medien, Bildung, Beratung) selten oder gar nicht auf. Weiterhin lässt sich eine Korrelation zwischen der Anzahl aktiver Chatter und der Beitragslänge feststellen. Dies kann man dadurch erklären, dass kurze Beiträge in überfüllten Chaträumen dazu geeignet sind, den Abstand zwischen aufeinander bezogenen Gesprächsschritten möglichst gering zu halten. Kurze Beiträge sind also in diesen Kontexten funktional und sinnvoll; die

Beitragslänge sagt in diesem Fall nichts über die Kompetenz der betreffenden Chatter zur Bildung komplexerer Sätze aus. Insgesamt wurde die Annahme einer Chat-Sprache, wie sie sich in den Medien, aber auch in mancher linguistischen Literatur zum Thema findet, nicht durch unsere Daten gestützt. Die Daten lassen aber neue Erkenntnisse darüber zu, welche Faktoren die sprachlichen Strukturen im Chat systematisch beeinflussen; diese können dann in neue und differenziertere Aussagen zur Chat-Kommunikation münden – in unserem Fall mit dem Erkenntnisinteresse, Faktoren herauszuarbeiten, mit denen sich Kohärenzprobleme und Missverständnisse in „seriösen" Anwendungskontexten (E-Learning, Beratung) vermeiden oder zumindest vermindern lassen.

Welche Tipps würden Sie Studierenden geben, die, motiviert durch ein korpuslinguistisches Seminar, selber mit Korpora arbeiten möchten?

Berman: Sich Zeit zu nehmen und die Anfragesprache zu lernen.
Braun: Es ist sehr wichtig, zu verstehen, dass die Korpuslinguistik lediglich Methoden zur Verfügung stellt, um Antworten auf eine ganz bestimmte Fragestellung zu finden. Deshalb muss man immer genau wissen, wofür man ein Korpus macht. Dieser Zweck bestimmt dann das Korpusdesign, die Marschrichtung beim Sammeln der Daten und bei der Abfrage. Gerade für ein Korpus gesprochener Sprache kann es sinnvoll sein, dass man erst einmal kleinere Pilotversuche macht. Wir haben z.B. nach den ersten Interviews immer noch einmal überlegt, wie wir unsere Fragen so stellen, dass wir die gewünschten Daten bekommen. In einem neuen EU-Projekt zum Aufbau von Jugendsprachkorpora wollen wir diese Erhebungsmethoden noch verbessern. Bei der Erstellung haben wir uns natürlich an den Standards orientiert, so dass Kollegen unsere Arbeit nachvollziehen und unsere Ergebnisse nutzen können. Die Inhalte haben wir aber selbst bestimmt. Was die Erschließung, Analyse und Abfrage eines Korpus betrifft, da sollte man gut recherchieren, was es da schon an Tools gibt. Es gibt heute sehr viel Open Source Software, die man benutzen oder anpassen kann. Das fängt an beim Wortartentagging und geht weiter mit XML-Annotationstools. Auch Transkriptionstools stehen zur Verfügung. Open Source sollte ohnehin der Trend sein, gerade im Bereich der Hochschulforschung – für Tools und für Korpora.
Ehrich: Was die korpuslinguistischen Seminare betrifft, die sollten für meine Begriffe nicht nur in der allgemeinen Sprachwissenschaft, sondern inbesondere auch in den Philologien zur Pflicht werden, und zwar aus zwei Gründen. Ein Grund ist der sprachwissenschaftliche, weil wir mit Daten wirklich Überraschungen erleben, die unsere Theorien modifizieren helfen können. Der zweite, aber auch ganz wichtige Grund ist, dass wir mit den Korpusuntersuchungen auch den Philologiestudenten ein gewisses technisch-analytisches

Know-how an die Hand geben, das sie für ein Leben nach der Universität auch dann gut gebrauchen können, wenn dieses Leben kein sprachwissenschaftliches Leben ist. Es gibt also gute Gründe dafür, Korpusarbeit zur Pflicht zu machen in der Lehre. Ich tue das bei fast jedem meiner Seminare und lasse gar keine Arbeiten mehr schreiben ohne Korpusuntersuchung. Was ich Studierenden, die mit Korpora arbeiten, gerne mitgeben würde, ist dies: 1. Man sollte immer mit einer klaren theoretischen Fragestellung an die Daten herangehen. 2. Man sollte nie vergessen, dass der Linguist, der mit Belegen arbeitet, diese immer auch verstehen bzw. interpetieren muss. Insofern sind Belegdaten interpretierte Daten.

Fellbaum: Die Studierenden sollten sich gut über die Beschaffenheit des Korpus und seine eventuellen Begrenzungen informieren (Größe des Korpus, Ausgewogenheit, Annotation, usw.). Diese Faktoren müssen bei der linguistischen Arbeit in Betracht gezogen werden, da sie die Ergebnisse beeinflussen können.

Hinrichs: Zunächst mal würde ich alle Studierenden, und nicht nur solche, die korpuslinguistische Seminare besucht haben, ermutigen, Korpora zu verwenden. Ein korpuslinguistisches Seminar kann sehr gut in die Werkzeuge einführen und auch an die Fragestellungen heranführen, die einen letztlich in den Stand versetzen, wirklich gut mit Korpora zu arbeiten.

Ich würde Studierenden raten, dass sie sich zunächst mal mit einem bestimmten Korpus anfreunden, d.h. dass sie ein Gefühl dafür bekommen, um welche Daten es sich handelt. Jedes Korpus, auch wenn es sehr groß ist, hat natürlich immer bestimmte Themen zum Gegenstand, und es ist sehr wichtig, dass man sich zunächst mal in diese Themen einarbeitet und dadurch ein Gefühl für die Daten bekommt. Es ist etwas ganz anderes, etwa mit Zeitungskorpora zu arbeiten als mit Korpora gesprochener Sprache. Man wird z.B. sehen, dass in Zeitungskorpora hypotaktische Strukturen sehr viel stärker vertreten sind als bei Korpora gesprochener Sprache, die eher parataktisch und häufig dialogisch strukturiert sind.

Beim Kennenlernen des Korpus geht es aber nicht nur darum, dass man die inhaltlichen Themen überschaut, sondern auch, wenn das Korpus annotiert ist, dass man sich die Annotation aneignet. Hier bietet sich an, dass man die entsprechenden Stylebooks, etwa zum Part of Speech Tagging, die bei Baumbanken eigentlich immer mitgeliefert werden, zunächst mal zur Kenntnis nimmt. Dann hat man die Spezifika der Annotation im Hinterkopf, wenn man nach Belegen sucht.

Lüdeling: Was ich immer wieder in meinen Seminaren sage, ist, dass die Fragestellung nicht aus dem Korpus kommen sollte, sondern aus einer Theorie oder aus einer Beobachtung. Dann sollte man sich überlegen, ob Korpusdaten helfen können, diese Frage zu beantworten. Manchmal muss man trotzdem mit Intuition arbeiten oder Experimente machen oder eine Fragebogen-

studie. Wenn man also entschieden hat, dass Korpusdaten für die Forschungsfrage geeignet sind, dann muss man noch ein Korpus finden, das die Daten enthält, die man benötigt. Dann muss man sehr viel selber angucken und darf sich nicht blind auf die linguistische Annotation verlassen. Eine weitere wichtige Botschaft ist: Wenn ihr nicht wisst, was ihr genau tut, dann arbeitet nicht quantitativ. Irgendetwas zu zählen, nur um etwas zu zählen oder eine nette Grafik zu haben, ist Zeitverschwendung. Man sollte dann lieber qualitativ arbeiten und nur eine beschreibende Statistik machen, wenn das wichtig ist und zum Vorhaben passt. Zu vielen Forschungsfragen kann man auch gut rein qualitativ arbeiten. Wenn man denn meint, dass man quantitativ arbeiten will, dann muss man eben die Zeit aufbringen, um sich mit statistischen Methoden und der Anwendung dieser Methoden ernsthaft zu beschäftigen. Ansonsten möchte ich empfehlen, dass man manche Dinge einfach mal ausprobieren und keine Angst haben sollte, mal etwas mehr Formales zu machen oder mit Corpus Query Tools zu arbeiten.

Klosa: Ich möchte Studierenden die folgenden Tipps geben:

- Man sollte sich vor der Korpusrecherche darüber klar sein, ob man das Korpus befragt, um die Bestätigung einer Vorannahme zu erhalten oder auch um geeignete Belege zu finden. Ein grundsätzlich anderer Zugang zum Korpus ist, sich von diesem überraschen zu lassen.
- Die Formulierung der Suchanfragen ist manchmal ganz einfach, manchmal aber auch das Ergebnis mehrerer Versuche, bis das optimale Ergebnis erreicht ist. Insofern ist es natürlich sinnvoll, sich vor der ersten Anfrage zu überlegen, was man genau wissen möchte, und die Suchanfrage entsprechend zu formulieren. Ein gewisses Maß an „Rumspielerei“ kann aber andererseits auch nicht schaden, um zu lernen, die Möglichkeiten voll auszuschöpfen.
- Vor allem aber sollte man nicht vorschnell sagen: *Das gesuchte Phänomen bzw. Wort ist in meinem Korpus nicht belegt*. Meine Erfahrung ist, dass das Korpus das Gesuchte höchstwahrscheinlich sehr wohl enthält, auf Grund der Suchanfragenformulierung aber das gewünschte Ergebnis nicht generiert werden kann. Um Korpora sinnvoll benutzen und ausnutzen zu können, bedarf es deshalb einer gründlichen Einarbeitung in die jeweiligen Recherche- und Analysewerkzeuge.
- Bei der Korpusauswertung sollte man weder den Korpusdaten noch seiner eigenen Sprachkompetenz blind vertrauen. So ist es eine zwingende Notwendigkeit, egal bei welcher Suchanfrage für welchen Zweck, immer den Weg bis in die Textbelege zu gehen, weil man sonst in die Irre gehen kann. Ein Beispiel: Aus einer Liste der Kookkurrenzpartner zum Suchwort *deutlich* ist das syntagmatische Muster *deutliche Worte* zu ent-

nehmen. Ob *deutlich* hierbei in der Lesart *gut wahrnehmbar* oder in der Lesart *eindeutig* vorliegt, können nur die Belege zeigen.

- Man sollte die Zusammensetzung des zugrunde liegenden Korpus nicht aus den Augen verlieren. Sowohl die Zusammensetzung nach Textsorten wie die chronologische Staffelung der Korpustexte wirkt sich auf die Ergebnisse aus und ist daher bei der Interpretation der Suchergebnisse zu berücksichtigen. Umgekehrt bedeutet dies auch, dass man sich für den eigenen Untersuchungszweck am besten das geeignete Korpus selbst zusammenstellt (sogenannte *virtuelle Korpora* können beispielsweise in den Korpora geschriebener Sprache am IDS erstellt werden).

Ansonsten kann ich nur dazu ermutigen, so viel wie möglich mit umfangreichen elektronischen Textkorpora zu arbeiten, weil dies ungeheuer überraschend, zum Nachdenken anregend und dadurch immer wieder für einen selbst wie für die Sache bereichernd ist.

Pittner: Wenn Studierende eigene Korpusanalysen erstellen wollen, bieten die vom Institut für Deutsche Sprache in Mannheim zur Verfügung gestellten Korpora und Suchwerkzeuge eine hervorragende, frei zugängliche Basis. Zudem kann man, je nach Untersuchungsziel, auch etwas kreativ sein beim Finden von Korpora. Es gibt eine Reihe von digitalisierten Textsammlungen, die nicht als linguistische Korpora gedacht oder aufbereitet sind und trotzdem hervorragend als solche verwendet werden können. Ich denke hier z.B. an die Digitale Bibliothek auf CDs, die auch einige rudimentäre Suchwerkzeuge mitbringt und daher für einfachere Korpusanfragen eine große Vielfalt an Texten von verschiedenen Gattungen und Zeiten bietet. Zum anderen ist es naheliegend und auch sehr spannend, das World Wide Web als Korpus zu benützen. Für bestimmte Arten von Anfragen sind Suchfunktionen der Suchmaschinen wie etwa Google durchaus ausreichend. Der Vorteil des Webs ist, dass man riesige Datenmengen an aktuellem Sprachmaterial durchsuchen kann.

Reis: Ich möchte Studenten zwei Tipps geben. Sie sollten technisch versiert sein und mit Statistik umgehen können. Korpora muss man zum Sprechen bringen durch statistische Auswertung. Es gibt wenige Fragestellungen, die ohne statistische Mittel beantwortbar wären. Zweitens sollte jede Korpusanalyse geleitet sein von ziemlich gutem linguistischen Vorwissen, damit man überhaupt weiß, wonach man guckt.

Storrer: Wenn es sich um zeitlich begrenzte Arbeiten (BA, MA, Staatsarbeiten) handelt, sollte man kleine Vorstudien machen, damit man ungefähr abschätzen kann, wie viel Daten im verfügbaren Zeitrahmen überhaupt analysiert werden können. Insbesondere wenn eigene Korpora erstellt werden, besteht nämlich die Gefahr, dass der Zeitaufwand für die Erstellung un-

terschätzt wird und die Zeit dann zur Analyse und zum Schreiben der Arbeit fehlt.

Tipp: Erst Ausschau halten, ob es schon passende Daten gibt. Dies gilt insbesondere für Untersuchungen an gesprochener Sprache, die erst noch transkribiert werden muss.

Tipp: Wenn man doch selbst Daten erheben und aufbereiten möchte/muss, dann sollte man möglichst früh an einer Probe eine Zeitabschätzung durchführen und dann noch einmal prüfen, ob die im gegebenen Zeitrahmen erfassbaren Datenmengen auch wirklich für die Untersuchungsfrage/das Thema ausreichen.

Tipp: Man sollte erst eine klare Vorstellung über die Fragestellung der Untersuchung gewinnen und dann nach einem geeigneten Korpus + Werkzeug suchen. Wer nach grammatischen Konstruktionen sucht, ist z.B. gut mit einer Baumbank und einem zugehörigen Werkzeug wie TigerSearch bedient. Allerdings erfordert dies mehr technische und auch linguistische Vorkenntnisse als z.B. der Umgang mit der schnell erlernbaren und intuitiv bedienbaren „einfachen“ Online-Schnittstelle zum DWDS-Korpus, dessen linguistische Aufbereitung (Lemmatisierung/ POS) ebenfalls interessante Untersuchungen zur Wortbildung oder zur Phraseologie erlaubt und das durch seine gleichmäßige Streuung über die Dekaden des 20. Jahrhunderts hinweg auch Aussagen über Wortschatzentwicklungen erlaubt.

Ein weiterer Tipp für zeitlich limitierte Arbeiten: Rechtzeitig in einer Vorstudie prüfen, ob das gewählte Korpus für die Fragestellung genügend Daten bereithält (bei zu vielen Daten kann man sampeln; bei zu wenig Daten scheitert die Untersuchung). An einem kleinen Pröbchen den Zeitaufwand für die geplanten Analysen abschätzen.

Insgesamt hielte ich es für sehr wünschenswert, wenn es auch für weniger computertechnisch versierte LinguistikstudentInnen einfach zu bedienende Werkzeuge zur Sortierung und Klassifikation der Belege gäbe. Bislang arbeiten diese meist mühsam und mit viel Handarbeit an Mischungen von Word- und Excel-Dateien.

Zwei weitere Tipps:

- Den Erfahrungsaustausch mit anderen Nutzern der Korpora und Tools suchen.
- Bei Problemen mit Tools sollte man sich nicht scheuen die Entwickler bzw. Anbieter zu kontaktieren. Diese sind meist interessiert an Feedback und helfen ihren Nutzern gerne weiter.