Sprachkorpora

Institut für Deutsche Sprache
Jahrbuch 2006

Sprachkorpora

Datenmengen und
Erkenntnisfortschritt

Herausgegeben von
Werner Kallmeyer und Gisela Zifonun

Walter de Gruyter · Berlin · New York

Redaktion: Franz Josef Berens

⊗ Gedruckt auf säurefreiem Papier,
das die US-ANSI-Norm über Haltbarkeit erfüllt.

ISBN 978-3-11-019273-5

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© Copyright 2007 by Walter de Gruyter GmbH & Co. KG, D-10785 Berlin.

Dieses Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Printed in Germany

Einbandgestaltung: Christopher Schneider, Berlin

Inhalt

Werner Kallmeyer / Gisela Zifonun: Vorwort	VII
Ludwig M. Eichinger: Linguisten brauchen Korpora und Korpora Linguisten – Wege zu wohl dokumentierten und verlässlichen Aussagen über Sprache	1
Christian Lehmann: Daten – Korpora – Dokumentation	9
Anke Lüdeling: Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik	28
Sam Featherston: Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie	49
Stefan Müller: Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus	70
Ulrike Demske: Das Mercurius-Projekt: Eine Baumbank für das Frühneuhochdeutsche	91
Annette Klosa: Korpusgestützte Lexikographie: besser, schneller, umfangreicher?	105
Jörg Asmussen: Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen	123
Gunter Senft: „Ich weiß nicht, was soll es bedeuten ...“ – Ethnolinguistische Winke zur Rolle von umfassenden Metadaten bei der (und für die) Arbeit mit Corpora	152
Alexandra N. Lenz: Zur variationslinguistischen Analyse regional- sprachlicher Korpora	169
Werner Kallmeyer: Möglichkeiten der maschinellen Unterstützung bei der Arbeit mit Interaktionskorpora	203
Thorsten Brants: Statistisch basierte Sprachmodelle und maschinelle Übersetzung	235
Michael Strube / Margot Mieskes / Christoph Müller: Gesprächs- protokolle auf Knopfdruck: Die automatische Zusammenfassung von gesprochenen Dialogen	249

Vorwort

Das Beschaffen, Dokumentieren und Archivieren von Sprachdaten gehört seit jeher zu den arbeitsaufwändigen, aber unverzichtbaren Tätigkeiten sprachwissenschaftlicher Forschung. Aber seit einer Reihe von Jahren widmen sich weltweit bedeutende wissenschaftliche Unternehmungen dem Aufbau großer Korpora. Die leitenden Gesichtspunkte sind dabei unterschiedlich, sie zielen u. a. auf die Sammlung und Dokumentation von schwer zu beschaffenden Sprachmaterialien, z. B. von bedrohten Sprachen (vgl. das entsprechende Programm der VW-Stiftung), auf die Weiterentwicklung der Theorie von Daten und Korpora ebenso wie auf die Entwicklung von Verfahren der Korpusanalyse (vgl. u. a. den SFB 441 „Linguistische Datenstrukturen“ in Tübingen und das Kooperationsprojekt „Nachhaltigkeit linguistischer Daten“ der SFBs 441, 538 und 632). So kann es auch nicht erstaunen, dass der linguistische Tagungskalender dieser Jahre mehrere Tagungen mit einer korpusorientierten Ausrichtung aufweist (so die Jahrestagung 2006 der Deutschen Gesellschaft für Sprachwissenschaft „Sprachdokumentation und Sprachbeschreibung“, die Jahrestagung 2006 des IDS „Sprachkorpora – Datenmengen und Erkenntnisfortschritt“ und der Kongress 2007 der International Pragmatics Association „Language data, corpora, and computational pragmatics“).

In dieser Intensivierung von Korpusaufbau und Korpusanalyse wird die technische Entwicklung sichtbar. Die Sprachwissenschaften verfügen heute über korpus technologische Werkzeuge, die einen relativ komfortablen Umgang mit großen Datenmengen erlauben. Mit der Verfügbarkeit von geschriebenen Texten in elektronischer Form wird es leichter, große Textkorpora zusammenzustellen – wenn man einmal von den teilweise dornigen Fragen des Copyrights absieht. Die Größe von schriftlichen Korpora, soweit sie sich auf öffentliche Texte konzentrieren, wächst geradezu explosionsartig im Vergleich mit den Korpora der 1970er Jahre. Für die gesprochene Sprache ist die Lage in dieser Hinsicht allerdings völlig anders. Die Aufnahme- und Speichertechnik für Ton und auch Video ermöglicht heute ebenfalls auf einfache Weise Sprachaufnahmen „im Feld“, d. h. in natürlichen Kommunikationssituationen. Außerdem wird öffentlich gesprochene Sprache von den Medien Rundfunk und Fernsehen in großer Menge verbreitet, in etwa vergleichbar dem öffentlichen Markt der gedruckten Texte. Hinzu kommt, dass die verfügbaren und finanzierbaren Speichermedien für Ton und Bild heute eine „massenhafte“ Speicherung solcher Daten erlauben. Aber die Korpora sind kaum

linguistisch auswertbar ohne Verschriftlichung des Gesprochenen. Die Transkription ist gleichsam das Nadelöhr, durch das mündliche Sprachdaten gezwängt werden müssen, um als Analysegrundlage dienen zu können. Angesichts des Arbeitsaufwandes, den eine brauchbare Verschriftlichung gesprochener Sprache erfordert, wundert es nicht, dass die mündlichen Korpora ungleich langsamer wachsen als die schriftlichen.

In der verstärkten Zuwendung zu korpuslinguistisch gestützten Untersuchungen zeigt sich teilweise auch eine Veränderung der theoretischen und methodischen Orientierung der Linguistik. Die Situation stellt sich in dieser Hinsicht allerdings für die verschiedenen linguistischen Teilbereiche unterschiedlich dar.

Die Beiträge des vorliegenden Jahrbuchs konzentrieren sich insbesondere auf theoretische und methodische Fragen zum Aufbau und zur Nutzung großer Sprachkorpora. Dabei treten sowohl grundlegende Fragen der linguistischen Arbeit mit Korpora hervor als auch bereichsspezifische Probleme.

Aus einer Perspektive der Grundlagenreflexion expliziert **Christian Lehmann** die Bedingungen für Linguistik als empirische Wissenschaft durch die Bestimmung von Daten, die Anforderungen an die Dokumentation und den produktiven Umgang mit Korpora. Die Auseinandersetzung mit der Kernfrage, ob, in welchem Sinne und wie die Linguistik eine empirische Wissenschaft ist, führt **Anke Lüdeling** fort mit der auf die Korpuslinguistik bezogenen Argumentation, wonach die Grundlage jeder quantitativen Analyse die qualitative Analyse oder Kategorisierung der Daten ist.

Für die grammatische Forschung erweisen sich zunehmend die in großen Korpora anzutreffenden Variationsphänomene und – zumindest auf den ersten Blick – auffälligen und irregulär erscheinenden Äußerungsstrukturen als wichtige Informationsquelle. Insofern hat sich eine Entwicklung hin zur korpusbasierten Arbeit etabliert. Die Erhebung von Grammatikalitätsurteilen ist ein wichtiges Instrument der grammatischen Forschung, wie **Sam Featherston** anhand der experimentellen Erhebung introspektiver Urteile in numerischer Form auf einer Intervallskala und ihrer Quantifizierung demonstriert. Auf der anderen Seite zeigt **Stefan Müller** am Beispiel von drei syntaktischen Phänomenen (Partikelverben, Extraposition und Subjazenz sowie mehrfache Vorfeldbesetzung) den Wert von Korpusabfragen, weil den Sprechern ihr sprachliches Wissen nicht immer zugänglich ist. Ein dritter Aspekt der neueren grammatischen Forschung ist die Arbeit mit syntaktisch annotierten Korpora (vgl. auch die Präsentation des TIGER-Korpus während der Tagung). Die Vorteile und Schwierigkeiten dieses Vorgehens für die historische Sprachforschung zeigt **Ulrike Demske** anhand des Aufbaus einer Baumbank für das Frühneuhochdeutsche.

Für die Lexikographie gehört es heute zum Standard, mit Unterstützung durch große maschinenlesbare Korpora zu arbeiten. Vor diesem Hintergrund diskutiert **Annette Klosa** anhand wichtiger Vorhaben korpusgestützter Lexiko-

graphie die Auswirkungen auf die Strukturierung der Wörterbuchartikel und die Wahl der Beispiele. **Jörg Asmussen** zeigt den Wert von korpuslinguistischen Verfahren für die Analyse von Bedeutungsparaphrasen in Wörterbüchern und ihre Optimierung am Beispiel eines korpusgestützten Wörterbuchs der dänischen Sprache.

Die Beiträge zur Untersuchung von Daten gesprochener Sprache gehen von einer ethnolinguistischen, soziolinguistischen beziehungsweise pragmatischen Perspektive aus. Bei der Arbeit mit kulturell fremden Daten tritt u. a. das Problem einer angemessenen Dokumentation des für die Analyse notwendigen kulturellen Hintergrundwissens hervor. Diese zusätzliche Anforderung gegenüber der Arbeit mit Daten aus vertrauten kulturellen Kontexten führt **Gunter Senft** an Beispielmateriale von Trobriand-Insulanern aus Papua-Neuguinea vor. Grundsätzlich, wenn auch nicht so zugespitzt, besteht dieser Bedarf an reichen Metadaten auch bei Feldforschungsdaten aus dem deutschen Sprachraum, z. B. bei regionalsprachlichen Untersuchungen, worauf auch **Alexandra Lenz** hinweist. Sie gibt einen Überblick über die verfügbaren regionalsprachlichen Korpora des Deutschen und demonstriert Möglichkeiten und gegenwärtig noch bestehende Beschränkungen für die variationslinguistische Auswertung. **Werner Kallmeyer** zeigt auf, welche besonderen Anforderungen sich bei Korpora sprachlicher Interaktion für die Entwicklung von Verfahren der computergestützten Volltextrecherche auf COSMAS-Basis stellen.

Die beiden folgenden Beiträge diskutieren Fortschritte und Aussichten der Korpusanalyse aus informatischer Sicht. Die Diskussion auf diesem Gebiet spiegelt noch einmal die in den linguistischen Beiträgen bereits aufscheinende Frage, ob und wann die Arbeit mit „reinen Daten“ oder die Anreicherung von Korpora mit Interpretationen (Annotation) sinnvoller ist und welche Möglichkeiten quantitative und wissensbasierte Auswertungsverfahren bieten. **Thorsten Brants** präsentiert Verfahren der statistisch basierten maschinellen Übersetzung, eine für Sprachwissenschaftler eher ungewohnte, aber anregende Perspektive. **Michael Strube**, **Margot Mieskes** und **Christoph Müller** schließlich locken mit der Erfüllung einer Wunschvorstellung von Protokollschreibern wie Gesprächsforschern, die diese aber immer wieder verdrängen mussten: „Gesprächsprotokolle auf Knopfdruck“.

Das Vortragsprogramm der Jahrestagung 2006 wurde durch einen Nachmittag mit korpustechnologischen Präsentationen ergänzt; Informationen dazu finden sich auf den Internetseiten des IDS unter der Adresse <www.ids-mannheim.de/org/tagungen/jt2006/presentation.html>.

Den Abschluss der Tagung bildete eine Podiumsdiskussion zum Thema „Varianten im Korpus“. Unter der Moderation von Norbert R. Wolf diskutierten Ulrike Hass, Christian Lehmann, Anke Lüdeling, Christian Mair und Gisela Zifonun miteinander und dem Publikum darüber, in welchem Umfang Variation in einem als standardsprachlich oder standardsprachennah konzipierten großen Korpus etwa des Deutschen, das für unterschiedliche Frage-

stellungen genutzt werden soll, zu berücksichtigen ist und in welcher Weise Architektur und Design des Korpus dem Faktor Varianz Rechnung tragen können. Ein solches Korpus kann, darin stimmten die Diskutanten überein, nicht repräsentativ für den Sprachgebrauch einer Sprachgemeinschaft sein. Wohl aber kann es als ausgewogenes oder exemplarisches Korpus eine begründete Auswahl aus der Vielfalt kommunikativer Gattungen vertreten, die in angemessenem Verhältnis zueinander berücksichtigt werden. In einem solchen Korpus wird Varianz, „nach der man nicht gefragt hat“, in jedem Fall auf der Ebene der Rohdaten vorhanden sein. Varianz – etwa in der phonetischen Realisierung – wird auf abstrakteren Ebenen der Datenrepräsentation gegebenenfalls reduziert werden (müssen). Bei Variation auf der Ebene der Morphosyntax oder der Satzverknüpfung stellt sich häufig die Frage, ob eine unter normativer Sehweise „anstößige“ Variante als Performanzfehler oder als zu Unrecht stigmatisierte Ausdrucksform oder gar als Indiz für Sprachwandelprozesse zu betrachten ist.

*Werner Kallmeyer
Gisela Zifonun*

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 sechszwanzigsten 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 Korpus-technologie 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 Ergebnisse 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. 181ff.; 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 Wissens einbettung 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 Nijmegen 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.

CHRISTIAN LEHMANN

Daten – Korpora – Dokumentation

Abstract

Der Begriff und die Rolle von Daten in einer Wissenschaft hängen eng mit ihrem Selbstverständnis zusammen. Als erstes ist zu überlegen, inwiefern Linguistik eine empirische Wissenschaft ist und also von Daten abhängt.

Während in den Philologien ein Korpus die Grundlage einer Disziplin abgibt, die ohne es nicht bestünde, ist in der Linguistik ein Korpus nur ein Weg, an Daten zu kommen. Hier ist zu diskutieren, welche relativen Meriten die alternativen Wege im Hinblick auf die angestrebten Ziele haben.

Während manches auf uns gekommene Korpus seine Sprache sicher nicht angemessen repräsentiert, könnte eine heute von Linguisten erstellte Dokumentation diesen Anspruch im Prinzip einlösen. Hier stellt sich die Frage, ob das – angesichts des infiniten Charakters der Sprache – überhaupt möglich ist und in wie weit die Repräsentativität wieder von den angestrebten Zielen abhängt.

Dies sind alles Fragen linguistischer Methodologie. Eine Zeitlang hat man in der Linguistik geglaubt, ohne Methodologie zu Theorien gelangen zu können. Seit sich das als irrig herausgestellt hat, ist die Entwicklung von Methoden ein fühlbares Desiderat geworden. Wie man repräsentative Daten erhebt, wie man ein Korpus zusammenstellt und nutzt, wie man eine Sprache dokumentiert, sind alles Fragen, die eigentlich in die Alltagsroutine einer Wissenschaft fallen müssten. Dass sie noch weitgehend ungeklärt sind, ist ein Symptom dafür, dass die Linguistik noch keine erwachsene Wissenschaft ist.¹

0. Einleitung

Das Thema dieser Tagung und dieses Beitrags ist in erster Linie ein methodologisches. Die Fragen nach der Natur von Daten, von Korpora und von Dokumentation, ihrer Rolle in unserer Wissenschaft und in der Praxis kommen zwar je für sich auch in anderen Disziplinen vor; in ihrer Kombination aber sind sie typisch für die Linguistik. Sie hängen eng mit der Frage zusammen, was für eine Wissenschaft Linguistik eigentlich ist, und scheinen deren Klärung vorauszusetzen. Tatsächlich ist dies aber kein einseitiges Voraussetzungsverhältnis, weil Entscheidungen über den Gegenstand und die Methodologie einer Wissenschaft ihr Wesen prägen.

¹ Ich danke Ulrich Ammon, Oddleif Leirbukt, Hans Uszkoreit und Herbert-Ernst Wiegand für hilfreiche Kommentare.

Die obersten Entscheidungen, die das Wesen einer Wissenschaft bestimmen, sind die folgenden (vgl. Abb. 1):

1. Da eine Wissenschaft eine menschliche Aktivität ist, ist sie in erster Linie durch ihr Ziel bestimmt. Da Wissenschaft das Streben nach objektiver Erkenntnis ist, ist jede einzelne Wissenschaft durch ihr spezifisches Erkenntnisinteresse bestimmt.
2. Das Erkenntnisinteresse konstituiert den Gegenstand einer Wissenschaft. Eine erste binäre Einteilung folgt dem Kriterium, ob sich der Gegenstand außerhalb des erkennenden Geistes befindet oder nicht. Im letzteren Falle ist es logische Erkenntnis. Im ersteren Falle unterscheidet das weitere Kriterium, ob der Gegenstand Produkt eines (ebenfalls erkennenden) Geistes sind oder nicht, zwischen hermeneutischer und empirischer Erkenntnis.

Abb. 1. Arten der Erkenntnis

Viele Disziplinen lassen sich nach diesen Kriterien zwanglos kategorisieren. Die Linguistik scheint hiernach eindeutig eine hermeneutische Wissenschaft zu sein. Das war auch durch den größten Teil ihrer Geschichte, vom Altertum bis in die Neuzeit, überwiegend so. Während der Neuzeit bildete sich jedoch für die Wissenschaften das Ideal objektiver Erkenntnis aus; und intern werden wissenschaftliche Aktivitäten danach bewertet, wie nahe sie diesem Ideal kommen.

Objektive Erkenntnis ist nun

- im Hinblick auf das erkennende Subjekt intersubjektiv
- und im Hinblick auf das Erkenntnisobjekt verallgemeinerbar.

In vollständig abgesicherter Weise wird dieses doppelseitige Objektivitäts-erfordernis nur in logischer Erkenntnis eingelöst, aber eben um den Preis, mit nichts zu tun zu haben, was sich außerhalb des erkennenden Geistes befände. Ziemlich erfolgreich in dieser Hinsicht ist auch noch die empirische Erkenntnis. Sie sieht von partikulären Eigenschaften ihres Gegenstandes ab und erreicht allgemeine Aussagen in Form von Hypothesen. An der Intersubjektivität muss sie Abstriche machen, denn die Hypothesen sind nicht verifizierbar, sondern nur falsifizierbar. Hermeneutische Erkenntnis schließlich ist in bezug auf beide Pole der Erkenntnis partikulär, insofern sie vom erkennenden Subjekt abhängt und insofern sie sich auf ein individuelles Objekt bezieht.

Während die jüngeren Fortschritte der Hermeneutik in Richtung einer objektiven Hermeneutik (vgl. z. B. Oevermann 2002) unbestritten sind, ergibt das Kriterium der Objektivität doch eine Hierarchie der Wissenschaftlichkeit unter den bestehenden Disziplinen: Ein unangefochtenes Prestige von Wissenschaftlichkeit haben nur die logischen Disziplinen. Die empirischen Disziplinen werden seit Hempel 1966, soweit eben möglich, nach dem Modell der logischen Disziplinen aufgezogen – z. B. die Physik nach dem Modell der Mathematik. Die hermeneutischen Disziplinen hinwiederum gelten überhaupt nicht als *sciences*, sondern lediglich als *humanities*. Manche von ihnen, darunter die Psychologie, die Soziologie und die Linguistik, wollen am liebsten keine hermeneutischen Disziplinen sein. Sie behandeln die Unterscheidung zwischen Erzeugnissen des Mitmenschen und anderen erfahrbaren Gegenständen als irrelevant und gehen nach Möglichkeit empirisch oder gar logisch vor.

Im Falle der Linguistik ist dies bis zu einem gewissen Grade durch die Natur des Gegenstandes gerechtfertigt. Sprachzeichen haben eine wahrnehmbare Seite, Sprechakte sind beobachtbare raumzeitliche Ereignisse. Man kann sie empirisch angehen. Ferner ist Sprache, in ihrer kognitiven Dimension, die äußere Form des Denkens und insofern logischer Erkenntnis zugänglich. Beides ändert freilich nichts daran, dass Sprachtätigkeit, in ihrer sozialen Dimension, Verständigung ist und dass diese hermeneutischer Methodik bedarf. Der Gegenstand Sprache hat wesentlich diese drei Aspekte, den hermeneutischen, den empirischen und den logischen. Jeglicher Versuch, einen davon zu eskamotieren oder zu verabsolutieren, ist steriler Reduktionismus.²

Dies vorausgeschickt, wird die Linguistik im folgenden überwiegend in ihrer Eigenschaft als empirische Disziplin behandelt. Dies ist einfach eine Folge der Themenstellung: Daten gibt es nur in empirischen Disziplinen.

1. Daten

1.1 Funktion von Daten

Ein Datum ist eine Repräsentation eines Phänomens aus dem Gegenstandsbereich einer Wissenschaft, die als gegeben angenommen wird (vgl. Lehmann 2004). Die Repräsentationsbeziehung ist nicht frei manipulierbar, sondern durch die wissenschaftliche Methodik geregelt. Das ist eben die Basis dafür, dass das Datum als gegeben angenommen wird. Es vertritt für die Zwecke wissenschaftlicher Argumentation das Phänomen selbst, welches i. a. nicht zur Hand ist. Daten spielen in einer Disziplin eine methodische Rolle in dem

² Die drei Aspekte sind in linguistischer Forschung miteinander verwoben. Aber in gewissen Forschungsrichtungen dominiert einer von ihnen. Z. B. ist formale Semantik überwiegend logische, statistische Korpuslinguistik überwiegend empirische und Konversationsanalyse überwiegend hermeneutische Wissenschaft.

Maße, in dem sie eine empirische Disziplin ist. In induktiver Methodik dient das Datum als Indiz und empirische Evidenz³, in deduktiver Richtung als Prüfstein im Test einer Theorie.

Eine Wissenschaft konstituiert sich gemäß ihrem Erkenntnisinteresse. Es gibt legitime oder jedenfalls in der Gesellschaft unangefochtene Erkenntnisinteressen, die nichts mit in der uns umgebenden Welt vorfindlichen und öffentlich beobachtbaren Phänomenen zu tun haben. Für die Linguistik aber gilt in dieser Hinsicht folgendes:

1. Eine Disziplin, die sich für eine empirische erklärt, muss ihren Gegenstandsbereich mit Bezug auf unabhängig von ihr selbst in der Welt vorfindliche und öffentlich beobachtbare Phänomene konstituieren. In dem Maße, in dem sie das nicht tut, ist es vielleicht eine logische oder hermeneutische Disziplin oder vielleicht auch Esoterik.
2. Es gibt in der uns umgebenden Welt vorfindliche und öffentlich beobachtbare Phänomene, nämlich Sprechakte im weitesten Sinne, an deren wissenschaftlicher Erforschung ein öffentliches Interesse und geradezu ein Auftrag besteht und für die keine andere Disziplin zuständig ist.
3. Die Linguistik selbst hat sich, in Gestalt herausragender Vertreter wie Leonard Bloomfield und Noam Chomsky, für eine empirische Disziplin (und gelegentlich gar, in rhetorischer Polarisierung, für eine Naturwissenschaft) erklärt. Das erfordert, dass sie ihren Gegenstandsbereich auf die in Nr. 1 beschriebene Weise konstituiert. Dessen vorgegebene Basis sind gerade die in Nr. 2 genannten Phänomene.
4. Jegliche Wissenschaft, die über reine Idiographie hinausgelangen will, arbeitet mit Abstraktionen. Das gilt jedenfalls für die Linguistik und a fortiori für die Systemlinguistik und ihren Gegenstand, das Sprachsystem. Von den notwendigen und angestrebten Abstraktionen sind zu allererst die Daten betroffen. Sie werden gewisser Eigenschaften entkleidet, denen das Erkenntnisinteresse nicht gilt. So entstehen sekundäre Daten, die oft eine sehr vermittelte Beziehung zu den Phänomenen haben.
5. Wie jegliche empirische Wissenschaft kann auch die Linguistik ihren Anspruch, eine solche zu sein, nur aufrecht erhalten, wenn sie die Beziehung ihrer Theorien und a fortiori ihrer Daten auf die Phänomene jederzeit methodisch kontrolliert und objektiv nachvollziehbar macht. In dem Maße, in dem sie das nicht tut, ist ihr Anspruch, eine empirische Disziplin zu sein, wissenschaftspolitische Rhetorik.
6. Gemäß dem in § 1 Gesagten ist die Linguistik sowohl eine empirische als auch eine logische als auch eine hermeneutische Disziplin. Sie hängt also

³ Die Forderung, die Daten, auf welche Generalisierungen sich stützen, verfügbar zu machen, wird immer häufiger erhoben. S. z. B. Corbett 2005, § 8 über „Reproduzierbarkeit“ von typologischen Ergebnissen dadurch, dass die Daten, auf denen sie basieren, zur Verfügung gestellt werden.

nicht ausschließlich von Daten ab, sondern gewinnt Erkenntnis auch auf andere Weise. Desto komplexer ist ihre Methodologie und desto größer die Gefahr der Manipulation von Ergebnissen.

Die Frage, welcher Art die Daten sind und welche Rolle sie in der Methodologie spielen, berührt das Selbstverständnis einer Disziplin auf das Intimste. Die Linguistik bestand jahrtausendlang nur in Form von logischen und hermeneutischen Disziplinen und wird erst seit dem 20. Jh. ernstlich als empirische Wissenschaft aufgefasst.⁴ Der Begriff des sprachlichen Datums ist deshalb noch unterentwickelt. Bis in die Gegenwart haben Linguisten in dem Bemühen, Evidenz für ihre Hypothesen anzuführen, immer wieder Beispielsätze als Daten ausgegeben. Nun haben das Datum und das Beispiel in der Tat so viel gemeinsam, dass beide Repräsentationen von Phänomenen des Gegenstandsbereichs sind. Aber sie haben völlig verschiedene Funktion und daher verschiedenen methodologischen Status. Ein Beispiel dient der Veranschaulichung eines Theorems. Es hat einen kommunikativen und oft gar didaktischen Zweck, denn es soll beim Leser Verständnis sichern. Dazu muss es solche Eigenschaften aufweisen, auf die es im gegebenen Zusammenhang ankommt; und andere Eigenschaften, die ablenken oder irreführen könnten, sollte es nicht aufweisen. Daten, die diese Anforderungen erfüllen, sind oft schwer oder überhaupt nicht zu finden. Daher ist es üblich, dass der Autor des Theorems auch das Beispiel selbst bildet. Dagegen ist nichts einzuwenden, solange niemand glaubt, das Beispiel sei ein Datum. Als solches müsste es nämlich vom Wissenschaftler unabhängig sein, sonst besteht keine Veranlassung, es als gegeben anzunehmen. Folglich können Daten ggf. als Beispiele dienen, aber nicht umgekehrt.

Weite Teile der Linguistik des 19. und 20. Jh. sind, was die Methodologie und insbesondere die Daten angeht, nahtlose Fortsetzungen jahrtausendalter Vorgeschichte. Die philosophische Grammatik des Altertums, z. B. Aristoteles, arbeitete ausschließlich mit ausgedachten Beispielsätzen. Dieser Brauch wurde von der Schulgrammatik übernommen. Der europäische Strukturalismus, z. B. F. de Saussure, folgte demselben Usus. Im amerikanischen Strukturalismus verfuhr Bloomfield und dann insbesondere Chomsky so.

Tatsächlich vorgekommene Äußerungen als Grundlage empirischer Linguistik haben natürlich ebenfalls eine Vorgeschichte. Da ist zunächst die Philologie, deren Aufgabe es ist, vorhandene Texte zu verstehen. In ihrem Gefolge sind z. B. die Beispielsätze, mit denen die historisch-vergleichende Sprachwissenschaft des 19. Jh. arbeitet, überwiegend aus dem überlieferten Korpus entnommen. In den USA begründet ab 1900 Franz Boas eine

⁴ Die Idee, die Linguistik sei eine Naturwissenschaft, findet sich zwar bereits im 19. Jh. Aber da überwog die wissenschaftspolitische Rhetorik wohl die Bereitschaft, ernstlich auf der Basis von Primärdaten empirisch zu arbeiten.

Tradition, die eine Sprachbeschreibung aus einem Korpus ableitet. Der Übergang von Korpora schriftlicher zu Korpora mündlicher Sprache ist dann einerseits eine Frage des technischen Fortschritts, andererseits aber auch eine Frage der linguistischen Methodik, die es erlaubt, Daten gesprochener Sprache zu repräsentieren und zu analysieren und die erst ausgearbeitet werden musste.

1.2 Abstraktionen aus Daten

Die Systemlinguistik interessiert sich für Sätze und andere sprachliche Konstruktionen derselben Art. Sätze kommen in der Wirklichkeit nicht vor. Man kann sie induktiv als Abstraktionen über Äußerungen gewinnen. Dazu sieht man von den raumzeitlichen Koordinaten der Äußerung und somit von der Situation, in welcher sie verwendet wurde, ab und fokussiert stattdessen auf ihre rein sprachliche Struktur. Sätze sind somit dekontextualisiert mindestens in dem Sinne, wo der außersprachliche Kontext gemeint ist, meist aber außerdem dekontextualisiert in bezug auf sprachlichen Kontext. Das bedeutet natürlich, dass der Sinn, in dem die Äußerung tatsächlich verwendet wurde, verloren geht. Daraus folgt übrigens die vielleicht wichtigste Beschränkung der Systemlinguistik: Ihr Gegenstand ist zwar letztlich die Verständigung; aber der Sinn, welcher in der Verständigung konstruiert wird, bleibt ihr unzugänglich.

Ein Satz steht also um eine Abstraktionsstufe höher als eine Äußerung. Er gewinnt seinen empirischen Status ausschließlich daraus, dass er als Äußerung verwendbar ist. Dies wird ausschließlich dadurch überprüft, dass er *de facto* als Äußerung verwendet wird. Methodisch kann man hier im Prinzip in zwei Richtungen vorgehen:

1. Man kann zunächst einen Satz bilden, diesen dann als Äußerung verwenden und überprüfen, ob die Äußerung erfolgreich war.
2. Man kann vorhandene erfolgreiche Äußerungen aufnehmen und durch methodisch kontrollierte Abstraktion in Sätze überführen.

Den beiden Methoden ist das Problem gemeinsam, dass man Kriterien des Erfolgs von Äußerungen braucht. Die erste Methode hat das zusätzliche Problem der Praktikabilität. Insofern ist es wesentlich sinnvoller, sich als Systemlinguist mit der Analyse solcher Sätze zu befassen, die ihren empirischen Status als Äußerung bereits unter Beweis gestellt haben – eben mit aus Korpora destillierten Sätzen – als zu versuchen, selbst gebildete Sätze als Äußerungen zu lancieren.

1.3 Kompetenzdaten und Performanzdaten

In den allerletzten Jahrzehnten haben verschiedene Richtungen der Linguistik den postulierten empirischen Status dieser Wissenschaft so interpretiert, dass die Datenerhebung endlich ernst genommen werden müsste. Die erste

Strömung dieser Art war in den 1970er Jahren die Variationslinguistik. Es folgten verschiedene funktionalistische Richtungen (vgl. etwa Givón 1984, S. 8) und schließlich die Korpuslinguistik. Mehrere der Postulate dieser Strömungen waren explizit gegen die jahrzehntelange Übung gerichtet, Datenbeschaffung durch Introspektion zu ersetzen.

In dem Maße, in dem diese Attacken von verschiedenen Seiten Wirkung zeigten, gerieten die sterilsten Varianten der Systemlinguistik in die Defensive. Die Verteidigungsstrategie ist seit einiger Zeit eine Unterscheidung zwischen Kompetenz- und Performanzdaten.⁵ Sie setzt die Unterscheidung von Kompetenz und Performanz voraus, die in Chomsky 1965, S. 4 eingeführt und seitdem nie wesentlich revidiert wurde. Die Unterscheidung hat bei Chomsky eine zentrale theoretische Funktion. Außerhalb seiner Linguistik wird die Unterscheidung höchstens als „a methodologically useful preliminary“ (Givón 1984, S. 8) akzeptiert. Insofern die Unterscheidung außerhalb der generativen Linguistik keinen theoretisch relevanten Status hat, entfaltet auch die Verteidigungsstrategie ihre Wirkung nur innerhalb des Modells.

So etwas wie Kompetenzdaten gibt es nicht. „Kompetenzdaten“ ist nur ein neuer Name für durch Introspektion gewonnene Beispielsätze und deskriptive Aussagen. Solche sind aber, wie gesagt, keine Daten im Sinne einer empirischen Wissenschaft. Kompetenzdaten im Wortsinne kann es auch gar nicht geben; der Ausdruck ist eine *contradictio in adiecto*. Denn die Kompetenz ist definiert als ein „tacit knowledge“, das nicht bewusst gemacht werden kann. Darüber kann es also *per definitionem* keine Daten geben.

Sinnvollerweise kann man dagegen in diesem Zusammenhang zwei Unterscheidungen machen:

1. Sätze vs. Äußerungen;
2. sprachnutzende vs. sprachreflexive Daten.

Was Unterscheidung 1 betrifft, so können Eigenschaften von Sätzen durch phonologische, grammatische und semantische Analyse, also mit Methoden der Systemlinguistik, herausgebracht werden. Eigenschaften von Äußerungen sind Phänomene wie Lautstärke, Reaktionszeiten oder Häufigkeiten im Korpus. Sie lassen sich mit den Methoden anderer Subdisziplinen, etwa der Phonetik, Psycho- oder Soziolinguistik, messen.

Was Unterscheidung 2 betrifft, so enthalten Korpora sowohl sprachreflexive als auch sprachnutzende Äußerungen, also solche, die von Aspekten von Sprache handeln und solchen, die das nicht tun. Hier kann man empirisch feststellen,

⁵ Die Begriffe stammen allerdings nicht aus der Linguistik, sondern der Professionalisierungsforschung. Obwohl sie erst vor wenigen Jahren in der Linguistik aufgetaucht sind, werden sie bereits uneinheitlich verwendet. Mit ‚Kompetenzdaten‘ meinen die einen (z. B. H. Krenn) solche Informationen in elektronischen Korpora, die die Analysten durch Annotation dazugehen haben. Die anderen (z. B. H. Uszkoreit) meinen vom Linguisten ersonnene Beispiele (im Gegensatz zu vorgefundenen Daten).

- was die Sprecher auf sprachnutzender Ebene tatsächlich tun,
- was sie auf sprachreflexiver Ebene darüber meinen
- und wie das tatsächliche Sprachhandeln und die Ansichten darüber aufeinander bezogen sind.

Auf Kompetenz vs. Performanz lassen sich diese beiden Unterscheidungen aber nicht zurückführen. Wenn wir zum Zwecke des Arguments die erstere Unterscheidung einmal akzeptieren, so gilt jedenfalls, dass jegliches linguistische Datum auf eine Äußerung, also auf Performanz zurückgeht.

3. Korpus

3.1 Begriffsklärung

Ein Textkorpus ist eine Menge von Texten, insofern sie in einer wissenschaftlichen Untersuchung genutzt wird. Der Begriff hat sich in den letzten Jahrzehnten in der Linguistik in verschiedener Hinsicht gewandelt. Traditionell verstand man unter einem Korpus die Gesamtheit von Texten einer bestimmten Kategorie, z. B. das Korpus der hethitischen Texte⁶ oder das Korpus der Schriften Platons. Für diesen Begriff spielt es noch keine Rolle, ob diese Menge als geschlossene Sammlung verfügbar ist. Dies ist erst eine linguistische Einengung des Begriffs. Und eine jüngere Einengung verlangt, dass ein Korpus in elektronischer Form vorliege.⁷

Das Erfordernis, dass ein Korpus als geschlossene Sammlung vorliege, brachte die Konnotation mit sich, dass das Korpus kompiliert wird. Damit fällt andererseits das Erfordernis der Exhaustivität. Korpus von Texten der Kategorie X heißt nun auch eine Sammlung, die nur eine ausgewählte Teilmenge von X umfasst.⁸

Eine weitere Ausweitung des Begriffs resultiert aus dem Verzicht auf das Erfordernis, dass die größten Untereinheiten eines Korpus Texte sein müssen. Es kann jetzt sprachliches Material jeglicher Art sein; man spricht auch von Korpora von Sätzen, von Informantenurteilen oder von Lexikoneinträgen.

Schließlich war es ein – wenn auch implizites – Korollar des herkömmlichen Korpusbegriffs, dass die enthaltenen Texte unabhängig vom Kompilator zustande gekommen waren. Der Kompilator stellte die Texte nicht her; er

⁶ In diesem Sinne spricht man auch von einer Korpus Sprache.

⁷ Eine andere Einengung, die sich in diversen zeitgenössischen Definitionen findet, ist, dass ein Korpus für linguistische Zwecke zusammengestellt sei. Diese Bedingung dürfte auf disziplinäre Beschränktheit ihrer Autoren zurückgehen. Korpora wurden und werden in erster Linie von Philologen genutzt, weiter von Juristen und vielen anderen, darunter auch Linguisten.

⁸ Hieraus entsteht dann das Problem der Beziehung des Korpus zur Grundgesamtheit. Ein Weg, es anzugehen, ist, das Korpus als Stichprobe im statistischen Sinne zu betrachten. Ein anderer Ansatz wird in § 4.3 gestreift.

sammelte sie bloß. Zeitgenössische Korpora dagegen umfassen auch solches sprachliche Material, welches eigens für das Korpus hergestellt wurde.

Ich regele hier den Sprachgebrauch so, dass ein Korpus eine geschlossen verfügbare Sammlung von Texten einer bestimmten Kategorie ist. Die Erfordernisse, dass es Texte sein und dass diese als geschlossene Sammlung vorliegen müssen, werden beibehalten; die Erfordernisse der Exhaustivität des Korpus und der vorgängigen Existenz der Texte werden fallengelassen. Was durch die Bedingungen ausgeschlossen wird, mag dann unter den Oberbegriff „linguistische Datensammlung“ fallen.

Natürlich erbte der herkömmliche Korpusbegriff ein weiteres implizites Korollar vom Textbegriff, nämlich dass es schriftliche Texte sein müssten. Ich nehme an, dass diese Festlegung mit dem Wesen der Begriffe ‚Text‘ und ‚Korpus‘ nie etwas zu tun hatte, sondern lediglich durch den Stand der Technik bedingt war. Immerhin wirft die Tatsache, dass auch beim heutigen Stand der Technik Texte nur dann wissenschaftlich bearbeitet werden können, wenn sie schriftlich repräsentiert sind, interessante methodische Probleme auf. Mindestens so viel muss klar sein, dass ein Textkorpus jedenfalls kein Ausschnitt aus dem Phänomenbereich, sondern eine Repräsentation davon ist. Es enthält einerseits Abstraktionen über den Rohdaten und andererseits Interpretationen des Kompilers. Es ist essentiell, dass diese methodisch kontrolliert sind. Sonst kann man sich die Arbeit mit Korpora gleich schenken; sie würde uns der Empirie keinen Schritt näher bringen.

3.2 Korpusanalyse vs. Lehnstuhllinguistik

Die traditionelle Grammatik schreibt z. T. seit Jahrhunderten Regeln über syntaktische Konstruktionen und den Gebrauch morphologischer Kategorien fort. Untersuchungen des tatsächlichen Gebrauchs anhand von Korpora geben bisweilen überraschende Resultate.

3.2.1 Hauptkonstituentenstellung

Die Idee, dass der einfache Aussagesatz aus nominalem Subjekt und Prädikat bestehe und dass eine typische Konstituenz des Prädikats ein transitives Verb mit einem nominalen direkten Objekt sei, ist in der Linguistik seit langem wohl etabliert. Beispielsätze des Typs von Sapirs *the farmer kills the duckling* finden sich schon in Grammatiken des 19. Jh. Greenbergs (1963) Grundwortstellungstypologie fußt ganz wesentlich auf Sätzen dieser Struktur. Auf sie sind wiederum weitreichende syntaktische Theorien gegründet worden.

Das Problem mit dieser Satzstruktur ist nur, dass sie in den Texten der meisten Sprachen höchst selten ist und in vielen Sprachen überhaupt nicht vorkommt. Auswertungen von Korpora haben ergeben, dass es pro einfachem Satz normalerweise höchstens ein lexikalisch-nominales Satzglied gibt. Der Satz mit lexikalisch besetztem Subjekt und direktem Objekt fristet sein Dasein hauptsächlich in Grammatikbüchern, und die empirische Basis der von

ihm handelnden Grundwortstellungstypologie wird somit teilweise fragwürdig. Andererseits zeitigt eine Korpusuntersuchung der Konstituentenstruktur von Verbalsätzen neue Erkenntnisse, z. B. über den Zusammenhang zwischen nominaler und pronominaler Vertretung von Subjekt und direktem Objekt einerseits und den Funktionen von Topic und Fokus in der Informationsstruktur andererseits (vgl. Du Bois 1987, Lambrecht 1994, Kap. 4.5.2). Konstruktionen wie Linksversetzung und Spaltsatz werden neu verstanden: sie dienen nicht nur der Hervorhebung von Topic bzw. Fokus, sondern sie gewährleisten auch die Einhaltung des Prinzips, dass es pro einfachem Satz höchstens ein nominales Satzglied gibt.⁹

3.2.2 Das französische Futur

Im Französischen hat man zum Ausdruck des Futurs die Wahl zwischen drei Konjugationskategorien, dem periphrastischen Futur, illustriert in B1, dem synthetischen Futur (B2) und dem Präsens (B3). Die Beispiele stammen aus einem Korpus des gesprochenen kanadischen Französisch, dem *Corpus du français parlé à Ottawa-Hull*.¹⁰

B1. Bien demain, tu vas aller au Bingo, tu vas gagner.

„Morgen wirst du zum Bingo gehen und gewinnen.“

B2. J'ai dit, „Laisse faire, on ira a messe demain matin.“

„Ich habe gesagt: ‚Lass gut sein, wir gehen morgen früh zur Messe.‘“

B3. Il dit, „J'y vas demain matin chez vous“.

„Er sagt: ‚Ich komme morgen früh zu euch.‘“

Was die Verteilung des synthetischen und des periphrastischen Futurs betrifft, so lehrt die traditionelle französische Grammatik, ersteres sei die Default-Variante, während letzteres modale Nuancen oder größere Proximität zum Ausdruck bringe. Bereits die Beispielsätze bestätigen das nicht. Eine Frequenzanalyse des Korpus bringt folgendes Ergebnis: Zukunftsbezug wird durch die drei Konjugationskategorien mit folgender Häufigkeit ausgedrückt:

Kategorie	N	%
periphrast. Futur	2.627	73
synthet. Futur	725	20
Präsens	242	7
Summe	3.594	100

Abb. 2. Verteilung der Varianten zum Ausdruck der Zukunft

⁹ Dies ist seinerseits eine Konsequenz der ‚one new idea‘-Beschränkung (Chafe 1992, S. 92–95).

¹⁰ Das folgende ist, einschließlich der Beispiele, ein Referat von Poplack 2001, S. 415–418.

Rein quantitativ betrachtet, ist also das periphrastische Futur die Default-Variante. Bei der Untersuchung der Faktoren, die das synthetische Futur begünstigen, stellt sich heraus, dass keiner der von den Grammatikern für ausschlaggebend gehaltenen Faktoren einen statistisch nachweisbaren Einfluss hat. Statt dessen ist der Kontextfaktor, der statistisch betrachtet mit Abstand der relevanteste ist, der durch B4 illustrierte:

- B4. Dire que dans quatre cents ans d'ici il va avoir encore des Asselin, puis ils vont encore parler français. Qu'ils parleront pas l'anglais.
 „Zu denken, dass es in 400 Jahren immer noch Asselins geben wird und dass sie immer noch Französisch sprechen werden. Dass sie nicht Englisch sprechen werden.“

In negativen Sätzen wird fast ausnahmslos das synthetische Futur verwendet. Keine deskriptive oder präskriptive Grammatik weiß etwas davon. Es ist durch diese Korpusanalyse festgestellt worden. Nun bleibt es der deskriptiven Grammatik vorbehalten, diesen Zusammenhang zu klären.

3.2.3 Der deutsche Konjunktiv II

In deutlicher Anlehnung an die lateinische Grammatik lehrte die deutsche Grammatik seit zwei Jahrhunderten, der Konjunktiv II drücke irrealer Modalität aus; und zwar drücke der Konjunktiv Imperfekt (*käme*) den Irrealis der Gegenwart, der Konjunktiv Plusquamperfekt (*wäre gekommen*) den Irrealis der Vergangenheit aus. Zu dieser Regel gestand der Duden (1984, S.158 f.) in einer Fußnote einen ausnahmsweisen Zukunftsbezug des Konjunktiv Plusquamperfekt zu.

Eine Untersuchung an Texten (Leirbukt 1991) ergab dagegen, dass Beispiele wie B5 (*Neue Zürcher Zeitung* 18.08.1984) für den tatsächlichen Sprachgebrauch repräsentativ sind:

- B5. [Kontext: nächsten Monat steht eine Volksabstimmung über den radikaleren von zwei Vorschlägen an.] Ein Baustopp nach Leibstadt allein hätte vermutlich in der Volksabstimmung keine schlechten Chancen gehabt. Hingegen wäre es sonderbar, wenn eine Mehrheit der Stimmenden dafür gewonnen werden könnte, die fünf bestehenden Kraftwerke ... ersatzlos stillzulegen.

Offensichtlich wird der Konjunktiv Imperfekt für potentielle Modalität, der Konjunktiv Plusquamperfekt dagegen für irrealer Modalität verwendet. Der Zeitbezug spielt dabei nur eine untergeordnete Rolle. In diesem Falle hat die nächste Auflage des Duden (2005, § 751) die empirische Erkenntnis aufgenommen, und die Beschreibung wurde verbessert.¹¹

¹¹ Leirbukt (2004) stellt sogar noch weitere Ausdehnung des Konjunktiv Plusquamperfekt fest.

3.3 Zwischenbilanz

In den genannten Fällen hat sich die Grammatikschreibung jahrhundertlang im Kreise gedreht. Man fühlt sich an Brechts Galilei erinnert, der die Mitglieder seiner Fakultät auffordert, durchs Fernrohr zu gucken, um die tatsächlichen Verhältnisse zur Kenntnis zu nehmen, die jedoch abwinken, weil Erkenntnisse, die mit den Lehren des Aristoteles unvereinbar sind, von keinem wissenschaftlichen Belang sind. Datenbasiertes, induktives Vorgehen hat spekulative Grammatikerlehre ad absurdum geführt.¹²

Rückblickend kann man sagen, dass wesentliche Einsichten in das Funktionieren menschlicher Sprachen erst errungen worden sind, seit man unvoreingenommen hingeguckt hat, was für Äußerungen wirklich vorkommen und wie sie verwendet werden. Solche Einsichten betreffen nicht nur, wie man erwarten würde, die Semantik und Pragmatik, sondern sogar die grammatische Struktur. Und sie betreffen auch nicht nur die angemessene Analyse von bekannten Phänomenen, sondern z. T. sogar genuin neuartige Phänomene, von denen die Grammatiken und Lexika nichts wissen. Die kann man nur finden, wenn man unvoreingenommen an Texte herangeht. Die Konfrontation mit nicht selbst erzeugten Daten ist eine Quelle der Inspiration für den Linguisten, der nach Erkenntnissen strebt.

Selbstverständlich ist Korpuslinguistik nur eine neben anderen Methoden und nicht alleinseligmachend. Z. B. setzt die Beantwortung vieler linguistischer Fragen das Verfügen über Minimalpaare und Paradigmen voraus, und die sind oft sehr viel leichter durch Informantenbefragung als durch Durchforstung von Korpora zu bekommen. Aber die darauf gegründete grundsätzliche Skepsis gegenüber Korpusarbeit verliert an Kraft, je umfangreicher die Korpora werden. Wenn eine Form in dem Korpus der Gesamtheit dessen, was ein Sprecher in seinem Leben gehört hat, nicht auftritt, dann bleibt ihm nichts anderes übrig, als sie in Analogie zu gehörten Formen zu bilden. Ebenso verhält sich der Linguist angesichts von endlichen Korpora; und je größer diese sind, desto wahrscheinlicher ist es, dass er im Einzelfall damit richtig liegt.

4. Dokumentation

4.1 Dokumentation und Beschreibung einer Sprache

Varietäten oder Genres einer Sprache werden schon seit langem dokumentiert, z. B. in umfassenden Editionen oder Anthologien. Die Idee, eine ganze Sprache zu dokumentieren, ist relativ neu. Sie ist eine Reaktion auf das erst in den achtziger Jahren des 20. Jahrhunderts erwachte Bewusstsein von der akuten Bedrohung der meisten Sprachen der Welt und von der Dringlichkeit,

¹² Sinclair (1991) schließt, dass Linguisten nun endlich dem Text mehr vertrauen können als ihren eigenen Intuitionen oder den auf ihnen basierenden allgemein akzeptierten grammatischen Beschreibungen.

sie umfassend zu dokumentieren, falls sie dem Gedächtnis der Nachwelt nicht endgültig verloren gehen sollen.

Innerhalb der umfassenden Darstellung einer Sprache sind die Dokumentation und die Beschreibung aufeinander zugeordnet (s. Lehmann 2001, 2002). Die Dokumentation besteht i. w. aus einem Textkorpus, das gewissen Bedingungen genügt, auf die ich in § 4.3 komme. Die Beschreibung hat das Sprachsystem und die Situation der Sprache zum Gegenstand. Die Beschreibung verhält sich also zur Dokumentation so wie die *langue* zur *parole*. Eine Einheit des Sprachsystems steht zur einer Einheit einer Äußerung im Verhältnis des Typs (engl. *type*) zum Vorkommen (engl. *token*). Die Darstellung einer Sprache kann versuchen, dieses logische Verhältnis im formalen Aufbau von Beschreibung und Dokumentation nachzubilden. Die Beschreibung enthält dann Mengen von Objekten, deren jedes eine sprachliche Einheit repräsentiert und eine Identifikationsnummer hat. Das Korpus enthält nicht, wie natürlich gewachsene Korpora, die buchstäbliche Repräsentation sprachlicher Einheiten, sondern lediglich Ketten von deren Identifikationsnummern. Das Ganze lässt sich z. B. als objektorientierte relationale Datenbank implementieren. Formal betrachtet, liegt die Schnittstelle zwischen Beschreibung und Dokumentation dann in der Identifikation der Nummern zwischen den beiden Teilen. Falls das Korpus einmal jemand ansehen will, werden online zu den enthaltenen Vorkommen anhand der Identifikationsnummern die betreffenden Objekte des Systems herausgesucht und in der gewünschten Form, z. B. orthographisch oder phonetisch, dargestellt.

4.2 Sprachliche Variation zwischen Beschreibung und Dokumentation

Eine solche Implementation stellt das Problem, wie sprachliche Variation zu erfassen ist, mit besonderer Klarheit. Denn die in den Rohdaten enthaltenen Vorkommen sprachlicher Einheiten sind selbstverständlich alle verschieden. Man kann aber nicht pro Textvorkommen einen Typ im System ansetzen, sonst wird das System infinit. Folglich muss man verschiedene Vorkommen auf einen Typ beziehen und nur diesen im Korpus repräsentieren, muss also von gewissen Unterschieden zwischen Vorkommen im Korpus abstrahieren. Das bedeutet aber, dass man zwei Hauptebenen der sprachlichen Variation unterscheidet:

- Die obere Variationsebene ist Bestandteil des Sprachsystems oder der Sprachnorm. Die Varianten – z. B. Allophone und Allomorphe – sind Elemente der Sprachbeschreibung.
- Die untere Variationsebene gehört der Äußerung an. Die Varianten – z. B. idiolektale phonetische Feinheiten – werden übergangen und nirgends repräsentiert.

Das Schema Abb. 3 illustriert das Gemeinte am Beispiel des deutschen /r/. Eine enge phonetische Transkription als Repräsentation eines Textes des Kor-

pus kann die Variation auf beliebigen Stufen der Feinheit wiedergeben. Tatsächlich wird aber eine Entscheidung gefällt, phonetische Variation unterhalb einer bestimmten Abstraktionsebene zu ignorieren. Man zieht also in die Variation eine Trennlinie ein.

Abb. 3. Allophonie von /r/

Die Variation oberhalb der Trennlinie ist Bestandteil des Systems (bzw., mit Coseriu, der Norm). Die Sprachbeschreibung enthält daher die beiden Allophone [ʀ] und [ʁ] (die z. B. an dritter Position in *Kurt* auftreten). Die Variation unterhalb der Trennlinie ist für das System irrelevant. Alle Varianten unterhalb der Trennlinie stehen in der Relation eines Vorkommens zu einem Typ des Systems, welcher ein Allophon der untersten Ebene ist. Die im Korpus tatsächlich vorgekommen [ʀ] werden als Vorkommen des Typs [ʁ] repräsentiert; dass sie tatsächlich [ʀ] waren, wird nie wieder jemand wissen.

Hieraus folgen zwei Grundsätze für die Edition eines Textkorpus:

- Vor Abschluss der Beschreibung der Sprache kann man nicht sicher sein, ob eine für irrelevant gehaltene Variation nicht doch eine Funktion hat oder in anderer Hinsicht theoretisch interessant ist. Deshalb ist es vorsichtiger, die Trennlinie möglichst niedrig anzusetzen, d. h. bei der Edition Variation nicht auszumerzen, sondern darzustellen. Es besteht keine Gefahr, den Wald vor lauter Bäumen nicht mehr zu sehen. Denn die Varianten, die man in der Korpusrepräsentation auseinandergehalten hat, sind in der Beschreibung des Systems auf eine Invariante bezogen. Jeder, der sich bloß für die Invariante interessiert, kann die Variation ausblenden. Aber wenn die Variation bereits durch indistinkte Repräsentation im Korpus ausgeblendet ist, ist sie der Forschung für immer verloren.
- Die Rohdaten, von denen das Korpus eine Repräsentation ist, sind aufzubewahren.

4.3 Repräsentativität und Qualität

Die beiden wichtigsten Anforderungen an ein Textkorpora, das zur Dokumentation einer Sprache dienen soll, sind Repräsentativität und Qualität (vgl. Lehmann 2001, § 5). Die Repräsentativität erfordert, dass die Variation auf folgenden Dimensionen im Korpus vertreten ist:

- Diastratik: Sprecher verschiedener Gruppen,
- Diatopik: Situationskontexte,
- Kommunikationsaufgaben und Textgenres,
- Themen,
- Medien.

Der letzte Punkt erfordert, dass mündliche Texte stärker vertreten sind als schriftliche, denn sie verwenden das fundamentale Kommunikationsmedium.¹³

Die genannten Dimensionen der Variation betreffen z. T. Außersprachliches, z. T. Pragmatisches und Semantisches. Der Systemlinguist will natürlich sicherstellen, dass auch strukturelle Variation herrscht, insbesondere, dass alles, was zum Sprachsystem gehört, im Korpus auch vertreten ist. Bei wohlbekannten Sprachen kann man auf Wege sinnen, dies gezielt sicherzustellen. Bei schlecht beschriebenen Sprachen ist es unmöglich, da man die Strukturen, die man repräsentiert sehen möchte, noch nicht kennt. Es besteht aber die Vermutung, dass, wenn man für funktionale Variation auf all den genannten Dimensionen sorgt, sich die strukturelle Variation von selbst einstellen wird, nach dem Prinzip ‚form follows function‘.

Die Qualität der Texte betrifft zunächst ihre Erzeugung, also die phonetische, grammatische, stilistische Kunst des Sprechers sowie den Inhalt der Nachricht. Dieses Erfordernis, das dem professionellen Linguisten seltsam vorkommen mag, wird den Philologen ebenso wie auch den Laiennutzer der Dokumentation wie eine Trivialität anmuten. Die Dokumentation wird für die Nachwelt gemacht; für minderwertige Produkte hat sie weder Interesse noch Ressourcen. Die Qualität betrifft aber ebenso die Arbeit des Kompilators des Korpus, also die Güte der Video- oder Audioaufnahme, der Transkription, der Annotation und der Metadaten. Qualität geht hier eindeutig vor Quantität.

5. Schlussfolgerungen

5.1 Normative Grammatik vs. tatsächlicher Sprachgebrauch

Seit der Antike war die Grammatik eine *ars*, also eine Kunst. Bis in die Neuzeit hinein gehörte Sprachwissenschaft zusammen mit den Philologien in die Faculty of Arts oder die Faculté des Lettres. Der normative Grammatiker

¹³ Die Frage, wie die diachrone Variation einzubeziehen ist, kann hier nicht angemessen gewürdigt werden.

beherrscht die Kunst des grammatischen Sprechens und Schreibens und vermittelt sie seinen Schülern so ähnlich, wie ein Klaviervirtuose diese Kunst seinen Schülern vermittelt. Diese Auffassung von Grammatik wird heute von keinem Sprachwissenschaftler mehr vertreten. Aber sowohl Grammatiken, die Richtschnuren für den korrekten Ausdruck liefern wollen, als auch Grammatiken, die vorgeben, deskriptiv zu sein, werden de facto bis auf den heutigen Tag von Grammatikern im Vollgefühl persönlicher Sprachbeherrschung geschrieben.¹⁴

Eigentlich hat sich der Anspruch des Grammatikers in den letzten 200 Jahren sogar zugespitzt. Ein Musiker oder ein Maler sind zwar in der Lage, ihre Kunst den Gesellen zu vermitteln, erheben aber nicht den Anspruch, die gesamte Musik oder die gesamte Malerei einer Gesellschaft vollgültig zu repräsentieren. Der Grammatiker der Antike und des Mittelalters war hierin jedem anderen Künstler vergleichbar. Gerade diesen Alleinvertretungsanspruch haben Linguisten während der Herausbildung der strukturalen Sprachwissenschaft wenn auch nicht explizit, so doch immer selbstverständlicher erhoben. Dabei ist gerade in den Jahrzehnten der generativen Grammatik, als subtilste Konstruktionsunterschiede den Unterschied zwischen grammatisch und ungrammatisch ausmachen und somit den Ausschlag in theoretischen Entscheidungen geben sollten, sehr deutlich geworden, dass zwischen Grammatikern in bezug auf die Sprache, die sie zu beherrschen glauben, überhaupt keine Einigkeit herrscht. Es gibt folglich keine Basis für die Annahme, der Grammatiker sei allein imstande, den herrschenden Sprachgebrauch zu beurteilen. Es ist heute völlig klar, dass empirische Sprachbeschreibung die Auswertung von Korpora erfordert.

Andererseits stellt sich die Frage neu, wie eine normative Grammatik zu begründen ist. Wenn der einzelne Linguist hierfür keine relevante Instanz ist und die Gemeinde der Linguisten kein gemeinsames Urteil hat, auf welcher Grundlage kann man dann überhaupt Regeln für richtiges Reden und Schreiben formulieren? Wie es scheint, hat man in dieser Frage bescheiden auf die Stufe der Antike zurückzukehren, wo eine einzelne Person oder Gruppe durch ihren Sprachgebrauch einigen als Vorbild dient, anderen jedoch nicht. Irgendein allgemein-normativer Anspruch ergibt sich für diese Person oder Gruppe daraus nicht.

5.2 Sprachbeherrschung

Die generative Grammatik (Chomsky 1965, S. 3 f.) hat den idealen Sprecher-Hörer postuliert, der seine Sprache vollkommen beherrscht und keinen Beeinträchtigungen der Performanz unterliegt. Dieser hat jedoch kein empirisches Korrelat. In der wissenschaftlichen Praxis sind die idealen Sprecher-Hörer überraschenderweise immer identisch mit den jeweiligen Linguisten. Der

¹⁴ S. Klein (2004) über Präskriptivität in modernen deskriptiven Grammatiken.

ideale Sprecher-Hörer entpuppt sich somit als methodischer Schachzug, um unter der Fahne empirischer Wissenschaft weiterhin normative Grammatik zu betreiben.

Tatsächlich gibt es in diesem Zusammenhang aber ein interessantes empirisches Feld, das viel zu wenig untersucht worden ist. Fasst man nämlich Kompetenz als Sprachbeherrschung auf, so wird es möglich, den Begriff zu operationalisieren. Man kann Kriterien für die Beherrschung einer Sprache definieren, ähnlich, wie das z. B. der Europarat (Council of Europe 2000, Goethe-Institut & Inter Nationes 2001) für die Beherrschung von Fremdsprachen getan hat. Diese Aufgabe ist in jeder Hinsicht anspruchsvoll, denn die Kriterien sind sehr heterogen und unterliegen Einstellungen und Bewertungen der Sprachgemeinschaft, die ihrerseits wieder Gegenstand empirischer Untersuchungen sind. Zudem ist das ganze Thema ein heißes Eisen. Denn während das Faktum, dass Menschen Fremdsprachen zu unterschiedlichen Graden beherrschen, jedem bekannt und auch in Ausbildungsinstitutionen als solches akzeptiert ist, wird über die unterschiedliche Beherrschung der Muttersprache i. a. geschwiegen. Deutschlehrer und Germanistikprofessoren (um nur diese zu nennen) ahnen zwar dumpf, dass dieser Begriff ein empirisches Korrelat hat. Aber erstens sind die Kriterien der Beurteilung nie systematisiert worden. Zweitens ist der Vergleichsmaßstab i. a. die jeweilige Sprachnorm, und ein reflektierender Mensch hat Zweifel, ob der schiere Vergleich von jemandes Sprachbeherrschung mit der Sprachnorm zu einer bedeutsamen oder gar gerechten Beurteilung der ersteren ausreicht. Und drittens spielt bei Bewertungen immer auch die Ätiologie mit hinein: Wenn jemand eine Sprache weniger gut beherrscht, liegt das dann an minderer Begabung, an minderer bzw. anders orientierter Motivation, am sozialen Umfeld, an schlechter Ausbildung oder vielleicht noch anderen Faktoren? Wer hierüber empirisch forschen will, begibt sich auf vermintes Gelände.

Kein Systemlinguist würde seine introspektiven Urteile öffentlich damit rechtfertigen, dass er die Sprache zu einem Grade beherrscht, der die Beziehung anderer Mitglieder der Sprachgemeinschaft erübrigt. De facto sind jedoch die Art und das Maß, in dem Linguisten ihre Sprache beherrschen, ihrerseits wieder empirischer Untersuchung zugänglich. Und wenn man die Frage unvoreingenommen angeht, ist das eine Untersuchung mit offenem Ausgang. Solange freilich diese Frage nicht empirisch untersucht ist, besteht für Linguisten kein Grund anzunehmen, ihre introspektiven Urteile könnten als Basis ihrer deskriptiven Aussagen dienen.¹⁵

¹⁵ In Kepser/Reis 2005 z. B. wird davon ausgegangen, es gäbe so etwas wie „introspective evidence“, aber diese Annahme wird an keiner Stelle plausibilisiert.

5.3 Chancen und Gefahren

Daten erfüllen eine unentbehrliche Funktion in einer empirischen Wissenschaft. Aber sie sind kein Selbstzweck. Viele Untersuchungen, die unter der Flagge der Wissenschaft segeln, beschränken sich auf die Sammlung und Sortierung von Daten. Das ist kein spezifisch linguistisches Phänomen, sondern passiert in allen empirischen Wissenschaften. Viele Naturwissenschaften ertrinken in Daten. Satelliten nehmen Strahlung sämtlicher Wellenlängen in beliebig feiner Auflösung auf und funken die Werte Computern zu, die sie digitalisiert abspeichern. Es werden Löcher Hunderte von Metern tief in den Meeresboden gebohrt und das Profil der Sedimente über die Länge der Bohrsäule verzeichnet. Es gibt Karten, Diagramme, Kurven über alles, was sich aufzeichnen lässt. Ständig werden Technologien entwickelt und von Forschungsinstituten gekauft und eingesetzt, um noch mehr Daten aufzunehmen. Wenn irgendetwas technisch machbar ist, wird es gemacht. Es werden mehr Daten aufgenommen, als man je auszuwerten und für die Konstruktion von Theorien zu nutzen hoffen kann.

Viele philologische Dissertationen vom 18. bis ins 20. Jh. sind wenig mehr als Sammlungen von Belegstellen zu einem bestimmten Thema. Viele Arbeiten, die kommunikationsanalytisch sein sollen, gehen kaum über die Transkription von Gesprächen und deren Interpretation mithilfe gesunden Menschenverstandes hinaus. Wenn wir Korpuslinguistik betreiben und uns dafür einsetzen, dass in der Linguistik echte Daten die selbstersonnenen Beispiele ersetzen, so laufen wir Gefahr, als Faktenhuber missverstanden zu werden. Tatsächlich wäre für die Linguistik nichts gewonnen, wenn sich Korpuslinguistik auf Faktenhuberei reduzierte; und die rhetorische oder wissenschaftspolitische Abdrängung der Korpuslinguistik in diese Ecke wäre ebenso schädlich, weil sie einer falsch verstandenen theoretischen Linguistik Auftrieb verschaffen würde. Es ist daher kein Zufall, dass die neue Zeitschrift der Disziplin sich *Corpus Linguistics and Linguistic Theory*¹⁶ nennt und in ihrem Ankündigungstext verlangt, Beiträge müssten nicht nur „corpus-based“, sondern auch „theoretically relevant“ sein. Korpuslinguistik läuft nicht auf eine Senkung der theoretischen, sondern auf eine Hebung der methodischen Ansprüche hinaus.

Literatur

Chafe, Wallace (1992): The importance of corpus linguistics to understanding the nature of language. In: Svartvik, Jan (Hg.): Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991. Berlin/New York: Mouton de Gruyter. (Trends in Linguistics). S. 79–97.

¹⁶ Herausgegeben von Stefan Th. Gries und Anatol Stefanowitsch, publiziert von Mouton de Gruyter. Die Herausgeber postulieren ein „commitment to the systematic and exhaustive analysis of naturally occurring language“.

- Chomsky, Noam (1965): Aspects of the theory of syntax. Cambridge, Mass.: MIT Press. (Special Technical Report).
- Corbett, Greville C. (2005): Suppletion in personal pronouns: Theory vs. practice, and the place of reproducibility in typology. In: *Linguistic Typology* 9, S. 1–23.
- Council of Europe (2000): A common European framework of reference for languages: learning, teaching, assessment. Cambridge: Cambridge University Press.
- Du Bois, John W. (1987): The discourse basis of ergativity. In: *Language* 63(4), S. 805–855.
- Duden (1984): Grammatik. Mannheim: Bibliographisches Institut.
- Duden (2005): Grammatik. Mannheim: Bibliographisches Institut (7. Aufl.).
- Givón, Talmy (1984): Syntax. A functional-typological introduction. Vol. I. Amsterdam/Philadelphia: J. Benjamins.
- Goethe-Institut Inter-Nationes et al. (Hg.) (2001): Gemeinsamer Europäischer Referenzrahmen für Sprache Lernen, lehren, beurteilen. München: Goethe-Institut (www.goethe.de/referenzrahmen).
- Hempel, Carl Gustav (1966): Philosophy of natural science. Englewood Cliffs, N.J.: Prentice-Hall. (Prentice-Hall foundations of philosophy series).
- Kepser, Stefan/Reis, Marga (2005): Evidence in linguistics. In: Kepser, Stefan/Reis, Marga (Hg.): *Linguistic evidence. Empirical, theoretical and computational perspectives*. Berlin: Mouton de Gruyter. (Studies in Generative Grammar, 85). S. 1–6.
- Klein, Wolf Peter (2004): Deskriptive statt präskriptiver Sprachwissenschaft!? Über ein sprachtheoretisches Bekenntnis und seine analytische Präzisierung. In: *Zeitschrift für germanistische Linguistik* 32, S. 376–405.
- Lambrecht, Knud (1994): Information structure and sentence form. Topic, focus and the mental representations of discourse referents. Cambridge: Cambridge University Press.
- Lehmann, Christian (2001): Language documentation: a program. In: Bisang, Walter (Hg.): *Aspects of typology and universals*. Berlin: Akademie Verlag. (Studia Typologica). S. 83–97.
- Lehmann, Christian (2002): Structure of a comprehensive presentation of a language. With particular reference to the interface between text, grammar and lexicon. In: Tsunoda, Tasaku (Hg.): *Basic materials in minority languages 2002*. Osaka: Osaka Gakuin University. (Endangered Languages of the Pacific Rim Publication Series). S. 5–33.
- Lehmann, Christian (2004): Data in linguistics. In: *The Linguistic Review* 21(3/4), S. 275–310.
- Leirbukt, Oddleif (1991): „Nächstes Jahr wäre er 200 Jahre alt geworden.“ Über den Konjunktiv Plusquamperfekt in hypothetischen Bedingungsgefügen mit Zukunftsbezug. In: *Zeitschrift für germanistische Linguistik* 19(2), S. 158–193.
- Leirbukt, Oddleif (2004): Über Konjunktiv Plusquamperfekt und würde + Infinitiv II als Ausdruck von Potentialität oder Irrealität in Konstruktionen mit Gegenwarts- oder Zukunftsbezug. In: Leirbukt, Oddleif (Hg.): *Tempus/Temporalität und Modus/Modalität im Sprachenvergleich*. Tübingen: G. Narr. (Eurogermanistik, 18). S. 205–230.
- Oevermann, Ulrich (2002): Klinische Soziologie auf der Basis der Methodologie der objektiven Hermeneutik – Manifest der objektiv hermeneutischen Sozialforschung. Institut für objektive Sozial- und Kulturforschung (www.ihs.de am 09.03.2006).
- Poplack, Shana (2001): Variability, frequency, and productivity in the irrealis domain of French. In: Bybee, Joan/Hopper, Paul (Hg.): *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: J. Benjamins. (Typological Studies in Language, 45). S. 405–428.
- Sinclair, J. M. (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

ANKE LÜDELING

Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik

Abstract

Es gibt viele linguistische Forschungsfragen, für deren Beantwortung man Korpusdaten qualitativ und quantitativ auswerten möchte. Beide Auswertungsmethoden können sich auf den Korpustext, aber auch auf Annotationssebenen beziehen. Jede Art von Annotation, also Kategorisierung, stellt einen kontrollierten und notwendigen Informationsverlust dar. Das bedeutet, dass jede Art von Kategorisierung auch eine Interpretation der Daten ist. In den meisten großen Korpora wird zu jeder vorgesehenen Annotationssebene, wie z. B. Wortart-Ebene oder Lemma-Ebene, genau eine Interpretation angeboten. In den letzten Jahren haben sich neben den großen, ‚flach‘ annotierten Korpora Korpusmodelle herausgebildet, mit denen man konfligierende Informationen kodieren kann, die so genannten Mehrerebenen-Modelle (*multilevel standoff corpora*), in denen alle Annotationsebenen unabhängig vom Text gespeichert werden und nur auf bestimmte Textanker verweisen. Ich argumentiere anhand der Fehlerannotation in einem Lernerkorpus dafür, dass zumindest Korpora, in denen es stark variierende Annotationsbedürfnisse und umstrittene Analysen geben kann, davon profitieren, in Mehrebenen-Modellen kodiert zu werden.

1. Einleitung

Korpora werden zur Beantwortung von linguistischen Fragestellungen immer wichtiger. In einigen Bereichen wie der historischen Linguistik oder der Dialektologie wird natürlich traditionell schon immer mit Korpora gearbeitet (es gibt ja oft keine andere Datenquelle), aber in den letzten Jahren finden Korpora als Datenquelle auch in eher theoretisch ausgerichteten Arbeiten immer häufiger Anwendung (siehe zum Beispiel Gries 2003; mehrere Artikel in Kepser & Reis 2005; Featherston, in diesem Band; Müller, in diesem Band). Neben der Verwendung von Korpora als eine Art ‚Beispielbank‘ und rein qualitativen Analysen wird auch immer häufiger quantitativ gearbeitet. Neben relativ einfachen beschreibenden statistischen Techniken werden auch statistische Tests, Kollokationsanalysen, multivariate Techniken und Verfahren aus dem maschinellen Lernen eingesetzt. Die mathematischen Verfahren und deren Voraussetzungen sind in vielen Arbeiten beschrieben worden (Biber 1988; Oakes 1998; Manning & Schütze 1999; Baayen 2001; Evert 2005; Gries 2003 und viele andere).

Im Gegensatz zu anderen empirischen Gebieten wie der Psycholinguistik befindet sich die Entwicklung von Standards für die quantitative Analyse von Korpusdaten noch in der Diskussion. Neben der Frage, *wie* gezählt und ge-

rechnet wird, ist hier die Frage entscheidend, *was* eigentlich gezählt wird. In diesem Papier möchte ich nicht auf die mathematischen Techniken selbst eingehen, sondern auf die Grundlage jeder quantitativen Analyse, nämlich die qualitative Analyse oder Kategorisierung der Daten.

Jede quantitative Analyse ist abhängig von einer vorherigen Kategorisierung. Allerdings ist eine solche Kategorisierung oft nicht unproblematisch und unumstritten – diese Tatsache ist bekannt und eigentlich trivial. Trotzdem findet man in sehr vielen Arbeiten, die Korpora als Datenquelle für quantitative Untersuchungen verwenden, kaum Hinweise auf die zugrunde liegende Kategorisierung – oft werden weder die verwendeten Kategorien (Tagsets) noch die Richtlinien zur Vergabe der Kategorien angegeben. Noch seltener findet man Angaben zum so genannten „Inter Annotator Agreement“, der Zuverlässigkeit bei der Vergabe von Kategorien. Ohne diese Angaben sind quantitative Ergebnisse eigentlich nicht zu bewerten.

Ich möchte in diesem Papier zunächst verdeutlichen, dass es unmöglich ist, nicht zu kategorisieren (Abschnitt 2), auch wenn keine explizite Kategorisierung vorliegt. Dann zeige ich exemplarisch anhand eines Lernerkorpus, wie stark Kategorisierungen die Zählungen beeinflussen können (Abschnitt 4) und welche Schwierigkeiten sich durch fehlende Evaluierungen ergeben. Das verwendete Lernerkorpus Falko wird kurz in Abschnitt 3 vorgestellt. Im letzten Teil (Abschnitt 5) möchte ich für eine Mehrebenenkorpusarchitektur argumentieren, in der die unterschiedlichen Hypothesen zumindest explizit gemacht werden können, so dass die Grundlage für jede quantitative Auswertung sichtbar ist.

2. Kategorisierung und quantitative Analyse

Ein Grunddatum der empirischen Linguistik ist die sprachliche Äußerung. Korpora sind immer interpretierte Repräsentationen solcher Äußerungen und unterscheiden sich als solches bereits von den ursprünglichen Äußerungen. So muss man unterscheiden zwischen einer sprachlichen Äußerung, die in einer Zeitung steht und einem Korpus, in dem diese Äußerung repräsentiert ist (siehe auch Moisl, erscheint). Zum einen unterscheiden sich die Reaktions- und Analysemöglichkeiten (so kann man z. B. keinen Leserbrief an das Korpus schreiben, aber evtl. statistische Auswertungen vornehmen), zum anderen unterscheiden sich die Kontextinformationen. Noch deutlich wird dieser Schritt, wenn man an elektronische Wiedergaben von historischen Manuskripten oder Transkriptionen gesprochener Sprache denkt. In jedem Fall werden verschiedene Formen im Original notwendigerweise auf gleiche Formen im Korpus zurückgeführt – selbst, wenn man sehr diplomatisch arbeitet.¹

¹ Bloomfield (1933) unterscheidet in diesem Zusammenhang zwischen ‚primary data‘ und ‚secondary data‘, Himmelmann (1998) zwischen ‚raw data‘, ‚primary data‘ und ‚secondary data‘. Siehe auch Lehmann (in diesem Band).

Die Auswertung von Korpusdaten erfolgt dann oft auf einer weiteren Interpretationsstufe, der Annotation.

In beiden Interpretationsstufen werden Entscheidungen getroffen. Wir haben es also nie mit uninterpretierten Daten zu tun. Die Interpretation von Korpusdaten stellt in jedem Fall eine Kategorisierung dar, also kontrollierten und erwünschten Informationsverlust. Dies gilt für die qualitative genauso wie für die quantitative Auswertung, für die Arbeit ohne explizite Annotationen genauso wie für die Arbeit mit expliziten Annotationskategorien, unabhängig davon, wie die Korpusdaten verarbeitet werden.

Für viele Fragestellungen genügen einige wenige Beispiele – man braucht keine quantitativen Aussagen. Wenn man zum Beispiel nur illustrieren möchte, dass Lerner des Deutschen als Fremdsprache Schwierigkeiten bei der Verwendung des Korrelat-*es* haben, indem sie ein *es* einfügen, wo keines hingehört (1a) oder ein obligatorisches *es* nicht setzen ((1b); für eine detaillierte Analyse siehe Hirschmann 2005), genügt es, einige solche Fälle zu finden – man verwendet das Korpus als eine Art ‚Beispielbank‘.² Wichtig ist aber auch in diesen Fällen, dass die Beispiele nach externen Kriterien ausgesucht und gruppiert werden.³ Man hat vorher festgelegt, welche Konstruktionen betrachtet werden und auch, wie man solche Konstruktionen finden kann. Die Kriterien müssen so explizit sein, dass für Fälle wie (1c), wo das *es* entweder als stilistisches Problem oder als grammatischer Fehler gewertet werden kann, eine eindeutige und nachvollziehbare Entscheidung getroffen werden kann.

(1a) *Z.B. wenn es das konventionelle Wort „Lehrer“ existiert, sollte man nicht das Wort „Unterrichter“ verwendet.* (Falko L2, Text 36)

(1b) *Eine Vielzahl syntaktischer Mittel, z. B. Relativsatz, Präpositionalphrase, Genitivergänzung, adjektivische Modifikationen, ermöglichen ein bestimmtes Mitglied dieser Klasse hervorzuheben.* (Falko L2, Text 17)

(1c) *Das zweite Paradox ist es, dass der Mann sich weigert hereinzutreten, obwohl das Tor offensteht, und den Befehl des Türhüters befolgt.* (Falko L2, Text 5)

Bei der quantitativen Auswertung muss genauso nach externen Kriterien festgelegt werden, was gezählt wird. In der Korrelatstudie von Hirschmann beispielsweise sind nur sehr wenige echte Korrelatfehler zu finden, es gibt aber einige Zweifelsfälle (analog zu (1c)). Wenn die Zweifelsfälle nicht als Fehler gewertet werden, kann man schließen, dass Lerner in dieser Konstruktion kaum mehr Fehler machen als Muttersprachler. Wenn sie hingegen als Fehler gewertet werden, könnte man schließen, dass Lerner mit Korrelatkonstruk-

² Die Beispiele stammen aus dem Lernerkorpus Falko, das in Abschnitt 3 eingeführt wird. Die Beispiele enthalten außer den Korrelatfehlern noch andere Fehler, die aber hier nicht betrachtet werden.

³ Dies gilt letztendlich auch für die sogenannten korpusgesteuerten (corpus driven) Ansätze (Xiao, erscheint) oder auch für Arbeiten wie die von Golcher (2006), in denen ohne explizites Lexikon oder andere sprachspezifische Interpretationen Muster in einer Sprache gefunden werden.

tionen Probleme haben. (auf die Schwierigkeiten der Fehlerkategorisierung komme ich in Abschnitt 4 zurück).

Dass die Kategorisierung bei kleinen Korpora die Ergebnisse wesentlich beeinflussen kann, ist einleuchtend. Kleine spezielle Korpora werden meistens von Hand vorverarbeitet. Wenn die Verarbeitungskriterien und -verfahren gut dokumentiert sind, sind die Entscheidungen nachvollziehbar.

Aber auch bei großen Korpora (hier meine ich zum Beispiel die über das IDS oder über die Berlin-Brandenburgische Akademie verfügbaren Korpora, aber auch Korpora, die in der Computerlinguistik eine Rolle spielen) spielt die Kategorisierung eine entscheidende Rolle. Große Korpora werden in der Regel automatisch vorverarbeitet, d. h. tokenisiert, getaggt, lemmatisiert oder auf anderen Ebenen annotiert. In den meisten Fällen geschieht die Verarbeitung von Korpora sequentiell. Zunächst wird tokenisiert, dann werden die so gefundenen Tokens lemmatisiert und mit Wortarten getaggt, danach werden größere Einheiten (wie Chunks oder Sätze) ausgezeichnet (Manning & Schütze 1999; Carstensen et al. 2004). Bereits das automatische Tokenisieren kann problematisch sein. In den meisten europäischen Sprachen werden graphische Wörter als Tokens angesehen.⁴ Dies ist in vielen Fällen problematisch, wenn man ‚Token‘ als ‚Wort‘ ansehen möchte (Evert & Fitschen 2001, Haß-Zumkehr 2002): Einerseits hat man den Fall dass mehrere Tokens als eine Einheit verarbeitet werden sollten (es gibt keinen linguistischen Grund, Namen mit Spatium wie *New York* oder *Weil der Stadt* anders zu behandeln als *Hamburg*; es gibt eine Reihe von linguistischen Gründen, nicht kompositionelle Sequenzen mit Spatium wie engl.: *pigeon hole* ‚Postfach‘ als Einheiten zu sehen). Andererseits gibt es Tokens, die als mehrere lexikalische Einheiten angesehen werden sollten (dazu gehören zum Beispiel Verschmelzungen wie *siehste*). In einer sequentiellen Architektur bauen alle weiteren Verarbeitungsschritte auf dieser Tokenisierung auf, d. h., alle Probleme werden ‚mitgeschleppt‘. Die Wortartzuweisung und die Lemmatisierung müssen irgendwie mit merkwürdigen Tokens umgehen. Die Strategien sind hier unterschiedlich. Ganz generell ist es aber so, dass bekannte Probleme (wie *New York* oder *zum*) einheitlich behandelt werden können, während seltenere Fälle nicht abgefangen werden können, wie die folgenden Beispiele⁵ zeigen.

⁴ Dies ist anders für viele nicht-europäische Sprachen, aber auch für historische Texte mit scripta continua (siehe Lüdeling, Poschenrieder & Faulstich 2005 oder Ostler, erscheint).

⁵ Alle hier verwendeten Korpora wurden mit dem DecisionTreeTagger (Schmid 1994, <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>) annotiert. Hier wird ein Tabellenformat verwendet (siehe Abschnitt 5). Jedes Token ist mit einem Wortarttag und einem Lemma versehen. Die Wortarttags entstammen dem Stuttgart-Tübingen Tagset, <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>. Die hier vorkommenden Tags sind ADJA für attributives Adjektiv, KOUS für unterordnende Konjunktion, ART für Artikel, NN für Nomen, NE für Name. Die Korpora sind unter <http://www2.hu-berlin.de/korpling/korpora/index.php> verfügbar (kostenlose Registrierung erforderlich).

(2a) *Ach/ITJ/lach,/\$/I, siehstelADJA/<unknown>!!\$/!!*

(aus dem Parlamentsredenkorpus)

(2b) *Rolf Sannwald, Vorstandsmitglied der 'Calwer Decken- und Tuchfabriken AG, Calw, und AR-Vorsitzender der Wolldeckenfabrik WeillKOUSIweil der!ARTId Stadt!NN!Stadt AG [...] ist mit dem Bundesverdienstkreuz 1. Klasse ausgezeichnet worden.*

(aus dem Bonner Zeitungskorpus; zur besseren Lesbarkeit wurden Wortart-Tags und Lemmanamen für alle Tokens außer den hier relevanten entfernt).

Die besten statistischen Wortarttagger für das Deutsche gehen bei Zeitungstexten von einer Korrektheit von zwischen 96 und 98 % aus (Carstensen et al. 2004). Allerdings sind die Fehler nicht gleichmäßig über alle Wortarten verteilt, sondern oft fast ‚strukturell‘. Lüdeling (2000) diskutiert zum Beispiel die Konsequenzen, die sich daraus ergeben, dass getrennt stehende Verbpunkteln von statistischen Taggern wegen ihrer Homographie zu anderen Kategorien unzuverlässig getaggt werden. Bosch, Rozario & Zhao (2003) zeigen, dass das Demonstrativpronomen *der* bei einem statistischen Tagger fast durchgängig falsch getaggt wird, weil es deutlich seltener vorkommt als das Relativpronomen *der* und daher in den Trainingsdaten kaum auftritt. Quantitative Analysen des Demonstrativpronomens sind also auf dieser Basis nicht möglich. Bei Ambiguitäten zwischen homographen Namen und Nomina kann es auch zu systematischen Effekten kommen, wie in (3) illustriert wird. Die Tagfolge CARD NN NN ist viel häufiger als die Tagfolge CARD NN NE, so dass ein Tagger bei dem mehrdeutigen *Kohl* immer auf diese Art entscheidet (man kann sich ja auch Situationen vorstellen, in denen man ziemlich lange das Gleiche essen muss).

(3) *Wir haben jetzt 16/CARD Jahre/NN Kohl!NN hinter uns!*

(aus dem Parlamentsredenkorpus)

Bei großen, weit verbreiteten Korpora kann man sich auf die jeweils vorliegende Vorverarbeitung stützen; wenn das Korpus nur über Internetabfragen zugänglich ist, hat man keine andere Wahl. Die Vorverarbeitungsprogramme sollten daher generell bekannt sein.

Hier unterscheiden sich kleine, spezialisierte Korpora, wie Dialektkorpora, historische Korpora, die meisten Korpora gesprochener Sprache und auch Lernerkorpora. Kann man mit spezialisierten Korpora genauso umgehen wie mit den ‚Standard‘korpora? Für solche Korpora gibt es bisher kaum Annotationsstandards. Je weniger ‚standardisiert‘ die Daten selbst sind, desto weniger kann man auf vorhandene Annotationsprogramme und -verfahren zurückgreifen, da diese ja meist statistische Komponenten enthalten und daher auf Regelmäßigkeiten angewiesen sind. Spezialisierte Korpora werden daher, wenn überhaupt, manuell annotiert. Da ergibt sich ein Problem: oft sind die Analysen selbst kontrovers. Daher stellen sich die folgenden Fragen: Wie geht man mit konfligierenden Analysen um? Wie kann die Korpusarchi-

tektur aussehen, so dass unterschiedliche Analysen repräsentiert werden können? Wie kann man bei konfligierenden Analysen zählen? Zählt jedes Vorkommen eines Phänomens gleich viel?

Ich möchte diese Fragen in Abschnitt 4 anhand eines Lernerkorpus diskutieren. Zunächst stelle ich das Korpus kurz vor.

3. Lernerkorpora

Lernerkorpora sind kontrollierte Sammlungen von Äußerungen von Lernern einer Fremdsprache.⁶ Lernerkorpora können – zusammen mit gezielten Erhebungen wie zum Beispiel Elizitationstests – zur Erforschung von Erwerbsprozessen verwendet werden. Die Erkenntnisse können dann in der Vermittlung eingesetzt werden. Für das Englische als Fremdsprache werden Lernerkorpora schon relativ häufig verwendet (Granger 2002, erscheint; Mukherjee 2002; Pravec 2002; Tono 2003; Nesselhauf 2005; Römer 2006 und viele andere), für das Deutsche als Fremdsprache haben sie sich dagegen noch nicht so stark durchsetzen können. Obwohl es an vielen Orten mehr oder weniger private Sammlungen von Lerneräußerungen gibt (wie zum Beispiel das nicht frei zugängliche Korpus, das in Belz 2004 beschrieben wird), stehen bisher noch kaum (fehlerannotierte) Lernerkorpora frei und allgemein zur Verfügung (Ausnahmen sind die ESF-Korpora⁷ mit gesprochenen Daten aus dem ungesteuerten Zweitspracherwerb und das tief prosodisch annotierte Leap-Korpus⁸).

Falko⁹ versucht, hier eine Lücke zu schließen. Falko enthält Texte von fortgeschrittenen Lernern des Deutschen und besteht aus unterschiedlichen Subkorpora. In diesem Artikel beziehe ich mich nur auf Daten aus dem so genannten Zusammenfassungskorpus, das Texte enthält, die nicht-deutschsprachige Germanistikstudierende an der Freien Universität Berlin im Rahmen einer obligatorischen Sprachstandsüberprüfung im Zusammenhang mit der Zwischenprüfung verfasst haben (zur Zeit ca. 37.000 Tokens; das Korpus wächst ständig). Dies sind Zusammenfassungen eines linguistischen oder eines literaturwissenschaftlichen Fachtextes; die Aufgabenstellung war „Fassen Sie den Text mit eigenen Worten zusammen“. Die Texte liegen zunächst

⁶ Lernerkorpora unterscheiden sich von reinen Fehlersammlungen (wie in Heringer 1995) dadurch, dass sie ganze Texte enthalten, so dass zum Beispiel auch untersucht werden kann, welche Konstruktionen, Wörter etc. ein Lerner schon beherrscht und quantitative Untersuchungen möglich sind.

⁷ http://www.mpi.nl/world/data/esf_archive/html/esf.html (Mai 2006).

⁸ <http://www.phonetik.uni-freiburg.de/leap/corpus.html> (Mai 2006), siehe auch Milde & Gut (2002).

⁹ Falko steht für fehlerannotiertes Lernerkorpus. Das Korpus wird in Kooperation der Humboldt-Universität und der Freien Universität in Berlin entwickelt und ist frei zugänglich. Informationen und Zugriff unter <http://www2.hu-berlin.de/korpling/projekte/falko/index.php>.

handschriftlich vor¹⁰ und werden dann digitalisiert.¹¹ Des Weiteren gibt es ein vergleichbares Kontrollkorpus mit Muttersprachlerdaten (zurzeit ca. 13.000 Tokens).

4. Lernerkorpora: Annotation und Auswertung

Ich möchte die in Abschnitt 2 diskutierten Fragen nun exemplarisch anhand von Falko-Daten diskutieren. Nach jedem Unterabschnitt werde ich kurz diskutieren, wie sich die hier besprochenen Probleme in anderen Korpusarten auswirken.

Lernerkorpora werden auf zwei Arten ausgewertet, durch Fehleranalyse (Error Analysis, EA) und durch kontrastive Analyse (Contrastive Interlanguage Analysis, CIA). In der Fehleranalyse werden Fehler klassifiziert und annotiert. In der kontrastiven Analyse werden die Lernertexte statistisch mit anderen Texten (Muttersprachlertexte oder Texte anderer Lernergruppen) verglichen. So kann man zum Beispiel feststellen, ob Lerner ein bestimmtes Wort, eine Klasse von Wörtern oder eine Konstruktion im Vergleich zu häufig (Übergebrauch, overuse) oder zu selten (Mindergebrauch, underuse) anwenden, ob sich die durchschnittliche Satz- oder Wortlänge unterscheidet und wie sich die Fehleranzahl der Lerner im Vergleich zur Fehleranzahl der Muttersprachler verhält.

Sowohl für die Fehleranalyse als auch für die kontrastive Analyse müssen die Lernerdaten kategorisiert werden. Während sich für die Lemmatisierung und die Wortartzuweisung noch relativ leicht gemeinsame Kategorien finden lassen, muss man in der Arbeit mit vom Standard abweichenden Daten noch deutlich stärker interpretieren.

4.1 Fehleranalyse

Zunächst muss beantwortet werden, was ein *Fehler* ist. Dies ist in der Literatur häufig und kontrovers diskutiert worden – ich kann hier nur kurz einige Grundzüge der Diskussion zusammenfassen (zur Diskussion siehe zum Beispiel Cherubim 1980; Corder 1981 oder Lennon 1991). Ein Fehler kann zum

¹⁰ Dadurch wird verhindert, dass die Studierenden automatische Rechtschreibkorrekturen etc. nutzen können.

¹¹ Durch die Klausursituation ist die Erhebung stark kontrolliert: die Zeitdauer ist vorgegeben, die Studierenden haben keinen Zugriff auf Hilfsmittel. Die Studierenden werden als ‚fortgeschritten‘ eingestuft, da sie die DSH-Prüfung bestanden haben und ihr Grundstudium an der FU absolviert haben. Dies ist sicher zu diskutieren – man braucht jedoch ein externes Kriterium, um dann die sprachlichen Eigenschaften der jeweiligen Varietät zu beschreiben (siehe die Diskussion in Walter & Grommes 2006). Zusätzlich zu dem hier beschriebenen Zusammenfassungskorpus gibt es noch ein Korpus mit Longitudinaldaten, das an der Georgetown University in Washington erhoben wurde und einige weitere kleinere Korpora. Zur Zeit wird ein kontrolliertes Korpus mit Essays erhoben.

einen als ein Bruch einer Regel aufgefasst werden und zum anderen als eine Abweichung von einer Norm. Die Begriffe ‚Regel‘ und ‚Norm‘ werden unterschiedlich verwendet – man könnte sie zum Beispiel folgendermaßen sehen: Eine Regel ist von der Sprache vorgegeben (‚im Deutschen sind Subjekt und Verb numeruskongruent‘), der Bruch einer Regel führt zu ungrammatischen Äußerungen. Eine Norm hingegen ist von außen gesetzt (Orthographie oder Ausdruckskonventionen), die Abweichung einer Norm führt zu unakzeptablen Äußerungen. Corder (1973) spricht hier von ‚breaches of code‘.

Wie auch immer eine theoretische Einteilung von Fehlern oder Abweichungen aussieht, in konkreten Äußerungen ist es nicht immer möglich, sauber zu kategorisieren. Wenn man zum Beispiel Satz (4) betrachtet, kann man kaum sagen, wo hier die Fehler sind. *Zweispältig* ist sicher ein Lexikonfehler, allerdings ist das Wort nach den Regeln der Grammatik an sich korrekt gebildet. *Verhältnis in seinem Wort* ist sicher nicht korrekt, aber was genau ist hier der Fehler? Sollte man es ganz weglassen? Wie umformulieren? Gegen welche Regel oder Norm wird hier verstoßen?

- (4) *Auf der anderen Seite hat Kunstmärchen ein zweispältiges Verhältnis in seinem Wort.*

(Falko L2, Text 46)

Dann muss man auch unterscheiden zwischen solchen Fehlern, die ein Lerner macht, weil er eine bestimmte Regeln/Norm noch nicht beherrscht (Kompetenzfehler, errors) und solchen, die er macht, weil er müde oder unkonzentriert ist, auch wenn er die relevanten Regeln schon kennt (Performanzfehler, mistakes, lapses). An der Oberfläche sind errors und mistakes allerdings nicht zu unterscheiden.

Man kann auch Übergebrauch und Mindergebrauch als ‚Fehler‘ auffassen, wenn man Lennons Fehlerdefinition zugrunde legt. Hier ist ein Fehler „a linguistic form, ... which, in the same context would in all likelihood not be produced by the learner’s native speaker counterparts.“ (Lennon 1991, S. 182). Um solche Fehler zu finden braucht man ein vergleichbares Referenzkorpus mit Muttersprachleräußerungen und muss operationalisieren, was ‚in the same context‘ und ‚in all likelihood‘ bedeuten würde. Selbst wenn man das könnte, ist es schwierig zu sagen, wo genau der Fehler liegt: bei Übergebrauch (zum Beispiel einer bestimmten Konjunktion) ist ja jede einzelne Verwendung der Konjunktion korrekt. Wie soll man entscheiden, welche nicht hätte gebraucht werden dürfen? Mindergebrauch ist noch schwieriger: An welcher Stelle in einem Text müssten Lerner denn ein (im Vergleich zu Muttersprachlern) zu selten gebrauchtes Wort einsetzen? Wo genau ist der Fehler?

Man kann also nicht systematisch Regel- von Normverstößen oder errors von mistakes unterscheiden. Außerdem ist eine Operationalisierung von Über- und Mindergebrauch kaum möglich.

Analoge Probleme ergeben sich in vielen anderen Korpusstypen. So ist es zum Beispiel nicht klar, welche Parameter getaggt werden sollen, wenn man Variation und kurzfristigen Wandel untersuchen möchte (Baumgarten et al. erscheint). Oder wie zum Beispiel Code-Switching zu behandeln wäre. Welche zugrundeliegenden Kategorien sollen annotiert werden, wenn man sich mit Informationsstruktur (Hinterhölzl, Petrova & Solf 2005) oder mit Diskursaufbau befasst?

4.2 Konfligierende Analysen: Zielhypothesen

Ich möchte noch auf ein weiteres Problem eingehen: um einen Fehler zu annotieren, muss man eine Hypothese darüber haben, wie die Äußerung hätte korrekt lauten müssen. In den meisten Lernerkorpora sind diese so genannten Zielhypothesen oder „reconstruction of those utterances in the target language“ (Ellis 1994, S. 54) implizit. Die folgenden Beispiele aus Weinberger (2002) sollen das verdeutlichen, wobei Weinbergers Arbeit stellvertretend für viele steht. In (5a) gibt es eine Kongruenzverletzung bei *diese Phänomen*. Weinberger nimmt für *diese* einen Genusfehler an. Ihre implizite Zielhypothese ist also *dieses Phänomen*. Wahrscheinlich ist diese Hypothese durch weiteren Kontext gestützt (der aber in der Arbeit nicht gegeben wird). Ohne Kontextinformation hätte man hier auch annehmen können, dass *diese Phänomene* gemeint wäre und ein Numerusfehler bei *Phänomene* auftritt. In (5b) ist das Problem schwieriger: Weinberger geht von einem phrasalen Ausdrucksfehler aus. In ihrer Arbeit gibt sie an, dass der Lerner hätte schreiben sollen *die Gesellschaft hat sich verändert*. Dies ist aber im Korpus nicht zu sehen – jemand, der nicht gleichzeitig die Arbeit vor sich hat und mit dem Korpus arbeitet, kann das nicht erschließen.¹²

(5a) *die Erklärung für <MoArInGn> diese Phänomen ist*

(Fehlertags: Mo – morphology, Ar – article, In – inflection, Gn – gender)

(5b) *<LxPhCh> Es gibt eine veränderte Gesellschaft und*

(Fehlertags: Lx – lexical, Ph – phrase, Ch – incorrect choice)

Hier sieht man, dass jede Fehlerannotation notwendigerweise eine Zielhypothese voraussetzt, egal ob diese implizit oder explizit (wie zum Beispiel bei Izumi, Uchimoto & Isahara 2005) angegeben ist. Zielhypothesen sind aber fast nie unumstritten, meistens sind mehrere Zielhypothesen für eine Lerneräußerung möglich. Je komplexer die betrachteten Sätze werden, desto unterschiedlicher sind die möglichen Zielhypothesen.

Unterschiedliche Zielhypothesen führen zu verschiedenen Fehlerzählungen. Betrachten wir Beispiel (6a) aus dem Falko-Korpus und zwei der möglichen

¹² Außerdem gibt es hier Probleme durch das Tabellenformat. Das Ende des phrasalen Fehlers ist nicht explizit kodiert (die Hervorhebung findet sich nicht im Korpus). Siehe dazu Lüdelling, Poschenrieder & Faulstich (2005).

Zielhypothesen (Fehler, die in beiden Zielhypothesen korrigiert werden wie *Jahrhunder* werden hier nicht betrachtet).

Für Zielhypothese 1 in (6b) wird das Komma ernst genommen und ein Relativsatz konstruiert. Das führt zu deutlichen Wortstellungsveränderungen und in einer Fehlerzählung zu mehreren Wortstellungsfehlern. Außerdem sind zwei Auslassungsfehler (*die* und *gegründet wurde*) festzustellen. Zielhypothese 2 in (6c) greift weit weniger in die Struktur ein und fügt ein adjektivisches Partizip ein. Hier sehen wir also einen Auslassungsfehler und keine Wortstellungsfehler, aber einen zusätzlichen Orthographiefehler (ein überflüssiges Komma). Allein an diesem Beispiel sieht man, wie stark sich Zählungen je nach Zielhypothese unterscheiden können (eine ausführlichere Auswertung des so genannten Inter-Annotator Agreement von Zielhypothesen und den daraus folgenden Fehlerzählungen findet sich in Lüdeling (2007)).

- (6a) Lerneräußerung: *Der Realismus ist eine im 19. Jahrhundert, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte.*
(Falko L2, Text 25)
- (6b) Zielhypothese 1: *Der Realismus ist eine literarische Richtung, die im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründet wurde und sich ...*
- (6c) Zielhypothese 2: *Der Realismus ist eine im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründete literarische Richtung, die sich ...*

Wenn die Zielhypothese so entscheidend für die Fehlerzählung und alle sich daraus ergebenden Konsequenzen für die Beschreibung von Lernabläufen ist, sollte sie explizit angegeben und zugänglich sein. Idealerweise sollte es möglich sein, konkurrierende Zielhypothesen für denselben Text aufzustellen. Darauf komme ich in Abschnitt 5 zurück.

Auch dies ist nicht nur für Lernerkorpora problematisch, sondern auf andere Korpusarten übertragbar. Es gibt in jedem Korpus klar ungrammatische Äußerungen, die nicht einmal unbedingt selten sein müssen, wie man am Beispiel der berühmten Rede von Trappatoni sieht, die oft zitiert wird, oder an der Tatsache, dass bestimmte Tippfehler, wie *immer* mit drei *m* häufig vorkommen. Wie mit solchen ‚Fehlern‘ umgegangen wird, hängt sicher mit der Forschungsfrage zusammen, aber es gibt sicher viele Linguisten, die keine Grammatik des Deutschen schreiben wollen, die *ich habe fertig* lizenziert. Das bedeutet, dass das Korpus nach externen Maßstäben gefiltert wird – implizit werden hier Hypothesen über korrekte Strukturen angenommen.

Auch historische Korpora oder Korpora gesprochener Sprache werden interpretiert. Abkürzungen werden expandiert, Hypothesen über kaum lesbare Buchstaben oder unverständliche Äußerungen gebildet. In solchen Fällen wird im Korpus nur eine ‚Ziel‘hypothese angegeben, oft ohne zu markieren, dass die Hypothese unsicher ist.

4.3 Tagsets und Evaluation

Ein weiterer Bereich, der für die Zählung wichtig ist, ist die Kategorisierung der Fehler selbst. Dafür müssen Tagsets entwickelt werden, analog zur Entwicklung von Tagsets für Wortarten oder andere Annotationsebenen. Dabei müssen mehrere Entscheidungen getroffen werden: (a) Fehlerexponent (das Wort oder die Phrase, an der der Fehlertag hängt), (b) Granularität und (c) Beschreibung des Fehlers. Alle drei Entscheidungen haben Auswirkungen auf die Zählung.

Zu (a): Ein Fehlertag kann (abhängig von der Korpusarchitektur, siehe auch Abschnitt 5) an ein Wort oder eine Phrase ‚andocken‘. Bei Wortstellungsfehlern kann es schwierig sein zu entscheiden, wo genau der Fehler ist. Wo sollen fehlende Elemente kodiert werden? In Beispiel (7) fehlt ein *es* nach dem *und*. In den meisten Architekturen müssen Tags an vorhandene Tokens gebunden sein – soll man das *es* am *und* oder am *ist* markieren? Oder vielleicht am *ist klar*, dem ja das Subjekt fehlt? Eine mögliche Zielhypothese wäre *Wenn ... und es klar ist,* Hier wäre also auch eine Umstellung von *ist* und *klar* zu markieren. Wieder stellt sich die Frage, wo genau die Umstellung zu markieren wäre. Je nachdem, wie man entscheidet, hat man einen Verbstellungsfehler oder nicht.

- (7) *Wenn man die Valenz eines Verbs kennt und ist klar, was die einzelnen Kasus bedeuten, [...].* (Falko L2, Text 60)

Zu (b) Dieser Punkt muss für jedes Tagset entschieden werden. Will man ein möglichst allgemeines Tagset, das zwar relativ grob, aber dann auch auf viele andere Korpora übertragbar ist, wie Degneaux et al. (1996) oder Izumi, Uchimoto & Isahara (2005)? Oder will man sehr spezifische, für ein kleines Problem oder auf eine spezielle theoretische Annahme aufbauende Tagsets, wie Lippert (2005)? Dass die Zählungen von der Granularität des Tagsets beeinflusst werden, liegt auf der Hand. Ein weiteres Problem ist die Tagging-zuverlässigkeit. Bei automatischer Annotation wird üblicherweise gegen einen so genannten Goldstandard evaluiert, Dazu wird ein Teil des Korpus manuell getaggt. Der Tagger soll diesen Teil dann möglichst gut reproduzieren. Die Grundannahme ist, dass der Goldstandard korrekt ist. Dies wäre nur dann der Fall, wenn (a) das Tagset so eindeutig ist, dass es für jedes Token nur genau eine Möglichkeit gibt und (b) die Annotatoren keine Fehler machen. Beide Grundannahmen sind problematisch.

Für manuelle Annotation gibt es ein anderes Verfahren: hier wird derselbe Text mit denselben Tags nach denselben Richtlinien von mehreren Annotatoren getaggt. Dann werden die Annotationen verglichen und ein so genannter Inter-Annotator-Agreement-Wert (manchmal auch Inter Rater Reliability oder Urteilerübereinstimmungswert genannt, siehe zum Beispiel Carletta 1996, Bortz, Lienert & Boehnke 2000) ermittelt. Mir sind keine Untersuchungen zum Urteilerübereinstimmungswert bei Fehlerannotationen bekannt. Es gibt je-

doch Papiere, die zeigen, dass der Wert bei Aufgaben wie der Lesartenzuweisung schlechter ist als Zufall (Veronis 2001). Ohne eine Angabe des Urteilerübereinstimmungswerts ist also – zumindest bei Kategorisierungen, die nicht völlig trivial sind – eine quantitative Analyse kaum aussagekräftig.

Zu (c) Die Erstellung des Tagsets ist wesentlich für die späteren Abfragemöglichkeiten: nur was kodiert ist, kann automatisch gesucht werden. Fehler können dabei unterschiedlich kategorisiert werden: man kann zum Beispiel nach formaler Art eines Fehlers (Einfügung, Auslassung, Vertauschung) kategorisieren oder die Hypothese über die Fehlerentstehung (zum Beispiel Interferenz der Muttersprache des Lernalers oder die mangelnde Beherrschung einer Regel) in das Fehlertag mit einfließen lassen. Man kann die Wortart des Fehlerexponenten mit kodieren oder in bestimmten linguistischen Bereichen arbeiten. Ein Problem ist hier, dass in den vorliegenden Tagsets oft mehrere Informationen in einem Tag zusammengefasst werden. Ein Beispiel hierfür sind die Tags in (8). Problematisch ist dies, wenn die unterschiedlichen Teile des Tags unterschiedlich sicher sind. Nehmen wir an, wir wissen, dass Satz (8) von einem Lerner mit der Muttersprache Polnisch geschrieben wurde. Dann könnte man ein Fehlertag setzen, das einerseits den fehlenden Artikel enthält und andererseits eine Annahme, dass dieser Fehler durch eine Interferenz mit der Muttersprache hervorgerufen wurde. Für die Suchmöglichkeiten wäre dies sicher vorteilhaft. Während sich allerdings wahrscheinlich die meisten Annotatoren sicher sind, dass hier ein Artikel fehlt, ist die Annahme über die Ursache des Fehlers nur eine Hypothese. Unterschiedlich sichere Informationen sollten nicht in einem Tag zusammengefasst werden.

(8) *In Grammatik wird es aber undenkbar dass [...].* (Falko L2, Text 63)

4.4 Korpusstruktur: Unterschiedliche Gewichtung von Vorkommen

Ich möchte an dieser Stelle noch ein weiteres Problem für die Zählung von Fehlern ansprechen, das zwar sehr speziell für das Lernerkorpus Falko erscheint, aber auch eine grundsätzliche Komponente hat. Das Falko-Zusammenfassungskorpus enthält, wie oben beschrieben, Zusammenfassungen von linguistischen und literaturwissenschaftlichen Texten. Eine mögliche Strategie bei Zusammenfassungen ist es, bestimmte als wichtig erkannte Wörter und Phrasen direkt zu übernehmen. Diese Strategie wird von Lernern und Muttersprachlern angewendet, von Lernern allerdings deutlich stärker. Beispiel (9) zeigt alle Übernahmen von zwei oder mehr Wörtern eines Ausschnitts des Lernertextes 33, daneben der relevante Ausschnitt des Vorlagentexts.¹³ Die Frage, die sich hier stellt, ist, wie die übernommenen Sequenzen gezählt werden können. Einige Übernahmen sind so, dass sie in einer Zusammenfassung des Ursprungstextes auftreten können, ohne dass man sie als Textübernahme

¹³ Die Übernahmen wurden automatisch von dem Plagiatserkennungsprogramm WCopyfind 2.6 <http://plagiarism.phys.virginia.edu/Wsoftware.html> markiert.

werten muss (*des Realismus, tritt er, ...*) – der Lerner hätte sie wahrscheinlich auch selbständig so produziert. Bei anderen ist es unklar, ob der Lerner sie selbst hätte produzieren können (*sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts*). Das Problem ist nicht einfach lösbar. Mir ist hier nur wichtig, dass man die übernommen Textpassagen kenntlich machen muss, so dass sich die Auswertungen darauf beziehen können.

(9)

Falko L2, 33, Ausschnitt, Übernahmen kursiv	Vorlage, Ausschnitt (aus: Sprengel, Peter (1998): III. Stile und Richtungen. 1. Realismus In: ders.: Geschichte der deutschsprachigen Literatur 1870–1900. Von der Reichsgründung bis zur Jahrhundertwende. München: Verlag C.H. Beck, S. 99–101)
<p><i>Der Realismus ist eine überragende geistige und künstlerische Tendenz des 19. Jahrhunderts. Er erstreckt sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts. In der Literatur tritt er unter verschiedenen Namen auf. Die Literatur- und Kunstgeschichte einigte sich erst im nachhinein über die Grenzen des Realismus. Anfangs haben sich die Naturalisten hauptsächlich als Realisten verstanden. Hieraus erkennt man, dass die genauere Bezeichnung des Realismus nach dem damaligen Sprachgebrauch noch nicht vorhanden war.</i></p>	<p>Der Realismus ist die überragende geistige und künstlerische Tendenz des 19. Jahrhunderts. Noch in der ersten Hälfte des Jahrhunderts einsetzend, und dort vor allem als Gegenbewegung zu klassisch-romantischen Kunstauffassungen begründet, erstreckt er sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts. Freilich tritt er dabei unter einer Reihe verschiedener Namen auf, die vielfach in Konkurrenz zu einem eigentlichen Realismus (im engeren Sinne) stehen, über dessen Grenzen sich Literatur- und Kunstgeschichte erst im nachhinein verständigt haben und z. T. auch heute noch uneins sind. So haben sich z. B. die Naturalisten zunächst und hauptsächlich als Realisten verstanden, was nicht nur einiges über das Selbstverständnis dieser Bewegung verrät, sondern auch beweist, daß der Begriff des Realismus nach dem damaligen Sprachgebrauch noch nicht besetzt, jedenfalls nicht mit der gleichen Verbindlichkeit wie heute als Bezeichnung für eine – im wesentlichen in Distanz zum Naturalismus verharrende – literatur- und kunstgeschichtliche Richtung (inner- oder unterhalb jenes oben umrissenen umfassenden Realismus) etabliert war.</p>

Dieses sicherlich sehr spezielle Problem deutet auf ein allgemeines hin: Kann man jedes Vorkommen eines Wortes/einer Kategorie etc. in einem Korpus zählen wie jedes andere auch? Diese Frage ist bisher in der Literatur eher stiefmütterlich behandelt worden, kann aber starke Auswirkungen auf die Analyse haben. Es ist bekannt, dass Wörter, die einmal in einen Diskurs ein-

geführt werden, oft aufgegriffen werden (der so genannte Clumpiness-Effekt). Quantitative Analysen der Lexik oder der morphologischen Produktivität (Baayen 2001) müssen darauf Bezug nehmen. Szmrecsanyi (2006) zeigt, dass Persistenz – also das Wiederholen von bereits im Diskurs vorgekommenen Strukturen – auch für morphosyntaktische Eigenschaften eine Rolle spielt.

In großen Korpora muss man damit rechnen, dass Texte oder Textteile mehrfach vorkommen, zum Beispiel Agenturmeldungen, die von verschiedenen Zeitungen wörtlich übernommen werden. Wenn in einer solchen Agenturmeldung ein Vorkommen der zu untersuchenden Kategorie (zum Beispiel eine seltene Wortbildung) auftaucht, das dann aber von fünf Zeitungen übernommen wird, zählt man es dann einmal oder fünfmal (siehe Clough & Gaizauskas, erscheint, für eine ausführliche Diskussion von Textwiederverwendung).

Mehrfachvorkommen des gleichen Textes ist für Webkorpora ein besonderes Problem – Verfahren zur duplicate detection spielen hier eine wichtige Rolle (Lüdeling, Evert & Baroni 2007; Baroni & Kilgarriff 2006).

Man kann das Problem sicher nicht generell lösen, da für unterschiedliche Fragestellungen und Phänomene unterschiedliche Verfahren eine Rolle spielen. Ein Desiderat wäre allerdings, dass Textübernahmen und Mehrfachvorkommen von Texten im Korpus gekennzeichnet sind, so dass für jede Analyse deutlich wird, was gezählt wird.

5. Architektur

In Abschnitt 3 sollte deutlich werden, dass alle Korpusauswertungen auf Interpretation der Daten beruhen. In 4 habe ich exemplarisch anhand des Lernerkorpus Falko dargestellt, dass die Faktoren (a) Zielhypothese/Interpretation, (b) Tagset/Vergaberichtlinien und (c) Korpusstruktur die Kategorisierung und daher auch alle darauf aufbauenden quantitativen Analysen beeinflussen.

In diesem Abschnitt möchte ich nun dafür plädieren, die Grundlagen für jede quantitative Analyse so transparent wie möglich darzustellen. Dazu gehört dass Tagsets, Richtlinien und Zuverlässigkeit veröffentlicht werden und, wenn möglich, das Korpus frei zur Verfügung steht. Dazu gehört aber auch, dass alternative und konfligierende Hypothesen und Kategorisierungen in einem Korpus dargestellt werden können. Dies ist in den ‚traditionellen‘ flachen Korpusarchitekturen nicht möglich, wohl aber in Mehrebenenarchitekturen.

In ‚flachen‘ Architekturen werden Kategorisierungen in einem ‚Tabellenformat‘ an Tokens gehängt (daher stammt auch der Begriff ‚positionelle Annotation‘; zu jeder Korpusposition wird eine festgelegte Anzahl von Annotationsebenen – oder Spalten zur Verfügung gestellt). Diese Architektur ist gut geeignet für sehr große, automatisch annotierte Korpora, da sie gut zu indizieren und schnell zu durchsuchen ist. In einer solchen Architektur können konfligierende Annotationen allerdings nicht dargestellt werden.

Außerdem ist es nicht einfach möglich, neue Ebenen einzufügen. Dies führt dann oft zu der oben erwähnten Zusammenfassung verschiedener Informationen in ein Tag.

In einer Tabellenarchitektur ist es auch nicht möglich, Informationen an eine Sequenz von Tags zu hängen. Wie oben diskutiert, beziehen sich viele Fehler (und andere Annotationen) aber auf Sequenzen und nicht auf einzelne Tokens. Daher werden heute oft Baummodelle (meist XML) verwendet, die die zu annotierenden Token durch Anfangs- und Endtags einschließen. Ein Beispiel aus einem englischen Lernerkorpus ist (10) (aus Izumi, Uchimoto & Isahara 2005).

(10) *I belong to two baseball* <n_num crr=„teams“>*team*</n_num>

In XML-Dateien können im Prinzip mehrere Tokens durch ein Tag eingeschlossen werden. Allerdings ist es auch in XML-Dateien schwer möglich, konkurrierende Analysen zu kodieren. Dies liegt zum einen an der Baumstruktur von XML, die konfligierende Hierarchien nicht zulässt (das Problem ist technisch lösbar) und zum anderen daran, dass Daten und Annotationen in derselben Datei gespeichert werden.

In den letzten Jahren wurden für multimodale Korpora Mehrebenenmodelle mit standoff Architektur entwickelt (Bird & Liberman 2001, Carletta et al. 2003, Dipper et al. 2004, Wörner et al. 2006, Wittenburg, erscheint), in denen alle Korpusebenen (Sprachsignal, Transkription, Gestik) und Annotationsebenen in getrennten Dateien gespeichert werden, die alle auf eine gemeinsame Referenzebene (timeline) verweisen.

Die Mehrebenenarchitektur ist ursprünglich nicht für die Modellierung von konfligierenden Analysen entwickelt worden, kann aber dafür gebraucht werden. Statt eines Zeitstrahls ist der Originaltext die Referenzebene. Alle anderen Ebenen sind dann unabhängig davon. Konfligierende Hypothesen und Tagsets können dargestellt werden. Ich möchte dies wieder am Beispiel von Falko erläutern (für eine technischere Darstellung siehe Lüdeling et al. 2005).

Falko wird zurzeit im Partitur-Editor EXMARaLDA kodiert.¹⁴ Jedes Token eines Lernertextes bekommt eine Position in der Partitur – dadurch wird

¹⁴ EXMARaLDA ist an der Universität Hamburg zur Darstellung von gesprochenen Korpora entwickelt worden und frei verfügbar <http://www1.uni-hamburg.de/exmaralda/>. EXMARaLDA hat zwar mehrere Ebenen, ist aber bisher nicht standoff. Bisher gibt es noch wenige frei verfügbare und gut zu bedienende Tools, mit denen echte standoff-Architekturen kodiert werden können. Solche befinden sich aber an vielen Orten in der Entwicklung (siehe z. B. Wörner et al. 2006 und die darin beschriebenen Tools für die linguistischen Sonderforschungsbereiche in Tübingen, Berlin/Potsdam und Hamburg und die Vorarbeiten für das DDD-Projekt in Dipper et al. 2004; Lüdeling, Poschenrieder & Faulstich 2005; Faulstich, Leser & Lüdeling 2005). Im Moment werden für solche Architekturen auch mächtige Such- und Darstellungstools entwickelt (Schmidt & Wörner 2005, Vitt 2005, Siemen, Lüdeling & Müller 2006).

der Zeitstrahl simuliert. Alle Annotationsebenen sind in getrennten Spuren gespeichert. Man kann so viele Annotationsebenen hinzufügen, wie man möchte. Das hat die Vorteile, dass man sich nicht von vorneherein auf die Menge und Granularität der Tagsets festlegen muss und dass man konfligierende Zielhypothesen darstellen kann. Alle Annotationsentscheidungen und Tags stehen im Netz zur Verfügung, jede quantitative Aussage kann daher überprüft werden.

Im Moment wird jeder Falko-Text zunächst automatisch mit dem Tree-Tagger mit Wortart und Lemma getaggt (dabei ist zu bedenken, dass die Qualität des Taggings gegenüber Zeitungstexten sinkt, weil durch die hohe Anzahl an Rechtschreibfehlern der Lexikonzugriff oft fehlschlägt und durch Wortstellungsfehler die statistische Komponente des Taggers nicht greifen kann, siehe dazu auch van Rooy & Schäfer 2003). Dann wird eine explizite Zielhypothese angegeben. Alle Fehler werden im Bezug auf diese Zielhypothese hin getaggt. Wir haben dann verschiedene Schichten für jeden Annotationsbereich (zurzeit Definitheit, Wortstellung, Kongruenz & Rektion; an weiteren Bereichen wird gearbeitet).

Beispiel (11) zeigt einen Lersatz in der Tabellendarstellung (die Tagsets kann ich aus Platzgründen nicht im Einzelnen erläutern, eine Dokumentation findet sich unter <http://www2.hu-berlin.de/korpling/projekte/falko/index.php>). Die erste Zeile ([word]) ist die Referenzzeile. Jedes Token bekommt eine Korpusposition. Unter [target_hypothesis] ist die Zielhypothese gespeichert. Dabei sind nur solche Tokensequenzen aufgeführt, die sich vom Originaltext unterscheiden.

Abweichende Zielhypothesen können jederzeit hinzugefügt werden, so dass unterschiedliche Analysen der Daten möglich sind. Im Unterschied zu vielen bisher vorliegenden Korpora sind alle Interpretationen sichtbar, die Ergebnisse werden so transparent und reproduzierbar. Auf eine ähnliche Art können andere spezialisierte Korpora kodiert werden.

Ein weiterer Vorteil einer Mehrebenenarchitektur besteht darin, dass dadurch das verteilte Arbeiten an denselben Daten möglich ist.

[illegible]

6. Zusammenfassung

Jede quantitative Analyse von Korpusdaten stützt sich auf eine vorhergehende Kategorisierung. In diesem Artikel habe ich gezeigt, wie stark die Kategorisierung die Ergebnisse beeinflusst. Dabei spielen die Tagsets und die Richtlinien für deren Vergabe genauso eine Rolle wie die Zielhypothese und die Zusammensetzung des Korpus. Ich habe dies exemplarisch anhand des Lernerkorpus Falko und der Fehlerannotation diskutiert.

Da es nicht möglich ist, eine Interpretation der Daten zu vermeiden, habe ich dafür plädiert, die getroffenen Entscheidungen, wo immer möglich, explizit zu machen, so dass sie für die weitere Arbeit nachvollziehbar sind. Für spezialisierte Korpora ist eine standoff Mehrebenenarchitektur eine Möglichkeit, das zu erreichen. Ich habe an Falko gezeigt, wie eine solche Mehrebenenarchitektur, die eigentlich für multimodale Korpora entwickelt wurde, aussehen kann.

7. Danksagung

Falko würde es ohne die Mitarbeit von vielen Kolleginnen, Kollegen und Studierenden nicht geben. Maik Walter und Karin Schmidt sind verantwortlich für die Akquisition der Zusammenfassungstexte und die Gesamtkonzeption des Korpus. Peter Siemen hat das Suchprogramm erstellt und alles kompetent technisch unterstützt. Hanna Acke, Eva Lippert und Seanna Doolittle haben kommentiert und annotiert. Allen meinen herzlichen Dank!

8. Literatur

- Baayen, R. Harald (2001): *Word Frequency Distributions*. Dordrecht: Kluwer.
- Baroni, Marco/Kilgarriff, Adam (2006): Large linguistically-processed web corpora for multiple languages. Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics), East Stroudsburg PA: ACL. S. 87–90.
- Baumgarten, Nicole/Herkenrath, Annette/Schmidt, Thomas/Wörner, Kai/Zeevaert, Ludger (erscheint 2006): Studying Connectivity with the Help of Computer-Readable Corpora: Some Exemplary Analyses from Modern and Historical, Written and Spoken Corpora. In: Hohenstein, Christiane/Pietsch, Lukas/Rehbein, Jochen (eds.): *Connectivity in Grammar and Discourse*. Amsterdam: Benjamins.
- Belz, Judy A. (2004): Learner Corpus Analysis and the Development of Foreign Language Proficiency. In: *System: An International Journal of Educational Technology and Applied Linguistics*. 32.4, S. 577–591.
- Biber, Douglas (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Bird, Steven/Liberman, Mark (2001): A formal framework for linguistic annotation. In: *Speech Communication* 33(1,2), S. 23–60.
- Bloomfield, Leonard (1933): *Language*. London: Allen & Unwin.
- Bortz, Jürgen/Lienert, Gustav A./Boehnke, Klaus (2000): *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer-Verlag.

- Bosch, Peter/Rozario, Tom/Zhao, Yufan: Demonstrative Pronouns and Personal Pronouns. German der vs. er. Proceedings of the EACL 2003. Budapest. Workshop on The Computational Treatment of Anaphora.
- Carletta, Jean (1996): Assessing agreement on classification tasks: the kappa statistics. In: *Computational Linguistics*, 22(2), S. 249–254.
- Carletta, Jean/Evert, Stefan/Heid, Ulrich/Kilgour, Jonathan/Robertson, Judy/Voormann, Holger (2003): The NITE XML Toolkit: flexible annotation for multi-modal language data. In: *Behavior Research Methods, Instruments, and Computers* 35(3), S. 353–363.
- Carstensen, Kai-Uwe/Ebert, C./Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (eds.) (2004): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg/Berlin: Spektrum Akademischer Verlag.
- Cherubim, Dieter (Hg.) (1980): *Fehlerlinguistik. Beiträge zum Problem der sprachlichen Abweichung*. Tübingen: Niemeyer.
- Clough, Paul/Gaizauskas, Robert (erscheint 2007): Corpora and text re-use. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Corder, Stephen Pit (1973): *Introducing Applied Linguistics*. Harmondsworth: Penguin.
- Corder, Stephen Pit (1981): *Error Analysis and Interlanguage*. Oxford: Oxford University Press.
- Degneaux, Estelle/Denness, Sharon/Granger, Sylviane/Meunier, Fanny (1996): *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain.
- Dipper, Stefanie/Faulstich, Lukas/Leser, Ulf/Lüdeling, Anke (2004): Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In: *Workshop on XML-based richly annotated corpora*. Lisbon, Portugal.
- Ellis, Rod (1994): *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Evert, Stefan (2005): *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Online verfügbar unter <http://www.collocations.de/phd.html>.
- Evert, Stefan/Fitschen, Arne (2001): Textkorpora. In: Carstensen, Kai-Uwe/Ebert, C./Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (Hg.): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg/Berlin: Spektrum Akademischer Verlag. S. 369–376.
- Featherston, Sam (in diesem Band): Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie.
- Golcher, Felix (2006): Statistical text segmentation with partial structure analysis. In: *Proceedings of KONVENS 2006*. Konstanz.
- Granger, Sylviane (2002): A bird's-eye view of learner corpus research. In: Granger, Sylviane/Hung, Joseph/Petch-Tyson, Stephanie (Hg.): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. S. 3–33.
- Granger, Sylviane (erscheint): Learner corpora. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Gries, Stefan Th. (2003): *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London/New York: Continuum Press.
- Grommes, Patrick/Maik, Walter (2006): Fortgeschrittene Lernervarietäten. Vortrag auf der 28. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Bielefeld. Folien unter <http://userpage.fu-berlin.de/~maik/dgfs06.html>.

- Haß-Zumkehr, Ulrike (2002): Das Wort in der Korpuslinguistik. Chancen und Probleme empirischer Lexikologie. In: Ágel, Vilmos/Gardt, Andreas/Haß-Zumkehr, Ulrike/Roelcke, Thorsten (Hg.): Das Wort. Seine strukturelle und kulturelle Dimension. FS für Oskar Reichmann zum 65. Geburtstag. Tübingen: Max Niemeyer Verlag. S. 45–70.
- Heringer, Hans Jürgen (1995): Aus Fehlern lernen. CD-ROM für Win9x/NT. Augsburg.
- Himmelman, Nikolaus (1998): Documentary and descriptive linguistics. In: *Linguistics* 36, S. 161–195.
- Hinterhölzl, Roland/Petrova, Svetlana/Solf, Michael (2005): Diskurspragmatische Faktoren für Topikalität und Verbstellung in der althochdeutschen Tatianübersetzung (9. Jh.). In: Ishihara, S./Schmitz, M./Schwarz, A. (Hg.): Approaches and Findings in Oral, Written and Gestural Language, Interdisciplinary Studies on Information Structure (ISIS) 3. Potsdam: Universitätsverlag Potsdam. S. 143–182.
- Hirschmann, Hagen (2005): Platzhalterphrasen bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. Staatsexamensarbeit, Humboldt-Universität zu Berlin.
- Izumi, Emi/Uchimoto, Kiyotaka/Isahara, Hitoshi (2005): Error Annotation for a corpus of Japanese Learner English. In: *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*.
- Kepser, Stephan/Reis, Marga (2005) (Hg.): *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter.
- Lennon, Paul (1991): Error and the very advanced learner. In: *International Review of Applied Linguistics* 29, S. 31–44.
- Lippert, Eva (2005): Probleme von Nichtmuttersprachlern mit der Definitheit von Nominalphrasen. Magisterarbeit, Humboldt-Universität Berlin.
- Lüdeling, Anke (2000): Particle verbs in NLP lexicons In: Heid, Ulrich/Evert Stefan/Lehmann, Egbert/Rohrer, Christian (Hg.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. IMS, Stuttgart. S. 625–630.
- Lüdeling, Anke (2007): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes, Patrick/Walter, Maik (Hg.): *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung*. Tübingen: Niemeyer.
- Lüdeling, Anke/Evert, Stefan/Baroni, Marco (erscheint 2006): Using Web data for linguistic purposes. In: Hundt, Marianne/Biewer, Caroline/Nesselhauf, Nadja (Hg.): *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi.
- Lüdeling, Anke/Poschenrieder, Thorwald/Faulstich, Lukas C. (2005): DeutschDiachron-Digital – Ein diachrones Korpus des Deutschen. In: *Jahrbuch für Computerphilologie* 2004, S. 119–136. Online verfügbar unter <http://computerphilologie.tu-darmstadt.de/jahrbuch/jb6-content.html>.
- Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005): Multi-level error annotation in learner corpora. In: *Proceedings of corpus linguistics 2005*. Birmingham. Online verfügbar unter <http://www2.hu-berlin.de/korpling/projekte/falko/FALKO-CL2005.pdf>.
- Manning, Christopher/Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Milde, Jan-Torsten/Gut, Ulrike (2002): A prosodic corpus of non-native speech. In: Bel, B./Marlien, I. (Hg.): *Proceedings of the Speech Prosody 2002 conference*, 11–13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage. S. 503–506.
- Moisl, Hermann (erscheint 2007): Clustering techniques. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Müller, Stefan (in diesem Band): Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus.

- Mukherjee, Joybrato (2002): *Korpuslinguistik und Englischunterricht. Eine Einführung*. Frankfurt: Peter Lang.
- Nesselhauf, Nadja (2005): *Collocations in Learner Corpora*. Amsterdam: John Benjamins.
- Oakes, Michael (1998): *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Ostler, Nicholas (2007): Corpora of less studied languages. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Pravec, N. A. (2002): Survey of learner corpora. In: *ICAME Journal* 26, S. 81–114. Available at <http://nora.hd.uib.no/icame/ij26/>.
- Römer, Ute (2006): Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for the Future. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), S. 121–134.
- van Rooy, Bertus/Schäfer, Lande (2003): Automatic POS tagging of a learner corpus: the influence of learner error on tagger accuracy. In: Archer, D./Rayson, P./Wilson, A./McEnery T. (Hg.) *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster University: UCREL. S. 835–844.
- Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. September 1994. Online verfügbar unter <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>.
- Schmidt, Thomas/Wörner, Kai: Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung Online* 2005 Online verfügbar unter <http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf>.
- Siemen, Peter/Lüdeling, Anke/Müller, Frank Henrik (2006): FALKO – Fehler-Annotiertes LernerKOrpus des Deutschen. In: *Proceedings of Konvens 2006*. Konstanz
- Szmrecsanyi, Benedikt (2006): Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis. Berlin/New York: Mouton de Gruyter.
- Tono, Yukio (2003): Learner corpora: design, development, and applications. In: *Pre-conference workshop on learner corpora, Corpus Linguistics 2003*, Lancaster.
- Veronis, Jean (2001): Sense tagging: does it make sense? Paper presented at the *Corpus Linguistics 2001 Conference*, Lancaster, U.K. Online verfügbar unter <http://www.up.univ-mrs.fr/veronis/pdf/2001-lancaster-sense.pdf>.
- Weinberger, Ursula (2002): *Error Analysis with Computer Learner Corpora. A corpus-based study of errors in the written German of British University Students*. MA thesis, Lancaster University.
- Wittenburg, Peter (erscheint 2007): Preprocessing of multimodal corpora. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Wörner, Kai/Witt, Andreas/Rehm, Georg/Dipper, Stefanie (2006): Modelling Linguistic Data Structures. In: *Proceedings of 'Extreme Markup Languages' 2006*, Montreal.
- Xiao, Richard (2007): Theory-driven corpus research: using corpora to inform aspect theory. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

SAM FEATHERSTON

Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie*

Abstract

In diesem Beitrag versuchen wir darzulegen, unter welchen Umständen introspektive Urteile objektive, quantifizierbare, und empirisch adäquate linguistische Daten sein können. Dazu skizzieren wir, wie die Resultate unserer experimentell erhobenen, relativen Urteilsstudien aussehen, und argumentieren, dass sie eine unverzichtbare Evidenzquelle für die Syntax bilden, weil sie Einsichten in die Architektur der Grammatik erlauben, die mit anderen Mitteln nicht möglich sind.

Die Jahrestagung 2006 des Instituts für Deutsche Sprache fand zum Thema ‚Sprachkorpora – Datenmengen und Erkenntnisfortschritt‘ statt. Das Projekt A3 ‚Suboptimale syntaktische Strukturen‘ im SFB 441 ‚Linguistische Datenstrukturen‘ in Tübingen betreibt Forschung zu Datentypen und Syntaxtheorie, wobei der Hauptakzent auf introspektiven Urteilen liegt. Durch die Erhebung der Urteile von Muttersprachlern des Deutschen zu grammatischen, marginalen sowie ungrammatischen Sätzen gewinnt man wichtige Indizien zu zentralen linguistischen Fragestellungen, wie zum Beispiel was es psycholinguistisch bedeutet, wenn eine Struktur als ‚gut‘ oder ‚schlecht‘ bewertet wird, und welche linguistisch relevanten und irrelevanten Faktoren bei dieser Differenzierung eine Rolle spielen. Für die Auswertung dieser Studien werden auch Frequenzdaten aus Korpora herangezogen, und deren Ähnlichkeiten mit und Unterschiede zu Urteilsdaten erforscht und ergründet. Gerade unser Projektschwerpunkt im Bereich der Datentypen erlaubt es, den Stellenwert von Frequenzdaten mit einer gewissen ‚Außensicht‘ zu kommentieren. Die Aufgabe unseres Beitrags in dem vorwiegend korpusorientierten Kontext dieses Bandes ist es deshalb, den Evidenzwert, sowie die Vor- und Nachteile von Frequenzdaten relativ zu anderen Datentypen zu kommentieren, um erstens die zum Teil komplementären Stärken und Schwächen von Frequenzdaten und Urteilen zu fokussieren, und zweitens die Einsichten darzustellen, die ausschließlich mithilfe von Urteilsdaten zu gewinnen sind.

* Diese Arbeit entstand im Rahmen des Projekts A3 ‚Suboptimale syntaktische Strukturen‘ im SFB 441 ‚Linguistische Datenstrukturen‘ in Tübingen. Die Unterstützung der Deutschen Forschungsgemeinschaft wird dankend anerkannt. Vielen Dank an Wolfgang Sternefeld und Tanja Kiziak aus dem Projekt und den Entwicklern von *WebExp*.

Dieses Papier hat deshalb zum Ziel, zwei Thesen aufzustellen und deren Wert zu belegen. Die Thesen lauten:

1. Introspektive Urteile können objektive, quantifizierbare Daten sein.
2. Introspektive Urteile sind auch für den Syntaktiker notwendig, um ein volles Verständnis der Syntax zu erlangen.

Im Folgenden werden wir zuerst die Gründe für die Zweifel anführen, die Sprachwissenschaftler zunehmend dazu bewegen, Misstrauen gegenüber Urteilen als Datentyp zu hegen, und auf andere Datentypen, wie zum Beispiel Frequenzdaten, auszuweichen. Wir werden die Motiviertheit dieser Zweifel nicht bestreiten, sondern darlegen, wie diese Schwächen des Datentyps gemildert und vermieden werden können. Im Wesentlichen geht es darum, statt Urteile als Einzelperson selbst abzugeben, die Daten von Informantengruppen unter Einhaltung verschiedener Gebote der experimentellen Kontrolle zu benutzen. Wir werden argumentieren, dass methodologisch einwandfrei erhobene introspektive Daten durchaus als empirisch adäquat gelten können und darüber hinaus spezifische Vorteile in ihrem Evidenzwert haben.

In einem zweiten Schritt werden wir auf die Befunde von Studien mit dieser Methode eingehen. Tatsächlich liefern diese *relativen Urteile* ein anderes Bild von der Funktionsweise der Syntax als herkömmlich angenommen wird. Statt eines kategorischen Modells der Grammatikalität unterstützen diese Daten vielmehr ein Modell der gradierten Wohlgeformtheit. Insofern stimmen die Erkenntnisse aus diesen experimentell erhobenen Urteilen mit dem Muster überein, das man von Frequenzdaten kennt. Es gibt aber auch Parameter, bei denen sich die zwei Datentypen unterscheiden. Insbesondere liefern Häufigkeiten und Urteile unterschiedliche Verteilungsmuster. Häufigkeitsdaten weisen ein gradiertes Bild der Wohlgeformtheit auf, weil sie eine Vielzahl von binären Entscheidungen summieren; jedes einzelne relative, in numerischer Form abgegebene Urteil dagegen beinhaltet Gradiertheit bereits in sich selbst. Aus dieser Tatsache kann man Schlüsse über die Architektur der Teile des menschlichen syntaxverarbeitenden Prozesses ziehen, die auch für das Zusammenwirken der verschiedenen Funktionen in der Grammatik wichtig sind. Zur vollen Einsicht in die Architektur der Grammatik kommt man nur, indem man sowohl Frequenzdaten als auch Urteilsdaten berücksichtigt. Daher ist unsere Schlussfolgerung, dass sowohl Frequenzen wie auch introspektive Urteile für die Theoriebildung in der Syntax nötig sind, jedoch erlauben Urteile einen leichteren Zugang zu den einzelnen syntaktischen Beschränkungen als Frequenzdaten.

Introspektive Urteile als Datentyp

In diesem Teil werden wir argumentieren, dass introspektive Urteile unter bestimmten Umständen als ein verlässlicher, objektiver und quantifizierbarer Datentyp gelten können.

Introspektive Urteile: die Nachteile

Es ist immer wieder in der Literatur kritisiert worden, dass introspektive Urteile als Datentyp deutliche Nachteile haben; zum Beispiel von Labov (1975) und Sampson (2001), siehe Schütze (1996) für eine ausführliche Diskussion. Es wird Urteilen erstens vorgeworfen, dass sie ungenau sind. Das ist natürlich für das einzelne Urteil richtig: stellt man verschiedenen Informanten die gleiche Frage, dann geben sie öfter verschiedene Antworten. Man kann auch derselben Person die gleiche Frage zu verschiedenen Zeitpunkten stellen und dann nicht selten unterschiedliche Ergebnisse erhalten. Zum Teil als Reaktion auf dieses Phänomen haben Syntaktiker daher oft zum einen mit der Idee der individuellen Idiogrammatik gearbeitet und zum anderen dazu tendiert, Nicht-Linguisten die Fähigkeit abzuspüren, die Feinheiten der Urteilsabgabe ausreichend zu beherrschen.

Ein weiterer Vorwurf ist die mangelnde Quantifizierbarkeit von Urteilen. Dies hat zur Folge gehabt, dass manche Autoren nur binär zwischen guten und schlechten Beispielen unterscheiden, während andere ‚gut‘, ‚marginal‘ und ‚schlecht‘ differenzieren, in extremen Fällen gibt es bis zu sieben abgestufte Grade der Wohlgeformtheit (zum Beispiel hat Müller 1995 fünf: Ø [keine Angabe = ‚vollgrammatisch‘], ?, ??, *?, und *. Lakoff 1973 sechs: Ø, ?, ??, ?*, * und **. Wurmbrand 2001 sogar sieben: Ø, #, %, ?, ?, ??, *). Sprecher haben ein intuitives Gefühl dafür, was absolut grammatisch oder ungrammatisch ist, aber man kann bezweifeln, ob das ??-Urteil von Lakoff dem ??-Urteil von Wurmbrand entspricht, wenn man bedenkt, dass ?? von Lakoff die drittbeste von sechs Stufen ist und ?? von Wurmbrand die zweit-schlechteste von sieben Stufen.

Diese Überlegung führt uns zu einem weiteren Nachteil introspektiver Urteile, nämlich dass sie subjektiv sind. Laut dieser Kritik ist jede einzelne Urteilsbefragung keine Beobachtung eines externen Ereignisses sondern die Aussprache einer persönlichen Meinung. Solche Daten lassen sich aber nicht unabhängig messen, also können sie nicht angefochten werden. Damit ist das wissenschaftliche Prinzip der Überprüfbarkeit und Replizierbarkeit von Feststellungen nicht gewährleistet.

Das dritte große Problem introspektiver Daten können wir als die Unsicherheit des gemessenen Konstrukts zusammenfassen. Urteile abzugeben ist weder produktive noch rezeptive Verarbeitung, sondern es scheint sich um eine unmotivierter Metakompetenz zu handeln. Es ist also nicht sofort eindeutig, dass gerade dieses Messinstrument als definierendes Kriterium taugt, ob eine Struktur zur Sprache gehört oder nicht. Vorkommensdaten liefern viel eindeutiger Ergebnisse: wenn Sprecher eine Struktur benutzen, muss sie zur Sprache gehören.

Zudem ist es trotz eingehender Studien (siehe Schütze 1996) noch unklar, welche Faktoren in welchem Verhältnis und mit welcher Interaktionsfunktion bei der Bildung eines Gesamturteils beteiligt sind. Fest steht nur, dass u. a.

viele syntaktisch irrelevante Faktoren involviert sind, wie zum Beispiel die Plausibilität des Inhalts, der Bekanntheitsgrad des benutzten Wortschatzes und ähnliches. Damit lässt sich nicht auseinanderhalten, welche Teile eines Urteiles tatsächlich theoretisch relevant sind und welche nicht. Dies muss als zusätzliches Hindernis gelten, wenn man die empfundene Wohlgeformtheit und nicht das Vorkommen als Kriterium für Grammatikalität anwenden will.

Experimentell erhobene Urteile

Diese immer wieder gegen introspektive Urteile vorgebrachten Vorwürfe sind sachlich richtig, betreffen aber nur die Urteile, die mit der Standardmethode der Selbstbefragung erhoben werden. Dagegen ist es aber möglich und aus empirischer Sicht sogar notwendig, seine Daten unter strengerer Kontrolle zu erheben.

Das erste Gebot der empirischen Adäquatheit ist es, keine einzelnen Urteile zu sammeln. Dazu bieten sich mehrere Vorgehensweisen an, von denen wir hier nur drei erwähnen werden. Die herkömmlichste Art, eine solche Studie auszuführen, ist wohl die Benutzung einer Fünf- (oder Sieben)-Punkte-Skala (z. B. Crain/Fodor 1987). Der Versuchsleiter gibt dem Informanten eine graphische Skala mit einer normalerweise ungeraden Zahl von darauf gekennzeichneten Intervallen (Abbildung 1).

Abb. 1: Eine Sieben-Punkt-Skala zur Erhebung von relativen Urteilen

In letzter Zeit ist dagegen die Methode *Magnitude Estimation* (Bard et al. 1996) beliebter geworden. Drei Unterschiede zur normalen Urteilserhebung prägen diese Methode. Erstens werden Urteile numerisch abgegeben. Zweitens werden nur relative Urteile erhoben: relativ zu einem Referenzsatz und zu den eigenen vorherigen Urteilen. Drittens ist die Skala ohne Endpunkte und ohne Mindestabstand. Die Aufgabe lautet dementsprechend: ‚Wenn der Referenzsatz zehn wert ist, wieviel geben Sie diesem Beispiel?‘. Diese Methode erlaubt es dem Informanten, alle Wohlgeformtheitsunterschiede auszudrücken, die er oder sie empfindet, ohne von einer vorgeschriebenen Skala eingeschränkt zu sein.

Die von uns bevorzugte Methode nennt sich *Thermometer-Urteile* (Featherston 2004). Dieser Ansatz unterscheidet sich von *Magnitude Estimation* darin, dass statt eines einzigen Referenzsatzes immer zwei angegeben werden, deren Referenzwerte vom Versuchsleiter festgesetzt werden. Der Grund für diese Neuerung ist, dass sich herausgestellt hat, dass Informanten gar keine proportionalen Urteile abgeben können. Sprecher haben gar keine Intuition, ob eine Struktur ‚doppelt‘ oder ‚halb‘ so gut ist wie eine andere, wie das *Magnitude Estimation* verlangt. Tatsächlich haben wir nur ein Gespür für die

Distanz zwischen den Wohlgeformtheitsgraden von einzelnen Strukturen, und Informanten benutzen in der Praxis eine lineare Skala, auch wenn sie dazu aufgefordert sind, eine proportionale Skala zu benutzen (Featherston 2004, Poulton 1989).

Für die Zwecke unseres Papiers müssen wir jedoch nicht scharf zwischen den Ergebnissen dieser drei Methoden unterscheiden. Das Wichtigste ist, *dass* ein experimenteller Ansatz verwendet wird, und nicht so sehr *welcher*. Tatsächlich liefern diese drei Methoden sehr ähnliche Ergebnisse, was die Robustheit der Befunde unterstreicht. Diese Methoden haben vier wichtige Gemeinsamkeiten:

- 25+ Informanten
- 10+ Lexikalisierungen
- Urteile in numerischer Form
- kontrolliertes linguistisches Experimentmaterial.

Wenn diese vier Anforderungen erfüllt sind, dann verlieren die Kritikpunkte gegen introspektive Urteile weitgehend an Kraft. Weshalb das so ist, werden wir im Folgenden erläutern.

Wenden wir uns wieder den Vorwürfen zu. Zuerst haben wir beanstandet, dass Urteile als Datentyp ungenau sind. Sie sind variabel über Informanten und über Zeit, sie sind unquantifiziert, und sie sind nicht statistisch erfassbar. Dies alles trifft natürlich auf die einzelnen Urteile eines alleine in seinem Büro arbeitenden Syntaktikers zu, nicht jedoch auf diese Resultate von Urteilsexperimenten. Dadurch dass die Urteile numerisch erhoben werden, sind die Schwankungen erfassbar. Mit Werkzeugen wie Mittelwert und Standardabweichung lassen sich empirisch gestützte Aussagen machen. Resultatsmuster wie in Abbildung 2 zeigen, dass Urteile einer gewissen Fehlervarianz unterliegen, daher die berechtigte Kritik bezüglich der Schwankung. Abbildung 2 belegt aber auch eindeutig, dass die Fehlervarianz nicht regellos ist.

Diese Graphik stellt ein typisches Resultat einer Studie mit relativen Urteilen dar. Die getesteten syntaktischen Bedingungen werden auf der horizontalen Achse angeordnet, die vertikale Dimension quantifiziert empfundene Wohlgeformtheit, wobei höhere Zahlen eine größere Natürlichkeit der getesteten Sätze bedeuten. Die Fehlerbalken zeigen für die jeweilige Bedingung ein 95 % Konfidenzintervall für den Mittelwert.¹ Die Urteile sind z-Werte, d. h. relativ zum individuellen Mittelwert und zur individuellen Standardabweichung in eine normalisierte Skala überführt, damit die Urteile aller Informanten sinnvoll graphisch dargestellt werden können.

Diese Graphik zeigt, dass die Schwankung der Urteile nicht beliebig sondern mehr oder weniger normalverteilt ist. Manche individuellen Urteile bewegen sich weiter weg vom Mittelwert, aber sie bleiben trotzdem in relativer

¹ Diese Graphik zeigt die Ergebnisse einer Studie zur Diskursgebundenheit (*discourse linking*) im Deutschen (Featherston 2005a).

Nähe dazu. Angenommen eine Bedingung ist auf einer Sieben-Punkte-Skala konsensuell eine Zwei-Wert, so fallen neben den vielen Zweien auch Einser- und Dreier-Urteile an, eventuell auch ein Vierer-Wert, aber Fünfer-, Sechser-, und Siebener-Urteile kommen nicht vor. Das heißt, und darum geht es in dieser Diskussion, dass einzelne Urteile vom Mittelwert abweichen mögen, aber im Schnitt sind sich alle Sprecher einig. Dieser Effekt wird aber erst dann sichtbar, wenn man eine genügend große Stichprobe von Sprechern befragt. Aus dieser Perspektive kann man auch gut verstehen, weswegen manche Linguisten die Zuverlässigkeit von Urteilen angezweifelt haben: stellt man zwei Informanten die Frage, wie sie eine gewisse Struktur einschätzen, kann es sehr gut sein, dass der erste mäßig nach oben variiert, während der zweite etwas nach unten tendiert. Die zwei Urteile scheinen disjunkt zu sein. Dieser Tatbestand erlaubt unter anderem folgende Schlüsse: dass Urteile schwankungsanfällig und daher unbrauchbar sind, oder dass die zwei Informanten unterschiedliche Idiogrammatiken haben. Beide Folgerungen werden manchmal gezogen, aber sie sind keineswegs notwendig, denn die Erhebung weiterer Daten lässt das wahre Bild der Fehlervarianz erscheinen. Sie wird durch die üblichen Maße des Mittelwerts und der Standardabweichung erfasst und kontrolliert.

Abb. 2: Die Resultate einer Studie mit relativen Urteilen auf einer Intervallskala. Fehlervarianz verteilt sich normal um einen Mittelwert.

Es dürfte klar sein, dass nicht nur Varianz innerhalb einzelner Bedingungen sondern auch zwischen den Bedingungen quantifiziert werden kann. Bei Urteilen auf einer Intervallskala kann man sich nützlicher statistischer Verfahren wie der Varianzanalyse bedienen. Damit kann man empirisch festhalten, welche Unterschiede signifikant sind.

Auch der Vorwurf, dass Urteile als Datentyp subjektiv sind, ist nicht haltbar, wenn sie systematisch erhoben werden. Die Befragung von mehreren Informanten macht Urteile automatisch zumindest intersubjektiv, denn die

Mittelwerte sind gerade durch die Effekte bestimmt, die in der Informantengruppe überwiegen. Das Auftreten eines Effekts in einer Gruppe von 25 Informanten erlaubt darüber hinaus ganz klar die Vorhersage, dass ein ähnlicher Effekt in einer weiteren Stichprobe von Sprechern zu finden wäre. Damit sind die Effekte externalisiert und objektiv, und sie lassen sich daher auch replizieren. Jedenfalls sind die Ergebnisse insoweit objektiv, als dass sie nicht von der Willkür und Voreingenommenheit des Syntaktikers abhängen können.

Einen weiteren Schritt zur Objektivierung dieses Datentyps kann man mit der Einbeziehung von Standardsätzen als Füllsätzen tun. Für das Deutsche haben wir Gruppen von jeweils fünf Sätzen ermittelt und zusammengestellt, die die gesamte Skala der empfundenen syntaktischen Wohlgeformtheit abdecken. Wenn man in jedem Experiment eine oder zwei solcher Gruppen als Füllsätze einbindet, hat man eine Vergleichsgrundlage. Diese erlaubt es, die Werte aus verschiedenen Experimenten sinnvoll miteinander zu vergleichen und sie würde, wenn die Standardbeispiele besser bekannt wären, Urteilswerte nahezu absolut werden lassen. Die Auswirkungen dieser Technik werden klarer, wenn man sich den Nutzwert einer anderen geläufigen Standardskala überlegt. Wie viel schwieriger wäre es, von Temperaturen zu sprechen, wenn es nicht die Celsius-Skala gäbe! Tatsächlich ist die Wahl des Gefrier- und Kochpunkts von Wasser eher willkürlich (siehe Fahrenheitskala, Kelvin-skala), aber der Effekt ist deutlich: ein bekanntes Vergleichsmaß macht die Vorstellbarkeit und Kommunizierbarkeit von Temperaturen viel einfacher.

In Abbildung 3 sieht man anhand eines Beispiels, wie Standardbeispiele eingesetzt werden. In dieser Studie wurden die Dritte Konstruktion und das Lange Passiv miteinander verglichen, und zwar über eine Hierarchie von Verben hinweg. Die Fragestellung betraf die Reaktion dieser zwei Konstruktionen auf die lexikalischen Merkmale des Matrixprädikats (Featherston, unveröffentlicht). Der Befund: die Werte für die Dritte Konstruktion sind immer besser als die für das Lange Passiv, sie werden aber dennoch als sehr unnatürlich eingeschätzt, wie man im Vergleich zu den Standardbeispielen sieht. In so einem Fall kann eine Konstruktion zwar ‚besser‘ sein als eine andere, aber sie ist deshalb nicht automatisch Teil der Sprache.

Die systematische Erhebung von Urteilen erlaubt es auch, den Vorwurf zu kontern, dass das gemessene Konstrukt bei Urteilen nicht klar definiert sei. Bei diesem Vorgehen lassen sich Einschränkungen präzisieren, was gemessen wird und was nicht. Dadurch dass wir mehrere Informanten befragen, können wir behaupten, dass wir eine Generalisierung über Sprecher messen. Dadurch dass wir mehrere Lexikalisierungen testen, können wir behaupten, dass wir eine Generalisierung über die Struktur messen. Es ist auch Teil der experimentellen Methode, verschiedene irrelevante Störfaktoren auszuschließen oder zumindest zu reduzieren. Wir kontrollieren im Experimentmaterial die Form (lexikalische Häufigkeit, Wortlänge usw.), den Inhalt (Plausibilität), den

Abb. 3: In dieser Studie wurde eine Gruppe von Standardbeispielen als Vergleichsmaß benutzt. Diese fünf bekannten Werte sieht man auf der linken Seite der Graphik. Die Werte für die Dritte Konstruktion sind immer besser als diejenigen für das Lange Passiv, sie werden aber im Vergleich zu den Standardbeispielen als sehr unnatürlich eingeschätzt.

Kontext (man kann entweder Kontext variieren und explizit angeben oder einfach weglassen, dann nehmen die Informanten den zugänglichsten Kontext), und den Kommunikationsbedarf (in einem Experiment abwesend). Dadurch entfallen viele Störfaktoren.

Es ist interessant, Chomskys (1965, S. 3) Präzision des Anwendungsgebiets der linguistischen Theorie mit den tatsächlichen Merkmalen experimentell erhobener Urteilsdaten zu vergleichen.

Linguistic theory is concerned primarily with an ideal speaker-listener in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by memory limitations, distractions, shifts of attention and interest, and errors (random and characteristic) in applying his knowledge in actual performance.

Aus diesem berühmten Zitat können wir entnehmen, dass Chomsky über individuelle und lokale Unterschiede, über Verarbeitungseffekte (das heißt zeitdruckabhängige Effekte), über inhalts- und kommunikationsbezogene Faktoren sowie über kontextbezogene Faktoren hinwegidealisiert. Das System an sich ist von Interesse, und nicht die Art, wie das System in konkreten Situationen angewendet wird. Nun entspricht diese Wunschliste weitgehend dem, was man mit experimenteller Kontrolle erreichen kann. Es ist selbstverständlich nicht der Fall, dass experimentelle Urteile Zugang zur

reinen Kompetenz erlauben, es bleibt noch viel Rauschen in den Daten, aber näher als mit experimentellen Urteilen kommt man beim jetzigen Stand der Technik nicht an Chomskys Definition. Schon allein dadurch sollten Urteile für Linguisten von Interesse sein.

Zusammenfassend stellen wir fest, dass experimentell erhobene, relative Urteile quantifizierbar, statistisch erfassbar, intersubjektiv, replizierbar, und verhältnismäßig frei von Störfaktoren sein können. Zudem haben sie zwei große Vorteile: sie erlauben eine höchst fokussierte Erhebung und trotzdem liefern sie eine sehr feine Differenzierung. Man kann genau die Daten sammeln, die man für einzelne Punkte der Theoriebildung benötigt, und sei die Struktur noch so obskur oder selten. Und trotz der Seltenheit des Phänomens bekommt der Linguist klare, empirisch einwandfreie und statistisch signifikante Ergebnisse. Das macht diese Methode zu einem sehr leistungsstarken Werkzeug für die Syntaxforschung.

Ein letzter Kommentar: gerade die Robustheit der Ergebnisse ermöglicht es, diese Methode nur teilweise anzuwenden, wenn relativ einfache Resultate genügen. Das bedeutet, dass man oft keine allzu aufwändige experimentelle Vorgehensweise benutzen muss. Individuelle Intuitionen von Wohlgeformtheit korrelieren mit Experimentergebnissen sehr stark: wenn ein Resultat für eine Einzelperson nicht nachvollziehbar ist, dann ist das ein Zeichen dafür, dass dieses Resultat nur mit größter Vorsicht zu behandeln ist. Die Experimente liefern zwar mehr Details als die Intuitionen eines einzelnen Sprechers, aber sie sind letzten Endes der gleiche Datentyp, den man nur mit einer höheren Auflösung beobachtet. Dies wiederum bedeutet, dass die Urteile einer einzelnen Person nicht unterbewertet werden sollten, allerdings darf diese Person nicht voreingenommen sein. Den Urteilen eines einzelnen Linguisten kann man und sollte man nur geringen Glauben schenken, aber wenn fünf Unbeteiligte das gleiche behaupten, dann ist das mit hoher Wahrscheinlichkeit ein ernstzunehmender Befund. Anders ausgedrückt: ein verhältnismäßig großer Anteil der begründeten Zweifel an introspektiven Urteilen hängt mit der unsystematischen und unkontrollierten Art ihrer Erhebung zusammen. Sie haben jedoch keine intrinsisch mindere Qualität als Evidenz.

Einblicke in die Syntaxtheorie

In diesem zweiten Teil werden wir darlegen, welches Bild von der Syntax wir mit experimentell erhobenen Urteilen erhalten. Obwohl diese Daten viele Eigenschaften der Standardmodelle bestätigen, werden einige wichtige Annahmen falsifiziert, was für unser Bild von der Architektur der Grammatik Erneuerungen erzwingt. Ich werde im Folgenden für zwei Thesen argumentieren, die sowohl für die Syntaxtheorie selbst wie auch für den Evidenzwert von Korpusdaten von Bedeutung sind.

Die Ergebnisse von Urteilsstudien sehen so aus wie in Abbildung 4. Wie in den vorigen Graphiken zeigt die y -Skala die empfundene Wohlgeformtheit

an, die die Informanten numerisch ausgedrückt haben. Die Werte für jede Bedingung werden mit einem Fehlerbalken angegeben, aus dem der Mittelwert und das 95% Konfidenz-Intervall abzulesen sind. In diesem Beispiel haben wir acht syntaktische Bedingungen, die aus je drei binären Parameter bestehen. Die Fehlerbalken der einzelnen Minimalpaare sind mit einer Linie verbunden. Auf der x -Achse haben wir die Werte für die drei binären Parameter angegeben. Wenn die Beschränkung erfüllt ist, bekommt der Parameter den Wert 1, wenn sie verletzt wird, bekommt der Parameter den Wert 0.

Wenn wir die Resultate betrachten, so stellen wir zunächst fest, dass die Beurteilungen nicht, wie ein binäres Modell der Wohlgeformtheit es vorhersagen würde, in zwei Gruppen erscheinen (grammatisch und ungrammatisch), sondern ein Kontinuum darstellen. Auch in keiner anderen Studie ließ sich eine Binarität der Urteile nachweisen. Anstelle von zwei Gruppen sehen wir, dass die Bewertungen direkt auf die Anzahl und Schwere der verletzten Beschränkungen reagieren. Die Verletzung einer Beschränkung hat eine konstante Auswirkung auf die Bewertung einer Struktur. Dies erkennen wir, wenn wir die Werte für Minimalpaare betrachten. Das Verhältnis zwischen den Minimalpaaren, die eine bestimmte Beschränkung erfüllen und verletzen, ist konstant. Das heißt, die Verletzung einer einzelnen bestimmten linguistischen Beschränkung hat immer die gleiche Auswirkung, unabhängig davon, wie viele Beschränkungen eine Struktur sonst noch verletzt. Diesen Befund sehen wir immer wieder: Informanten benutzen kein binäres Modell der Wohlgeformtheit, wenn es ihnen nicht aufgezwungen wird. Man sollte im Auge behalten, dass dieser Befund nicht durch Performanzfaktoren wegerklärt werden kann, da diese in einer experimentellen Umgebung gerade kontrolliert werden.

Abb. 4: Diese Graphik zeigt exemplarisch, wie die Resultate von Experimenten mit relativen Urteilen aussehen. Diese Ergebnisse zeigen Wohlgeformtheit als Kontinuum und quantifizierbare, kumulative Beschränkungsverletzungskosten. Es gibt kein Anzeichen für ‚harte‘, kategorische Beschränkungen.

Anhand solcher Daten müssen wir konstatieren, dass Beschränkungen nicht kategorisch zu sein scheinen. Eine kategorische Beschränkung könnte sich bei diesem Datentyp in zwei möglichen Verhaltensweisen zeigen. Die eine Möglichkeit wäre, dass die verletzende Struktur so schlecht wird, dass sie durch zusätzliche Verletzungen nicht noch schlechter gemacht werden kann. Das würde dem Zustand der absoluten Unwohlgeformtheit entsprechen, denn was ausgeschlossen ist, kann nicht weiter bestraft werden. Diesen Zustand findet man jedoch nie. Strukturen können immer noch schlechter gemacht werden, denn das Bild der strukturellen Wohlgeformtheit, das wir von diesen Daten bekommen, verhält sich kumulativ (Keller 2000).²

Eine zweite Möglichkeit, eine ‚harte‘ Beschränkung zu erkennen, wäre, dass jede sie verletzende Struktur auf ein gleiches Niveau fällt, egal wie gut sie ohne diese Verletzung eingeschätzt wird. Wenn das eintreten würde, wäre die Quantifizierbarkeit der Verletzungskosten nicht mehr gegeben, und die Annahme wäre begründet, dass diese Beschränkung eine verletzende Struktur kategorisch als Teil der Sprache ausschließt. Aber auch dies kommt nicht vor. Die Kosten in Form von verminderter Wohlgeformtheit sind bei einer bestimmten Verletzung immer gleich und deshalb quantifizierbar. Man kann Beschränkungen mit stärkeren und andere Beschränkungen mit schwächeren Verletzungskosten erkennen, wie auch hier in der Abbildung, aber unser Datentyp unterstützt nicht die Annahme, dass es kategorische Beschränkungen in der Sprache gibt.

Bis jetzt haben wir gesehen, dass Informanten, wenn sie die freie Wahl der Skala haben, keine harten Beschränkungen erkennen, sondern immer nur quantifizierte, kumulative Verluste in der empfundenen Wohlgeformtheit. Dies erzwingt ein Bild der Wohlgeformtheit als Kontinuum. Wir sollten jedoch noch eine weitere Feststellung notieren, nämlich, dass wir kein Anzeichen von Verletzbarkeit im Sinne der *violability* der Optimalitätstheorie (OT, Prince/Smolensky 1993) ausmachen können. In der OT haben Beschränkungen immer nur dann eine Auswirkung, wenn sie zwischen den übrigen strukturellen Kandidaten differenziert. EVAL kann nicht zählen, nicht einmal bis eins, was bedeutet, dass diese Funktion zwischen dem Fall, bei dem alle noch aktuellen Kandidaten eine Beschränkung verletzen, und dem Fall, bei dem keiner der aktuellen Kandidaten eine Beschränkung verletzt, nicht unterscheidet. In keinem dieser beiden Fälle hat die Beschränkung eine Auswirkung auf die laufende Evaluation. Das bedeutet aber, dass die Beschränkung keine Anwendung findet. Diesen Tatbestand findet sich jedoch nicht in unseren Urteilen: alle Beschränkungen werden immer angewendet, alle Verletzungen haben immer eine Auswirkung auf die Urteile. Es spielt für eine Beschränkung keine Rolle, ob eine ansonsten perfekte Struktur vorliegt,

² Dies betrifft nur *strukturelle* Wohlgeformtheit, wohl gemerkt. Wenn Informanten einen Satz nicht mehr verstehen, wird das Bild verschwommener. Aber in unseren Studien geht es um die syntaktische Form, nicht um den Inhalt von Strukturen.

oder ob die Struktur bereits aus unabhängigen Gründen schlecht ist: die Beschränkung selbst findet immer Anwendung. Dementsprechend stützt dieser Datentyp auch nicht die Konzepte der probabilistischen oder optionalen Anwendung von Beschränkungen. Im Gegenteil deuten diese Daten darauf hin, dass die Anwendung von Beschränkungen blind und automatisch erfolgt, wie dies in der generativen Syntax traditionell angenommen wurde.

Diese Befunde sind in relativen Urteilen konstant, robust und unvermeidbar. Aber unsere Schlussfolgerungen für die Natur der Wohlgeformtheit stimmen nicht mit den gängigen Annahmen überein, die ebenfalls aufgrund von Urteilen entwickelt worden sind. Wie kann es sein, dass introspektive Urteile zwei so widersprüchliche Bilder abgeben können? Der Grund liegt darin, dass wir zwei Sorten von Urteilen unterscheiden müssen: relative und kategorische.

Relative Urteile und kategorische Urteile

In unseren empirischen Studien haben wir herausgefunden, dass relative Urteile, die wir mit Sorgfalt unter strenger Kontrolle erheben, ganz andere Merkmale aufweisen als traditionell angenommen wurde. Gleichzeitig aber gibt es verhältnismäßig starke Evidenz, dass die traditionell binären Urteile auch psychologisch reell sind. Jeder Sprecher kann mit dem Konzept ‚absolute Grammatikalität‘ etwas anfangen. Tatsächlich entspricht dieser Ausdruck einer intuitiv vorhandenen Kategorie, die wir haben, obwohl diese in unseren relativen Urteilen überhaupt nicht sichtbar ist. Daher sind wir gezwungen, zwei Typen von Urteilen zu unterscheiden.

Relative Urteile, die quantifizierbare, kumulative, beschränkungsspezifische Verletzungskosten auf einer Kontinuumsskala aufweisen, geben den Komputationsaufwand der Struktur wieder. Mit Komputationsaufwand meinen wir so etwas wie die Umwandlung von ungeformten Botschaften in Wortsequenzen (cf. Culicover/Nowak 2003 und Referenzen dort). Kategorische Urteile dagegen sind eine Aussage, ob eine Struktur gut genug ist, um vorzukommen. Unsere Hypothese ist es deshalb, dass diese zwei Urteilstypen auf unterschiedliche Faktoren reagieren. Relative Urteile quantifizieren den Denkaufwand, den eine Struktur verursacht. Dass wir auf die Menge kognitiver Arbeitsleistung empfindlich reagieren, und dass dies auch dem Bewusstsein zugänglich sein kann, ist aus anderen Kontexten klar; schließlich wundert sich niemand wenn wir wissen, ob eine Rechnung ‚leicht‘ oder ‚schwer‘ ist. Es ist entspannender, einen Krimi zu lesen als ein Werk schöngeistiger Literatur, weil ersteres weniger Denkmühe verursacht; wir sind uns dessen auch bewusst.

Kategorische Urteile dagegen sind ein ganz anderes Maß, denn sie drücken die Meinung aus, ob eine Struktur vorkommen würde. Dabei kann man sich zwei Evidenzquellen vorstellen, aufgrund deren diese Aussage gemacht werden könnte: das interne Korpus und der Vergleich mit anderen Kandidaten. Im ersten Fall ist das Vorgehen eher direkt. Der Informant sucht sein

internes Korpus daraufhin ab, ob er Formen dieser Struktur (oder Strukturverbindung) gehört hat. Falls ja, ist die Struktur grammatisch.

Der zweiten Fall ist komplexer und verlangt zuerst eine Erklärung, was wir unter Äußerungsauswahl (*output selection*) verstehen. Man muss davon ausgehen, dass der Sprecher im Produktionsmodus zwischen verschiedenen strukturellen Varianten auswählen muss, welche er produzieren wird. Diese Auswahl wird manchmal zwischen gleich guten Alternativen sein, manchmal zwischen besseren und schlechteren Varianten. In diesen Fällen gehen wir davon aus, dass der Sprecher normalerweise die ‚bessere‘ auswählt. Die Auswahl kann lexikalisch sein (*Mach die Tür auf!* vs. *Öffne die Tür!*), strukturell (*Er sagt, dass er mich liebt* vs. *Er sagt, er liebt mich*) oder auch thematisch (*Ich belade den Wagen mit Heu* vs. *Ich lade das Heu auf den Wagen*), aber fest steht, dass wir diese Fähigkeit haben, zwischen Alternativen von Äußerungen auszuwählen. Wenn wir ein binäres Urteil abgeben, kann es sein, dass wir mit diesem Auswahlverfahren die zu bewertende Struktur daraufhin abschätzen, ob man nicht eine andere Version bevorzugt hätte. Falls nicht, dann ist die Struktur ‚gut genug, um vorzukommen‘.

Es ist vielleicht bemerkenswert, dass Syntaktiker manchmal eine zuge-spitzte Version dieses zweiten Verfahrens einsetzen, die einen string-technischen Begriff der Grammatikalität zu Grunde legt. Sie scannen einen Satz, um festzustellen, ob ein bekannter Verletzungstyp darin vorkommt. Wenn nicht, dann geben sie der Struktur das Prädikat ‚grammatisch‘. Tatsächlich bildet dieses Vorgehen einen dritten Urteilstyp, dessen Wohlgeformtheitsdefinition etwa ‚diese Struktur enthält keine Verletzung, die normalerweise mit Vorkommen inkompatibel ist‘ lauten könnte.

Wir haben nun drei verschiedene Arten von Urteilen differenziert, die alle in der Linguistik benutzt werden. Unsere Hypothese ist: Relative Urteile reagieren auf die komputationelle Denklast, während kategorische Urteile aussagen, ob eine Struktur vorkommen würde. Die dritte, stringtechnische Urteilsart sagt aus, ob eine bekannte schwerwiegende Beschränkungsverletzung in einem Satz identifiziert werden kann. Soweit zu Urteilsdaten. Nun werden wir das Verhältnis zwischen unseren bevorzugten relativen Urteilen und Vorkommensfrequenzen erläutern.

Relative Urteile und Häufigkeitsdaten

Ein häufiger Befund in der Syntaxforschung ist, dass die Ergebnisse von Studien mit Urteilen und solchen mit Korpusfrequenzen übereinstimmen, wenn es darum geht, die beste Form aus einer Reihe von Kandidaten zu identifizieren (Featherston 2004, 2005b, Kempen/Harbusch 2005). Diese zwei Datentypen liefern jedoch kontrastierende Bilder, wenn man das Verhalten der schwächeren Kandidaten betrachtet. In Frequenzdaten kommen die Verlierer kaum oder gar nicht vor. In relativen Urteilen dagegen werden Strukturvarianten, die zu schlecht sind, um jemals natürlich vorzukommen, genau

so differenziert und gestaffelt bewertet wie vorkommende Strukturen. Unter der Annahme unserer Hypothese, dass relative Urteile eine Bewertung von komputationeller Komplexität darstellen, ist dies auch erklärbar. Korpusdaten können notwendigerweise immer nur Effekte bei Strukturen messen, die auch vorkommen. Frequenzen von Strukturen, die allesamt nicht vorkommen, werden nicht unterschieden. Der Denkaufwand einer noch so schlechten Struktur wird dagegen sehr wohl gemessen, und fließt daher auch in unsere relativen Urteile ein. Das bedeutet, dass relative Urteile Informationen enthalten, die in Frequenzen nie erscheinen können.

Abb. 5: Die Auswirkung des Eisbergeffekts: Frequenzen können uns immer nur ein Bild von guten Strukturen geben. Was nicht gut genug ist, um vorzukommen, ist unsichtbar. Vergleiche mit dem Bild der relativen Urteile in Abbildung 4.

Wir werden uns in diesem Kontext einer Metapher bedienen. Sprachliche Daten sind wie ein Eisberg: es gibt Strukturen die vorkommen: sie ähneln dem sichtbaren Teil des Eisbergs, der aus dem Wasser ragt. Es gibt noch viel mehr Strukturen, die nie vorkommen: sie bilden den versteckten Teil des Eisbergs unter der Wasseroberfläche. Diese Strukturen sind anhand von Frequenzdaten nicht zu erkennen; man kann lediglich feststellen, dass sie nicht vorkommen.

In diesem Kontext beschränken wir uns darauf, nur ein Ergebnis zu erwähnen, das bei Urteilsdaten zu ungrammatischen Strukturen anders ausfällt als bei Frequenzdaten. Bei Frequenzdaten ist inhärent der Eindruck gegeben, dass die Verletzung einer Beschränkung das Vorkommen der verletzenden Struktur verhindert. Diese Idee des kategorischen Ausschlusses durch Verletzungskosten ist auch natürlich, denn Frequenzdaten bestehen aus vielen individuellen Einzelentscheidungen, ob eine bestimmte Struktur benutzt wird oder nicht. Daher sind die Bausteine von Korpusdaten binär, obwohl das Gesamtbild dieser kategorischen Einzelentscheidungen oft gradierte Formen annimmt. Wenn man mit Frequenzdaten arbeitet, ist es also eine sehr natürliche

Annahme, dass linguistische Beschränkungen kategorischer Natur sind. In Abbildung 5 sieht man skizzenhaft das Bild der Datenlage, das man mit Frequenzdaten bekommt: das obere Glied des Minimalpaares ist sichtbar, das untere oft nicht. Um die Verletzungskosten einer Beschränkung zu quantifizieren, müsste die Distanz zwischen den Elementen eines Minimalpaares sichtbar sein. Man kann die Verletzungskosten einer einzelnen Beschränkung mit Frequenzdaten also nicht messen, wenn das untere Glied eines Paares fehlt. Dies ist immer dann gegeben, wenn eine Verletzung in der Praxis bedeutet, dass eine verletzende Struktur effektiv nicht vorkommt. Dieser Tatbestand vermittelt den Eindruck, dass Strukturen entweder gut sind, oder gar nicht vorkommen (siehe Abbildung 5).

Diese Annahme wird aber durch relative Urteile widerlegt. Tatsächlich sehen die Verhältnisse zwischen Strukturvarianten, die nicht vorkommen, genau so aus wie die zwischen belegten Strukturvarianten oberhalb der Wasseroberfläche. Beschränkungen, die sowohl oberhalb wie auch unterhalb zur Anwendung kommen, benehmen sich überall gleich. Auch die Beschränkungen, die normalerweise als kategorisch eingeschätzt werden, haben quantifizierbare Verletzungskosten. Das legt folgendes Fazit nahe: Obwohl es den Frequenzdaten nach so aussieht, schließen Beschränkungsverletzungen eine Struktur nie automatisch aus. Die Vorkommengrenze an sich ist für relative Urteile nämlich völlig unsichtbar. Es muss festgestellt werden, dass das Verhältnis zwischen empfundener Wohlgeformtheit, wie sie in relativen Urteilen gemessen wird, und Vorkommen, gemessen anhand von Frequenzen, nicht einfach ist.

Bedeutung für die Architektur der Grammatik

Unsere These ist, dass die Sprachverarbeitungseinheit, die als Grammatik verstanden wird, tatsächlich aus zwei Modulen besteht, die unterschiedliche Funktionsweisen haben. Das erste stellt Kandidaten zur Verfügung, wie die Grammatik dies traditionell tut, und das zweite wählt in einem getrennten Schritt unter den möglichen Strukturen aus.

Sehr vereinfacht dargestellt werden im ersten Modul die verschiedenen linguistischen Beschränkungen auf die Form der Struktur angewendet (siehe Featherston 2005b). Beschränkungen sind nicht gerankt, noch werden sie geordnet angewendet, noch blockieren die einen die Anwendung der anderen, wie die OT dies vorsieht, sondern die Beschränkungen verhalten sich zueinander wie die klassische generative Syntax dies immer vorgesehen hat: blind, simultan, und automatisch. Dieses Modul stellt auch die Verletzungskosten fest: eine Struktur, die eine Beschränkung verletzt, wird um einen beschränkungsspezifischen Wert in ihrer empfundenen Wohlgeformtheit herabgesetzt. Werden zwei oder mehr Beschränkungen verletzt, wird die Struktur noch schlechter, denn diese Verletzungskosten sind kumulativ. Dieser Teil der Grammatik schließt keine Strukturvarianten aus, sondern gibt alle

Varianten mit einer Art Wohlgeformtheitsgewichtung weiter. Wir nennen diesen Teil die Beschränkungsanwendung (*constraint application*). Die Existenz dieses Modul ist notwendig, um die Form der relativen Urteilsdaten zu erklären.

Die zweite Funktion heißt Äußerungsauswahl (*output selection*). Hier wird zwischen den Kandidaten ausgewählt, denn nur eine Form kann tatsächlich produziert werden. Als Auswahlkriterium wird die Wohlgeformtheit berücksichtigt. So kommt es, dass generell nur ‚gute‘ Strukturen produziert werden. Korpusdaten messen die Ausgabe dieses Moduls, denn sie zählen, was tatsächlich produziert wird.

Beschränkungsanwendung: *constraint application*

- wendet Beschränkungen an,
- stellt Verletzungen fest,
- weist Verletzungskosten zu (Wohlgeformtheitsgewichtung),
- liefert Kandidatenstrukturen mit Gewichtungen zur Auswahl.

Äußerungsauswahl: *output selection*

- wählt aus, welche Strukturvariante produziert wird,
- benutzt Wohlgeformtheitsgewichtungen als Kriterium,
- funktioniert probabilistisch.

Diese Thesen bieten eine Erklärung dafür, weshalb relative Urteile und Frequenzdaten unterschiedliche Ergebnismuster liefern. Die Urteile messen die Wohlgeformtheitsgewichtung aus dem ersten Modul, Frequenzdaten messen das Vorkommen. Da die Konkurrenz um das Vorkommen auf der Basis des Wohlgeformtheitsstatus stattfindet, stimmen Urteilsdaten und Frequenzdaten in ihrer Identifikation der besten Strukturen überein. Suboptimale und schlechte Strukturen dagegen bekommen sehr wohl Wohlgeformtheitswerte, und können deshalb in den Urteilen unterschieden werden.

Diese schwächeren Kandidaten kommen aber in der Praxis nicht vor, denn sie gewinnen nie den Wettbewerb in der Äußerungsauswahl. Sie werden also in den Frequenzdaten nicht unterschieden. Dieser Erklärungsversuch basiert also direkt auf den empirischen Befunden.

Schlussfolgerungen für die Syntaxtheorie

In diesem letzten Teil gehen wir auf zwei Vorteile ein, die sich aus der Unterscheidung der beiden Grammatikmodule ergeben. Der erste betrifft die empirische Adäquatheit der Syntaxtheorie; der zweite die Relevanz der Korpusdaten für die Theorienbildung. Die Syntaxtheorie leidet gerade unter mangelnder Entscheidungsfähigkeit. Nehmen wir die kontroverse Rolle der Konkurrenz in der Theorie als Beispiel. Syntaktiker benutzen mehrere in ihrer Architektur völlig inkompatible Grammatikmodelle. Die *Government and Binding*-Gruppe von Syntaxmodellen (Chomsky 1981) enthält gar keine Kon-

kurrenz. Strukturen sind inhärent entweder gut oder schlecht. Beschränkungen werden blind und automatisch angewendet. Head-driven Phrase Structure Grammar (Pollard/Sag 1994) beinhaltet auch nur ansatzweise eine Konkurrenz-Funktion, in der *obliqueness hierarchy* in Kapitel 6. Hier muss ein Argument weniger oblique sein, um ein anderes binden zu können. Im Minimalistischen Programm (Chomsky 1993) spielt das economy principle eine größere Rolle. Hier konkurrieren Strukturvarianten eindeutig gegeneinander; deren Wohlgeformtheit ist somit relativ zu einer Vergleichsgruppe. In der OT (Prince/Smolensky 1993) findet die Konkurrenz in der Syntax seine Apotheose. In der OT gibt es gar keine absolute Wohlgeformtheit, sondern nur relative. Auch Beschränkungen haben keine sichere Existenz in der Sprache, denn nur diejenige haben eine Auswirkung, die zwischen Kandidaten unterscheiden. Die Anwendung von Beschränkungen ist deshalb nicht blind und automatisch sondern bedingt. Dieser Aufstieg der Konkurrenz in der Syntax wird von Müller/Sternefeld (2001) mit Weitblick beschrieben.

Diese Modelle unterscheiden sich zum Teil recht stark, aber sie werden alle als Erklärungen für den mehr oder weniger gleichen Evidenzbestand angeboten. Wir haben nur ihre Architekturunterschiede in Bezug auf einen einzigen Parameter dargestellt, nämlich den Platz, der dem Konzept der Konkurrenz eingeräumt wird. Dass Syntaktiker unterschiedliche Analysen spezifischer Strukturen bevorzugen, sollte nicht verwundern, dass sie aber so völlig inkompatible Beschreibungsrahmen annehmen können, muss als besorgniserregend gelten.

Auf den ersten Blick lässt dies zwei Erklärungen zu: entweder betrachten Syntaktiker die Evidenz nicht, die zwischen den Modellen entscheiden könnte, oder aber die Evidenzbasis der Syntax ist so dünn, dass man nicht einmal die groben Züge der Architektur der Syntax erkennen kann. Keine der beiden Möglichkeiten gibt ein sehr positives Bild unseres Feldes.

Akzeptiert man aber unser Modell der zwei Grammatik-Module, so eröffnet sich eine neue Perspektive. Das erste Modul, *constraint application*, funktioniert blind, automatisch und kumulativ, ohne jegliche Konkurrenz. Die Ausgabe dieses Modul, so haben wir argumentiert, ist uns durch relative Urteile zugänglich. Diese Funktionsweise entspricht in etwa den Annahmen der *Government and Binding*-Theorie. Das zweite Modul, *output selection*, ist eine Konkurrenzfunktion; Strukturen konkurrieren aufgrund ihrer Wohlgeformtheit um das Vorkommen. Die Ausgabe dieses Moduls sind Frequenzdaten. Die Funktionsweise entspricht der der Optimalitätstheorie. Es scheint also, dass die starken Abweichungen in den Grammatikarchitekturen jeweils einen Teil der Realität der Funktionsweisen der Grammatik widerspiegeln. So kann es kommen, dass Grammatikmodelle mit und ohne Konkurrenz-elemente einigermaßen empirisch adäquat sein können. Korpusdaten beinhalten Beweise dafür, dass die Produktionsweise von Äußerungen unter anderem auch kompetitiv arbeitet. Wenn man Wohlgeformtheit über Vorkommen definiert, dann scheint ein Konstrukt der relativen Wohlgeformtheit

sinnvoll und empirisch begründet. Verwendet ein Linguist eher (nichtexperimentelle) Urteile, bekommt er kein Indiz für Konkurrenz, sondern für die blinde Funktionsweise des ersten Moduls der Beschränkungsanwendung, die absolute Wohlgeformtheit erzeugt.

Wir müssen daher weder die Schlussfolgerung ziehen, dass Syntaktiker das Beweismaterial ignorieren, noch dass die Evidenzlage zu mager ist, um gesicherte Erkenntnisse zu erlauben. Die derart unterschiedlichen Grammatikmodelle können nicht alle richtig sein, aber sie erfassen alle einen Teil des Gesamtbildes. Dieses ist jedoch etwas komplexer als die einzelnen Grammatikmodelle annehmen. Sie treffen auf Schwierigkeiten, weil sie versuchen, die unterschiedlichen Merkmale von zwei unterschiedliche Funktionsweisen auf dieselbe Weise in einem einzigen Modell zu erfassen.

So viel zum ersten Vorteil unserer Unterscheidung von zwei Grammatikmodulen. Der zweite Vorteil betrifft den Evidenzwert von Korpusbelegen. Findet man eine Struktur in einem Korpus, so bedeutet dies, dass diese Struktur im Wettbewerb gegen andere Strukturvarianten gewonnen hat. Nun hat aber dieser Wettbewerb, wie wir auch oben erwähnt haben, einen probabilistischen Charakter, was zur Folge hat, dass manchmal auch die zweitbeste oder sogar die drittbeste Struktur, gemessen an ihrer Wohlgeformtheit, erscheinen kann.

Dieses Merkmal der Arbeitsweise der Äußerungsauswahlfunktion ist durchaus empirisch begründet und anerkannt. Viele korpusbasierte Studien untersuchen, welche Faktoren die Auswahl der einen oder der anderen Strukturvariante bevorzugen (z. B. Bader/Häussler 2006, Bresnan et al. 2005). Dabei ist es aber wesentlich, dass diese Korpusdaten kein kategorisches Gesamtbild zeigen. Nicht Vorkommen und Abwesenheit werden verglichen, sondern höhere und niedrigere Frequenzen. In jedem einzelnen Produktionsfall ist es deshalb nur eine Frage von Wahrscheinlichkeiten, ob der Sprecher zum Beispiel *Den Studenten hat der Vortrag gefallen* oder *Der Vortrag hat den Studenten gefallen* sagt, denn die beitragenden Faktoren halten sich in diesem Fall ungefähr die Waage (Subjekte im Vorfeld werden bevorzugt aber belebte Experienter im Vorfeld werden auch bevorzugt). Damit ist aber auch vorgegeben, wofür wir argumentieren: die Äußerungsauswahl ist eine probabilistische Kompetition auf der Basis von Wohlgeformtheit.

Wenn dieser Schluss einmal akzeptiert ist, gilt es die Konsequenzen zu untersuchen. Natürlich wird die ‚beste‘ Strukturvariante am häufigsten vorkommen. Dass auch nur etwas weniger ‚gute‘ Kandidaten ebenfalls vorkommen, wenn auch mit minderer Frequenz, leitet sich aus der probabilistischen Funktionsweise der Äußerungsauswahl ab.

Aber auch deutlich weniger ‚gute‘ Strukturen werden zwar selten aber regelmäßig vorkommen, denn die Probabilistik sagt voraus, dass auch unwahrscheinliche Ereignisse hin und wieder stattfinden. Für uns Linguisten bedeutet dies, dass wir in Korpusdaten selten aber vorhersehbar mehr oder weniger ‚schlechte‘ Strukturen finden werden, ohne dass dieser Befund irgend-

welche Schlussfolgerungen für die Grammatik nach sich ziehen muss. Man kann aus dieser Evidenz nichts über die kausalen Faktoren schließen, die die verschiedenen Beschränkungen im Kern der Grammatik bedingen, denn es ist die probabilistische Auswahlfunktion (output selection), die das Vorkommen dieser ‚schlechten‘ Beispiele bewirkt. Ihr Vorkommen in einem Korpus ist daher keine Evidenz, dass sie als ‚gut‘ einzuschätzen sind. Ihr Erscheinen in Korpusdaten lässt uns höchstens schließen, dass sie nicht allzu viel schlechter sind als die Strukturvarianten, die normalerweise vorkommen.

Wenn Linguisten Belege von Strukturen in einem Korpus finden, so nehmen sie dies als Evidenz, dass diese Struktur grammatisch ist. Unsere Hypothese zum bimodularen Charakter der Grammatik stellt dies in Frage. Der erste Teil enthält die Beschränkungen, ihre Anwendung, ihre Interaktion, und die Feststellung der Wohlgeformtheit von Strukturen: somit entspricht er dem Kernteil der Grammatik. Ein gesonderter zweiter Teil wählt zwischen den Varianten aus, und zwar unter Berücksichtigung der im ersten Modul zugewiesenen Wohlgeformtheitsgewichtung. Damit ist gewährleistet, dass allgemein immer nur die besten Strukturen in Korpusdaten erscheinen. Aber da diese Auswahlfunktion probabilistisch arbeitet, können manchmal auch suboptimale Varianten die Auswahlkonkurrenz gewinnen, was bedeutet, dass sie dann in Korpora zu finden sind. Findet der Linguist also wider Erwarten seltene Belege für eine suboptimale Struktur, bedeutet dies nicht, dass er sein Grammatikmodell anpassen muss, um diesem Befund Rechnung zu tragen (kontra z. B. Müller 2003).

Zusammenfassung

In diesem Beitrag haben wir dargelegt, unter welchen Umständen introspektive Urteile objektive, quantifizierbare und empirisch adäquate linguistische Daten sein können. Hierzu ist ein experimenteller Ansatz notwendig. Wenn man eine Vielzahl von Informanten befragt, werden auch introspektive Urteile effektiv objektiv, denn sie lassen Vorhersagen zu, die falsifiziert werden können. Die Erhebung von Urteilen in numerischer Form auf einer Intervallskala erlaubt die Quantifizierung von Urteilen, die Ermittlung von deren Mittelwerten und Standardabweichung, sowie die Anwendung nützlicher statistischer Testverfahren. Wenn man auch verschiedene lexikalische Varianten der syntaktischen Bedingungen testet, dann lässt sich die Behauptung machen, dass die Resultate frei von lexikalischen Effekten sind. Durch den experimentellen Kontext werden viele Störfaktoren ausgeschlossen oder deren Auswirkung abgemildert. Damit erfüllen diese experimentell erhobenen, relativen Urteile die Anforderung der empirischen Adäquatheit soweit das im Rahmen des Möglichen erreicht werden kann. Weil sie eine sehr genau fokussierte Erhebung erlauben und zudem noch klare, robuste Ergebnisse liefern, sind diese Urteile eine sehr attraktive Datenquelle für die Syntax.

Im zweiten Teil haben wir beschrieben, wie die Resultate solcher Studien aussehen. Sie zeigen ein Kontinuum der Wohlgeformtheit, ohne binäre Unterscheidung zwischen ‚gut‘ und ‚schlecht‘. Sie zeigen, dass alle Beschränkungen immer angewendet werden und dass die Verletzung einer Beschränkung eine quantifizierbare, beschränkungsspezifische Reduzierung der empfundenen Wohlgeformtheit zur Folge hat. Die Grenze zwischen Strukturen, die vorkommen, und solchen, die nicht vorkommen, ist dagegen unsichtbar. Daraus schließen wir, dass relative Urteile etwas anderes messen als das Vorkommen.

Um diesen Tatbestand zu erklären, haben wir unsere Hypothese von den zwei Grammatikmodulen entwickelt. Der erste Teil wendet linguistische Beschränkungen an und ermittelt Wohlgeformtheit. Der zweite ist eine Konkurrenzfunktion und wählt die Struktur aus, die produziert wird. Wir haben argumentiert, dass diese Architektur eine Lösung zu zwei Problemen in der Syntaxforschung bietet. Zum ersten kann sie erklären, weshalb Grammatikmodelle mit völlig unterschiedlichen Architekturen empirisch erfolgreich sein können. Die einen spiegeln die Funktionsweise des ersten Moduls wider, die anderen die Funktionsweise des zweiten Moduls. In einem weiteren Schritt haben wir darauf hingewiesen, dass die Architektur voraussagt, dass selten aber regelmäßig sub-optimale Strukturen in Korpusdaten erscheinen werden. Diese Belege erzwingen aber keine Änderungen in der Grammatik, denn sie erscheinen nur wegen des probabilistischen Charakters des Auswahlverfahrens. Nur ein relativ häufiges Vorkommen hat Implikationen für die Wohlgeformtheit.

Bibliographie

- Bader, Markus/Häussler, Jana (2006): Word-order variation: Why corpus and judgment data do not go hand in hand! Poster bei der Tagung ‚Linguistic Evidence‘, Tübingen.
- Bard, Ellen/Robertson, Dan/Sorace, Antonella (1996): Magnitude estimation of linguistic acceptability. In: *Language* 72 (1), S. 32–68.
- Bresnan, Joan/Cueni, Anna/Nikitina, Tatiana/Baayen, Harald (2005): Predicting the Dative Alternation. Erscheint in: Royal Netherlands Academy of Science. Workshop on Foundations of Interpretation proceedings.
- Chomsky, Noam (1981): Lectures on Government and Binding. The Pisa Lectures. Berlin: Mouton de Gruyter.
- Chomsky, Noam (1995): The Minimalist Program. Cambridge, Massachusetts: MIT Press.
- Crain, Steven/Fodor, Janet (1987): Sentence matching and overgeneration. In: *Cognition* 26, S. 123–169.
- Culicover, Peter/Nowak, Andrzej (2003): Markedness, antisymmetry and complexity of constructions. In: Pica, Pierre/Rooryk, Johann (Hg.): *Variation Yearbook*. Amsterdam: Benjamins.
- Featherston, Sam (2004): Judgements in syntax: Why they are good, how they can be better. Vortrag bei der DGfS Jahrestagung 2004.
- Featherston, Sam (2005a): Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. In: *Lingua* 115/11, S. 1525–1550.

- Featherston, Sam (2005b): The Decathlon Model: Design features for an empirical syntax. In: Reis, Marga/Kepser, Stephan (Hg.): *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter. S. 187–208.
- Keller, Frank (2000): *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Disseration, University of Edinburgh.
- Kempen, Gerard/Harbusch, Karin (2005): The relationship between grammaticality judgements and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Reis, Marga/Kepser, Stephan (Hg.): *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. S. 329–350. Berlin: Mouton de Gruyter.
- Labov, William (1975): What is a linguistic fact? In: Austerlitz, Robert (Hg.): *The Scope of American Linguistics*. Lisse: Peter de Ridder. S. 77–133.
- Lakoff, George (1973): Fuzzy grammar and the performance/competence terminology game. In: *Chicago Linguistics Society* 9, S. 271–291.
- Müller, Gereon (1995): *A-bar Syntax. A Study in Movement Types*. *Studies in Generative Grammar* 42. Berlin/New York: de Gruyter.
- Müller, Gereon/Sternefeld, Wolfgang (2001): The rise of competition in syntax: A synopsis. In: Müller, Gereon/Sternefeld, Wolfgang, (Hg.): *Competition in Syntax*. Berlin: Mouton de Gruyter.
- Müller, Stefan (2003): Mehrfache Vorfeldbesetzung. In: *Deutsche Sprache* 31, S. 29–62.
- Pollard, Carl/Sag, Ivan (1994): *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
- Poulton, Edward (1989): *Bias in Quantifying Judgments*. Hove & London: Lawrence Erlbaum.
- Prince, Alan/Smolensky, Paul (1993): *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Center for Cognitive Science Technical Report 2.
- Sampson, Geoffrey (2001): *Empirical Linguistics*. London/New York: Continuum.
- Schütze, Carson (1996): *The Empirical Basis of Linguistics*. Chicago: University of Chicago Press.
- Wurmbrand, Susanne (2001): *Infinitives: Restructuring and Clause Structure*. Berlin/New York: Mouton de Gruyter.

STEFAN MÜLLER

Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus¹

Abstract

In diesem Aufsatz diskutiere ich drei syntaktische Phänomene, die für die Grammatikforschung von zentraler Bedeutung sind. Ich zeige, dass Introspektion als Stütze von Theorien nicht ausreicht und entwickle Korpusanfragen für die diskutierten Fälle. Der Aufsatz schließt mit Anmerkungen zu den Grenzen der Korpuslinguistik.

1. Einleitung

Zu den Zielen der theoretischen Linguistik gehört es, möglichst starke Aussagen über Sprache bzw. Sprachen zu machen. Ein Beispiel für eine solche Aussage ist die Behauptung, dass sich Subjekte in allen Sprachen von Objekten unterscheiden, wobei ein Unterschied darin bestehen soll, dass man Wortgruppen, die zu Subjekten gehören, nicht voranstellen kann (siehe z. B. Grewendorf 1989, S. 35, S. 41, Kaufmann 1995, S. 194, G. Müller 1998, S. 32, S. 163, Sabel 1999, S. 98, Fanselow 2001, S. 422).² Eine andere Behauptung ist, dass man bei nicht-lokalen Umstellungen nie mehr als zwei Maximalprojektionen bestimmter Kategorien (z. B. vollständige Nominalphrasen) kreuzen kann, wobei für Bewegung nach links die Art der Kategorien sprachspezifisch bzw. sprachklassenspezifisch ist (Baltin 1981, S. 262).

Eine Aussage ist stark, wenn sie leicht widerlegbar ist. Die Behauptung, dass ein Prinzip für alle Sprachen gilt, ist stärker, als die Annahme eines Prinzips für nur eine Sprache, da zur Widerlegung eines sprachübergreifenden Prinzips Material aus einer beliebigen Sprache verwendet werden kann. In jedem Fall ist ein Prinzip für Äußerungen mit bestimmten Eigenschaften

¹ Ich möchte mich bei den Organisatoren der IDS-Jahrestagung für die Einladung bedanken. Hartmut und Renate Schmidt danke für Kommentare zu diesem Aufsatz, und Gisbert Fanselow, Sam Featherston, Detmar Meurers, George Smith und den Teilnehmern der IDS-Jahrestagung danke ich für Diskussion. Meinem Vortrag zur Jahrestagung lag das Tiger-Korpus von 2003 zugrunde, für diesen Aufsatz habe ich z. T. das Tiger-Korpus von 2005 verwendet, das gegenüber der 2003er Version korrigiert und erweitert wurde.

² Bei Voranstellung und mitunter auch bei Nachstellung von Material spricht man auch von Extraktion. Je nach Theorie geht man davon aus, dass es eine bestimmte Grundkonfiguration gibt, aus der eine bestimmte Wortgruppe ‚extrahiert‘ wird.

formuliert und rein logisch reicht für die Widerlegung eine einzelne Äußerung, die nicht den Vorhersagen entspricht, aus. Ich möchte das gleich einschränken, denn, wenn man genau ein Beispiel hat, das einer Generalisierung widerspricht, so kann dieses immer noch auf Sprachproduktionsfehler oder – wenn es sich um geschriebene Sprache handelt – ganz einfach auf typographische Fehler zurückzuführen sein. Außerdem kann auch eine idiomatische Wendung vorliegen, die nicht den normalen grammatischen Gesetzmäßigkeiten entspricht. Man sollte also schwächer formulieren: Wenn es eine bestimmte Klasse von Gegenbeispielen gibt, kann eine Generalisierung oder ein Prinzip nicht mehr als richtig bzw. sinnvoll gelten.

Dies soll noch am Beispiel der Diskussion um die Extraktion aus Subjekten verdeutlicht werden. Haider (1993, S. 159) diskutiert die Beispiele in (1):³

- (1) a. [Wessen Beispiele]_i hat [_i analysieren zu müssen] dich mehr frustriert – seine oder meine?
 b. [Diese Beispiele analysieren zu müssen] hat dich mehr frustriert.

Im a-Satz ist eine Teilphrase aus dem Subjektsinfinitiv vor das finite Verb gestellt worden. Der b-Satz zeigt dagegen den gesamten Infinitiv im Vorfeld. Akzeptiert man Sätze wie den a-Satz, ist die Diskussion um die Möglichkeit der Extraktion aus Subjekten beendet, denn solche Beispiele zeigen, dass sie möglich ist. Es ist dann logisch nicht legitim zu behaupten – wie das auch in den 80er Jahren gemacht wurde – dass es aber viele Fälle gibt, in denen die Extraktion aus Subjekten eindeutig ungrammatisch ist. Vielmehr muss man davon ausgehen, dass die Syntax ein Gerüst zur Verfügung stellt, das Extraktion aus Subjekten prinzipiell zulässt und muss dann zusätzlich dazu Beschränkungen finden, die die Extraktion aus Subjekten genau in den Fällen ausschließt, in denen sie auch wirklich unmöglich ist.

Bis in die 50er Jahre des letzten Jahrhunderts wurde in der Sprachwissenschaft mit Belegen gearbeitet, die zum großen Teil aus den Werken anerkannter Schriftsteller stammen. Ende der 50er Jahre änderte sich das in einem Teilbereich der Linguistik und zwar so radikal, dass ausschließlich introspektiv gewonnene Beispiele verwendet wurden. In den Anfangsjahren der generativen Linguistik war das unproblematisch, da Sätze wie *Das Buch schenkt Peter seinem Bruder zu Weihnachten* behandelt wurden, deren Akzeptabilität außer Frage steht. In den 70er und 80er Jahren des letzten Jahrhunderts erreichte die Theorie jedoch eine beachtliche Komplexität, und die jeweiligen Analysen wurden immer noch durch introspektiv konstruierte Beispiele gestützt bzw. auf Grund introspektiv konstruierter Beispiele verworfen. Dabei entstand eine unbefriedigende Situation, da mitunter der Einfluss eines einzelnen Autors über Kernbereiche einer syntaktischen Theorie entschied, ohne dass die Akzeptabilitätsurteile des Autors bzw. dessen Behauptungen in Bezug auf

³ Zur Extraktion aus Subjektsnominalphrasen siehe Engel 1970, S. 78, Oppenrieder 1991, S. 79, Haider 1993, S. 173 und De Kuthy 2001, S. 50.

nicht vorkommende Muster je überprüft worden wären. Argumente gegen eine Theorie wurden mitunter mit der Begründung „man teile die Beurteilung nicht“ einfach beiseite gewischt (zu diesen Punkten siehe auch Fanselow 2004). In jüngster Zeit gibt es Bestrebungen, dies zu ändern. Es stehen größere Korpora und bessere Korpus-Werkzeuge für die Allgemeinheit zur Verfügung (Belica 2006; Bodmer 2005; F.H. Müller 2004a; Ule 2004; Brants u. a. 2004; Geyken, erscheint) und auch die Grundlagenforschung zur experimentellen Erhebung von Sprachdaten hat große Fortschritte gemacht (einige wichtige empirieorientierte Arbeiten findet man in Sorace/Keller 2005; Featherston 2005; Bresnan u. a. 2005; Kepser / Reis 2005; Barbiers u. a. 2005; Charles u. a. 2006 und Fanselow u. a. 2006).

Im Folgenden möchte ich anhand einiger Fallstudien zeigen, dass uns unsere grammatischen Fähigkeiten mitunter nicht zugänglich sind und dass wir uns deshalb selbst misstrauen sollten. Ich werde auch zeigen, wie einfach es mitunter ist, mit einer Korpusanfrage Klarheit in Bezug auf grammatische Phänomene zu gewinnen. Die Vorfeldfähigkeit von Verbzusätzen wird Gegenstand von Abschnitt 2 sein. Ich zeige, dass sich auch in unannotierten Korpora (wie z. B. dem World Wide Web) Belege für die Vorfeldbesetzung finden lassen. Abschnitt 3 beschäftigt sich mit der Frage, über wieviele Wortgruppengrenzen hinweg Konstituenten umgestellt werden können. Für das Auffinden relevanter Korpusdaten ist hier ein annotiertes Korpus sehr hilfreich, und ich zeige, wie die Anfragen formuliert werden können. Als drittes beschäftige ich mich mit der scheinbar mehrfachen Vorfeldbesetzung, die ein wichtiges Phänomen darstellt, da die Existenz solcher Mehrfachbesetzungen die Annahme widerlegt, dass die Vorfeldbesetzung durch eine Wortfolge hinreichende Information über deren Konstituentenstatus liefert. Die dritte Fallstudie unterscheidet sich von den ersten beiden dadurch, dass man annotierte Korpora braucht, denn ohne Information über Abhängigkeitsverhältnisse ist die Suche nach entsprechenden Belegen extrem aufwendig.

2. Partikelverben

Die Mehrzahl der Grammatiker ist bis vor wenigen Jahren davon ausgegangen, dass eine Verbpartikel nicht ins Vorfeld gestellt werden kann.⁴ Die Ausnahmen, die gesondert behandelt wurden, sind solche, in denen die Partikel prädikativ ist, d. h. auch in Kopulakonstruktionen verwendet werden kann (vergleiche *das Fenster aufmachen* und *Das Fenster ist auf.*) und Zifonuns „Partikelverben im weiteren Sinne“ (1999, S. 223): Bildungen in denen die Partikel auch durch eine Präpositionalphrase ersetzt werden könnte.

Die Behauptung zur Vorfeldbesetzung ist nicht korrekt, wie z. B. die folgenden Daten zeigen:

⁴ Siehe jedoch Reis 1976, S. 68; Lötscher 1985, S. 211; Dürscheid 1989, S. 21; Hoberg 1997, S. 1621 und Lüdelling 1997 zum Deutschen und Hoeksema 1991 und Bennis 1991 zum Niederländischen.

- (2) a. Ob er seine Strafe dort absitzen muß, war gestern ebenso unklar wie die Frage, ob er die gesamten elf Monate weggeschlossen wird. *Vor hat* er das jedenfalls.⁵
 b. *Los ging* es schon in dieser Woche.⁶
 c. Sein Vortrag wirkte [...] ein wenig arrogant, nicht zuletzt wegen seiner Anmerkung, neulich habe er bei der Premiere des neuen „Luther“-Films in München neben Sir Peter Ustinov und Uwe Ochsenknecht gesessen. Gut *an kommt* dagegen die Rede des Jokers im Kandidatenspiel: des Thüringer Landesbischofs Christoph Kähler.⁷

Weitere Daten und eine ausführliche Diskussion aller Behauptungen in Bezug auf die Vorfeldfähigkeit von Partikeln findet man in Müller 2002a, b, 2005.

Aus dem in der Einleitung Gesagten ergibt sich, dass Belege wie die in (2) ausreichen, um Theorien/Analysen, die auf der falschen Annahme der Nicht-voranstellbarkeit aufbauen, zu widerlegen.

Liest man die Literatur über Partikelverben sorgfältig, so findet man in Aufsätzen von Autoren, die behaupten, dass Partikeln nicht ins Vorfeld gestellt werden können, Sätze wie die folgenden:

- (3) a. Hinzukommt, daß sich unabhängig von den gesondert zu behandelnden Topikalisierungsphänomenen eine Reihe von Beispielen anführen läßt, die [...]
 b. Hinzukommt, daß Partikel-Verb-Kombinationen durchaus produktiv sind, [...]
 c. Hinzukommt, daß dieses methodologische Argument ohnehin zu relativieren ist, da [...]

Diese Sätze sollten Verbzweitsätze sein. Sowohl nach der alten als auch nach der neuen Rechtschreibung sollte *hinzukommen* getrennt geschrieben werden.⁸ Zählt man diese Fälle zu den Partikelverben, widerlegen sie direkt die Aussage des Autors, dass die Partikel nicht vorfeldfähig ist. Mir liegt es fern, die Orthographie in einem wissenschaftlichen Aufsatz zu kritisieren, aber ich denke, dass die Beispiele in (3) zeigen, dass uns unsere sprachlichen Fähigkeiten nicht immer bewusst sind.

Allein die Tatsache, dass sehr gute Grammatiker behauptet haben, dass Verbpartikeln nicht vorfeldfähig sind, zeigt, dass den Menschen ihre grammatischen Fähigkeiten nicht unmittelbar zugänglich ist, selbst wenn sie sich darum bemühen. Das ist insbesondere bei Phänomenen so, die relevant für die Informationsstruktur sind, denn in solchen Fällen muss man sich besondere Äußerungskontexte überlegen, die weniger häufige Stellungsmuster ermög-

⁵ taz, 15. 07. 1999, S. 19.

⁶ taz, 11. 10. 1995, S. 4.

⁷ taz, 4. 11. 2003, S. 3.

⁸ Der Duden gibt Beispiele für Partikelvoranstellung mit *auffallen* (Duden 1991) bzw. *fehlschlagen* (Duden 1996).

lichen bzw. sogar erzwingen. Probiert man nur einige Beispiele im Kopf durch, gelangt man eventuell zu falschen Schlüssen. Viel einfacher ist es oft, das Korpus zu befragen. Man bekommt dann zusammen mit entsprechenden Belegen einen natürlichen Kontext gleich mitgeliefert.

Im Folgenden möchte ich zeigen, wie einfach es mitunter ist, Behauptungen zu überprüfen bzw. zu widerlegen. Für den gerade diskutierten Fall gibt es eine sehr einfache Suchmethode: Man gibt die Verbpartikeln, die einem einfallen, gefolgt von einem Verb/Hilfsverb in die Suchmaske einer Suchmaschine im World Wide Web ein. Da Partikel und Verb bei Verbletzstellung meist zusammen geschrieben werden, bekommen wir Belege für Partikeln im Vorfeld. Der Vorteil dieser Methode ist, dass man kein annotiertes Korpus benötigt und somit auf riesige Datenbestände zugreifen kann. Der Nachteil ist, dass viel Handarbeit erforderlich ist, denn man muss sich die zu suchenden Partikelverben ausdenken, sie per Hand eingeben und bekommt obendrein noch Ergebnisse, die man nicht verwerten kann, da die Suchergebnisse auch Präpositionen vor dem Verb enthalten. So bekommt man z. B. Kombinationen von *aus* und *kommt* wie die in (4), in denen das *aus* keine Verbpartikel ist.

(4) Von Berlin aus kommt man überall hin.

Verwendet man dagegen ein annotiertes Korpus wie z. B. Tiger (Brants u. a. 2004) kann man Anfragen wie die folgende stellen:

```
[pos="PTKVZ"] . [pos=finite]
```

Im Tiger-Korpus sind alle Wörter in Bezug auf ihre Wortart (*pos* steht für *part of speech*) annotiert. Im obigen Beispiel werden zwei Wörter beschrieben: zum einen eins mit der Wortart Partikel/Verbzusatz (PTKVZ) und zum anderen ein finites Verb. Der ‚.‘ sagt aus, dass die beiden beschriebenen Wörter adjazent sein müssen.

Mit einer solchen Anfrage bekommt man falsche Ergebnisse nach dem Muster von (4) nicht, da die Annotation sicherstellt, dass wir wirklich Partikelverben finden. Allerdings liefert die Anfrage auch Verbletztsätze wie den Relativsatz in (5):

(5) dem Anfang der neunziger Jahre Hohn und Spott zuteil wurde (s37155)

zuteil wurde als Verbpartikel annotiert, das Verb zuteil werden wird allerdings auseinander geschrieben. Man kann die Anfrage präzisieren und dadurch auch Ergebnisse wie (5) ausschließen.

```
#s:[cat="S"] > #part:[pos="PTKVZ"] &
#part . [pos=finite] &
#s >@1 #leftcorner &
#leftcorner:[pos!= prorel] &
#leftcorner:[pos!= prointer] &
#leftcorner:[pos!= conjunction]
```

Diese etwas komplexere Anfrage verwendet Variablen, um mehrfach auf bestimmte Knoten in einem Baum zu verweisen. So bezieht sich #s auf einen Knoten im Baum mit der Kategorie S (Satz). Der Satz muss eine Partikel (#part) dominieren. Die Dominanzrelation wird durch das >-Zeichen ausgedrückt. Die zweite Zeile besagt, dass die Partikel an ein finites Verb grenzen muss. Der >@l-Operator in der dritten Zeile sucht das Element, das am weitesten links im Baum steht. Für dieses Element wird der Bezeichner #left-corner vergeben. In den Zeilen vier bis sechs wird verlangt, dass dieses Element nicht von der Wortart Relativpronomen, Interrogativpronomen bzw. Konjunktion ist. Die Negation wird dabei durch das Ausrufezeichen repräsentiert. Durch die Beschränkung der Wortart des am weitesten links im Baum stehenden Elements werden die meisten Verbletztsätze ausgeschlossen, Sätze wie (6) werden jedoch nicht erfasst:

(6) auf das wir uns sehr gefreut haben

In (6) steht das Relativpronomen nicht an erster, sondern an zweiter Stelle. Dieses Beispiel zeigt, dass es viel schwieriger ist, Verbletztsätze genau zu beschreiben, aber eine absolut genaue Beschreibung ist glücklicherweise gar nicht nötig. Bei Korpusanfragen reicht es aus, gerade so genau zu sein, dass man sich bei der Auswertung der Ergebnisse Arbeit erspart. Wenn die Korpusanfrage keine oder nur sehr wenige unerwünschte Ergebnisse liefert, war die Anfrage ausreichend genau formuliert. Ungenauigkeit kann sogar von Vorteil sein, da eine ungenaue Beschreibung von gesuchten Bäumen unter Umständen aufgrund ihrer geringeren Komplexität schneller vom Computer verarbeitet werden kann.

Mit der oben angegebenen Anfrage findet man im Tiger bzw. im Negra-Korpus z. B. die folgenden Sätze:

(7) a. *Fest steht*, daß dort 580 der insgesamt 4650 Arbeitsplätze wegfallen.

(Tiger:s30056)

b. *Los geht* es am Samstag nach der offiziellen Eröffnung mit einem gemütlichen Beisammensein. (Negra:s11241)

c. *Entgegen kamen* sich Koalition und Opposition in der Frage um die Verkehrsberuhigung der Titusstraße. (Negra:s11514)

Man sieht, dass man mit relativ geringem Aufwand schnell Belege für ein Muster finden kann, von dem oft behauptet wird, dass es nicht existiert.

3. Extraposition und Subjazenzenz

Die sogenannte Subjazenzenz spielt in der Linguistik Chomskyscher Prägung eine große Rolle. Es wird behauptet, dass bestimmte Umstellungen nur eine beschränkte Reichweite haben. So diskutieren zum Beispiel Baltin (1981) und Chomsky (1986, S. 40) Daten wie die in (8) und behaupten, dass der extraponierte Relativsatz mit Bezug auf die einbettende NP interpretiert werden

muss, d. h. die Sätze sind nicht äquivalent zu Sätzen, in denen der Relativsatz an der durch *t* markierten Stelle steht, sondern zu solchen, in denen er an der *t'*-Stelle steht.

- (8) a. [NP Many books [PP with [stories *t*]] *t'*] were sold [that I wanted to read].
 b. [NP Many proofs [PP of [the theorem *t*]] *t'*] appeared [that I wanted to think about].

Dies wird durch die Annahme des Subjazenz-Prinzips erklärt, das das Kreuzen von mehr als einer Maximalprojektion einer Hauptkategorie (NP, PP, VP, AP) verbietet (Baltin 1981, S. 262).

Die Bezugnahme auf Subjazenz als wichtiges Prinzip in der Grammatik ist weit verbreitet und wird auch gegenwärtig in Lehrmaterialien vertreten (Haegeman 1994, S. 422, Klenk 2003, S. 96, Baltin 2006). Die Aussage, dass Extraposition lokal stark beschränkt ist, findet man für das Deutsche auch bei Grewendorf (1988, S. 281), Haider (1991, 1996), Rohrer (1996) und Kiss (2005, S. 316–317). Z. T. arbeiten diese Wissenschaftler in anderen Frameworks (LFG, HPSG) bzw. machen andere Annahmen in Bezug auf die syntaktische Analyse der Extraposition (Basisgenerierung vs. Bewegung).

Subjazenz ist eine zentrale Säule in der Chomskyschen Linguistik (mit einigen wenigen Ausnahmen wie z. B. von Stechow und Sternefeld 1988, S. 93). Baltins generalisierte Subjazenz erhebt den Anspruch, Ross' Bedingungen für Extraktionen (1967) zu ersetzen.

Im Folgenden möchte ich zeigen, dass die Subjazenzbedingungen, die in der Literatur vorgestellt wurden, für die Extraposition im Deutschen nicht gelten. (9) zeigt drei Beispiele für Relativsatzextraposition.⁹ Dabei wird die Anzahl der Einbettungen von Beispiel zu Beispiel erhöht. Durch die Bedeutung des Relativsatzes ist nur die Bezugnahme auf Frau möglich.¹⁰

- (9) a. Karl hat mir [ein Bild [einer Frau *_i*]] gegeben,
 [die schon lange tot ist]_i.
 b. Karl hat mir [eine Fälschung [des Bildes [einer Frau *_i*]]] gegeben,
 [die schon lange tot ist]_i.
 c. Karl hat mir [eine Kopie [einer Fälschung [des Bildes [einer Frau *_i*]]]]
 gegeben, [die schon lange tot ist]_i.

⁹ Siehe auch Müller 1999, S. 211 bzw. Müller 2004b.

¹⁰ Wollte man zeigen, dass die Umstellung über viele Phrasengrenzen unmöglich ist, müsste man die Beispiele so konstruieren, dass wie in (9) aus semantischen Gründen nur ein Bezug auf das am tiefsten eingebettete Nomen möglich ist. Die entsprechenden Sätze müssten dann immer unakzeptabel sein. Die Beispiele in (8) sind für diesen Zweck nicht geeignet, da hier mehrere Anbindungsmöglichkeiten zur Verfügung stehen und die Tatsache, dass es nur eine Lesart gibt, auf Verarbeitungspräferenzen zurückzuführen sein könnte. Chomsky argumentiert aber für eine Trennung von Kompetenz und Performanz. Bei der Zusammenstellung von Beispielen für Experimente müssen also Inferenzen mit Performanzphänomenen so weit wie möglich ausgeschlossen werden.

Diese Art von Einbettung könnte man noch weiter fortsetzen, wenn nicht irgendwann die Nomina ausgingen, die eine semantisch sinnvolle Einbettung gestatten.

In den Beispielen in (9) wurde jeweils ein Relativsatz extraponiert. Diese Relativsätze werden als Adjunkte behandelt, und es gibt Analysen, die davon ausgehen, dass extraponierte Adjunkte nicht durch Umstellung an ihren Platz gelangt sind, sondern basisgeneriert werden und die Koreferenz/Koindizierung über spezielle Mechanismen sichergestellt wird (Kiss 2005). Für Vertreter solcher Analysen wären die Beispiele in (9) im Rahmen der Subjanzdiskussion irrelevant, da es sich eben nicht um Bewegung handelt und das Subjanzprinzip nur über Bewegungen etwas aussagt. Die Extraposition über Phrasengrenzen hinweg ist jedoch nicht auf Relativsätze beschränkt, auch Satzkomplemente können extraponiert werden:

- (10) a. Ich habe [von [der Vermutung _i]] gehört,
 [dass es Zahlen gibt, die die folgenden Bedingungen erfüllen]_i.
 b. Ich habe [von [einem Beweis [der Vermutung _i]]] gehört,
 [dass es Zahlen gibt, die die folgenden Bedingungen erfüllen]_i.
 c. Ich habe [von [dem Versuch [eines Beweises [der Vermutung _i]]]]
 gehört,
 [dass es Zahlen gibt, die die folgenden Bedingungen erfüllen]_i.

Da es auch Nomina gibt, die *zu*-Infinitive bzw. Präpositionalphrasen einbetten und da diese wie die Sätze extraponiert werden können, muss sichergestellt werden, dass die syntaktische Kategorie des nachgestellten Elements der Kategorie entspricht, die vom entsprechenden Nomen verlangt wird. Das heißt, es muss irgendeine syntaktische Beziehung zwischen regierendem Nomen und extraponiertem Element geben.

Bisher habe ich gezeigt, dass man die Behauptung in Bezug auf die Lokalität der Extraposition mit konstruierten Beispielen widerlegen kann. Im Folgenden soll gezeigt werden, wie man mit Hilfe von Korpora Beispiele für die Extraposition über mehrere Phrasengrenzen hinweg finden kann. Verfügt man nicht über ein annotiertes Korpus, kann man dennoch Belege finden. Man kann zum Beispiel im World Wide Web oder – wenn man Wert auf bestimmte Textsorten legt – auch im COSMAS-Korpus nach Nomina suchen, die einen *dass*-Satz einbetten können und auch wirklich zusammen mit *dass* vorkommen. Der Nachteil dieser Methode ist, dass man sehr viele unerwünschte Belege bekommt, z. B. solche wie (11), in denen der *dass*-Satz nicht extraponiert ist.

- (11) Ja, es kam so weit, daß einige Kantone in der Auffassung, daß nur ein Geisteskranker zum Mörder werden könne, oder daß kein Richter das Recht habe, einen Mitmenschen zum Tode zu verurteilen, die Todesstrafe in ihren kantonalen Strafrechten ausmerzten.¹¹

¹¹ St. Galler Tagblatt, 06.02.1999; Das letzte Blutgericht in Lichtensteig.

Dadurch, dass in (11) mehrere *daß* auftreten, hilft es nicht, Beschränkungen in Bezug auf den Wortabstand zwischen Nomen und *daß* zu formulieren.

Im Jahr 2000 – als es noch keine annotierten Korpora für das Deutsche gab – habe ich mit dieser Suchmethode nach einer Stunde auch den Beleg in (12) im COSMAS-Korpus gefunden.

- (12) Für das Volk der Deutschen Demokratischen Republik ist dabei [die einmütige Bekräftigung [der Auffassung _i]] wichtig, [daß es die Interessen des Friedens und der Sicherheit erfordern, daß [. . .]];¹²

Wesentlich effizienter ist dagegen die Suche in einem syntaktisch annotierten Korpus. Mit der folgenden Anfrage findet man Belege im Tiger-Korpus:

```
#xp:[cat="NP"] >OC [ ] &
[cat=("NP"|"PP")] > #xp &
discontinuous(#xp)
```

Mit der Anfrage sucht man nach einem NP-Knoten (#xp), der einen Objektsatz (OC) unmittelbar dominiert. Die Art der Dominanzrelation ist in den Bäumen annotiert und mit Operatoren wie >OC kann man nach entsprechenden Bäumen suchen. Zeile zwei drückt aus, dass #xp selbst unmittelbar von einem Knoten dominiert werden muss, der eine NP oder PP ist. NPen und PPen werden im Tiger-Korpus als flache Strukturen annotiert, so dass eine einfache Dominanz in Zeile zwei ausreicht. In PPen befinden sich das Kopfnomen, der Determinator der NP und die Präposition auf einer Ebene. In der dritten Zeile der Anfrage wird verlangt, dass der Knoten #xp diskontinuierlich ist. Normalerweise ist die Extraposition des Objektsatzes der Grund für die Diskontinuität von Nominalphrasen mit Objektsätzen.

Diese Anfrage liefert 12 Treffer in der 2003er Version des Tiger-Korpus (40018 Sätze). Das Ergebnis bekommt man in 1,1 Sekunden zuzüglich der Zeit, die man zur Eingabe der Anfrage braucht. (13) zeigt einen der Belege:

- (13) [...] die Erfindung der Guillotine könnte [NP die Folge [NP eines verzweifelten Versuches des gleichnamigen Doktors] gewesen sein, [seine Patienten ein für allemal von Kopfschmerzen infolge schlechter Kissen zu befreien]. (s16977)

Diese Anfrage lässt sich leicht so modifizieren, dass man statt der extraponierten Objektsätze extraponierte Relativsätze findet: Man muss nur die Dominanzbeschränkung in der ersten Zeile der Anfrage durch >RC ersetzen. Sätze mit extraponierten Relativsätzen, die noch eine weitere Maximalprojektion kreuzen, kann man mit der folgenden Anfrage finden:

```
#xp:[cat="NP"] >RC [ ] &
discontinuous(#xp) &
#yp:[cat=("NP"|"PP")] > #xp &
[cat=("NP"|"PP")] > #yp
```

¹² Neues Deutschland, 06. 12. 1969, S. 1, COSMAS.

Durch die Suche nach einem zusätzlichen NP- bzw. PP-Knoten, der die erste NP bzw. PP (#yp genannt) dominiert, stellt man sicher, dass das Nomen mit dem Relativsatz unter mindestens zwei Nominalphrasen eingebettet ist. Man findet dann Belege wie (14):

- (14) Der 43jährige will nach eigener Darstellung damit [NP den Weg [PP für [NP eine Diskussion [PP über [NP den künftigen Kurs [NP der stärksten Oppositionsgruppierung]]]]]] freimachen, [die aber mit 10,4 Prozent der Stimmen bei der Wahl im Oktober weit hinter den Erwartungen zurückgeblieben war]. (s27639)

Nach der Beschäftigung mit der Subjazenzen möchte ich mich nun der scheinbar mehrfachen Vorfelddbesetzung zuwenden.

4. Mehrfache Vorfelddbesetzung

Die dritte Fallstudie setzt sich mit einem weiteren Phänomen auseinander, das gängigen Vorstellungen darüber, wie die deutsche Sprache zu sein hat, zuwiderläuft. Es handelt sich dabei um die scheinbar mehrfache Vorfelddbesetzung.

Normalerweise steht im Deutschen eine Konstituente vor dem finiten Verb (Erdmann 1886, Kapitel 2.4; Paul 1919, S. 69, S. 77): (15a) zeigt ein Beispiel mit einer Nominalphrase im Vorfeld, im abweichenden Beispiel in (15b) gibt es dagegen zwei Nominalphrasen vor dem Finitum.

- (15) a. Maria stellt Max Peter vor.
b. *Maria Max stellt Peter vor.

Der Test auf Vorfelddfähigkeit von Wortfolgen ist ein wichtiger Konstituententest, und mitunter wird die Vorfelddfähigkeit sogar in der Definition des Begriffes Satzglied verwendet (siehe z. B. Bußmann 1983, in Bußmann 1990 wurde der Eintrag jedoch korrigiert).

Schaut man sich natürlich vorkommende Sprache an, so stellt man jedoch fest, dass die Vorfelddbesetzung allenfalls als Indiz für den Konstituentenstatus verwendet werden kann, keinesfalls jedoch als hinreichendes Kriterium. Belege für scheinbare Mehrfachbesetzung des Vorfeldds zeigt (16):

- (16) a. [Trocken] [durch die Stadt] kommt man am Wochenende auch mit der BVG.¹³
b. [Gezielt] [Mitglieder] [im Seniorenbereich] wollen die Kendoka allerdings nicht werben.¹⁴
c. [Wenig] [mit Sprachgeschichte] hat der dritte Beitrag in dieser Rubrik zu tun, [...] ¹⁵

¹³ taz berlin, 10. 07. 1998, S. 22.

¹⁴ taz, 07. 07. 1999, S. 18.

¹⁵ Zeitschrift für Dialektologie und Linguistik, LXIX, 3/2002, S. 339.

- d. [Ein wenig Farbe] [ins triste (Menü-)Grau] bringt die Setup-Einstellung zum Wechseln der Balkenfarbe, aber einfacher zu bedienen macht das die Kamera nicht.¹⁶

Ähnlich problematische Fälle wurden auch von van de Velde (1978), Grubacic (1965), Lötscher (1985) und Eisenberg (1994, S. 412) besprochen. Eine ausführliche Datendiskussion findet man in Müller 2003 und eine aktualisierte Belegsammlung auf der in der Literaturliste angegebenen Web-Seite. In der in Müller 2005 vorgestellten Analyse solcher Sätze gehe ich davon aus, dass das Material im Vorfeld eine Konstituente bildet, die von einem leeren verbalen Kopf abhängt, der einem Verb im Satz entspricht (siehe auch Fanselow 1993, Hoberg 1997, S. 1634 und G. Müller 1998, Kapitel 5.3 zu ähnlichen Vorschlägen). Somit liegt letztendlich doch wieder ein Satz vor, in dem sich das finite Verb an zweiter Stelle befindet. Trotzdem bleiben die Daten für Konstituententests problematisch, denn die Konstituententests sollen uns ja dabei helfen, herauszufinden, ob *gezielt*, *Mitglieder* und *im Seniorenbereich* in (17) eine Konstituente bilden.

- (17) Die Kendoka wollen allerdings nicht gezielt Mitglieder im Seniorenbereich werben.

Sowohl *gezielt* als auch *im Seniorenbereich* beziehen sich aber auf *werben* und sollten nicht ohne weiteres einer gemeinsamen Konstituente zugeordnet werden, die auch ein Argument des Verbs *werben* enthält.

Es ist klar, dass solche scheinbar mehrfachen Vorfeldbesetzungen strengen informationsstrukturellen Restriktionen unterliegen. Wie diese genau aussehen, ist gegenwärtig nicht ausreichend erforscht. Da diese Varianten stark restringiert und auch sehr kontextabhängig sind, neigt man als Grammatiker dazu, sie ganz auszuschließen. Konstruierte Beispiele wie (15b) sind sehr schlecht. Introspektion hilft an dieser Stelle also nur weiter, wenn man das Phänomen schon kennt und auch eine Vorstellung darüber hat, wie die pragmatischen Bedingungen für solche Äußerungen aussehen müssen. Dass uns dieses Wissen nicht bewusst ist, sollen die beiden folgenden Sätze zeigen, die in einem wissenschaftlichen Text über Partikelverben direkt nacheinander zu finden sind:

- (18) [Eher] [Probleme] bekommt er mit den Sätzen, in denen das gesamte Partikelverb topikalisiert wurde. In diesem Fall würden zwei Konstituenten vor dem finiten Verb eines Hauptsatzes stehen, was im Deutschen ja nicht so ohne weiteres möglich ist.

Im Folgenden soll untersucht werden, ob Korpusanfragen helfen können, Sätze mit scheinbar mehrfacher Vorfeldbesetzung zu finden.

Sieht man sich die Belege in Müller 2003 an, so erkennt man, dass Material sehr verschiedener Kategorien im Vorfeld stehen kann. Deshalb kann man

¹⁶ <http://www.digitalkamera.de/Info/News/27/17.htm>. 08.09.2005.

ohne ein annotiertes Korpus ganz schlecht nach Belegen suchen. Selbst wenn man Information über die Wortart hat, hilft das nicht, denn man kann ohne Dependenzinformation Belege wie (16a) bzw. (19a) nicht von (19b) unterscheiden:

(19) a. [Stolz auf sein 30jähriges Bestehen] ist das Drei-Sterne-Hotel Hübsch am Bischofsheimer Griesterweg 12:

[.]¹⁷

b. [Empört] [auf die Pläne] reagierte der Fahrgastverband „Pro Bahn“.¹⁸

In (19a) hängt die *auf*-PP von *stolz* ab, *Stolz auf sein 30jähriges Bestehen* bildet also eine Konstituente und ist somit kein Beleg für die scheinbar mehrfache Vorfeldbesetzung. In (19b) dagegen hängt die PP vom Verb *reagieren* ab und nicht vom Adjektiv. Man kann sich das verdeutlichen, indem man den Sätzen in (19) die Kopulasätze in (20) gegenüberstellt:

(20) a. Das Drei-Sterne-Hotel ist stolz auf sein 30jähriges Bestehen.

b. *Der Fahrgastverband ist empört auf die Pläne.

Genauso lässt sich die Abfolge in (21a) nicht von der in (21b) unterscheiden:

(21) a. [Dem Votum Polens] schlossen sich die anderen Nationen an.

b. [Dem Zeitgeist] [Rechnung] tragen im unterfränkischen Raum die privaten, städtischen und kommunalen Musikschulen.¹⁹

In (21a) bilden *dem Votum* und *Polens* eine Konstituente, in (21b) hängt aber weder *Zeitgeist* von *Rechnung* noch *Rechnung* von *Zeitgeist* ab. An der morphologischen Markierung von *Rechnung* kann man nicht erkennen, ob *Rechnung* ein Genitiv ist, der zu einer Nominalgruppe gehört oder ein Akkusativ, der von einem anderen Element im Satz abhängt. Das heißt, wenn man mit einem unannotierten Korpus Belege wie (21b) finden will, muss man sich durch hunderttausende von Belegen nach dem Muster von (21a) arbeiten.

Verfügt man über ein entsprechend annotiertes Korpus, kann man Anfragen wie die folgende stellen:

```
#s:[cat="S"] >HD #fin:[pos=finite] &
#s >@l #sleftcorner &
#s > #vf1 &
#vf1 >@l #sleftcorner &
#vf1 >@r #vf1rightcorner &
#s > #vf2 &
#vf2 >@l #vf2leftcorner &
#vf2 >@r #vf2rightcorner &
#vf1rightcorner . #vf2leftcorner &
#vf2rightcorner . * #fin
```

¹⁷ Frankfurter Rundschau, 24. 09. 1999, S. 2.

¹⁸ taz, 28. 07. 1999, S. 1.

¹⁹ Fränkisches Volksblatt, zitiert nach Spiegel, 24/2002, S. 234.

Mit dieser Anfrage sucht man nach einem Satzknoten ($\#s$), der ein finites Verb ($\#fin$) und zwei weitere Konstituenten $\#vf1$ und $\#vf2$ dominiert. Von $\#vf1$ wird verlangt, dass es an $\#vf2$ grenzt, und $\#vf2$ soll vor dem Finiten Verb $\#fin$ stehen. Die Bedingung, dass $\#vf1$ an $\#vf2$ grenzen muss, lässt sich leider nicht so einfach ausdrücken, wie die Adjazenzbedingung für Partikel und Verb, die im Abschnitt 2 diskutiert wurde. Das liegt daran, dass es im Tiger-Korpus Bäume²⁰ mit diskontinuierlichen Konstituenten gibt, so dass sich unmittelbare Präzedenz nicht so ohne Weiteres repräsentieren lässt. In der obigen Anfrage wird deshalb der >@l' - bzw. >@r' -Operator verwendet. Die Beschränkung in Zeile zwei besagt, dass das am weitesten links im von $\#s$ dominierten Baum stehende Wort $\#sleftcorner$ ist. Für den Knoten $\#vf1$ wird in Zeile vier dasselbe verlangt. Damit erreicht man, dass $\#vf1$ am linken Rand von $\#s$ beginnt. $\#vf1$ kann dabei ein Wort oder auch eine komplexe Konstituente sein. In der fünften Zeile wird verlangt, dass das am weitesten rechts im von $\#vf1$ dominierten Teilbaum stehende Wort $\#vf1rightcorner$ ist. Analog bestimmt man die Knoten, die den am weitesten links bzw. rechts stehenden Wörtern innerhalb von $\#vf2$ entsprechen (Zeile 7–9). Dann kann man den Adjazenz-Operator verwenden und verlangen, dass das am weitesten rechts stehende Wort von $\#vf1$ an das am weitesten links stehende Wort von $\#vf2$ grenzen muss ($\#vf1rightcorner \#vf2leftcorner$). Für das am weitesten rechts in $\#vf2$ stehende Wort wird nicht verlangt, dass es direkt an das finite Verb grenzen muss, da das Belege wie (17) und (22) mit mehr als zwei Konstituenten im Vorfeld ausschließen würde.

- (22) a. [Endlich] [Ruhe] [in die Sache] brachte die neue deutsche Schwulenzbewegung zu Beginn der siebziger Jahre.²¹
 b. Auf Internet-Auktionen werden nur rund fünf Prozent der Zahlungen per Kreditkarte abgewickelt. [Bei Online-Auktionären ebenso wie bei Internet-Shoppern] [hoch] [im Kurs] stehen dagegen Überweisungen, Zahlung per Nachnahme oder Barzahlung.²²
 c. [Wenig] [mit Politik] [am Hut] hat auch der Vorarbeiter, der sich zur Aussage hinreißen lässt, „daß der Sausgruber das falsche anhat“.²³
 d. [Wenig] [mit den aktuellen Ereignissen] [im Zusammenhang] steht die Einstellung der Produktion bei der Montlinger Firma Mega-Stahl AG auf Ende November.²⁴

²⁰ Im Tiger-Handbuch wird zwischen Syntax-Bäumen und Syntax-Graphen unterschieden, wobei Syntax-Graphen im Gegensatz zu Bäumen diskontinuierliche Konstituenten zulassen. Syntax-Bäume sind unterschiedlich definiert worden, und ich zähle wie z. B. Ojeda (1988) auch Graphen mit kreuzenden Kanten zu den Bäumen.

²¹ taz, 07.11.1996, S. 20.

²² c't, 16/2002, S. 93.

²³ Vorarlberger Nachrichten, 03.03.1997, S. A5.

²⁴ St. Galler Tagblatt, 26.10.2001; Sparsam auf bessere Zeiten wartend.

Statt dessen wird nur verlangt, dass #vf2rightcorner vor #fin steht. Der ,*‘ hinter dem ,.‘ erlaubt einen beliebigen Abstand > 0.

Leider liefert die obige Anfrage auch Verbletztsätze, in denen zwei oder mehr Konstituenten vor dem finiten Verb stehen. Ein Beispiel zeigt (23):

- (23) kennt niemanden, [der] [nicht] glaubt, daß Perot als Präsident eine absolute Katastrophe wäre (s18)

Man muss also noch wie im Abschnitt 2 gezeigt ausschließen, dass Relativ- bzw. Interrogativpronomina oder subordinierende Konjunktionen am Anfang von #s stehen. Leider reicht das einfache Verbot von Relativ- bzw. Interrogativpronomina am Satzanfang nicht aus, da diese auch an zweiter Stelle stehen können, wenn sie wie in (24) in einer Präpositionalphrase eingebettet sind.

- (24) Aber in einer Zeit, in der keine Partei eine Mehrheit hat, [...] (s2854)

Zusätzlich zu der obigen Anfrage braucht man also noch die folgenden Zeilen:

```
#sleftcorner:[pos!= prorel]      &
#sleftcorner:[pos!= prointer]    &
#sleftcorner:[pos!=conjunction] &
#sleftcorner:[ ] . #sleftcorner_2:[pos != prorel] &
                               #sleftcorner_2:[pos != prointer] &
```

Diese Anfrage ist schon relativ kompliziert, wäre Information über topologische Felder im Korpus enthalten, könnte man sehr viel einfachere Anfragen stellen. Außer den unerwünschten Verbletztsätzen bekommt man aber noch Belege wie (25) mit Adverbien wie *indes*, *jedoch*, *wiederum*, die es ebenfalls auszuschließen gilt:

- (25) Hier wiederum mangelt es an Opferbereitschaft. (s3110)

Da es sich um eine relativ kleine Klasse von Adverbien handelt, die in dieser Konstellation vorkommen (Engel, 1977, S. 215), kann man sie durch Aufzählung ausschließen:

```
#vf2:[! (word= ("aber" | "also" | "auch" |
"allerdings" | "dagegen" |
"freilich" | "hingegen" |
"indes" | "jedenfalls" |
"jedoch" | "nämlich" |
"schließlich" | "wiederum" | "wohl"))]
```

Diese Beschränkung sagt, dass #vf2 keins der durch Disjunktion ,|‘ verknüpften Wörter sein darf. Diese Bedingung ist auch dann erfüllt, wenn #vf2 kein Wort sondern eine komplexe Konstituente ist.

Auch hier ist es nicht notwendig mit einer exakten Anfrage zu arbeiten und alle in dieser Position möglichen Adverbien aufzuzählen: Wenn keine uner-

wünschten Beispiele gefunden werden, muss man auch keine ausschließen. Es reicht also aus, die Wörter aufzuzählen, die im Korpus vorkommen.

Mit dieser Anfrage bekommt man die folgenden Belege:

- (26) a. [Besonders] [in Mitleidenschaft] geriet das Amt für „islamische Führung“, nach offiziellen Angaben gingen hier Tausende von Koranexemplaren in Flammen auf. (s362)
- b. [Ihm] [auf den Fersen] sind hierzulande die Haniel-Familie (6,4 Milliarden), die „Aldi“-Brüder Albrecht (5,1 Milliarden), die Waschmittel-Dynastie Henkel (4,3 Milliarden) und die Familie Quandt, die vor allem BMW ihr eigen nennt und auf vier Milliarden Dollar taxiert wird. (s2310)
- c. [Am schwersten] [mit der Selbstkritik] tat sich Jürgen Kocka. (s11838)
- d. [Negativ] [auf den Gewinn] wirkten sich vor allem Wechselkurschwankungen aus. (s28309)
- e. [Prinzipiell] [zu diesem Lager] rechnet sich auch Stuttgarts SPD-Umweltminister Harald B. Schäfer, der die Windenergie als Alternative zu Kernkraftnutzung forcieren will. (s28970)
- f. [Glanz] [in das einsame Tal] brachten einst die österreichische Kaiserin Sissi, seine k. u. k. Hoheit Thronfolger Franz Ferdinand und andere Vertreter des europäischen Hochadels. (s31606)
- g. [Neu] [in den SPD-Vorstand] kamen außer den beiden Jüngeren eine Vertreterin der Senioren-Arbeitsgemeinschaft „60 plus“, Liesel Schäfer, die hessische Ministerin Christine Hohmann-Dennhardt, die Berliner Senatorin Christine Bergmann, der Bonner Fraktionsgeschäftsführer Wolf-Michael Catenhusen, der nordrhein-wesfälische Minister Wolfgang Clement und der SPD-Fraktionschef im Schweriner Landtag, Gottfried Timm. (s33427)

Diese Sätze passen gut in das Muster der Daten, die in der 30-seitigen Materialsammlung (Müller, 2003) diskutiert werden. Für sinnvolle Generalisierungen in Bezug auf das Muster reicht die Anzahl der Belege jedoch nicht aus. Das Problem ist, dass die gegenwärtig verfügbaren annotierten Korpora nicht groß genug sind. Der Grund hierfür liegt darin, dass die Annotation sehr aufwendig und teuer ist. Es gibt zwar bereits große automatisch annotierte Korpora, doch sind diese für die gegenwärtige Frage nicht genau genug annotiert, da z. B. possessive Genitive nicht als Teil einer größeren Nominalphrase annotiert sind, sondern als eigenständige NP. Damit ist nicht mehr zu unterscheiden, ob zwei unabhängige NPen im Vorfeld stehen oder aber zwei NPen, die eigentlich zu einer intern komplexen NP gehören.

Bis sich die Menge der reich annotierten Sätze so erhöht hat, dass man auch die relativ seltenen Belege für die mehrfache Vorfeldbesetzung in ausreichender Zahl finden kann, wird noch einige Zeit ins Land gehen. Allgemein kann man sagen, dass bei sehr seltenen Phänomenen eventuell nicht einmal die Vergrößerung des Korpus etwas nützt (Zipfs Gesetz). In solchen Fällen bleibt also

nur, die Technik zu benutzen, die schon in den letzten Jahrhunderten benutzt wurde: aufmerksam lesen.

5. Zusammenfassung

In diesem Aufsatz habe ich gezeigt, dass uns unser sprachliches Wissen nicht immer zugänglich ist. In vielen Fällen lassen sich Behauptungen über Sprache mittels einfacher Suchanfragen im World Wide Web oder in großen unannotierten Korpora wie dem COSMAS-Korpus überprüfen. Zu den Belegen, die man dann findet, bekommt man automatisch den Kontext dazu, was insbesondere bei Phänomenen wichtig ist, die mit der Informationsstruktur interagieren.

Das Arbeiten mit unannotierten Korpora kann aber je nach Problemlage sehr ineffizient sein, da man unter Umständen Ergebnisse bekommt, die nicht genau dem entsprechen, was man sucht. Ich habe gezeigt, wie sich mit Baumbeschreibungssprachen einfach Korpusanfragen formulieren lassen, die innerhalb von Sekunden die gewünschten Ergebnisse liefern, falls das Phänomen im Korpus belegt ist. Die dritte Fallstudie hat gezeigt, dass dabei jedoch das Problem existiert, dass man seltene Phänomene unter Umständen nicht oder nicht ausreichend im Korpus findet. Es ist klar, dass man aus dem Nicht-Vorkommen von Mustern nicht vorschnell auf deren Unmöglichkeit schließen darf.

Literatur

- Baltin, Mark (1981): *Strict Bounding*. In: Baker, Carl Lee/McCarthy, John J. (Hg.): *The Logical Problem of Language Acquisition*, Cambridge: Massachusetts, London: England: The MIT Press.
- Baltin, Mark (2006): *Extraposition, the Right Roof Constraint, Result Clauses, Relative Clause Extraposition, and PP Extraposition*. Draft of 23.09.2001. In: *The Syntax Companion* (<http://www.uilots.let.uu.nl/syncom/>), Utrecht Institute of Linguistics, <http://www.nyu.edu/gsas/dept/lingu/people/faculty/baltin/papers/extrapos.pdf>. 16. 12. 2005.
- Barbiers, Sief/Bennis, Hans/Vogelaer, Gunther De/Devos, Magda/van der Ham, Margreet/Haslinger, Irene/van Koppen, Marjo/ Craenenbroeck, Jeroen Van/den Heede, Vicky Van (2005): *Syntactic Atlas of the Dutch Dialects (SAND): Volume I: Pronouns, Agreement and Dependencies*. Amsterdam: Amsterdam University Press.
- Belica, Cyril (2006): http://www.ids-mannheim.de/kl/projekte/cosmas_I/gesamtkonzept.html. 09.05.2006.
- Bennis, Hans (1991): *Theoretische Aspekten van Partikelvooropplaatsing*. In: *TABU Bulletin voor Taalwetenschap* 21(3), S. 89–95.
- Bierwisch, Manfred (1963): *Grammatik des deutschen Verbs*. studia grammatica II, Berlin: Akademie Verlag.
- Bodmer, Franck (2005): *COSMAS II. Recherchieren in den Korpora des IDS*. In: *Sprachreport* 3, S. 2–5.
- Brants, Sabine/Dipper, Stefanie/Eisenberg, Peter/Hansen-Schirra, Silvia/König, Esther/Lezius, Wolfgang/Rohrer, Christian/ Smith, George/Uszkoreit, Hans (2004): *TIGER*:

- Linguistic Interpretation of a German Corpus. In: *Research on Language and Computation* 2(4), S. 597–620.
- Brants, Sabine/Dipper, Stefanie/ Hansen, Silvia/Lezius, Wolfgang/Smith, George (2002): The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, www.bultreebank.org/proceedings/paper03.pdf.
- Bresnan, Joan/Cueni, Anna/Nikitina, Tatjana/Baayen, Harald (2005): Predicting the Dative Alternation. In *Knaw Academy Colloquium: Cognitive Foundations of Interpretation* 27–28. Oktober 2004, Amsterdam.
- Büring, Daniel/Hartmann, Katharina (2001): The Syntax and Semantics of Focus-Sensitive Particles in German. In: *Natural Language and Linguistic Theory* 19(2), S. 229–281. <http://www.linguistics.ucla.edu/people/buring/>.
- Bußmann, Hadumod (1983): *Lexikon der Sprachwissenschaft*. Kröners Taschenausgabe, Nr. 452, Stuttgart: Alfred Kröner Verlag.
- Bußmann, Hadumod (1990): *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kröner Verlag.
- Chomsky, Noam (1973): Conditions on Transformations. In: Anderson, Stephen R./Kiparski, Paul (Hg.): *A Festschrift for Morris Halle*. New York: Holt, Rinehart & Winston. S. 232–286.
- Chomsky, Noam (1986): *Barriers*, Band 13 von *Linguistic Inquiry Monographs*. Cambridge: Massachusetts, London: England: The MIT Press.
- Chomsky, Noam (1993): *Lectures on Government and Binding – The Pisa Lectures*. *Studies in Generative Grammar*, Nr. 9, Berlin/New York: Mouton de Gruyter. 7. Auflage.
- Clifton, Charles/Fanselow, Gisbert/Frazier, Lyn (2006): Amnestying Superiority Violations: Processing Multiple Questions. *Linguistic Inquiry* 37(1), S. 51–68.
- De Kuthy, Kordula (2001): Splitting PPs from NPs. In: Meurers, Walt Detmar/Kiss, Tibor (Hg.): *Constraint-Based Approaches to Germanic Syntax* In: *Studies in Constraint-Based Lexicalism*, Nr. 7, S. 31–76, Stanford: CSLI Publications, <http://cslipublications.stanford.edu/site/1575863049.html>. 18.08.2002.
- Duden (1991): *Die deutsche Rechtschreibung*. Band 1. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.
- Duden (1996): *Die deutsche Rechtschreibung*. Band 1. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.
- Duden (2005): *Grammatik der deutschen Gegenwartssprache*. Band 4. Mannheim, Leipzig, Wien, Zürich: Dudenverlag, 7. Auflage.
- Dürscheid, Christa (1989): *Zur Vorfeldbesetzung in deutschen Verbzweit-Strukturen*. FOKUS, Nr. 1. Trier: Wissenschaftlicher Verlag.
- Eisenberg, Peter (1994): *Grundriß der deutschen Grammatik*. Stuttgart, Weimar: Verlag J. B. Metzler, dritte Auflage.
- Eisenberg, Peter (1999): *Grundriß der deutschen Grammatik*. Band 2. Der Satz. Stuttgart, Weimar: Verlag J. B. Metzler.
- Engel, Ulrich (1970): *Regeln zur Wortstellung*. *Forschungsberichte des Instituts für deutsche Sprache* 5. Institut für deutsche Sprache, Mannheim.
- Engel, Ulrich (1977): *Syntax der deutschen Gegenwartssprache*. Band 22 von *Grundlagen der Germanistik*. Berlin: Erich Schmidt Verlag.
- Engel, Ulrich (1994): *Syntax der deutschen Gegenwartssprache*. Band 22 von *Grundlagen der Germanistik*. Berlin: Erich Schmidt Verlag. 3., völlig überarbeitete Auflage.
- Erdmann, Oskar (1886): *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung*, Band 1. Stuttgart: Verlag der J. G. Cotta'schen Buchhandlung, Neudruck: Hildesheim: Georg Olms Verlag, 1985.

- Eroms, Hans-Werner (2000): *Syntax der deutschen Sprache*. de Gruyter Studienbuch, Berlin: Walter de Gruyter Verlag.
- Fanselow, Gisbert (1987): *Konfigurationsalität*. Studien zur deutschen Grammatik Nr. 29. Tübingen: Gunter Narr Verlag.
- Fanselow, Gisbert (1991): *Minimale Syntax*. Groninger Arbeiten zur Germanistischen Linguistik 32.
- Fanselow, Gisbert (1993): Die Rückkehr der Basisgenerierer. In: *Groninger Arbeiten zur Germanistischen Linguistik* 36, S. 1–74.
- Fanselow, Gisbert (2001): Features, q-Roles, and Free Constituent Order. In: *Linguistic Inquiry* 32(3), S. 405–437.
- Fanselow, Gisbert (2004): Fakten, Fakten, Fakten! In: *Linguistische Berichte* 200. <http://www.ling.uni-potsdam.de/~fanselow/download/files/FanselowNeu.pdf>. 26.11.2004.
- Fanselow, Gisbert/Fery, Caroline/Schlesewsky, Matthias/Vogel, Ralf (Hg.) (2006): *Gradience in Grammar: Generative Perspectives*. Oxford: Oxford University Press.
- Featherston, Sam (2005): Universals and Grammaticality: Wh-constraints in German and English. In: *Linguistics* 4(43), S. 667–711.
- Geyken, Alexander (erscheint): *The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century*. In: Fellbaum, Christiane (Hg.): *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*, London: Continuum Press.
- Grewendorf, Günther (1988): *Aspekte der deutschen Syntax. Eine Rektions-Bindungs-Analyse*. Studien zur deutschen Grammatik, Nr. 33. Tübingen: Gunter Narr Verlag.
- Grewendorf, Günther (1989): *Ergativity in German*. Studies in Generative Grammar, Nr. 35. Dordrecht: Holland, Providence: U.S.A.: Foris Publications.
- Grewendorf, Günther (1990): Verb-Bewegung und Negation im Deutschen. In: *Groninger Arbeiten zur Germanistischen Linguistik* 30, S. 57–125.
- Grubac, Emilija (1965): *Untersuchungen zur Frage der Wortstellung in der deutschen Prosadichtung der letzten Jahrzehnte*. Dissertation, Philosophische Fakultät, Zagreb.
- Haegeman, Liliane (1994): *Introduction to Government and Binding Theory*. Blackwell Textbooks in Linguistics, Nr. 1, Oxford, UK/Cambridge, USA: Blackwell Publishers, zweite Auflage.
- Haider, Hubert (1982): *Dependenzen und Konfigurationen: Zur deutschen V-Projektion*. In: *Groninger Arbeiten zur Germanistischen Linguistik* 21, S. 1–60.
- Haider, Hubert (1990): Topicalization and other Puzzles of German Syntax. In: Grewendorf, Günther/Sternefeld, Wolfgang (Hg.): *Scrambling and Barriers*. In: *Linguistik Aktuell/Linguistics Today*, Nr. 5, S. 93–112. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Haider, Hubert (1991): *Fakultativ kohärente Infinitivkonstruktionen im Deutschen*. Arbeitspapiere des SFB 340 Nr. 17, IBM Deutschland GmbH, Heidelberg.
- Haider, Hubert (1993): *Deutsche Syntax – generativ*. Vorstudien zur Theorie einer projektiven Grammatik. *Tübinger Beiträge zur Linguistik*, Nr. 325. Tübingen: Gunter Narr Verlag.
- Haider, Hubert (1996): Downright Down to the Right. In: Lutz/Pafel (1996), S. 245–271.
- Haider, Hubert (1997a): Precedence among Predicates. In: *The Journal of Comparative Germanic Linguistics* 1, S. 2–41.
- Haider, Hubert (1997b): Projective Economy. On the Minimal Functional Structure of the German Clause. In: Abraham, Werner/van Gelderen, Elly (Hg.): *German: Syntactic Problems – Problematic Syntax*. *Linguistische Arbeiten*, Nr. 374. Tübingen: Max Niemeyer Verlag. S. 83–103.

- Haider, Hubert/Olsen, Susan/Vikner, Sten (1995): Introduction. In: Haider, Hubert/Olsen, Susan/Vikner Sten (Hg.): *Studies in Comparative Germanic Syntax*. Band 31 von *Studies in Natural Language and Linguistic Theory*. Dordrecht, Boston, London: Kluwer Academic Publishers. S. 1–45.
- Hoberg, Ursula (1997): Die Linearstruktur des Satzes. In: Eroms, Hans-Werner/Stickel, Gerhard/Zifonun, Gisela (Hg.): *Grammatik der deutschen Sprache*. Band 7.2 von *Schriften des Instituts für deutsche Sprache*. Berlin/New York: Walter de Gruyter. S. 1495–1680.
- Hoeksema, Jack (1991): Theoretische Aspekten van Partikelvooropplaatsing. In: *TABU Bulletin voor Taalwetenschap* 21(1), S. 18–26.
- Jacobs, Joachim (1986): The Syntax of Focus and Adverbials in German. In: Abraham, Werner/de Meij, S. (Hg.): *Topic, Focus, and Configurationality. Papers from the 6th Groningen Grammar Talks*, Groningen, 1984. In: *Linguistik Aktuell/Linguistics Today*, Nr. 4, S. 103–127, Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Kathol, Andreas (1996): Discontinuous Lexical Entries. In: Blache, Philippe (Hg.): *Third International Conference on HPSG–Abstracts*. 22–24 May 1996. Marseille. S. 43–44.
- Kaufmann, Ingrid (1995): Konzeptuelle Grundlagen semantischer Dekompositionsstrukturen. Die Kombinatorik lokaler Verben und prädikativer Elemente. *Linguistische Arbeiten*, Nr. 335. Tübingen: Max Niemeyer Verlag.
- Kepser, Stephan/Reis, Marga (Hg.) (2005): *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. Berlin/New York: Mouton de Gruyter.
- Kiss, Tibor (1994): Obligatory Coherence: The Structure of German Modal Verb Constructions. In: Nerbonne, John/Netter, Klaus/Pollard, Carl J. (Hg.): *German in Head-Driven Phrase Structure Grammar*. In: *CSLI Lecture Notes*, Nr. 46, S. 71–108. Stanford: CSLI Publications.
- Kiss, Tibor (2005): Semantic Constraints on Relative Clause Extraposition. In: *Natural Language and Linguistic Theory* 23(2), S. 281–334.
- Klenk, Ursula (2003): *Generative Syntax*. Narr Studienbücher. Tübingen: Gunter Narr Verlag.
- Lötscher, Andreas (1985): Syntaktische Bedingungen der Topikalisierung. In: *Deutsche Sprache* 13, S. 207–229.
- Lüdeling, Anke (1997): Strange Resultatives in German: New Evidence for a Semantic Treatment. In: Blight Ralph C./Moosally Michelle J. (Hg.): *Texas Linguistic Forum 38: The Syntax and Semantics of Predication. Proceedings of the 1997 Texas Linguistics Society Conference*. S. 223–233. Austin, Texas: University of Texas Department of Linguistics.
- Lutz, Uli/Pafel, Jürgen (Hg.) (1996): On Extraction and Extraposition in German. In: *Linguistik Aktuell/Linguistics Today*, Nr. 11. Amsterdam: Benjamins.
- Müller, Frank Henrik (2004a): *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen, <http://www.sfb441.uni-tuebingen.de/a1/Publikationen/stylebook-04.pdf>.
- Müller, Gereon (1998): *Incomplete Category Fronting. A Derivational Approach to Remnant Movement in German*. *Studies in Natural Language and Linguistic Theory*, Nr. 42. Dordrecht/Boston/London: Kluwer Academic Publishers.
- Müller, Stefan (1999): *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. *Linguistische Arbeiten*, Nr. 394. Tübingen: Max Niemeyer Verlag, <http://www.cl.uni-bremen.de/~stefan/Pub/hpsg.html>. 13.05.2006.

- Müller, Stefan (2002a): Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German. *Studies in Constraint-Based Lexicalism*, Nr. 13, Stanford: CSLI Publications, <http://www.cl.uni-bremen.de/~stefan/Pub/complex.html>. 13. 05. 2006.
- Müller, Stefan (2002b): Syntax or Morphology: German Particle Verbs Revisited. In: Dehé, Nicole/Jackendoff, Ray S./McIntyre, Andrew/Urban, Silke (Hg.): *Verb-Particle Explorations, Interface Explorations*, Nr. 1, S. 119–139, Berlin/New York: Mouton de Gruyter, <http://www.cl.uni-bremen.de/~stefan/Pub/syn-morph-part.html>. 13. 05. 2006.
- Müller, Stefan (2003): Mehrfache Vorfeldbesetzung. In: *Deutsche Sprache* 31, S. 29–62, <http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-ds.html>. 13. 05. 2006.
- Müller, Stefan (2004b): Complex NPs, Subjacency, and Extraposition. *Snippets* 8, 10–11, <http://www.cl.uni-bremen.de/~stefan/Pub/subjazen.html>. 13. 05. 2006.
- Müller, Stefan (2005): Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. In: *Linguistische Berichte* 203, S. 297–330, <http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-lb.html>. 13. 05. 2006.
- Neeleman, Ad/Weermann, Fred (1993): The Balance between Syntax and Morphology: Dutch Particles and Resultatives. In: *Natural Language and Linguistic Theory* 11, S. 433–475.
- Ojeda, Almerindo E. (1988): A Linear Precedence Account of Cross-Serial Dependencies. In: *Linguistics and Philosophy* 11, S. 457–492.
- Olsen, Susan (1997a): Prädikative Argumente syntaktischer und lexikalischer Köpfe: Zum Status der Partikelverben im Deutschen und Englischen. In: *Folia Linguistica* 31(3–4), S. 301–329.
- Olsen, Susan (1997b): Zur Kategorie Verbpartikel. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 119, S. 1–32.
- Oppenrieder, Wilhelm (1991): Von Subjekten, Sätzen und Subjektsätzen. *Linguistische Arbeiten*, Nr. 241, Tübingen: Max Niemeyer Verlag.
- Paul, Hermann (1919): *Deutsche Grammatik. Teil IV: Syntax*, Band 3. Halle an der Saale: Max Niemeyer Verlag, 2. unveränderte Auflage 1968, Tübingen: Max Niemeyer Verlag.
- Reis, Marga (1976): Reflexivierungen in deutschen A.c.I-Konstruktionen. Ein transformationsgrammatisches Dilemma. In: *Papiere zur Linguistik* 9, S. 5–82.
- Rohrer, Christian (1996): Fakultativ kohärente Infinitkonstruktionen im Deutschen und deren Behandlung in der Lexikalisch Funktionalen Grammatik. In: Harras, Gisela/Bierwisch, Manfred (Hg.): *Wenn die Semantik arbeitet*. Klaus Baumgärtner zum 65. Geburtstag. Tübingen: Max Niemeyer Verlag. S. 89–108.
- Ross, John Robert (1967): *Constraints on Variables in Syntax*. Dissertation, MIT, reproduced by the Indiana University Linguistics Club.
- Sabel, Joachim (1999): Das Passiv im Deutschen. Derivationale Ökonomie vs. optionale Bewegung. In: *Linguistische Berichte* 177, S. 87–112.
- Schütze, Carson T. (1996): *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago/London: The University of Chicago Press.
- Sorace, Antonella/Keller, Frank (2005): Gradience in Linguistic Data. In: *Lingua* 115(11), S. 1497–1524.
- Steyer, Katrin (2003): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hg.): *Wortverbindungen – mehr oder weniger fest*, *Jahrbücher des Instituts für Deutsche Sprache*. Berlin/New York: Walter de Gruyter. S. 87–116.
- Stiebels, Barbara (1996): Lexikalische Argumente und Adjunkte: Zum semantischen Beitrag verbaler Präfixe und Partikeln. *studia grammatica XXXIX*. Berlin: Akademie Verlag.

- Stiebels, Barbara/Wunderlich, Dieter (1992): A Lexical Account of Complex Verbs. Arbeiten des SFB 282 Nr. 30. Seminar für Allgemeine Sprachwissenschaft. Universität Düsseldorf.
- Ule, Tylman (2004): Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen, <http://www.sfs.uni-tuebingen.de/tuppp/dz/markupmanual.pdf>.
- Uszkoreit, Hans/Brants, Thorsten/Duchier, Denys/Krenn, Brigitte/Konieczny, Lars/Oepen, Stephan/Skut, Wojciech (1998): Studien zur performanzorientierten Linguistik: Aspekte der Relativsatzextraposition im Deutschen. Kognitionswissenschaft. Themenheft Ressourcenbeschränkungen 7(3), S. 129–133.
- van de Velde, Marc (1978): Zur mehrfachen Vorfeldbesetzung im Deutschen. In: Conte, Maria-Elisabeth/Giacalone Ramat, Anna/Ramat, Paolo (Hg.): Wortstellung und Bedeutung: Akten des 12. Linguistischen Kolloquiums, Pavia 1977. Linguistische Arbeiten, Nr. 61. Tübingen: Max Niemeyer Verlag. S. 131–141.
- von Stechow, Arnim/Sternefeld, Wolfgang (1988): Bausteine syntaktischen Wissens. Ein Lehrbuch der Generativen Grammatik. Opladen/Wiesbaden: Westdeutscher Verlag.
- Webelhuth, Gert/Ackerman, Farrell (1999): A Lexical-Functional Analysis of Predicate Topicalization in German. In: American Journal of Germanic Linguistics and Literatures 11(1), S. 1–61.
- Wunderlich, Dieter (1984): Zur Syntax der Präpositionalphrase im Deutschen. In: Zeitschrift für Sprachwissenschaft 3(1), S. 65–99.
- Zifonun, Gisela (1999): Wenn mit alleine im Mittelfeld erscheint: Verbpartikeln und ihre Doppelgänger im Deutschen und Englischen. In: Wegener, Heide (Hg.): Deutsch kontrastiv. Typologisch-vergleichende Untersuchungen zur deutschen Grammatik. Tübingen: Stauffenburg Verlag. S. 211–234.

ULRIKE DEMSKE

Das Mercurius-Projekt: Eine Baumbank für das Frühneuhochdeutsche

Abstract

Wer sich mit grammatischen Phänomenen historischer Sprachstufen beschäftigt, kann seine empirischen Daten bekanntermaßen nicht auf der Grundlage von Sprecherurteilen gewinnen, sondern muss zunächst Korpusrecherchen betreiben. Die Größe des auszuwählenden Korpus ist sehr stark phänomenabhängig: So reicht es im Bereich der Syntax in der Regel nicht aus, kleinere Textausschnitte aus verschiedenen Textquellen zu einem Korpus zusammenzufügen, vielmehr müssen vollständige Texte nicht nur nach raum-zeitlichen Koordinaten, sondern auch textsortenabhängig ausgewählt werden, um ein repräsentatives Korpus für eine spezifische Sprachstufe zu erstellen. Da eine manuelle Sichtung dieser doch recht großen Korpora sich sehr zeitaufwändig gestaltet, bietet sich gerade im Bereich der historischen Syntax der Einsatz von syntaktisch annotierten, digitalen Korpora an. Im folgenden Beitrag wird der Aufbau einer solchen Baumbank für das Frühneuhochdeutsche einschließlich der verfügbaren Recherchemöglichkeiten vorgestellt.

1. Einleitung

Die rasch zunehmende Verfügbarkeit digitalisierter Texte prägt seit einer Reihe von Jahren maßgeblich den methodischen Zugang in ganz unterschiedlichen Wissenschaftsdisziplinen, darunter auch den zu sprachwissenschaftlichen Fragestellungen. Die sprachwissenschaftliche Nutzung von Korpora erstreckt sich vor allem auf lexikologische Unternehmungen wie die Erstellung von Wörterbüchern oder von Kollokationslisten, wie sie beispielsweise durch die am Institut für Deutsche Sprache entwickelten Suchwerkzeuge von ‚Cosmas‘ und den dort verfügbaren Korpora ermöglicht werden. Erst in jüngerer Zeit erlauben Entwicklungen in der Computerlinguistik eine weiter gehende linguistische Interpretation größerer Korpora und damit die elektronische Suche nicht mehr allein nach Wörtern oder Wortketten, sondern nach syntaktischen Strukturen. Es handelt sich hier um eine Form der Empirie, die von der primär gegenwartssprachlich ausgerichteten Syntaxforschung erst in den letzten Jahren wirklich ernst genommen wird. Diachron orientierte Forschung dagegen ist durch das Fehlen negativer Evidenz *per se* korpuslinguistisch fundiert, da für jede historisch ausgerichtete Fragestellung ein der Fragestellung angemessenes Textkorpus zusammengestellt und ausgewertet werden muss. So unterstreicht Ebert (1986, S. 23), dass jede Unter-

suchung von Sprachwandelphänomenen auf einem repräsentativen Korpus beruhen muss, das es erlaubt, den Einfluss von zeitlichen, geographischen, sozialen und stilistischen Faktoren auf den fraglichen Wandel zuverlässig einschätzen zu können. Es liegt auf der Hand, dass sich diese Forderung nach Repräsentativität¹ umso schwieriger erfüllen lässt, desto geringer die Auftretensdichte des zu untersuchenden Phänomens ist. Ein Beispiel möge dies veranschaulichen: Um die Flexionsmorphologie des Frühneuhochdeutschen in ihren wesentlichen Grundzügen zu beschreiben, war es nach Solms/Wegera (1998) ausreichend, unter Berücksichtigung von zeitlichen, geographischen und stilistischen Faktoren Texte nicht zwangsläufig in ihrer Gesamtheit auf die relevanten sprachlichen Eigenschaften hin auszuwerten, vielmehr konnte diese Auswertung auf Textausschnitte von etwa 12.000 laufenden Wortformen beschränkt werden. Aufgrund einer deutlich geringeren Belegdichte kann die Untersuchung von syntaktischem Wandel im Unterschied dazu jedoch ausschließlich von Volltexten ausgehen. Und hier erlaubt die zunehmende Digitalisierung historischer Texte die Sichtung weit größerer Datenmengen als es bislang durch die rein manuelle Auswertung möglich war. Solange die vorliegenden Korpora aber reine Textkorpora ohne Anreicherung durch syntaktische Informationen sind, ist für die historische Syntax nicht viel gewonnen. Eine Auswertung solcher Korpora für syntaktische Fragestellungen würde nur dann einen Vorteil gegenüber ausschließlich in Papierform vorliegenden Texten erbringen, wenn es um syntaktische Phänomene ginge, die zwingend einen sichtbaren Reflex in der Morphologie haben (etwa der Kasusmorphologie bei der Suche nach Argumentstrukturen). Das gilt jedoch nur für eine äußerst geringe Zahl von syntaktischen Phänomenen. Für die Mehrzahl setzt die Suche nach syntaktischen Mustern voraus, dass die digitalisierten Texte mit syntaktischen Informationen angereichert sind, wie das im Rahmen von sog. Baumbanken geleistet wird.² Für das Gegenwartsdeutsche sind mit ‚Negra‘ und ‚Tiger‘ solche Baumbanken vorgelegt worden (Skut et al. 1998, Brants et al. 1999, Brants et al. 2002). Für historische Sprachstufen des Deutschen bestehen bislang keine vergleichbaren Korpora. Hier soll das seit 2003 an der Universität des Saarlandes etablierte Mercurius-Projekt Abhilfe schaffen: Ziel ist der Aufbau einer Baumbank auf der Grundlage von Zeitungstexten der Frühen Neuzeit. Da Baumbanken Informationen

¹ Wie Biber (1993) ausführt, ist die Größe eines Korpus nur mittelbar ein Faktor für seine Repräsentativität. Repräsentativität bezieht sich vielmehr auf das Maß, zu dem ein Korpus den vollen Umfang von Varianten in einer Sprachgemeinschaft einschließt. Und diese Varianten können außer- oder innersprachlich bedingt sein. Vor der Zusammenstellung eines Korpus müssen deshalb Situationsparameter und linguistische Merkmale identifiziert werden, um dem Ziel der Repräsentativität eines Korpus möglichst nahe zu kommen; zum Aufbau strukturierter Korpora s. auch Wegera (2000). Von Lüdeling (in diesem Band) wird die Auffassung vertreten, dass von Repräsentativität nur bei Kenntnis der Grundgesamtheit gesprochen werden darf.

² Zur Einführung s. Lezius (2001).

auf unterschiedlichen Ebenen der syntaktischen Analyse enthalten, sind sie im Unterschied zu anderen Ergebnisdatenbanken nicht phänomenorientiert, sondern lassen sich mittels geeigneter Abfragen nach immer neuen syntaktischen Mustern durchsuchen.

Im vorliegenden Beitrag werde ich über den Aufbau der Mercurius-Baumbank berichten sowie auf einige Probleme eingehen, die sich bei der Anreicherung historischer Texte mit syntaktischen Informationen stellen. Abschließend werden beispielhaft die zur Verfügung stehenden Recherchemöglichkeiten vorgestellt.

2. Die Mercurius-Baumbank

2.1 Textgrundlage

Erster und namengebender Baustein der Baumbank ist ein Zeitungstext aus dem Jahr 1667, der wöchentlich in Hamburg erschienene ‚Nordische Mercurius‘, herausgegeben von Georg Greflinger.³ Folgende Gründe sprechen dafür, den Aufbau eines syntaktisch annotierten Korpus mit einem Text aus der Frühen Neuzeit zu beginnen: Ein solcher Text bietet sich für eine Erprobung der Annotationswerkzeuge schon deshalb an, weil wir es hier mit der neuhochdeutschen Standardsprache *in statu nascendi* zu tun haben, die zunächst eine ausschließlich geschriebene Varietät des Deutschen mit entsprechend komplexen syntaktischen Strukturen darstellt (Glaser 2003, von Polenz 1994). Die aus der Literatur hinreichend bekannte Komplexität von Satzstrukturen im Frühneuhochdeutschen (Admoni 1980, Ebert 1986) mag folgendes Beispiel aus dem Mercurius illustrieren, in dem sechs Teilsätze in einem Satzgefüge verschachtelt sind. Satzgefüge wie unter (1) stellen im Mercurius eher die Regel als die Ausnahme dar.

- (1) Von dem Ober-Ungar. General sind dieser Tagen Briefe eingekommen/wie daß die Türcken in Siebenbürgen 2. grosse Machten/ als eine bey dem Schloß Jenona/die andere aber bey der festen Insul Julia versamleten/und daß sie muthmassentlich des Vorhabens seyen/ entweder den Fürsten Abaffi/ weiln er die der Pforten placidire Contributiones nicht abführen kan/ absetzen/ oder aber etliche Berg-Städte/wo der beste Siebenbürgische Wein wächst/ in der Türcken devotion zu bringen/ (M 8.8)⁴

Es wird deshalb angenommen, dass bei der syntaktischen Anreicherung der Texte sehr schnell die möglichen Grenzen eines Annotationswerkzeugs zu Tage treten und an entsprechenden Modifikationen des Werkzeugs gearbeitet werden kann, die dann für alle Stufen der deutschen Sprachgeschichte nutzbar sind. Gleichzeitig muss berücksichtigt werden, dass die bestehenden Werkzeuge für das Gegenwartsdeutsche entwickelt worden sind. Aufgrund dessen

³ Für Details s. Gieseler/Kühnle-Xemaire (1995) sowie Böning (2005).

⁴ Angegeben ist jeweils die Seiten- und Zeilenzahl des Beleganfangs.

empfiehlt es sich aus rein technischen Gründen, mit der Annotation von Texten zu beginnen, die in ihrem Lexikon und ihrem grammatischen System noch eine starke Übereinstimmung mit dem Gegenwartsdeutschen aufweisen. Das wird im nächsten Abschnitt deutlich werden, wenn die Annotationschritte im Einzelnen vorgestellt und erläutert werden.

2.2 Annotation

Vor der syntaktischen Annotation der Texte steht ihre Digitalisierung: Im Fall des Mercurius wurde diese Digitalisierung auf der Grundlage des Jahrgangs 1667 vorgenommen.⁵ Einen typischen Ausschnitt aus der auf Korrespondenzen basierenden Ausgabe zeigt Abb. 1:

Frankreich.

Paris vom 28. Octob. Nachdem der Herr
Bischoff von Münster mit Cur Colln über der
Coadjutourschafft von Münster in Differentien
kommt/so haben Sie beyde ihre Gesandten hieher
geschickt/ umb daß unser König sich bey dem Pab-
ste wegen ihrer Interesse interponiren möchte.
Der Graf von Fürstenberg Bischoff zu Mez
negotiiert vor unsern König bey den Deutschen
Fürsten. Zu Toulouse arbeiten täglich 800. Mañ/
umb einen tieffen Haven daselbst zu haben / und
grosse Schiffe zu legen. Zu Carcassone wird an
grossen Bley Minen/so jüngst entdeckt worden/
gearbeitet/so soll auch grosse Arbeit an einer Kupf-
fer Mine geschehen / welche man beyde zu dieses
Reiches grossen Vortheil erachtet. Die recrui-
tation zur Verstärckung der alten Regimente
geschehen mit ungemeinem Eifer. Umb das Fest
S. Hubert wol zu feyren / geschicht grosse Zuri-
ftung/die Dames lassen sich alle/ gleich Amazones/

Franckreich.

Pariß vom 28. Octob. Nachdem der Herr
Bischoff von Münster mit Cur Colln über der
Coadjutourschafft von Münster in Differentien
kommt/ so haben Sie beyde ihre Gesandten hieher
geschickt/ umb daß unser König sich bey dem Pab-
ste wegen ihrer Interesse interponiren möchte.
Der Graf von Fürstenberg Bischoff zu Mez
negotiiert vor unsern König bey den Deutschen
Fürsten. Zu Toulouse arbeiten täglich 800. Mañ/
umb einen tieffen Haven daselbst zu haben/ und
grosse Schiffe zu legen. Zu Carcassone wird an
grossen Bley Minen/ so jüngst entdeckt worden/
gearbeitet/ so soll auch grosse Arbeit an: einer Kupf-
fer Mine geschehen/ welche man beyde zu dieses
Reiches grossen Vortheil erachtet. Die recrui-
tation zur Verstärckung der alten Regimente
geschehen mit ungemeinem Eifer. Umb das Fest
S. Hubert wol zu feyren/ geschicht grosse Zuri-
ftung/ die Dames lassen sich alle/ gleich Amazones/

Abb. 1: Digitalisierung von Textquellen

Der erste Schritt auf dem Weg zu einem syntaktisch annotierten Korpus umfasst die Zerlegung des Textes in Satzeinheiten, d. h. seine Tokenisierung. Bereits bei dieser Arbeitsaufgabe manifestiert sich ein entscheidender Unterschied zwischen den im Rahmen der Negra- und Tigerprojekte bearbeiteten Texten aus der Gegenwart einerseits und dem Text aus der Frühen Neuzeit andererseits: Während für die Texte aus dem heutigen Deutsch im Wesentlichen davon ausgegangen werden kann, dass Satzeinheiten durch Punkte abzugrenzen sind, markieren Satzzeichen im Frühneuhochdeutschen Satzgrenzen nur sehr unzuverlässig (Stolt 1990). Dieses Problem möchte ich an

⁵ Die Digitalisierung des Nordischen Mercurius war Teil des DFG-Projekts ‚Entstehung und Entwicklung der Zeitungssprache im 17. Jahrhundert‘, das in den Jahren 1987 bis 1992 unter der Leitung von Prof. Erich Straßner und Prof. Gerd Fritz an der Universität Tübingen durchgeführt wurde.

einem Textausschnitt illustrieren: Der folgende Ausschnitt enthält insgesamt drei Satzgefüge, von denen allein der Umfang des letzten Satzgefüges unstrittig ist. Der Textteil vor dem Absatz ist durch Punkte in zwei Satzeinheiten gegliedert: Der Beginn des ersten Satzes fällt mit dem Beginn des Textausschnitts zusammen (= *Es*), der zweite Satz beginnt laut Interpunktion und Majuskel mit dem Pronominaladverb *worüber*. Inhaltlich bilden beide Sätze eine Einheit. Unter syntaktischen Gesichtspunkten wäre eine Satzgrenze allerdings eher vor dem ebenfalls hervorgehobenen *Mons(ieur)* zu setzen, da hier mit einem V2-Satz eine syntaktisch unabhängige Einheit beginnt, wohingegen das mit einer Majuskel beginnende *worüber* einen V-letzt-Satz einleitet. Im Gegenwartsdeutschen sind diese weiterführenden Nebensätze abhängige Satzteile größerer Satzgefüge.

- (2) Es sind viel/ welche vermeynen/ daß man allhier nicht die beste Opinion gegen Franckreich habe/ aus Ursachen/ weil sie in der Levante die unsrige nicht zum besten meynen sollen **Mons.** d'Almeras soll damahls/ als die Veränderung in Portugall vorgegangen/ mit vielen Frantzösischen Schiffen auff der Revier von Lisabona gewesen seyn. **Worüber** auch seltsame Gedancken entstehen.
Heute hat die Königl. Societät die Überbringung des Blutes aus einem Schafe in einen Menschen experimentirt, welches wol succedirte vor die quantität von 12 à 14 Untzen. (M 813.27)

Die Schlussfolgerung, die meines Erachtens aus Textbeispielen wie (2) zu ziehen ist, lautet, dass die Tokenisierung historischer Texte anders als im Gegenwartsdeutschen nicht automatisch zu leisten ist, sondern entweder manuell oder halbautomatisch zu erfolgen hat. Auf jeden Fall müssen vor der Tokenisierung die einschlägigen (syntaktischen) Kriterien für Satzwertigkeit genauestens festgeschrieben werden (Demske-Neumann 1990, Demske et al. 2004).

Die nächsten Arbeitsschritte beinhalten die Anreicherung der Satztokens mit syntaktischen Informationen; im Einzelnen sind das die Wortartkodierung⁶ sowie die Auszeichnung von Phrasentypen (runde Umrahmung) und Funktionen von Phrasen (eckige Umrahmung). Wie das im Detail aussieht, wird im Folgenden am Beispiel der Analyse eines Teilsatzes aus dem Textausschnitt (2) illustriert.

Wie die Abbildung zeigt, sind die syntaktischen Strukturen sehr flach und einfach aufgebaut: Argumente und Modifikatoren werden nicht positionell, sondern über eine Funktionsbezeichnung unterschieden (OA = Akkusativobjekt, MO = Modifikator). Das erklärte Ziel besteht hier in der Vermeidung von Ambiguitäten beim Aufbau von Strukturen (Brants et al. 2002). Unbestreitbarer Vorteil des verwendeten Annotationsschemas ist die Möglichkeit, Abhängigkeitsverhältnisse zwischen Konstituenten auch durch sich kreuzende Äste im Baumdiagramm darzustellen. Damit kann dem häufigen Auftreten

⁶ Zur Kodierung der Wortarten verwendet „Annotate“ das Stuttgart/Tübinger Tagset (STTS) mit Modifikationen, s. Smith/ Eisenberg (2000).

von diskontinuierlichen Konstituenten im Frühneuhochdeutschen Rechnung getragen werden, wie hier am Beispiel eines extraponierten Relativsatzes gezeigt wird.

Abb. 3: Diskontinuierliche Konstituenten

Eine weitere Annotationsebene wird durch die sog. sekundären Kanten eingeführt, mit deren Hilfe die Zuordnung von Argumenten in Koordinationsstrukturen korrekt dargestellt werden kann. In einer früheren Version des Werkzeugs konnten Argumente nur mit dem nächstliegenden verbalen Kopf verknüpft werden; in der aktuellen Version können die genauen Abhängigkeitsverhältnisse durch sekundäre Kanten angezeigt werden, vgl. Abb. 4. Diese Annotationsebene ist für die Analyse der Satzstrukturen im ‚Nordischen Mercurius‘ schon deshalb besonders interessant, weil die Dichte von komplexen Koordinationsstrukturen sehr hoch ist. Für weitere technische Details der Annotation sei auf Brants et al. (2002) verwiesen.

Die Auszeichnung erfolgt interaktiv mit Hilfe des Werkzeugs ‚Annotate‘, einer graphischen Benutzeroberfläche, die für die Erstellung der Baumbanken ‚Negra‘ und ‚Tiger‘ entwickelt respektive weiterentwickelt worden ist. Interaktiv ist das Verfahren deshalb, weil auf der Grundlage von bereits annotiertem Material sukzessive Vorschläge für Wortarten, Phrasentypen und -funktionen gemacht werden, die der Annotator dann akzeptiert, modifiziert oder verwirft und einen eigenen Vorschlag zur Analyse von Wörtern oder Phrasen eingibt. Die Vorschläge für jede Annotationsebene beruhen auf statistischen Voraussagen für diese Ebene, einem Markov-Modell (Brants

3.1 Formulierung von Suchanfragen

Für Suchanfragen stehen zwei Anfragesprachen zur Verfügung: (i) eine logische Abfragesprache (beruhend auf booleschen Ausdrücken und Operationen) und (ii) eine graphische Abfragesprache. Wie diese beiden Abfragesprachen funktionieren, macht man sich am Besten an einem Beispiel klar, das ausreichend komplex ist. Ein solches Beispiel wäre etwa die Suche nach Nominalphrasen, die als Teilkonstituenten einen Artikel, das Kopfnomen und einen attributiven Satz enthalten („arity“ = 3). Zusätzliche Bedingung soll sein, dass der Attributivsatz diskontinuierlich auftritt. In der logischen Abfragesprache wäre dann wie folgt zu formulieren,

- (3) *NP: Artikel, Kopfnomen und Attributivsatz (diskontinuierlich)*
 #n1: [cat="NP"] > [cat="S"] &
 #n1 > [pos="NN" | pos="NE"] &
 #n1 > [pos="Art"] &
 arity(#n1,3) &
 discontinuous(#n1)

wobei '>' für die Dominanzrelation zwischen zwei Knoten steht und '|' alternativ einen Gattungsnamen oder einen Eigennamen als Kopf der Nominalphrase erlaubt. Suchabfragen können gespeichert werden, so dass die Ergebnisse der Suchanfrage problemlos reproduzierbar sind. In der graphischen Abfragesprache gestaltet sich die identische Anfrage so, dass Dominanz- und Präzedenzrelationen zwischen Knoten durch lineare Verbindungen einer bestimmten Art angezeigt werden:

Abb. 5: Graphische Suchanfrage

Es ist ohne weiteres möglich, graphische Suchanfragen in die logische Abfragesprache übersetzen zu lassen; umgekehrt funktioniert das allerdings nicht.

3.2 Präsentation der Ergebnisse

Die Ergebnisse der Suchanfrage werden unabhängig von der Abfragesprache in graphischer Form, d. h. Baumdiagrammen ausgegeben, zwischen denen hin- und hergeblättert werden kann. Die Baumdiagramme in der Ergebnisdatenbank können einzeln oder insgesamt exportiert und ausgedruckt werden. Grundsätzlich besteht bei der Darstellung der Baumdiagramme die Wahl, ob das gesamte Satzgefüge dargestellt werden soll oder ob nur der für die Suchanfrage relevante Ausschnitt gezeigt wird. Für unsere Suchanfrage ergeben sich somit zwei mögliche Darstellungen: In Abb. 6 findet sich ein Satzgefüge, in dem das gesuchte syntaktische Muster – bestehend aus einer Nominalphrase mit Artikel und diskontinuierlichem Attributivsatz – enthalten und farbig hervorgehoben ist. In Abb. 7 ist die Fundstelle reduziert auf das einschlägige syntaktische Muster, ohne dass sprachliches Material interveniert.

Die Ergebnisausgabe kann auch in Form von Tabellen geschehen, eine Form der Ausgabe, die besonders dann interessant ist, wenn Frequenzen eine Rolle spielen. Ganz abgesehen davon, dass für die Bestimmung der Produktivität eines syntaktischen Musters in einer gegebenen historischen Sprachstufe des Deutschen eine solche Abfragemöglichkeit unabdingbar ist, kann die Abfrage von Frequenzen auch dazu beitragen, die Qualität der syntaktischen Annotation zu verbessern und eine konsistente Annotation zu unterstützen. Das lässt sich etwa an der Analyse von Präpositionalphrasen demonstrieren, für die das Annotationsschema drei mögliche Funktionen vorsieht: (i) Adjunkt, (ii) Komplement und (iii) Teil eines Funktionsverbgefüges. Neben Kriterien, die für Sprecher des Gegenwartsdeutschen oft nur schwierig auf historische Texte anwendbar sind (mögliche Pronominalisierung, Erweiterbarkeit, Erfragbarkeit etc.), ist ein Kriterium wie die Reihenbildung sicher hilfreich bei der Abgrenzung der einzelnen Funktionen.

- (4) a. *Budoa und Castel novo nahe Ragusa sollen gantz in die See weggesuncken seyn.* (M 285.21)
 b. Die aus Engeland/ Franckreich und Holland melden von allen Seiten grosser Zurüstung/ ehist mit mächtigen Flooten wiederumb in die See zu kommen. (M 195.23)

Wie die nachstehende Tabelle zeigt, liegen bei den Verbindungen der Verben *gehen*, *kommen* und *bringen* mit der Präpositionalphrase *in die See* Reihenbildungen vor, während *wegsinken* wie die Verben *laufen* und *begeben* nur einmal in Verbindung mit der fraglichen Präpositionalphrase erscheint. Eine Analyse der Präpositionalphrase als Modifikator im ersten Fall sowie als Teil eines Funktionsverbgefüges im zweiten Fall (in dem das Verb vor allem aspektuelle und Aktionsartmerkmale trägt, etwa im Sinne von ‚lossegeln‘),

kann deshalb durch eine statistische Auswertung unterstützt werden. Solche Auswertungen liefern so die Grundlage für Listen von Funktionsverbgefügen (oder von Verben mit präpositionalen Objekten), die die Arbeit der Annotatoren weiter erleichtern können.

Abb. 6: Graphische Darstellung der Ergebnisse

Abb. 7: Graphische Darstellung der Ergebnisse

Frequency	Feature 1	Feature 2	Feature 3	Feature 4
	#t1_word	#t2_word	#t3_word	#t4_word
10	in	die	See	gehen
5	in	die	See	kommen
3	in	die	See	bringen
1	in	die	See	laufen
1	in	die	See	begeben
1	in	die	See	wegsinken

Tab. 1: Häufigkeitsverteilung der Präpositionalphrase *in die See* + Verb

4. Zusammenfassung und Ausblick

Im vorliegenden Beitrag habe ich über die Mercurius-Baumbank berichtet, ein mit syntaktischen Informationen angereichertes Zeitungskorpus aus dem späten 17. Jahrhundert. Die individuellen Ebenen der Annotation wurden beschrieben (Wortarten, Typen und Funktionen von Phrasen), Abgrenzungsprobleme von Sätzen, wie sie typischerweise in historischen Texten auftreten, thematisiert. Im letzten Abschnitt wurden schließlich die Rechercheoptionen vorgestellt, die zukünftig dazu beitragen sollen, dass in digitalisierten Texten nicht nur nach morphologischen Markern, sondern auch nach syntaktischen Mustern gesucht werden kann. Es ist offensichtlich, dass die vorliegende Baumbank die eingangs formulierten Kriterien der Repräsentativität für ein historisches Textkorpus bislang nicht erfüllt. Allerdings ist hervorzuheben, dass die Zeitungen der Frühen Neuzeit zumindest hinsichtlich der Text-

sortenabdeckung eine angemessenere Datengrundlage liefern als andere Textquellen, weil sie eine Reihe ganz unterschiedlicher Textsorten wie Berichte, Reportagen, Dokumente, Briefe etc. umfassen.⁷ Dass die Textmenge, die der Mercurius-Baumbank zugrunde liegt, unbedingt der Erweiterung bedarf, um wirklich verlässliche Aussagen über die Produktivität und Veränderungen in der Produktivität syntaktischer Muster zu erhalten, versteht sich von selbst. Deshalb ist vorgesehen, baldmöglichst mit der Annotation weiterer Zeitungstexte zu beginnen. Darüber hinaus sollen die momentan abfragbaren Informationen um solche zu morphologischen Merkmalen und Lemmatisierung ergänzt werden. Auch hierzu ist bereits ein entsprechendes Werkzeug entwickelt worden („TigerMorph“). Die Dringlichkeit dieser Forschungsaufgabe liegt auf der Hand. Das ist andernorts längst erkannt worden, weshalb beispielsweise zum Englischen entsprechende Korpora für einzelne historische Sprachstufen bereits vorliegen (Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English, Penn-Helsinki Parsed Corpus of Middle English).

Literatur

Quellen

[M] = Mercurius 1667. Nordischer Mercurius. Welcher kürztlich vorstellet/was in diesem 1667. Jahre an Novellen aus Europa einkommen ist. Hamburg 1667.

Sekundärliteratur

- Admoni, Vladimir G. (1980): Zur Ausbildung der Norm der deutschen Literatursprache im Bereich des neuhochdeutschen Satzgefüges (1470–1730). Ein Beitrag zur Geschichte des Gestaltungssystems der deutschen Sprache. Berlin. (Bausteine 56/IV).
- Biber, Douglas (1993): Representativeness in corpus design. In: *Literary and Linguistic Computing* 8, S. 243–257.
- Böning, Holger (2005): Wie man Leser glücklich macht. In: *Die Zeit* Nr. 25, S. 88.
- Brants, Sabine/Dipper, Stefanie/Hansen, Silvia/Lezius, Wolfgang/Smith, George (2002): The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories* Sozopol/Bulgarien. S. 24–41.
- Brants, Thorsten (1999): Tagging and Parsing with Cascaded Markov Models – Automation of Corpus Annotation. Saarbrücken: DFKI, Saarbrücken Dissertations in Computational Linguistics and Language Technology Bd. 6.
- Brants, Thorsten/Skut, Wojciech/Uszkoreit, Hans (1999): Syntactic Annotation of a German Newspaper Corpus. In: *Proceedings of the ATALA Treebank Workshop*. Paris, S. 69–76.
- Demske, Ulrike (1996): Bestandsaufnahme zum Untersuchungsbereich „Syntax“. In: Fritz, Gerd/Straßner, Erich (Hg.) (1996): *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*. Tübingen. S. 70–125.
- Demske, Ulrike/ Frank, Nicola/Laufer, Stefanie/Stierner, Hendrik (2004): Syntactic Interpretation of an Early New High German Corpus. In: Kübler, Sandra/Nivre, Joakim/

⁷ Vgl. dazu vor allem Schröder (1995).

- Hinrichs, Erhard/Wunsch, Holger (eds.) (2004): *Proceedings of the Third Workshop on Treebanks and Linguistics Theories (TLT 2004)*. Tübingen. S. 175–182.
- Demske-Neumann, Ulrike (1990): Charakteristische Strukturen von Satzgefügen in den Zeitungen des 17. Jahrhunderts. In: Betten, Anne (Hg.) (1990): *Neuere Forschungen zur historischen Syntax des Deutschen*. Tübingen. S. 239–252.
- Ebert, Robert P. (1986): *Historische Syntax des Deutschen II: 1300–1750*. Bern.
- Gieseler, Jens/Kühnle-Xemaire, Elke (1995): Der Nordische Mercurius – eine besondere Zeitung des 17. Jahrhunderts? Eine sprachwissenschaftliche Untersuchung der Hamburger Zeitung. In: *Publizistik* 40, 2.
- Glaser, Elvira (2003): Zu Entstehung und Charakter der neuhochdeutschen Schriftsprache: Theorie und Empirie. In: Berthele, Raphael/Christen, Helen/Germann, Sybille/Hove, Ingrid (Hg.) (2003): *Die deutsche Schriftsprache und die Regionen: Entstehungsgeschichtliche Fragen in neuer Sicht*. Berlin. S. 57–78.
- König, Esther/Lezius, Wolfgang/Voormann, Holger (2003): *TigerSearch User's Manual*. University of Stuttgart.
- Lezius, Wolfgang (2001): Baumbanken. In: Carstensen, Kai-Uwe/Ebert, Christian/Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (Hg.) (2001): *Computeringuistik und Sprachtechnologie – eine Einführung*. Heidelberg. S. 377–385.
- Lezius, Wolfgang (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Dissertation. Universität Stuttgart. (Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung 8,4).
- von Polenz, Peter (1994): *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart*. Bd. II: 17. und 18. Jahrhundert. Berlin.
- Schröder, Thomas (1995): *Die ersten Zeitungen. Textgestaltung und Nachrichtenauswahl*. Tübingen.
- Skut, Wojciech/Brants, Thorsten/Krenn, Brigitte/Uszkoreit, Hans (1998): A Linguistically Interpreted Corpus of German Newspaper Text. In: *Proceedings of the Conference on Language Resources and Evaluation LREC-98*. Granada. S. 705–711.
- Smith, George/Eisenberg, Peter (2000): *Kommentare zur Verwendung des STTS im NEGRA-Korpus*. Manuskript. Universität Potsdam.
- Solms, Hans-Joachim/Wegera, Klaus-Peter (1998): Das Bonner Frühneuhochdeutsch-korpus. Rückblick und Perspektiven. In: Bergmann, Rolf (Hg.) (1998): *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung am 1. und 2. November 1996*. Stuttgart. S. 22–39.
- Stolt, Birgit (1990): Redeglieder, Informationseinheiten: Cola und Commata in Luthers Syntax. In: Betten, Anne (Hg.) (1990): *Neuere Forschungen zur historischen Syntax des Deutschen*. Tübingen. S. 379–392.
- Wegera, Klaus-Peter (2000): Grundlagenprobleme einer mittelhochdeutschen Grammatik. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hg.) (2000): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. Berlin. S. 1304–1320.

ANNETTE KLOSA

Korpusgestützte Lexikographie: besser, schneller, umfangreicher?

Abstract

In diesem Beitrag geht es einerseits um eine Definition dessen, was korpusgestützte Lexikographie ist, und andererseits um eine Bestandsaufnahme der gegenwärtigen Praxis korpusgestützter Lexikographie. Dabei wird ein Schwerpunkt gelegt auf allgemeinsprachige Wörterbücher der Gegenwartssprache, deren Inhalt die Beschreibung von Bedeutung und Verwendung von Lexemen ist. Außerdem liegt die Einschätzung zugrunde, dass die Auswertung elektronischer Korpora die Wörterbucharbeit weitgehend positiv beeinflusst und verändert, vorausgesetzt, dass zugrunde gelegte Korpus wurde für das geplante Wörterbuch so gut wie möglich in Umfang und Zusammensetzung eingerichtet.

John Sinclair, ein erfahrener Lexikograph und Korpuslinguist, hat vor kurzem bekannt, dass er zu Beginn seines COBUILD-Wörterbuch-Projektes im Jahr 1980 angenommen hatte, dass der Einsatz eines Korpus Genauigkeit und Reichhaltigkeit verbessern und den Prozess der Lexikographie [...] beschleunigen würde. Sinclair stellt weiter fest, dass sich manches davon bewahrheitete, er aber vor allem den Effekt der neuen Informationen, die das Korpus lieferte, grob unterschätzt habe:

At the start of the Cobuild project in 1980 I assumed that the use of a corpus would improve accuracy and comprehensiveness, and it would speed up the process of lexicography because of the clarity of the descriptions and the organising power of the computer. Some of this proved to be correct, but I grossly underestimated the effect of the new information that the corpus supplied, and in particular the total lack of fit between the evidence coming from the corpus and the accepted categories of English lexicography. (Sinclair 2004, S. 9)

Vor diesem Hintergrund ist zu fragen: Ist korpusgestützte Lexikographie wirklich besser, schneller, umfangreicher, in dem Sinn, dass sie schneller zu besseren und umfangreicheren Ergebnissen kommt, wobei „besser“ als wissenschaftlich angemessener zu verstehen ist? Bei der Beantwortung dieser Fragen gehe ich von meinen praktischen Erfahrungen als Lexikographin aus und zeige Beispiele aus der Wörterbucharbeit, um zu einer Bestandsaufnahme der gegenwärtigen Praxis korpusgestützter deutschsprachiger Lexikographie zu kommen. Dabei wird auch zu überlegen sein, ob es für die Lexikographie tatsächlich „heute zum Standard [gehört], mit Unterstützung durch große maschinenlesbare Korpora zu arbeiten – allerdings, bedingt durch die verfü-

baren Korpora, praktisch exklusiv mit schriftlichen Texten“, wie es im Ausschreibungstext für die Jahrestagung 2006 des IDS formuliert wurde. Bevor jedoch die Wertungsfragen beantwortet werden können, muss im ersten Teil des Beitrags differenzierter dargestellt und abgegrenzt werden, was hier unter korpusgestützter Lexikographie verstanden wird.

1. Was heißt korpusgestützte Lexikographie?

a) Lexikographie und (Sprach-)Wörterbücher

Mit Wiegand (1998) kann man ein Sprachwörterbuch (im Folgenden immer nur noch: Wörterbuch) als ein „Nachschlagewerk“ verstehen, „dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann“ (Wiegand 1998, S. 58). Dabei stehen hier Nachschlagewerke für menschliche Benutzer im Zentrum der Betrachtung. Offen bleibt noch die Frage, vor allem im Bereich der korpusgestützten Lexikographie, ob die Daten automatisch oder von Menschen erstellt sind.

Hier sollte man (mit Müller-Spitzer 2003) differenzieren zwischen automatisch erstellten Wortschatzinformationssystemen bzw. lexikographisch bearbeiteten Wörterbüchern. Unter lexikographischer Bearbeitung wird dabei „jede Art der reflektierten menschlichen Bearbeitung der automatisch erstellten Daten verstanden, vom Überprüfen über das Umsortieren bis hin zum Kommentieren“ (Müller-Spitzer 2003, S. 150). Solche lexikographisch bearbeiteten, gedruckten oder elektronischen Wörterbücher stehen hier im Vordergrund. Automatisch erstellte Wortschatzinformationssysteme werden aber auch kurz erwähnt, weil sie erst im Zusammenhang mit dem fortschreitenden Ausbau umfangreicher elektronischer Textsammlungen und der Entwicklung korpus technologischer und informationstechnologischer Verfahren möglich wurden bzw. werden.

b) Korpus

Von den verschiedenen möglichen und vorgeschlagenen Definitionen für „Korpus“ soll hier Folgende gelten: Ein Korpus ist eine Sammlung von Texten, von denen man annimmt, sie sei repräsentativ¹ für eine bestimmte Sprache. Diese Sammlung wurde zum Zweck der linguistischen Analyse zusammengestellt. Daneben gelten außerdem die Annahmen, dass es sich bei den Texten um natürlichsprachliche Äußerungen (häufig in schriftlicher, aber auch in mündlicher) Form handelt und dass die Sammlung nach bestimmten Auswahlkriterien für einen bestimmten Zweck kompiliert wurde. Dabei wird der Anspruch erhoben, dass diese Sammlung authentische Sprache repräsentiert (vgl. Tognini Bonelli 2001, S. 2).

¹ „Repräsentativität“ wird hier nicht im statistischen Sinne verwendet.

Wichtig ist außerdem, dass heute mit Korpus häufig implizit eine in elektronischer Form zur Verfügung stehende Textsammlung gemeint ist, die mithilfe von elaborierter Recherche- und Analysesoftware erschlossen werden kann. Zurzeit darf man für lexikographische Zwecke von einer solchen Software mindestens folgende Funktionalitäten erwarten (im Einzelnen vgl. hierzu Sinclair 1991, S. 21–37):

- Die Software kann zwischen Lemma und Wortformen unterscheiden; sie verfügt über ein Lemmatisierungsprogramm.
- Die Software erstellt Frequenzlisten zu Lemmata oder Wortformen der unterschiedlichsten Art (z. B. alphabetisch oder nach Häufigkeit sortiert).
- Die Software bietet Fundstellenkonkordanzen in Form von Key-Word-in-Context-Zeilen (= KWICs) an, die je nach Bedarf nach unterschiedlichen Kriterien definiert und sortiert werden können (z. B. alphabetisch nach dem folgenden oder dem vorausgehenden Wort).
- Die Software bietet eine Kookkurrenzanalyse an, die auf der Basis mathematisch-statistischer Verfahren das Aufdecken von signifikanten Regelmäßigkeiten bei der Verwendung von Wortkombinationen ermöglicht, wie sie beispielsweise am IDS entwickelt wurde (vgl. Belica 1995 und <http://www.ids-mannheim.de/kt/misc/tutorial.html>).

Im Folgenden wird unter Korpus – zumindest im lexikographischen Kontext – ein elektronisches Korpus verstanden. Im lexikographischen Zusammenhang ist außerdem wichtig, dass im Idealfall für ein bestimmtes Wörterbuch gezielt ein eigenes Korpus zusammengestellt wird (vgl. Landau 2001, S. 323).

c) Andere Wörterbuchquellen

Neben einem elektronischen Korpus können auch andere primäre Quellen im Sinne von Wiegand (1998, S. 140) einem Wörterbuch zugrunde liegen, und zwar traditionellerweise Belegzettel. Diese sind nach bestimmten Vorgaben angefertigte Exzerpte aus Texten; diese Texte als Materialbasis für ein Wörterbuch werden wiederum ebenfalls nach bestimmten Kriterien (z. B. Ausgewogenheit hinsichtlich der Textsorten) zusammengestellt. Alle Belegzettel sind in einem Belegzettellarchiv bzw. lexikographischen Zettelarchiv zusammengefasst. Beispiel hierfür ist etwa das Belegarchiv des „Deutschen Fremdwörterbuchs“ von Schulz/Basler, das am IDS neu bearbeitet wird:

Das DFWB beruht, vor allem um der Darstellung seiner Wort- und Bedeutungsgeschichten ein tragfähiges Fundament zu geben, auf einer breiten Material- und Quellenbasis. Den entsprechenden Grundstock auch für die Neubearbeitung bildet das Schulz/Baslische Quellen- und Belegmaterial: Die Sammlung, die ca. 2 Mio. Belegzettel umfasst, beruht in erster Linie auf einer systematischen Exzerption von mehr als 10000 gedruckten Quellen, die sich auf den ganzen neuhochdeutschen Zeitraum von etwa 1450 bis 1980/90 und auf ein breit gefächertes Textsortenspektrum erstrecken. (<http://www.ids-mannheim.de/lexik/fremdwort/quellen.html>)

Ein solches Belegzettellarchiv kann neben der klassischen Zettelform außerdem in digitaler Form fortgeführt werden, wie dies z. B. in der Dudenredaktion geschieht:

Auf ganz traditionelle Weise ‚durchkämmen‘ wir Texte außerdem per Hand, besser gesagt mit dem Auge. Diese recht zeitintensive Tätigkeit, an der wir wegen der Qualität der Ergebnisse seit Jahrzehnten festhalten, übernehmen erfahrene Sprachbeobachter für uns. [...] Die auf diese Art vorselektierten Fundstellen – einzelne Sätze oder kürzere Textpassagen – werden in die Duden-Sprachkartei aufgenommen. Sie ist inzwischen auf über 3 Millionen Belege angewachsen und wird seit 1998 elektronisch geführt. (<http://www.duden.de>)

Außerdem können Zettelbelegarchive auch retrodigitalisiert werden, was aber z. B. weder für die Duden-Sprachkartei noch für andere Belegzettellarchive realisiert wurde². Zwar erhöht eine digitale Form eines Belegarchivs den Nutzwert, indem beispielsweise Volltextsuchen über alle Belegtexte möglich werden oder einzelne Belege verschiedenen Lemmata zugewiesen werden können. Doch ist ein elektronisches Belegarchiv offensichtlich nicht das Gleiche wie ein Korpus im eben definierten Sinn. Belegtexte sind für die linguistische Analyse nur noch bedingt repräsentativ, auch wenn die zugrunde liegende Textauswahl dies war, weil menschliche Exzerptoren eher dazu neigen, das Ungewöhnliche als das Normale auf Belegzetteln festzuhalten.

Bei der Zusammenstellung der primären Quellen für ein Wörterbuch ist das Problem der geeigneten Auswahl sowohl für Belegarchive wie für Korpora zu berücksichtigen; beide sollen „ein zuverlässiges Abbild des Lexembestands und Lexemgebrauchs der Objektsprache bilden“ (Schlaefer 2002, S. 104) – und das ist nicht immer leicht zu erreichen. In der deutschsprachigen Lexikographie werden zurzeit verschiedene Varianten der Korpuszusammenstellung speziell für lexikographische Zwecke neben der immer noch andauernden Pflege existierender Belegarchive praktiziert:

- Die Anlage eines eigenen lexikographischen Korpus für genau ein Wörterbuch, z. B. für das „Mittelhochdeutsche Wörterbuch“, das in Trier und Göttingen neu erarbeitet wird:

Den Kern seiner Quellenbasis bildet ein Corpus von philologisch gesicherten Texten aller Textsorten der Periode. Auf der Grundlage dieses Quellenkorpus ist ein digitales Textarchiv angelegt und aus diesem durch computergestützte Lemmatisierung ein digitales Belegarchiv erstellt worden, das bei der Ausarbeitung des Wörterbuches zugrunde gelegt und durch Nachexzerption ständig ergänzt wird. (http://www.adwmainz.de/2005/index.php?sektion=vorh_gsk&ID=38)

² Auch nicht für das „Deutsche Rechtswörterbuch“, wie Schlaefer (2002, S. 107) schreibt; hierzu liegt eine anders lautende E-Mail-Auskunft von Ingrid Lemberg vom 15. 02. 2005 vor.

- Die Anlage eines lexikographischen Korpus als primäre Quelle für verschiedene lexikographische Produkte aus einer Bearbeitungsstelle, z. B. für die Wörterbücher aus der Dudenredaktion:

Wir bauen in der Dudenredaktion gerade ein eigenes elektronisches Korpus auf, das genau auf unsere Zwecke zugeschnitten sein wird: das ‚Duden-Korpus‘. Es wird zunächst ca. 500 Millionen Wortformen umfassen und aus unterschiedlichsten Textsorten (Romanen, Zeitungsartikeln, Gebrauchsanleitungen etc.) zusammengesetzt sein. Darüber hinaus ist es ‚annotiert‘, was bedeutet, dass jede der 500 Millionen Wortformen mit besonderen sprachlichen Informationen angereichert ist. (<http://www.duden.de>)

- Die Zusammenstellung eines virtuellen lexikographischen Korpus aus einem größeren, allgemeinen Korpus, z. B. für das Projekt *lexiko* am IDS in Mannheim:

Um für die Auswertung sprachlicher Daten eine gute empirische Basis zugrunde legen zu können, wurde nach formalen und inhaltlichen Kriterien aus den gesamten IDS-Korpora ein umfangreiches digitales Textkorpus zusammengestellt. Dieses Korpus ist ein dynamisches Korpus (ein sogenanntes Monitorkorpus), welches regelmäßig erweitert und aktualisiert wird, um die jeweils neuesten Entwicklungen verfolgen und damit aktuelle Beschreibungen liefern zu können. Derzeit (Dezember 2005) umfasst es ca. 1,3 Milliarden Textwörter. (<http://www.lexiko.de/Korpusbasiertheit.html>)

Neben Korpora und Belegarchiven als primäre Quellen greifen Lexikographen bei ihrer Arbeit auch auf elektronische Textsammlungen im weiteren Sinn zu. Dies können z. B. Gesamtausgaben bestimmter Autoren auf CD-ROM, digitale Zeitungsarchive auf CD-ROM und nicht zuletzt auch Texte im Internet sein, die mithilfe einer Suchmaschine erschlossen werden. Solch ein Verfahren wird etwa im Folgenden beschrieben:

Zusätzlich zur Recherche im Duden-Korpus suchen wir punktuell auch in anderen elektronischen Quellen nach neuen oder bislang noch nicht verzeichneten Wörtern. Allen voran ist hier natürlich das Internet zu nennen, aber auch diverse Wirtschaftsdatenbanken oder Korpora anderer Institute. (<http://www.duden.de>)

Bei solchen Textsammlungen und insbesondere dem Internet handelt es sich aber nicht um Korpora, denn sie genügen wichtigen Kriterien nicht: Sie wurden nicht zum Zweck linguistischer Analyse nach bestimmten Kriterien zusammengestellt und sie sind nicht mithilfe von elaborierter Recherche- und Analysesoftware auswertbar. Wie Belica/Perkuhn (2006, S. 4) erklären, haben Internetsuchanfragen auch eine andere und „ganz spezielle Funktion“: Sie dienen nämlich ausschließlich dem Aufspüren von Dokumenten im Internet. Allerdings kann man natürlich aus solchen elektronischen Textsammlungen, z. B. dem Internet, Korpora zusammenstellen (vgl. hierzu z. B. Thelwall 2005). Werden solche elektronischen Textsammlungen bei der lexikographischen Arbeit benutzt, um beispielsweise gezielt nach Belegen für neue Wörter zu suchen, dann stellen sie nur eine Form von primärer Quelle dar mit dem Nachteil, nicht unbedingt in die ursprüngliche Quellenkonzeption für das Wörterbuch eingebunden zu sein.

Überhaupt ist zumindest für die deutsche allgemeinsprachige Lexikographie der Gegenwart festzustellen, dass (noch) überwiegend eine Kombination verschiedener primärer Quellen als Wörterbuchbasis dient, wie etwa das Wörterbuch „Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen“ zeigt: Primärquellen dieses Wörterbuches sind einerseits das virtuelle, so genannte neo-Korpus, das aus den IDS-Korpora der geschriebenen Sprache zusammengestellt wurde, und andererseits eine projekteigene Wortkartei mit durch Exzerption gewonnenen Belegen „aus den verschiedensten gedruckten Texten, [daneben] Hörbelege aus Fernseh- und Rundfunksendungen des Erfassungszeitraumes sowie Internetbelege“ (Herberg/Kinne/Steffens 2004, S. XVI).

Das Projekt *lexiko*, das ebenfalls am IDS beheimatet ist, verfolgt dagegen den Ansatz, ein Wörterbuch auf der Basis eines eigenen Korpus vollständig neu zu erarbeiten – ein Ansatz, der in der englischsprachigen Lexikographie schon 1987 mit dem „Collins COBUILD English Language Dictionary“ von John Sinclair und seinem Team realisiert wurde. Dieses Wörterbuch wird deshalb wohl zu Recht als „ein erstes beeindruckendes Ergebnis strikt korpusbasierter Lexikographie“ (Engelberg/Lemnitzer 2001, S. 206) bezeichnet. Für die englischsprachige Lexikographie ist mit Sicherheit Hunston (2002, S. 96) zuzustimmen, dass Korpora das Schreiben von Wörterbüchern in solchem Ausmaß revolutioniert haben, dass es großen Verlagen heutzutage praktisch unmöglich ist, beispielsweise ein Lernerwörterbuch zu veröffentlichen, das nicht beanspruchen kann, auf einem Korpus zu basieren:

Corpora have so revolutionised the writing of dictionaries and grammar books for language learners (or rather, for learners of English) that it is by now virtually unheard-of for a large publishing company to produce a learner's dictionary or grammar reference book that does not claim to be based on a corpus. As a result, this is probably the application of corpora that is most far-reaching and influential, in that even people who have never heard of a corpus are using the product of corpus investigation. (Hunston 2002, S. 96)

d) korpusgestützt – korpusgebunden

Was heißt korpusgebunden? Und was ist der Unterschied zu korpusgestützter Lexikographie? Ein korpusgebundenes Wörterbuch wird ausschließlich auf der Basis des Wörterbuchkorpus und ohne Hinzuziehung anderer primärer Quellen, aber auch ohne Hinzuziehung sekundärer und/oder tertiärer Quellen erarbeitet. Solche sekundären Quellen sind nach Wiegand (1998, S. 140) „alle Wörterbücher, die nach dem Instruktionsbuch entweder obligatorisch oder fakultativ [als Hilfsmittel bei der lexikographischen Arbeit] konsultiert werden sollen, und zu den tertiären Quellen gehören alle sonstigen Sprachmaterialien, die benutzt werden sollen, wie z. B. linguistische Monographien und Grammatiken [...]“. Ein korpusgebundenes Wörterbuch bildet genau die Sprachwirklichkeit ab, die das zugrunde gelegte Korpus repräsentiert – aber nicht mehr und auch nicht weniger. Dies wäre z. B. bei einem Autorenwörterbuch oder einem Wörterbuch zu einem bestimmten Text der Fall.

In der korpusgestützten Lexikographie, die sich auch ausschließlich auf ein Wörterbuchkorpus als primäre Quelle stützt, ist es dagegen möglich, neben der Auswertung des Wörterbuchkorpus noch sekundäre und/oder tertiäre Quellen hinzuzuziehen. Dieses Vorgehen wird beispielsweise an der Beschreibung der Praxis der Lesartendisambiguierung im „Cambridge International Dictionary of English“ (kurz: CIDE) deutlich:

Lexicographers considered how to split up a polysemic word among a number of guide-words, using their own understanding and a study of other dictionaries, and testing this understanding against the picture emerging from the data. Through using the corpus, CIDE [= Cambridge International Dictionary of English] lexicographers often found that previous dictionaries defined quite rare senses of words but missed important, common ones. (Baugh/Harley/Jellis 1996, S. 41)

Ein korpusgestütztes Wörterbuch bietet also keine 1:1-Abbildung der Sprachwirklichkeit des zugrunde gelegten Korpus, sondern ergänzt, wo nötig, das Bild, das aus der Korpusanalyse gewonnen wurde. Mit korpusgestützter Lexikographie soll aber noch mehr gemeint sein: Hier genügt es nicht, dass ein Korpus (im oben definierten Sinne einer speziell zusammengestellten Sammlung elektronisch aufbereiteter Texte) als primäre Quelle zugrunde liegt, sondern dieses muss mithilfe korpuslinguistischer Verfahren erschlossen werden. Insofern ist etwa das oben erwähnte „Mittelhochdeutsche Wörterbuch“ kein Vertreter korpusgestützter Lexikographie, denn dort wird aus dem Korpus unter Einsatz eines Lemmatisierungstools ein elektronisches Belegarchiv erstellt, das wiederum die Grundlage für die Artikelbearbeitung ist. Andere korpuslinguistische Verfahren wie z. B. Kontextanalysen kommen aber (noch) nicht zum Einsatz. Der Unterschied zwischen korpusgestützt erarbeiteten Wörterbüchern und nicht korpusgestützt erarbeiteten Wörterbüchern ist also hauptsächlich ein methodischer.

Zurzeit wird in der mit Korpora arbeitenden Lexikographie in verschiedenen Ländern allgemein dem korpusgestützten Vorgehen gegenüber dem korpusgebundenen Vorgehen der Vorzug gegeben. Als Beispiel hierfür können folgende Zitate, die sich auf das „Longman Dictionary of Contemporary English“ in der dritten Auflage von 1995 beziehen, gelten:

The corpus is a massively powerful resource to aid the lexicographer, which must be used judiciously. Our aim at Longman is to be corpus-based, rather than corpus-bound. (Summers 1996, S. 262)

[...] in dictionary making editorial judgment is of paramount importance, because blindly following the corpus, no matter how carefully it may be constructed to represent the target language type accurately, can lead to oddities. (Summers 1996, S. 266)

e) korpusvalidierend – korpusgesteuert

Innerhalb der korpusgestützten Lexikographie selbst können wiederum zwei unterschiedliche Ansätze verfolgt werden: der korpusvalidierende (engl. *corpus based*) Ansatz und der korpusgesteuerte (engl. *corpus driven*) Ansatz.

Unter *korpusvalidierend* versteht man eine Methode, die von einem Korpus Gebrauch macht, um Theorien und Beschreibungen darzulegen, zu testen oder zu veranschaulichen:

[...] the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study. (Tognini Bonelli 2001, S. 65)

Im Gegensatz dazu wird beim korpusgesteuerten Ansatz ein Korpus ganz anders als für die Suche nach Beispielen, die bestimmte linguistische Argumente unterstützen oder theoretische Annahmen validieren sollen, genutzt³:

In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back preexisting theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. (Tognini Bonelli 2001, S. 84)

An Beispielen aus der Lexikographie sollen die unterschiedlichen Ansätze im Folgenden deutlich werden. Als Vertreter eines korpusvalidierenden Ansatzes kann das schon zitierte „Cambridge International Dictionary of English“ gelten. Im oben genannten Zitat von Baugh/Harley/Jellis (1996, S. 41) wird das Vorgehen deutlich: Etwas überspitzt formuliert geht der Erarbeitungsweg von Introspektion über Konsultation von Sekundärquellen schließlich hin zum Vergleich mit der Korpusauswertung. Oder, in einer Formulierung zu dem Wahrig-Wörterbuch „Fehlerfreies und gutes Deutsch“ von 2003:

Empirisch abgesichert sind alle Schreibungen und Wortverwendungen durch einen Abgleich [Unterstreichung hinzugefügt] mit dem WAHRIG Textkorpusdigital, einer rd. 500 Mio. Wörter umfassenden, digitalen Dokumentation der deutschen Gegenwartssprache, die den aktuellen Sprachgebrauch authentisch widerspiegelt. (Wahrig 2003, S. VII)

Als Beispiel für den überwiegenden Einsatz des korpusgesteuerten Verfahrens in der Lexikographie kann das ebenfalls schon genannte „Collins COBUILD

³ In Steyer (2004, S. 93) werden für die beiden genannten empirischen Prinzipien die Termini Konsultationsparadigma (für den korpusvalidierenden Ansatz) und Analyseparadigma (für den korpusgesteuerten Ansatz) vorgeschlagen.

English Language Dictionary“ gelten. Das Lexikographenteam schätzt seine Arbeit als Antwort auf die Frage „Was zeichnet wahre Korpuslexikographie aus?“ folgendermaßen ein:

What characterizes true corpus lexicography? Each entry in a dictionary represents a detailed examination of the corpus evidence. [...] The entry reflects what is found – and what our users are likely to find in the real text world – rather than what was believed to be the case. (Clear u. a. 1996, S. 308f.)

An einem Beispiel aus der lexikographischen Arbeit im Projekt *lexiko* soll deutlich werden, wie korpusvalidierender und korpusgesteuerter Ansatz in korpusgestützter Lexikographie sinnvoll verbunden werden können. Durch den Einsatz der statistischen Kookkurrenzanalyse⁴ werden im Projekt korpusgesteuert erste Kandidatenwörter für paradigmatische Relationen gewonnen, ohne dass sich die Lexikographen auf die eigene Intuition stützen. Wichtig ist, dass dies nicht bedeutet, „dass man auf eine lexikografische Auswertung der dort extrahierten Ergebnisse verzichten darf, denn Sinnrelationen müssen identifiziert, bestätigt, zugeordnet und illustriert werden“ (Storjohann 2005 a, S. 254).

Mithilfe des korpusgesteuerten Ansatzes kommt es nämlich nicht immer zu einer umfassenden Beschreibung sinnverwandter Wörter, sodass in einem zweiten Schritt sekundäre Quellen, z. B. Synonymwörterbücher, konsultiert werden, in denen weitere Kandidatenwörter zu finden sind. Diese werden dann gezielt im Korpus gesucht, das damit nicht Ausgangspunkt der lexikographischen Beschreibung ist, sondern der Rückprüfung dient. Die mithilfe dieses korpusvalidierenden Ansatzes gefundenen paradigmatischen Partnerwörter müssen allerdings den gleichen Aufnahmekriterien für den Wortartikel genügen (Frequenz, Streuung über mehrere Quellen und mehrere Jahrgänge) wie die mithilfe des korpusgesteuerten Ansatzes gewonnenen Partner.

Kurz zusammengefasst kann man sagen: Korpusgestützte Lexikographie erarbeitet Wörterbücher auf der Grundlage elektronischer, gezielt zusammengestellter Textsammlungen, die (bezogen auf die Gegenwartssprache) authentische Sprache repräsentieren sollen und die mithilfe geeigneter Recherche- und Analysesoftware erschlossen und ausgewertet werden. Alle Daten werden redaktionell auch unter Einbezug sekundärer und/oder tertiärer Quellen geprüft und bewertet, wobei verschiedene Wörterbücher entweder vom Korpus selbst ausgehen (korpusgesteuert) oder das Korpus zur Rückprüfung benutzen (korpusvalidierend) oder beide Methoden kombinieren. Die meisten der bisher gezeigten Beispiele waren gegenwartssprachliche Wörterbücher, deren Gegenstand zumeist Bedeutungs- und Verwendungsbeschreibung ist. Korpusgestützte Lexikographie hat sich genau in diesem Bereich, besonders aber auch für Lernerwörterbücher entwickelt. Eine Aus-

⁴ Am IDS durch Cyril Belica entwickelt und seit 1995 als Teil der Korpusrecherche- und Analysewerkzeuge der COSMAS-Plattform angeboten.

weitung der Methoden z. B. auf fachsprachliche oder anderssprachliche Wörterbücher, auf dialektale oder regionale Wörterbücher, oder auf Antonymwörterbücher usw. steht weitgehend noch aus.

2. Wie wirkt sich Korpusgestützte auf Wörterbücher aus? – Zwei Beispiele

In ihrer Darstellung zum „Cambridge International Dictionary of English“ beschreiben Baugh/Harley/Jellis (1996) detailliert, wie sie Korpusdaten sowie Korpusrecherche- und -analysemethoden eingesetzt haben, um dieses Wörterbuch zu erarbeiten. Sie gehen dabei auf die Entscheidung über die Aufnahme von Stichwörtern, die Lesartendisambiguierung, die Arbeit mit Belegen, die Angabe von Kollokationen, die Erarbeitung von grammatischen und morphologischen Angaben und weiteres mehr ein. Von diesen Punkten wird hier nur die besonders relevante Frage des Stichwortansatzes und der Stichwortliste sowie die Angabe von Wortverbindungen bzw. der Einbezug des Kontextes berücksichtigt. Dabei soll auch überlegt werden, ob die Benutzung von Korpora diese Arbeitsvorgänge tatsächlich schneller und die Ergebnisse besser und umfangreicher macht.

a) Die Frage des Stichwortansatzes und der Stichwortliste

Korpusauswertung spielt bei der Frage des Stichwortansatzes eine Rolle und ist außerdem bei der Frage der Lemmaliste in zweierlei Hinsicht relevant: Für schon bestehende Wörterbücher kann für Überarbeitungen und Neuauflagen eine systematische Korpusauswertung dabei helfen, Kandidaten für neu aufzunehmende Stichwörter zu ermitteln und so genannte Wörterbuchleichen aufzuspüren. Für neu zu erstellende Wörterbücher können vollständig neue Lemmalisten durch systematische Korpusauswertungen erarbeitet werden.

Bevor überhaupt eine Stichwortliste erarbeitet werden kann, ist aber zu klären, in welcher Form die im Wörterbuch zu beschreibenden Wörter in diese Liste aufgenommen werden sollen. Traditionellerweise geschieht dies in der so genannten Nennform, Grundform oder Zitierform. In Textkorpora kommen solche Grundformen allerdings in den seltensten Fällen vor, sondern in Texten liegen flektierbare Wörter natürlich in verschiedenen Wortformen vor, die entweder intellektuell oder unter Einsatz eines Lemmatisierungsprogrammes auf eine Wortform zusammengeführt werden können. John Sinclair nennt hierzu die Schwierigkeiten:

Lemmatization looks fairly straightforward, but is actually a matter of subjective judgement by the researcher. There are thousands of decisions to be taken. Also, it is not yet understood how meanings are distributed among forms of a lemma, and a new branch of study is looming – the interrelationships of a lemma and its forms. (Sinclair 1991, S. 41)

Sinclair überlegt weiter, ob es tatsächlich so sein muss, dass die Grundform eines Wortes als Stichwort im Wörterbuch dient, selbst wenn diese Grund-

form in einem Korpus fast nicht belegt ist. Könnte nicht stattdessen die am meisten belegte Form das Lemma werden?:

Another interesting question in this area is how to decide what the physical form of a lemma should be. Traditionally, the 'base', or uninflected, form is used even when that form is hardly ever found on its own, or hardly ever found at all. But a case could be made for any of a number of alternatives, for example, that the most frequently-encountered form should be used for the lemma; and the first-stage evidence from the computer can provide a good basis for planning new methods of access to the word-forms for the language. (Sinclair 1991, S. 42)

Aber selbst das COBUILD-Wörterbuch, oben schon als das erste tatsächlich korpusgestützte Wörterbuch bezeichnet, geht diesen Weg nicht, sondern bleibt bei der traditionellen Form des Stichwortansatzes nach der Grundform. Dies hängt zum einen sicherlich mit der Frage der Benutzerfreundlichkeit zusammen, aber auch mit den allgemein üblichen Schritten, wie Stichwortlisten erstellt werden.

Bei der vollständig neuen Erarbeitung einer Stichwortliste für ein Wörterbuchprojekt kann man entweder „bereits vorhandene Stichwortlisten auswählen, auswerten und schließlich mehr oder minder modifiziert übernehmen“ oder „man kann sie auf der Basis von Korpora komplett neu erstellen“ (Schnörch 2005, S. 73). Für das Projekt *lexiko*, das hier als Beispiel für korpusgestützte Lexikographie in diesem Bereich dienen soll, wurde der zweite Weg gewählt, dabei aber zusätzlich der Abgleich mit anderen Stichwortlisten als Ergänzung und Korrektiv eingesetzt. So ist zunächst und in kurzer Zeit eine Liste an Stichwortkandidaten entstanden, die mit ca. 320.000 Einträgen sehr umfangreich war⁵. Die redaktionelle Bearbeitung dieser Stichwortkandidatenliste hin zu einer wirklichen Stichwortliste dauerte dagegen viel länger; in Summe waren damit drei Lexikographen jeweils mehrere Monate beschäftigt, und noch immer überprüfen studentische Hilfskräfte einzelne unklare Stichwortkandidaten.

Warum ist aber eine redaktionelle Überprüfung der Stichwortkandidatenliste überhaupt nötig? Zunächst einmal entspricht solch ein redaktioneller Überarbeitungsgang grundsätzlich dem korpusgesteuerten Verfahren, bei dem ja der Lexikograph das Bild, das sich ihm aus dem Korpus darstellt, betrachtet, analysiert, bewertet, sortiert und schließlich beschreibt. So wurde in *lexiko* also jeder mithilfe der Korpustechnologie vorgeschlagene Stichwortkandidat angesehen, nach bestimmten Kriterien (wie „richtig“, „unklar“, „falsch“) bewertet, in verschiedene Teilmengen einsortiert und dann in Form der öffentlich zugänglichen Stichwortliste⁶ von *lexiko* dokumentiert. Doch

⁵ Natürlich muss die korpusgestützte Erstellung einer Stichwortliste nicht zu solch einem Umfang führen, denn auch aus einem viele Millionen Wortformen enthaltenden Wörterbuchkorpus kann prinzipiell eine deutlich kleinere Stichwortliste ermittelt werden, wenn man etwa andere Frequenzen als Kriterien vorgibt.

⁶ Im Internet zu erreichen unter <http://www.lexiko.de>.

ist dies nicht nur wegen der angewandten Methode nötig gewesen, sondern auch weil die Zusammenführung verschiedener Wortformen auf eine Grund-, Nenn- bzw. Zitierform mithilfe eines automatischen Lemmatisierers natürlich nicht immer fehlerfrei funktioniert und teilweise nach lexikographischen Gesichtspunkten revidiert werden muss⁷.

Als Fazit ist festzuhalten: Es geht korpusgestützt schnell, die Stichwortkandidaten ermitteln zu lassen, und man kann sehr umfangreiche Kandidatenlisten erzeugen, ob aber die lexikographische Überprüfung dieser Kandidaten und die Erarbeitung der wirklichen Stichwortliste schneller geht als etwa das Kompilieren aus schon existierenden Wörterbüchern, ist schwer zu beurteilen. Außerdem ist zu berücksichtigen, dass zumindest dann, wenn einem Wörterbuch ein dynamisches Korpus zugrunde liegt, die Stichwortliste in regelmäßigen Abständen aktualisiert werden muss. Ob Stichwortlisten besser werden, wenn sie korpusgestützt erarbeitet sind, ist ebenfalls differenziert zu beantworten. Sie bilden zunächst einmal direkt das Korpus ab, und wenn dieses für den Wörterbuchzweck gut zusammengestellt wurde, dann ist auch die Stichwortliste gut. Wenn aber das zugrunde gelegte Korpus nicht wirklich für den Wörterbuchzweck geeignet ist, muss die Stichwortliste u. U. redaktionell ergänzt werden, damit sie gut wird. Werden Stichwortlisten aus anderen Listen kompiliert, stellt sich natürlich die Frage der Auswahlkriterien und der Vollständigkeit – eine gute und aktuelle Stichwortliste zu erhalten, ist mit dieser Methode keineswegs leichter.

Die Ergänzung und Überprüfung einer schon erarbeiteten Stichwortliste ist der dritte Punkt, der hier betrachtet werden soll. Die Unterstützung für den Lexikographen durch Korpusauswertung ist hierbei wirklich enorm:

Sprache bleibt bekanntlich nicht stehen; neue Wörter kommen hinzu, alte erhalten neue Bedeutungen. Was früher in langer, ebenso akribischer wie mühsamer Handarbeit einzeln zusammengesucht werden musste, kann nun schnell und systematisch auf der Grundlage eines großen Korpus der deutschen Gegenwartssprache ermittelt werden. (<http://www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html>)

Tatsächlich bedeutet der Einsatz von Korpora in diesem Bereich eine Verbesserung hinsichtlich der Aktualität. Dann kann gelten, was für das COBUILD-Projekt schon 1987 formuliert wurde: „The main criterion for inclusion or exclusion of headwords and senses was the strength of the corpus evidence“ (Sinclair 1987, S. 65).

Auf der anderen Seite bedeutet es redaktionellen Aufwand, die aus dem Korpus generierten Vorschläge für Neuaufnahme- oder Streichkandidaten zu überprüfen. Aus den etwa 5.000 „interessanten neuen Wörtern“, die aus dem zugrunde gelegten Korpus für das Wahrig-Wörterbuch der „Deutschen Rechtschreibung“ als Kandidaten ermittelt wurden, sind beispielsweise dann doch nur etwa 1.100 tatsächlich in die Neuauflage des genannten Wörter-

⁷ Zu einzelnen Beispielen für Fehllemmatisierungen vgl. Schnörch (2005, S. 77f.).

buches eingegangen⁸ – und solch ein Arbeitsgang kostet (wenn auch lohnende) Zeit. Denn nicht immer sind Wörter, die als sehr frequent auffallen, auch wirklich für ein Wörterbuch interessant, und nicht immer muss ein niedrig frequentes Wort tatsächlich aus dem Wörterbuch gestrichen werden. Im Einzelfall hängt dies sehr vom einzelnen Wörterbuchgegenstand und -ziel ab. Kriterien für den Ausschluss eines Stichwortes trotz hoher Frequenz können z. B. die fehlende Streuung über verschiedene Quellen oder über längere Zeiträume sein, aber auch das jeweilige Register, die Varietät oder die Periode, denen das Wort zuzuordnen ist (vgl. hierzu Baugh/Harley/Jellis 1996, S. 49).

Als Fazit für die Überarbeitung bestehender Stichwortlisten ist zu ziehen: Systematische Korpusauswertung vereinfacht und beschleunigt die Ermittlung von Aufnahme- oder Streichkandidaten. Neben der Frequenz können weitere Korpusdaten (z. B. Zahl der Quellen oder Jahrgänge, in denen das Stichwort belegt ist) dem Lexikographen bei der Entscheidung über Aufnahme oder Streichung helfen – ihm die Entscheidung abnehmen können sie aber nicht. Trotzdem ist dieses korpusgesteuerte Verfahren besser als eine rein redaktionelle Belegsuche nach zufällig aus anderen Quellen, z. B. bei der eigenen Lektüre einer Tageszeitung, entdeckten Aufnahmekandidaten. Durch den Einsatz des korpusgesteuerten Verfahrens nimmt der Umfang an korrigierenden Eingriffen in eine Stichwortliste eher zu, und zumindest in der Verlagslexikographie wird mit jeder Neuaufnahme auch geworben:

Mit rund 5000 neu aufgenommenen Wörtern, wie beispielsweise Billigflieger, Dosenpfand, Fotohandy, Genmais, Ich-AG, LAN-Party, Minijob und Sars, ist das Wörterverzeichnis auf den aktuellen Stand gebracht. (Duden 2004, S. 5)

Offensichtlich ist die Zeit gekommen, in der für viele Wörterbücher dem korpusgesteuerten Verfahren bei der Überarbeitung und Aktualisierung der Stichwortliste der Vorzug gegeben wird.

b) Wortverbindungen und Kontext

Korpusgestützte Lexikographie ist in besonderer Weise durch den Einbezug des Kontextes und die Berücksichtigung von Wortverbindungen gekennzeichnet, wie bereits Teubert (1999, S. 312) beobachtet:

Längst hat mit der Verfügbarkeit größerer Korpora und fortgeschrittener Korpussoftware die Korpuslinguistik auch hierzulande Eingang in die allgemeinsprachige Lexikographie gefunden. Man wendet sich größeren Bedeutungseinheiten zu und bezieht vermehrt den Kontext in Bedeutungsbeschreibungen ein. Mittlerweile setzt sich die Überzeugung durch, dass die nächste Wörterbuchgeneration, die einsprachige ebenso wie die zweisprachige, zumindest korpusvalidiert, wenn nicht korpusbasiert sein muss. (Teubert 1999, S. 312)

In der dritten Auflage des „Longman Dictionary of Contemporary English“ von 1995 tauchen erstmals so genannte „lexical units“ mit eigenen Bedeu-

⁸ Vgl. <http://www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html>.

tungspunkten unter einem Einwort-Stichwort auf, wofür Summers (1996, S. 263) den Wörterbucheintrag *to be liable* als Beispiel bietet:

1 *be liable to do something* to be likely to do or say something or to behave in a particular way, especially because of a fault or natural tendency: *The car is liable to overheat on long trips.* **2** [not before the noun] legally responsible for the cost of something: [+for] *Tenants have legal liability for any damage they cause.* **3** likely to be affected by a particular kind of problem, illness etc: [+to] *You're more liable to injury when you don't get regular exercise.* [...] (Summers 1996, S. 263)

Summers erklärt, dass „Phraseologie“ ein wichtiges Merkmal für die Neuauflage des Wörterbuches werden sollte, weil fortgeschrittene Lerner des Englischen als Zielgruppe galten. Allerdings hatte die Redaktion nicht damit gerechnet, wie weitreichend diese Entscheidung sich erweisen würde – für das Aussehen des Wörterbuches, für den Aufbau der Einträge und für die Form der Bedeutungserläuterungen selbst (vgl. Summers 1996, S. 262). Zumindest wird diese Entscheidung der Tatsache gerecht, dass Wörter in Texten eben nicht isoliert, sondern mit anderen Wörtern zusammen auftreten.

Häufig treten aus Korpora gerade mehr oder weniger feste Wortverbindungen hervor. Für Lexikographen stellt sich natürlich die Frage, wie mit ihnen umzugehen ist. Wie im Falle des „Longman Dictionary“ können sie in Einzelwortartikel integriert werden. Dieses Vorgehen entspricht dabei durchaus dem korpusgesteuerten Verfahren, weil ja durch die Korpusanalyse die Wortverbindung zunächst als solche hervortritt und erst dann der Lexikograph die Entscheidung trifft, ob und wie sie im Wörterbuch zu behandeln ist. Dieses Vorgehen bleibt aber natürlich ein lexikographisches. Anders ist der korpuslinguistische Ansatz, der nämlich solche Wortverbindungen gar nicht mehr im Rahmen eines am einzelnen Wort interessierten Wörterbuches beschreiben will, sondern ganz herauslöst aus diesem Kontext, wie dies am IDS etwa im Projekt „Usuelle Wortverbindungen“ geschieht:

Im Projekt UWW werden rekurrente Muster des Sprachgebrauchs, so wie sie sich in Kookkurrenzprofilen vor allem im hochfrequenten Bereich manifestieren, aus den IDS-Korpora herausgefiltert und nach linguistischen Kriterien systematisiert und beschrieben. Dazu werden die am IDS entwickelten mathematisch-statistischen Analyse- und Clusteringverfahren, speziell die statistische Kookkurrenzanalyse, systematisch angewendet und einer ständigen linguistischen Reflexion unterzogen. Langfristiges Ziel ist die Visualisierung solcher Gebrauchsmuster in online abrufbaren Wortverbindungsnetzen für unterschiedliche Nutzungssituationen. (<http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen/>)

Wolfgang Teubert stellte fest, dass der Kontext in korpusgestützter Lexikographie bei den Bedeutungsbeschreibungen eine größere Rolle als zuvor spielt. Dies bestätigen die Autoren des COBUILD-Wörterbuches:

The corpus also, more indirectly, influenced the definition style which we developed and which has been widely imitated. We were continually confronted by the phraseology and lexicogrammatical patterns associated with individual words and meanings, and we realized that the clearest way to communicate these was to build them into the definition

text, and to define an item **in its context**, rather than as disembodied entities. (Clear et al. 1996, S. 309)

Im Projekt *lexiko* werden Definitionen auch in ganzen Sätzen formuliert, unter anderem deshalb, weil dies erlaubt, „die am Gebrauch des Wortes beteiligten typischen Beziehungspartner [...] in die semantische Paraphrase] aufzunehmen“ (Storjohann 2005b, S. 190). Ergänzend dazu sind zwei weitere Angabetypen in der *lexiko*-Artikelstruktur zu sehen: die typischen Verwendungsmuster und die so genannte Mitspielerangabe. Beide Angaben sind innerhalb der Artikelstruktur im Bereich „Bedeutung und Verwendung“ verankert, und gemeinsam mit der Beschreibung von Verwendungsspezifika bilden sie zusammen mit der Bedeutungsparaphrase den Korpusbefund zu diesem Bereich ab:

Die Angaben zu ‚Semantischer Umgebung und Mitspielern‘ [...] sollen einen bestimmten Typ kognitiver Assoziationen zwischen dem Stichwort in einer Lesart und Wörtern in dessen Textumgebung darstellen. Dabei ist der empirische Zugang zur Textumgebung qua Korpus entscheidend. Es hat sich gezeigt, dass die Listen der signifikanten Kookkurrenzpartner [...] Lexeme enthalten, die sich als Realisierung semantischer Rollen auffassen lassen, und zwar unabhängig von der Wortart. (Haß 2005, S. 228)

Mit diesem Angabetyp liegt also ein Resultat korpusgestützter Lexikographie vor: Mithilfe intelligenter Korpusrecherche- und -analysesoftware ermittelte Informationen werden nicht etwa verworfen oder möglicherweise zwanghaft in übliche Wörterbuchstrukturen integriert, sondern bekommen einen neuen Platz im Wortartikel zugewiesen. Natürlich unterliegt das Projekt *lexiko* wegen der Publikation im Internet dabei keinen Platzbeschränkungen und kann vor diesem Hintergrund leicht einen neuen, Raum beanspruchenden Angabetyp anbieten.

Ist korpusgestützte Lexikographie, was diesen Bereich betrifft, besser, schneller, umfangreicher? Gerade hier führt das Prinzip der Korpusgestüttheit zu einem deutlichen Qualitätsgewinn von allgemeinsprachigen Bedeutungswörterbüchern, besonders aber auch Lernerwörterbüchern. Bedeutungsparaphrasen sind so eng und aktueller am tatsächlichen Sprachgebrauch orientiert, sie können mit frequenten, aktuellen Wortverbindungen ergänzt werden; Wortverbindungen können in sinnvoller Weise in Wortartikel integriert werden, wenn sie nicht alleiniger Gegenstand von Untersuchungen und Darstellungen werden. Dabei sind die Lexikographen nicht mehr auf Introspektion angewiesen, sondern Usuelles, Zentrales wird aus dem Wörterbuchkorpus ermittelt und beschrieben. Dass eine Ausweitung von Angabetypen (wie am Beispiel von *lexiko* gezeigt) natürlich gleichzeitig eine längere Bearbeitungszeit bedeutet, ist zu berücksichtigen. Auf der anderen Seite ermöglicht intelligente Korpussoftware, dass Wörterbuchbeschreibungen den Sprachgebrauch in einer bislang nicht gekannten Nähe abbilden können, weil sie bei der Wahl zwischen Signifikantem und Nicht-Signifikantem hilft.

Auch im Vergleich zu automatisch erstellten wortbezogenen Informationssystemen ist der Qualitätsaspekt wichtig. Diese können zwar in relativ kurzer

Zeit eine große Menge an lexikalischen Informationen zusammentragen, doch haben diese Informationen, weil sie nicht durch Lexikographen aufbereitet sind, nicht die gleiche Qualität wie die Angaben in Wörterbüchern.

3. Ausblick

Die Arbeit mit elektronischen Korpora wirkt sich in vielerlei Hinsicht auf die gängige lexikographische Praxis aus. Dies gilt bereits im Rahmen der Fortführung traditioneller Wörterbucharbeit, die von Einzelwörtern in ihrer Grundform ausgeht. Es bleibt dabei, zu diesen Wörtern Lesarten anzugeben, doch wird inzwischen auch teilweise dazu übergegangen, diese nach ihrer Frequenz im Korpus anzuordnen und nicht beispielsweise nach ihrem etymologischen Zusammenhang. Es bleibt dabei, diese Lesarten zu paraphrasieren und neben grammatischen Angaben z. B. typische Verwendungen zu nennen, sowie auch illustrierende Textbelege zu bringen. Geschieht dies alles korpusgestützt und unter bevorzugtem Einsatz des korpusgesteuerten Verfahrens, dann ist dies an sich ein deutlicher Qualitätsgewinn gegenüber nicht-korpusgestützter Lexikographie, da die Beschreibungen dem durch das Korpus repräsentierten Sprachgebrauch näher sind.

In gewisser Weise bleibt solch ein Vorgehen aber trotzdem traditionell, indem es das Einzelwort betrachtet und beschreibt. Wolfgang Teubert formuliert im Kontrast dazu eine andere Perspektive:

Dass die Bedeutung der Wörter Textbelegen zu entnehmen ist, war seit Adelung und Campe das Grundprinzip der klassischen Lexikographie, das von der Korpuslinguistik aufgegriffen wird. Indessen unterscheidet sie sich von der klassischen Lexikographie in bestimmten Einzelheiten. Einmal wertet die Korpuslinguistik das Korpus systematisch und nicht nur exemplarisch aus. Zum anderen verzichtet sie [...] darauf, die Bedeutung eines Wortes [...] in Isolation zu beschreiben. Drittens bemüht sie sich, auf unterschiedlichen Kontexten beruhende Gebrauchsunterschiede zu verzeichnen und nicht, wie die klassische Lexikographie, ein Wort [...] in Bedeutungen zu zerlegen. Viertens geht es der Korpuslinguistik weniger um das Einzelwort als vielmehr um die Wechselwirkung zwischen Textelement und Kontext. (Teubert 1999, S. 302)

So ist zu erwarten, dass sich Lexikographie noch stärker verändern wird, wenn die Arbeit mit Korpora weiteren Einfluss nimmt⁹. John Sinclair beschreibt im Rückblick auf die COBUILD-Arbeit (2004, S. 9), dass eben nicht nur Details, sondern Beschreibungskategorien und schließlich ganze theoretische Annahmen revidiert werden mussten. Und von der Konzentration auf das Wort als Träger der lexikalischen Bedeutung bewegte sich Sinclair hin zur Vorstellung des „lexical item“, das mehrere Wörter lang sein kann und der Hauptträger der lexikalischen Bedeutung ist. Wie wird sich das erst auf die Lexikographie auswirken? Und wie auf die Benutzer? Dies bleibt abzuwarten.

⁹ An Beispiel des Projektes *ellexiko* sind diese Veränderungen zumindest teilweise schon zu beobachten.

Offen ist auch, ob es dabei bleiben kann, unter Lexikographie das redaktionelle Schreiben von Wörterbüchern zu verstehen. Es entstehen neue Arten von automatisch erstellten wortbezogenen Informationssystemen, die vorzugsweise im Internet publiziert werden. Ein Beispiel hierfür ist das Projekt „Digitales Wörterbuch der deutschen Sprache“, das an der Berlin-Brandenburgischen Akademie der Wissenschaften entsteht und in dem Korpuslinguisten die retrodigitalisierte Fassung eines traditionellen Wörterbuches um automatisch aus dem Korpus generierte Wortinformationen (z. B. automatisch ermittelte Belege) anreichern und erweitern. Wortbezogene Informationssysteme können aber auch ohne Lexikographen oder Korpuslinguisten erarbeitet werden, wie das Projekt „Wortschatz-Lexikon“ zeigt, das in der Hand von Informatikern an der Universität Leipzig liegt und ausschließlich automatisch ermittelte Informationen (z. B. Belege, Angaben zu Synonymen) zu einzelnen Wörtern anbietet. Auch hierbei bleibt abzuwarten, was von den Benutzern in welchen Benutzungssituationen angenommen werden wird und was sich damit durchsetzen wird.

Insofern befindet sich die Lexikographie in einer hoch spannenden Phase; sie sollte sich auch weiterhin mit den Methoden und Möglichkeiten der Korpuslinguistik auseinander setzen, wie auch die Korpuslinguistik von der Auseinandersetzung mit den Methoden und Möglichkeiten der Lexikographie profitieren kann.

4. Literatur

- Baugh, Simon/Harley, Andrew/Jellis, Susan (1996): The Role of Corpora in Compiling the Cambridge International Dictionary of English. In: *International Journal of Corpus Linguistics* 1/1 1996, S. 39–59.
- Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethoden. Mannheim.
- Belica, Cyril/Perkuhn, Rainer (2006): Korpuslinguistik – Das unbekannte Wesen. Oder: Mythen über Korpora und Korpuslinguistik. In: *Sprachreport* 1/2006, S. 2–8.
- Clear, Jeremy/Fox, Gwyneth/Francis, Gill/Krishnamurthy, Ramesh/Moon, Rosamund (1996): COBUILD: The State of the Art. In: *International Journal of Corpus Linguistics* 1/2 1996, S. 303–314.
- Duden (2004): Die deutsche Rechtschreibung. 23., völlig neu bearbeitete und erweiterte Auflage. Herausgegeben von der Dudenredaktion. Auf der Grundlage der neuen amtlichen Rechtschreibregeln. Mannheim/Leipzig/Wien/Zürich.
- Engelberg, Stefan/Lemnitzer, Lothar (2001): Lexikographie und Wörterbuchbenutzung. Tübingen.
- Haß, Ulrike (2005): Semantische Umgebung und Mitspieler. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. *ellexiko – das Online-Informationssystem zum deutschen Wortschatz*. S. 227–234.
- Herberg, Dieter/Kinne, Michael/Steffens, Doris (2004): Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Unter Mitarbeit von Elke Tellenbach und Doris al-Wadi. Berlin/New York. (Schriften des Instituts für Deutsche Sprache Band 11).
- Hunston, Susan (2002): Corpora in Applied Linguistics. Cambridge.
- Landau, Sidney I. (2001): Dictionaries: The art and craft of lexicography. New York.

- Müller-Spitzer, Carolin (2003): Ord nende Betrachtung zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: Lexicographica 19/2003, S. 140–168.
- Schlaef er, Michael (2002): Lexikologie und Lexikographie. Eine Einführung am Beispiel deutscher Wörterbücher. Berlin.
- Schnörch, Ulrich (2005): Die *exlexiko*-Stichwortliste. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. *exlexiko* – das Online-Informationssystem zum deutschen Wortschatz. S. 71–90.
- Sinclair, John (1987): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London.
- Sinclair, John (1991): Corpus Concordance Collocation. Oxford.
- Sinclair, John (Hg.) (2004): How to Use Corpora in Language Teaching. Amsterdam/Philadelphia.
- Steyer, Kathrin (2004): Kookkurrenz. Korpusmethode, linguistisches Modell, lexikografische Perspektiven. In: Kathrin Steyer (Hg.) 2004: Wortverbindungen – mehr oder weniger fest. S. 87–116.
- Storjohann, Petra (2005a): Paradigmatische Relationen. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. *exlexiko* – das Online-Informationssystem zum deutschen Wortschatz. S. 249–264.
- Storjohann, Petra (2005b): Semantische Paraphrasen und Kurzetikettierungen. In: Ulrike Haß (Hg.) 2005: Grundfragen der elektronischen Lexikographie. *exlexiko* – das Online-Informationssystem zum deutschen Wortschatz. S. 182–203.
- Summers, Della (1996): Computer lexicography: the importance of representativeness in relation to frequency. In: Jenny Thomas/Mick Short (Hg.) 1996: Using Corpora for Language Research. Studies in the Honour of Geoffrey Leech. S. 260–266.
- Teubert, Wolfgang (1999): Korpuslinguistik und Lexikographie. In: Deutsche Sprache 4/1999, S. 292–313.
- Thelwall, Mike (2005): Creating and using Web corpora. In: International Journal of Corpus Linguistics 10/4 2005, S. 517–541.
- Tognini Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam/Philadelphia.
- Wahrig (2003): Fehlerfreies und gutes Deutsch. Das zuverlässige Nachschlagewerk zur Klärung sprachlicher Zweifelsfälle. Gütersloh/München.
- Wiegand, Herbert Ernst (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York.

5. Internetquellen (zuletzt eingesehen März 2006)

http://www.adwmainz.de/2005/index.php?sektion=vorh_gsk&ID=38
<http://www.duden.de/>
<http://www.dwds.de>
<http://www.exlexiko.de/Korpusbasiertheit.html>
<http://www.ids-mannheim.de/kt/misc/tutorial.html>
<http://www.ids-mannheim.de/lexik/fremdwort/quellen.html>
<http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen/>
<http://www.uni-saarland.de/verwalt/presse/campus/2002/2/07-C3-PO.html>
<http://www.wortschatz.uni-leipzig.de/>

JÖRG ASMUSSEN

Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen

Abstract

In diesem Beitrag wird untersucht, wie mithilfe korpuslinguistischer Verfahren Erkenntnisse über den Aufbau von Bedeutungsparaphrasen in Wörterbüchern gewonnen werden können. Diese Erkenntnisse sollen dazu genutzt werden, den Aufbau von Bedeutungsparaphrasen in Wörterbüchern umfassend und systematisch zu beschreiben, z. B. im Hinblick auf eine Optimierung der Bedeutungsparaphrasen für so genannte elektronische Wörterbücher oder für die Extraktion lexikalisch-semantischer Information für NLP-Zwecke.

1. Vorbemerkungen

Als Folge des Bestrebens, Sprachgebrauch lexikographisch immer präziser ermitteln und beschreiben zu können, ist die Verwendung von Korpora als Quelle bei der Erarbeitung insbesondere allgemeinsprachlicher Bedeutungswörterbücher der Gegenwartssprachen nicht mehr wegzudenken.

Unter *Korpus* ist in einem solchen Zusammenhang eine sehr umfangreiche, digitalisierte Sammlung von Texten oder Textauszügen zu verstehen, die als angenommene repräsentative Stichprobe eines durch verschiedene Parameter bestimmten Sprachgebrauchs dient, im allgemeinsprachlich-lexikographischen Kontext meist als intendierte Stichprobe der Allgemeinsprache schlechthin, vgl. zur Definition von *Korpus* in diesem Sinne einschlägige Beschreibungen und Diskussionen z. B. in Kennedy (1998), McEnery u. Wilson (2001), Tognini Bonelli (2001) oder Lemnitzer u. Zinsmeister (2006).

Wegbereiter für die Verwendung solcher Korpora in der lexikographischen Arbeit dürfte allen voran das englische *COBUILD*-Projekt gewesen sein, das 1980 als Kooperation zwischen der University of Birmingham und dem Verlag Collins begann, und dessen Ergebnis (*COBUILD*: Sinclair u. a. 1987) somit als das erste korpusgestützte Wörterbuch angesehen werden muss. Einige der korpuslinguistischen Verfahren, die diesem Wörterbuch zugrunde liegen, sind in Sinclair (1991) beschrieben.

Das erste und bisher einzige korpusgestützte Wörterbuch der dänischen Sprache ist *Den Danske Ordbog* (DDO: Hjorth u. a. 2005). Bei diesem Wörterbuch handelt es sich um ein sechsbändiges allgemeinsprachliches Bedeutungswörterbuch, vgl. Lorentzen (2004), das 1991–2005 von der Gesellschaft

für dänische Sprache und Literatur (*Det Danske Sprog- og Litteraturselskab*, DSL) von Grund auf neu erarbeitet und herausgegeben wurde.

Die Hauptquelle des DDO bildete das eigens zu diesem Zweck von der DSL zusammengestellte Textkorpus *Den Danske Ordbogs Korpus* (DDOC), vgl. Norling-Christensen u. Asmussen (1998). Es umfasst etwa 40 Millionen Textwörter in über 43 000 Texten bzw. Textausschnitten einer Vielzahl von unterschiedlichen Textsorten aus dem Zeitraum 1983–1992; und es enthält detaillierte bibliographische sowie kommunikativ und soziolinguistisch orientierte Annotationen auf Textebene. Das DDOC diente bei der Erstellung der Stichwortliste für das DDO, der Ermittlung von Lesarten und ihrer Häufigkeiten, dem Auffinden von Kollokationen oder illustrativen Textbelegen sowie bei der Ermittlung orthographischer und morphologischer Varianz.

Die dem DDO zugrunde liegenden, durch Korpusevidenz weitgehend gesteuerten Verfahren dürften mittlerweile eine weite Verbreitung innerhalb der auf Gegenwartssprachen bezogenen (nicht ausschließlich kommerziell orientierten) lexikographischen Arbeit gefunden haben, so auch in deutschsprachigen Wörterbuchprojekten wie dem *lexiko* am Institut für Deutsche Sprache, vgl. Storjohann (2005), oder dem *DWDS* an der Berlin-Brandenburgischen Akademie der Wissenschaften, vgl. Klein (2004).

Bislang weniger verbreitet in der praktischen lexikographischen Arbeit ist die Verwendung von Korpora und korpuslinguistischen Verfahren bei der Untersuchung des Bedeutungsparaphrasenvokabulars oder des strukturellen Aufbaus von Bedeutungsparaphrasen. Dieser Beitrag wird sich daher nicht mit Einsatzmöglichkeiten von Korpora als *Quelle* für die lexikographische Sprachbeschreibung auseinander setzen, sondern wird vielmehr die Bedeutungsparaphrasen des Wörterbuchs selbst als ein spezielles Korpus betrachten, über das sich mithilfe allgemein üblicher korpuslinguistischer Verfahren Einblicke in den Aufbau von Bedeutungsparaphrasen gewinnen lassen. Das längerfristig angestrebte Ziel dabei ist es, eine umfassende systematische Beschreibung des Aufbaus von Bedeutungsparaphrasen zu erarbeiten. Auf einer solchen Grundlage ließen sich traditionelle, für das Buchmedium konzipierte Wörterbücher wie das DDO optimieren, indem sie den funktionalen Möglichkeiten des Mediums Computer besser angepasst werden könnten unter gleichzeitiger Berücksichtigung einer potenziellen Verwendbarkeit als NLP-Ressourcen.

2. Wörterbücher im digitalen Kontext

2.1 „State of the Art“

Im Laufe der Neunziger haben so genannte elektronische, also durch das Medium Computer vermittelte, Wörterbücher zunehmend an Verbreitung gewonnen. So sind mittlerweile von den meisten kommerziell produzierten Wörterbüchern auch digitale Versionen verfügbar. Das „Elektronische“ vieler dieser Wörterbücher besteht jedoch häufig lediglich darin, dass die her-

kömmliche Buchversion in eine inhaltlich und funktional völlig identische digitale Version umgesetzt wurde, die sich auf dem PC einsetzen lässt. Die Vorteile solcher elektronischer Wörterbücher beschränken sich daher in aller Regel auf Geschwindigkeitsgewinne beim Nachschlagen; qualitative Vorteile gegenüber der Buchversion, die außerdem das Funktionspotenzial des Mediums Computer wirklich ausschöpfen würden, scheint es bislang nur ansatzweise zu geben.

So verfügen einige englische Lerner-Wörterbücher in ihren jüngeren Versionen über Funktionalitäten, die scheinbar über diejenige herkömmlicher Buchversionen hinausreichen. Im *Longman Dictionary of Contemporary English* (LDOCE: Bullon u. a. 2003) gibt es die Funktion *Dictionary Search*, die es ganz ähnlich auch im *Macmillan English Dictionary* (MED: Rundell u. a. 2002) unter der Bezeichnung *SmartSearch* gibt. Beide erlauben es, scheinbar begriffsorientierte Recherchen durchzuführen, die bei näherer Untersuchung jedoch einige Defizite aufweisen. So ergibt beispielsweise eine Suche nach „yellow fruit“ im LDOCE – neben Wörtern, die wie zu erwarten eine gelbe Frucht denotieren – verblüffenderweise auch das Verb *bear*, während die Anfrage „instrument NOT string“ im MED neben solchen Wörtern, die tatsächlich saitenlose Instrumente bezeichnen, auch *banjo* zu Tage fördert. In Wirklichkeit wird keine semantische Suche vollzogen, sondern es wird lediglich untersucht, ob bei der Anfrage „yellow fruit“ die Zeichensequenzen *yellow* und *fruit* in ein und demselben Artikel vorkommen, während bei „instrument NOT string“ das Wort *instrument* in der Bedeutungsparaphrase vorkommen muss, das Wort *string* hingegen nicht vorkommen darf. Die Bedeutungsparaphrase des Lemmas *banjo* lautet *a musical instrument like a guitar but with a smaller round body*, weshalb es als vermeintlich saitenloses Instrument im Ergebnis der Suche erscheint. Bei näherer Betrachtung entpuppt sich die scheinbar begriffsorientierte Recherchefunktion als reine Volltextsuche, deren Skopus auf einen Artikel (LDOCE) bzw. bestenfalls auf die Bedeutungsparaphrasen (MED) beschränkt ist.

Trotz der Defizite zeigen die beiden Beispiele, dass ein Potenzial des Mediums Computer erkannt worden ist, nämlich die prinzipielle Möglichkeit einer onomasiologisch orientierten Recherche. Allerdings scheint das Datenmaterial der gedruckten Wörterbücher diesen neuen Möglichkeiten kaum (oder wahrscheinlich gar nicht) angepasst worden zu sein. Die Beispiele scheinen aber auch anzudeuten, dass eine mögliche Diskrepanz bestehen könnte zwischen dem Streben einer Wörterbuchredaktion nach verständlichen, allzu viel Redundanz und semantische Dekomposition vermeidenden Bedeutungserklärungen einerseits und der Notwendigkeit andererseits, entweder logisch exakte und stringente Bedeutungsparaphrasen in einem digitalen Kontext zu verwenden oder andere Verfahren zu entwickeln, die eine semantische Navigation in den herkömmlichen Wörterbuchdaten ermöglichen.

2.2 Ein Konzept für digitale Wörterbücher

Im Rahmen des DSL-Projektes *ordnet.dk* soll in einer ersten Phase 2004–2009 das ursprünglich nur für das Buchmedium konzipierte DDO in ein digitales lexikalisches Nachschlagewerk umgesetzt werden unter bewusster Berücksichtigung der Möglichkeiten für eine erweiterte Recherche, die das Computermedium prinzipiell bereitstellt. Insbesondere soll als eines der Ergebnisse der mit diesem Projekt verbundenen Entwicklungsarbeiten ein begriffsorientierter, onomasiologischer Zugriff auf den Datenbestand des Wörterbuchs ermöglicht werden.

Ausgangspunkt bei der Umsetzung ist die Annahme, dass ein „elektronisches Wörterbuch“ im medialen Sinne kein Buch ist, sondern eine nach bestimmten funktionalen Kriterien organisierte lexikalische Datensammlung, weshalb man mit der Bezeichnung *digitales lexikalisches Nachschlagewerk* das Produkt vermutlich präziser beschreibt als mit der Bezeichnung *elektronisches Wörterbuch*.

Unterschiede zwischen einem Wörterbuch und einem digitalen lexikalischen Nachschlagewerk lassen sich in der Strukturierung des Datenmaterials und der Funktion, d. h. seiner intendierten und implizit gegebenen Verwendungsmöglichkeiten, feststellen.

So werden durch das Buchmedium vorgegebene strukturelle Bindungen aufgehoben, einmal hinsichtlich des Platzes, der prinzipiell nicht mehr beschränkt ist, zum anderen hinsichtlich der Artikelstruktur selbst, bei der eine Organisation des Materials nach funktionalen, oft sprachtechnologisch determinierten Kriterien für die Behandlung im Computer adäquater zu sein scheint, als eine unmittelbar auf den intendierten menschlichen Rezipienten ausgerichtete.

Weiter werden funktionale Bindungen des Buchmediums aufgehoben: Es muss nicht mehr vorab festgelegt werden, welche Funktionen ein geplantes Wörterbuch zu erfüllen hat. Dies eröffnet zumindest prinzipiell die Möglichkeit, die zugrunde liegenden lexikalischen Daten multifunktional anzulegen, z. B. auch im Hinblick auf eine eventuelle sprachtechnologische Verwendung. Erst bei der Präsentation wird das Datenmaterial in die jeweils entsprechende Form gebracht. Auch die typische funktionale Bindung eines ausschließlich alphabetisch indexierten, semasiologischen Zugriffs auf die Einträge wird prinzipiell aufgehoben.

2.3 Beispiel: Onomasiologischer Zugriff

Es soll somit also auf dem Hintergrund aufgehobener struktureller und funktionaler medialer Bindungen insbesondere der Frage nachgegangen werden, wie ein herkömmlicher lexikalischer Datenbestand mittels korpuslinguistischer Verfahren bearbeitet werden kann, sodass ein onomasiologischer Zugriff auf das Wörterbuch möglich wird. Dabei werden im Rahmen dieses Beitrags nur Substantive mit konkreten Lesarten berücksichtigt. Diese Beschränkung auf

die *1st order entities*, vgl. Lyons (1977), ist notwendig, um nicht den Rahmen dieses Beitrags zu sprengen, zumal die prinzipiell notwendige und interessante Erweiterung der Untersuchungen auf andere Wortarten und den *2nd* oder *3rd order entities* den Komplexitätsgrad der Beschreibung wesentlich erhöhen würde und den Fokus dieses Beitrags fort von den korpuslinguistischen Verfahren und hin zu der strukturellen Beschreibung von Bedeutungsparaphrasen verschieben würde.

Wenn ein semasiologisch angelegtes Wörterbuch um die Funktion eines onomasiologischen Zugriffs erweitert werden soll, bietet sich insbesondere die Analyse bestehender Bedeutungsparaphrasen an, um eine strukturelle Grundlage für einen solchen Zugriff zu erarbeiten. In diesem Zusammenhang sollte eine umfassende, generelle, systematische Beschreibung des strukturellen Aufbaus von Bedeutungsparaphrasen angestrebt werden; eine Beschreibung, die sich zumindest teilweise auf korpuslinguistische Verfahren stützen lässt, und die ein erster Schritt in Richtung auf eine eigentliche Formalisierung der lexikalisch-semantischen Information der Bedeutungsparaphrasen wäre als Voraussetzung für eine brauchbare algorithmische Prozessierung dieser Information.

In diesem Zusammenhang stellen sich bezüglich des DDO die folgenden beiden Fragen:

1. Welche korpuslinguistischen Verfahren ermöglichen einen Einblick in den Aufbau der Bedeutungsparaphrasen?
2. Welche Optimierungen der semantischen Beschreibung sind notwendig, um einen onomasiologischen Zugriff zu ermöglichen?

Vor der weiteren Erörterung dieser beiden Fragen in den Abschnitten 4 und 5 soll im Folgenden zunächst der Aufbau der semantischen Beschreibung im DDO näher beschrieben werden.

3. Semantische Beschreibung im DDO

3.1 Ein Beispiel

Im Folgenden wird der Aufbau der semantischen Beschreibung im DDO kurz dargelegt. Der Begriff *semantische Beschreibung* bezieht sich dabei auf denjenigen Teil der Mikrostruktur, in welchem insgesamt semantische Angaben zu einem Lexem gemacht werden.

Abbildung 1 zeigt einen Ausschnitt der semantischen Beschreibung des Lexems *mus* ‚Maus‘, nämlich das Semem *Nagetier*. Der hierarchische – in den Wörterbuchdaten durch XML ausgezeichnete – Aufbau der Mikrostruktur ist hier als etwas vereinfachte Baumstruktur wiedergegeben, wo die funktionalen lexikographischen Textelemente, vgl. Wiegand (1989a), die Blätter bilden, deren Funktion und Platzierung im strukturellen Gefüge aus den ihnen übergeordneten Knoten hervorgehen. Durch die Angabe eines Fragezeichens, Asteriskus oder Pluszeichens nach dem Knotennamen wird die mögliche Anzahl der von diesem Knoten ausgehenden Kanten angegeben,

wobei hier die für reguläre Ausdrücke üblichen Konventionen gelten. Fehlt die Angabe solcher Quantoren, kann von dem betreffenden Knoten genau eine Kante ausgehen. Sollten von einem Knoten null Kanten ausgehen, ist der Knoten obsolet. Solche Knoten, die also von der Strukturbeschreibung des Wörterbuchs her zwar mögliche, aber nicht realisierte Textelemente oder Teilstrukturen bezeichnen, sind in dieser Darstellung ausgelassen. In dem angeführten Beispiel sind bei weitem nicht alle von der formalen Mikrostrukturbeschreibung des Wörterbuchs her möglichen Angaben realisiert, sondern nur diejenigen, die für die semantische Beschreibung des Semems *Nagetier* im Artikel *mus* ‚Maus‘ relevant sind.

Abb. 1: Semantische Beschreibung im Artikel *mus* ‚Maus‘

Unter dem Knoten *Markierung* gibt *Domäne* eine fachsystematische, für redaktionelle Zwecke genutzte und im Buch nicht mitgedruckte Domänenzuordnung des Semems an, hier *Zoologie*.

Die Angabe der Bedeutungsparaphrase *BP* besteht aus dem eigentlichen Textelement der im Wörterbuch gedruckten Bedeutungsparaphrase sowie unter dem Knoten *GenProx* aus der Angabe eines (nicht mitgedruckten) lexikalisierten Ausdrucks für das in der Bedeutungsparaphrase angegebene genus proximum.

Semantische Relationen wie Synonymie, Antonymie etc. werden unter dem Knoten *SemRel* gesammelt, im Beispiel eine durch die unter dem Knoten

Kommentar durch das Textelement *Familie* kommentierte Angabe eines Synonyms, nämlich der lateinischen Familienbezeichnung *Muridae*.

Unter dem Knoten *Enzykl* stehen enzyklopädische Angaben, in diesem Fall der Zusatz *omfatter bl.a. husmus, skovmus, rotter og dværgmus* ‚umfasst u. a. Hausmäuse, Waldmäuse, Ratten und Zwergmäuse‘, der sich auf die Synonymangabe zu beziehen scheint und somit übrigens einen Verstoß gegen das in diesem Wörterbuch verwendete XML-Auszeichnungsprinzip darstellt, wonach Daten von ihrer möglichen Repräsentation grundsätzlich zu trennen sind und somit der Inhalt einzelner Strukturelemente sich nicht durch eine präsupponierte typographische Abfolge auf diejenigen anderer beziehen sollte. Würde man z. B. beschließen, in einer modifizierten Version dieses Wörterbuchs – aus welchen Gründen auch immer – die Synonymangaben zu unterdrücken, ergäbe die hier im Strukturelement *Enzykl* gemachte Angabe nämlich keinen Sinn mehr.

Unter dem Knoten *Kookk* sind Angaben zu korpusstatistisch ermittelten Kookkurrenzen verzeichnet. Diese sind im gegebenen Beispiel wiederum unterteilt in attributive Kookkurrenzen unter dem Knoten *Attribut*, im Beispiel *lille (mus)* ‚kleine (Maus)‘ sowie *Gefüge* wie *(mus) og rotter* ‚(Mäuse) und Ratten‘ und *forsøg med (mus)* ‚Versuche an/mit (Mäusen)‘.

Schließlich sind unter dem Knoten *Beleg* authentische, aus dem zugrunde liegenden Quellkorpus übernommene Verwendungsbeispiele angegeben, gliedert in die Angabe des *Zitats* sowie seiner *Quelle*.

Von den relativ vielfältigen funktionalen Textelementen der semantischen Beschreibung im DDO sollen im Weiteren nur die eigentliche Bedeutungsparaphrase sowie die dazu gehörende Angabe des verwendeten *genus proximum* einer näheren Betrachtung unterzogen werden.

3.2 Die Bedeutungsparaphrase (BP)

3.2.1 Grundmuster: *genus proximum* und *differentia specifica*

Verglichen mit seinem Vorgänger, dem 28-bändigen, *Ordbog over det danske Sprog* (ODS: Dahlerup u. a. 1956), das mindestens sechs verschiedene Formen der Bedeutungsparaphrase (fortan auch BP) in einer Vielzahl struktureller Varianten verwendet, selbst in ein und demselben Artikel, vgl. Asmussen (2003), fallen die Bedeutungsparaphrasen im DDO wesentlich einheitlicher aus und gliedern sich in die Haupttypen: synonymische Paraphrasen, Verwendungsangaben sowie Paraphrasen nach dem klassischen Definitionsschema der Angabe des *genus proximum* sowie der *differentia specifica*. Die überwiegende Mehrheit der Bedeutungsparaphrasentypen – gegeben durch eine hohe Anzahl konkreter Substantive im Wörterbuch – ist nach diesem Muster aufgebaut.

Es sei an dieser Stelle darauf hingewiesen, dass das klassische Definitionsschema der Angabe eines *genus proximum* und der *differentia specifica* historisch dem Zweck der Bildung einer vorsprachlichen Ontologie diene und

nicht der Erklärung der Bedeutung sprachlicher Ausdrücke, vgl. Wiegand (1989b). Aufgrund seiner traditionellen Verankerung in der Lexikographie wird es im Folgenden zur Beschreibung eines Teilbereichs der lexikographischen Paraphrasierpraxis im DDO allerdings ganz bewusst verwendet, wobei ein *genus proximum* in diesem Kontext immer als ein Hyperonym zu verstehen ist, für den es (in der Sprache der Bedeutungsparaphrasen) einen *lexikalisierten Ausdruck* (in aller Regel in Gestalt eines Wortes) gibt: so meint die fortan verwendete Bezeichnung GP einen lexikalisierten Ausdruck für ein *genus proximum*.

Entsprechendes gilt für die *differentia specifica* (im Folgenden kurz DS), die hier zu verstehen ist als *sprachlicher Ausdruck* (bzw. eine Abfolge von sprachlichen Ausdrücken), der ein spezifizierendes Merkmal (bzw. ein Bündel solcher Merkmale) angibt.

Die Bedeutungsparaphrasen im DDO wurden nicht nach einem vorgegebenen semantischen Beschreibungsmodell abgefasst, in dem z. B. festgelegt gewesen wäre, welche semantischen Merkmale und Relationen bei welchen Wörtern z. B. in Abhängigkeit ihrer grundsätzlichen Semantik, z. B. den *entities* i. S. v. Lyons (1977), bzw. ihrer Wortart in der DS-Angabe beschrieben werden sollten. Auch wurde weder ein abgegrenzter Wortschatz für die Bedeutungsparaphrasen verwendet, noch gab es Vorgaben für ihren strukturellen Aufbau. Das Abfassen von Bedeutungsparaphrasen war somit dem einzelnen Redakteur relativ frei überlassen. Während es für die meisten anderen Elemente der Mikrostruktur explizit festgelegte Regeln gab, fehlten solche für das Abfassen von Bedeutungsparaphrasen fast völlig.

Auch wenn diese „pragmatisch“ orientierte Vorgehensweise beim Abfassen von Bedeutungsparaphrasen in der praktischen Lexikographie traditionell verankert ist und ihre Zweckmäßigkeit selten hinterfragt wird, so hat sie dennoch schwerwiegende Implikationen für die Nutzung solcher Bedeutungsparaphrasen in einem digitalen Kontext. Dies gilt insbesondere für die Identifikation der unterschiedlichen Elemente der *differentia specifica*, wo gleiche DS-Angaben sehr unterschiedlich zum Ausdruck kommen und auch sehr unterschiedlich in der Abfolge platziert sein können.

3.2.2 Das *genus proximum* (GP)

Die teilweisen Hyponymiestrukturen im DDO, die sich aus seinen in den Bedeutungsparaphrasen verwendeten *genera proxima* ermitteln lassen, besitzen auf der Grundlage der im vorangegangenen Abschnitt gegebenen Definition keine unbedingte Affinität zu einer vorsprachlichen Ontologie. Die Verwendung eines ausdrucksseitigen Kriteriums bei der Festlegung des GP ist hier motiviert durch die Notwendigkeit, GP-Angaben relativ zuverlässig algorithmisch bestimmen und verarbeiten zu können.¹

¹ Auch die Einheiten, aus denen Korpora bestehen, sind zunächst nur ausdrucksseitig beschrieben (bzw. beschreibbar), wodurch sie algorithmisch prozessierbar werden.

Die Wahl eines geeigneten GP der BP lag bei der Ausarbeitung der DDO-Artikel stets im Ermessen des jeweils zuständigen Redakteurs, wobei man sich von einem nirgendwo näher explizitierten redaktionellen Ideal einer „allgemein verständlichen“ BP leiten ließ. Die Verwendung einer vollends aus dem Querkorpus eruierten Hyponymiestruktur wäre untersuchenswert, obwohl die allgemein hierfür verwendeten Methoden, vgl. z. B. Widdows (2003), eine Reihe von Unzulänglichkeiten aufweisen.

Zwei typische Bedeutungsparaphrasen aus dem DDO sind:

(1)

a. *grammofon* ‚Plattenspieler‘:

[[GP apparat][DS til at afspille grammofonplader med]]BP

≈ Gerät zum Abspielen von Schallplatten

b. *cd-afspiller* ‚CD-Player‘:

[[GP apparat][DS til at afspille cd'er med]]BP

≈ Gerät zum Abspielen von CDs

In beiden Bedeutungsparaphrasen ist als GP *apparat* ‚Apparat‘ bzw. ‚Gerät‘ angegeben, während die DS-Angabe sich – wie bei der Beschreibung von Artefakten sinnvoll – auf die Funktion bezieht, also die telische Rolle in der Qualiastruktur, vgl. Pustejovsky (1996), angibt. So auch im *Duden – Deutsches Universalwörterbuch* (DUW: Wermke u. a. 2003), das die folgenden Bedeutungsparaphrasen gibt:

(2)

a. *Plattenspieler*: [[GP Gerät][DS zum Abspielen von Schallplatten]]BP

b. *CD-Player*: [[GP Abspielgerät][DS für Compact Discs]]BP

Das eine Gerät dient also beiden Wörterbüchern zufolge dem Abspielen von Schallplatten, das andere dem Abspielen von CDs. Die Beispiele unter (1) und (2) zeigen, dass das DDO identische GP-Angaben in den Bedeutungserklärungen dieser beiden semantisch eng miteinander verwandten Begriffe hat, während das DUW in einen Fall das GP *Gerät*, im anderen das GP *Abspielgerät* verwendet. Zwar lässt sich dafür argumentieren, dass *Abspielgerät* lediglich eine lexikalisierte Variante der Phrase *Gerät zum Abspielen* ist, dass also in beiden Bedeutungsparaphrasen dasselbe GP in zwei sprachlichen Realisationen auftritt: das eine als *Wort* und das andere als *Phrase*. Da Phrasen laut hier verwendeter GP-Definition (vgl. Abschnitt 3.2.1) als GP allerdings nicht in Frage kommen, da sie sich algorithmisch nicht sicher genug ermitteln lassen, ist hier von zwei unterschiedlichen GP im DUW auszugehen, nämlich *Gerät* und *Abspielgerät*, wobei Letzteres ein Hyponym zu *Gerät* darstellen dürfte und deshalb selber das GP *Gerät* haben müsste. Im DUW fehlt allerdings der Eintrag *Abspielgerät*, sodass sich streng ausdrucksseitig-algorithmisch keine Hyperonyme hierzu im DUW ermitteln lassen.

Die identische GP-DS-Struktur in den Beispielen unter (1) deutet auf den ersten Blick darauf hin, dass das DDO ein höheres Maß an Konsistenz von Form und Inhalt der Bedeutungsparaphrasen erwarten lassen dürfte. Eine relative Konsistenz des Aufbaus der BP wurde im DDO dadurch erzielt, dass es weitgehend nach Bedeutungsfeldern redigiert wurde, d. h. dass sich typisch (aber nicht konsequent) nur *ein* Redakteur der Beschreibung der Ausdrücke eines semantischen Bereichs, z.B. aller Geräte, annahm und somit ein in sich einigermaßen schlüssiges Muster für die Bedeutungsparaphrasen konsequent verwenden konnte.

Trotzdem ist die auf den ersten Blick festzustellende Konsistenz keineswegs immer gewährleistet, wie folgende Beispiele zeigen:

(3)

a. *kanin* ‚Kaninchen‘:

[[_{GP} gnaver][__{DS} der ligner en lille hare, men har forholdsvis kortere ører og bagben og graver gange i jorden]]_{BP}

≈ Nagetier, das einem kleinen Hasen ähnelt, aber verhältnismäßig kürzere Ohren und Hinterbeine hat und Gänge in der Erde gräbt

b. *hare* ‚Hase‘:

[[_{DS} gråbrunt][__{GP} pattedyr][__{DS} med hvid bug, to store fortænder i over- og undermund, lange ører og kraftige bagben]]_{BP}

≈ graubraunes Säugetier mit weißem Bauch, zwei großen Schneidezähnen im Ober- und Unterkiefer, langen Ohren und kräftigen Hinterbeinen

c. *gnaver* ‚Nagetier‘:

[[_{GP} pattedyr][__{DS} med to store fortænder i over- og undermund der er særligt egnede til at gnave med]]_{BP}

≈ Säugetier mit zwei großen Schneidezähnen im Ober- und Unterkiefer, die zum Nagen besonders geeignet sind

Im Beispiel (3-a) wird *kanin* ‚Kaninchen‘ mit dem GP *gnaver* ‚Nagetier‘ beschrieben, gleichzeitig wird die Ähnlichkeit mit *hare* ‚Hase‘ in der BP angegeben. *Hare* ‚Hase‘ selbst im Beispiel (3-b) erhält dahingegen das GP *pattedyr* ‚Säugetier‘ und in der DS-Angabe ein Merkmal, das auch Nagetiere auszeichnet (*store fortænder* ‚große Schneidezähne‘). *Gnaver* ‚Nagetier‘ im Beispiel (3-c) ist schließlich ebenfalls als *pattedyr* ‚Säugetier‘ beschrieben. Genauso wie im Beispiel (2-b) ließe sich im Beispiel (3-b) semantisch-logisch durchaus für ein phrasal ausgedrücktes GP *pattedyr med to store fortænder i over- og undermund* ‚Säugetier mit zwei großen Schneidezähnen im Ober- und Unterkiefer‘ als Äquivalent zum lexikalisierten *gnaver* ‚Nagetier‘ argumentieren. Es entspricht allerdings wiederum nicht der hier verwendeten Definition eines GP und muss von daher verworfen werden. In diesem Fall

würde sich die algorithmische Identifikation außerdem wesentlich schwieriger gestalten als im Beispiel (2-b), da die GP-Phrase in der BP unterbrochen wird durch eine DS-Angabe im Unterschied zur ganz entsprechenden GP-Phrase im Beispiel (3-c), die zusammenhängend (allerdings um einen Relativsatz erweitert) angegeben ist.

Diese Beispiele dürften zeigen, dass ein GP-Begriff, der außer lexikalisierten Einheiten auch Phrasen zuließe, sich in einem algorithmischen Kontext nur schwer operationalisieren lässt. Auf diesem Hintergrund stellt die Angabe des fernereren Hyperonyms *pattedyr* ‚Säugetier‘ (mit eventuell einschränkenden Merkmalen in den DS-Angaben) als GP für *hare* ‚Hase‘ unter gleichzeitiger Angabe des nächsten Hyperonyms *gnaver* ‚Nagetier‘ als GP für *kanin* ‚Kaninchen‘ eine Inkonsistenz in der Hyponymiestruktur dar. Solche Hyponymie-Inkonsistenzen sollten nach Möglichkeit im Wörterbuch bereinigt werden.

Abbildung 2 zeigt einen kleinen Ausschnitt aus der Hyponymiestruktur des DDO mit dem Ausgangspunkt im GP *gnaver* ‚Nagetier‘.

Abb. 2: Ausschnitt aus der impliziten Hyponymiestruktur des DDO

Abgesehen von der bereits besprochenen Hyponymie-Inkonsistenz geht aus der Abbildung weiter hervor, dass auch *muldyr* ‚Maultier‘ wahrscheinlich zu hoch in der Hierarchie angesiedelt wurde, obwohl andererseits das Hyperonym *dyr* ‚Tier‘ in der Allgemeinsprache in diesem Fall durchaus seine Berechtigung haben kann – und es ließe sich mit ziemlicher Wahrscheinlichkeit auf der Grundlage des Querkorpus auch so belegen. Dieses Beispiel zeigt die Diskrepanz zwischen einer (fachsprachlichen) Taxonomie einerseits und (allgemeinsprachlichen) semantischen Beschreibungen im Wörterbuch andererseits – ein Umstand, dem bei einer algorithmischen Nutzung des Materials Rechnung getragen werden muss.

Aus dem abgebildeten Ausschnitt aus der Hyponymiestruktur im DDO geht weiter hervor, das in der BP von *skadedyr* ‚Schädling‘, das GP *dyr* ‚Tier‘ angegeben wurde. Nun ist ein Schädling zwar ein Tier und insofern ist das GP *dyr* durchaus gerechtfertigt, dennoch gehört es keiner bestimmten Art an und ist insofern auch kein Kohyponym zu *hvirveldyr* ‚Wirbeltier‘, *bløddyr* ‚Weichtier‘ oder gar *muldyr* ‚Maultier‘; wäre dem so, würde dies unsinnigerweise ausschließen, dass ein Wirbeltier ein Schädling sein kann. Die Hyponymiebeziehung zwischen *skadedyr* ‚Schädling‘ und *dyr* ‚Tier‘ scheint demnach also eine qualitativ andere zu sein, als diejenige zwischen *hvirveldyr* ‚Wirbeltier‘ und *dyr* ‚Tier‘. Dieser Unterschied sollte für eine algorithmische Nutzung des DDO expliziert werden.

Weitere Schwierigkeiten für die algorithmische Nutzung des DDO-Materials ergeben sich aus dem Umstand, dass die GP-Angaben stets die Gestalt eines sprachlichen Ausdrucks haben, vgl. die Definition in Abschnitt 3.2.1, und nicht die eines (formal angegebenen) Inhalts, wie die Beispiele unter (4) und (5) belegen:

(4)

a. *svævefly* ‚Segelflugzeug‘:

[[_{DS} motorløs][_{GP} flyvemaskine][_{DS} der skal trækkes op i luften (..) og derefter holder sig svævende (..)]]_{BP}

≈ motorloses Flugzeug, das (..) in die Luft hochgezogen wird und sich danach (..) in der Schwebelage hält

b. *propelfly* ‚Propellerflugzeug‘:

[[_{GP} fly][_{DS} som drives frem vha. én el. flere propeller]]_{BP}

≈ Flugzeug, das durch einen oder mehrere Propeller vorangetrieben wird

(5)

a. *milt* ‚Milz‘:

[[_{DS} aflangt mørkerødt][_{GP} organ][_{DS} i bughulen hvis vigtigste funktion er (..)]]_{BP}

≈ längliches, dunkelrotes Organ in der Bauchhöhle, dessen wichtigste Funktion es ist (..)

b. *magistrat* ‚Magistrat‘:

[[_{GP} organ][_{DS} som varetager den kommunale forvaltning i visse større byer (..)]]_{BP}

≈ Organ, das die kommunale Verwaltung in gewissen größeren Städten wahrnimmt (..)

In den beiden Beispielen unter (4) werden jeweils die synonymischen GP *flyvemaskine* und *fly* ‚Flugzeug‘ angegeben, was dazu führt, dass nach der rein ausdrucksseitigen GP-Auffassung hier zwei scheinbar unterschiedliche GP

verwendet werden, deren identischer Inhalt sich nicht unmittelbar algorithmisch ermitteln lässt. Interessanterweise wird dabei ausgerechnet bei *svævefly* ‚Segelflugzeug‘ die GP-Variante mit der Betonung der Motorisiertheit *flyve-maskine* wörtlich ‚Flugmaschine‘ angegeben.

In den Beispielen unter (5) wird das GP *organ* in zwei unterschiedlichen Lesarten verwendet, welches nach der rein ausdrucksseitigen GP-Auffassung dazu führt, dass *milt* ‚Milz‘ und *magistrat* in Übersichten wie in Abbildung 2 als Kohyponyme zu *organ* auftreten würden.

Die Beispiele zeigen somit, dass Synonymie und Polysemie bei GP-Angaben bei einer rein ausdrucksseitigen algorithmischen Nutzung des Wörterbuchmaterials zu Unregelmäßigkeiten führen werden, weshalb Synonymie und Polysemie bei GP-Angaben in einem digitalen Wörterbuch vermieden werden sollten.

Abschließend zeigt (6) ein Beispiel für ein GP mit nur einem Hyponym im ganzen Wörterbuch:

(6)

ildtang ‚Feuerzange‘:

[[GP *jerntang*]_{DS} der bruges til at arrangere brænde i et ildsted]]_{BP}

≈ Eisenzange, die verwendet wird, um die Feuerung in einer Feuerstelle zu-rechtzurücken

Das verwendete GP *jerntang* ‚Eisenzange‘ hat nur das Hyponym *ildtang* ‚Feuerzange‘, ist also einmalig im gesamten DDO. Prinzipiell steht der algorithmischen Verwendung eines solchen Hapaxlegomenons nichts im Wege, solange man für das verwendete GP selbst wieder ein allgemeineres GP im Wörterbuch findet. In diesem Fall gibt es allerdings für *jerntang* ‚Eisenzange‘ ebenso wenig wie für *Abspielgerät* oben in Beispiel (2-a) einen eigenen Eintrag im Wörterbuch. *Ildtang* liegt somit außerhalb jeglicher Hyponymiestrukturen im DDO. Auch dieses müsste für eine algorithmische Nutzung des Materials behoben werden.

3.2.3 Die *differentia specifica* (DS)

Während die GP-Angaben durch ein gesondertes Element in der Artikelstruktur expliziert werden, das zu Systematisierungszwecken, z. B. zur Bildung von teilweisen Begriffshierarchien wie in Abbildung 2, verwendet werden kann, lassen sich DS-Angaben aus den bestehenden Bedeutungsparaphrasen wesentlich schwieriger systematisieren.

Aus den Beispielen unter (1) geht hervor, dass die DS-Angaben sowohl von ihrer Semantik (in diesem Fall eine Funktionsbeschreibung) als von ihrem sprachlichen Ausdruck (*til at* ‚um zu‘ mit Infinitivphrase) her gleichermaßen aufgebaut sind. Dies legt die Vermutung nahe, dass sämtliche Bedeutungsparaphrasen hinsichtlich der DS-Angaben relativ einheitlich aufgebaut sind. Beispiel (7) zeigt allerdings, dass dem nicht so ist. Hier wird ebenfalls ein Lexem mit dem GP *apparat* beschrieben:

(7)

fjernsyn ‚Fernseher‘:

[[DS-1 kasseformet][GP apparat][DS-2 der kan modtage tv-signaler og omsætte dem til bevægelige billeder på en skærm og tilhørende lyd i apparatets højttalere]BP

≈ kastenförmiges Gerät, das TV-Signale empfangen und sie zu bewegten Bildern auf einem Schirm und dazugehörigem Ton in den Lautsprechern des Gerätes umsetzen kann

Unmittelbar fällt auf, dass in dieser BP auch Angaben zum allgemeinen Erscheinungsbild des Geräts gemacht werden. Der BP zufolge scheint die Kastenförmigkeit offenbar eine nennenswerte Eigenschaft eines Fernsehers zu sein, und offenbar scheint dieses Gerät im Gegensatz zu jenen im Beispiel (1) auch eine wesentlich detaillierte Funktionsbeschreibung verdient zu haben, was wenig plausibel erscheint. Darüber hinaus wird das GP in der DS in einer genitivischen Form wieder aufgenommen – was zur Komplexität dieser BP weiter beiträgt.

Es sind hier also zwei unterschiedliche Arten von *differentiae specifica* im Spiel: eine, die das Aussehen des Definiendums beschreibt, ausgedrückt als attributives Adjektiv, und eine weitere, die seine telische Rolle beschreibt und zwar in Gestalt einer finiten Form des Modalverbs *kunne* ‚können‘ mit zwei Infinitivphrasen *modtage tv-signaler* ‚TV-Signale empfangen‘ und *omsætte dem* ‚sie [TV-Signale] umsetzen‘. Es kann auf dieser Grundlage die Annahme gemacht werden, dass viele Artefakt-Bedeutungsparaphrasen nach diesem Muster aufgebaut sind, also die fakultative Angabe einer Eigenschaft (DS-1) sowie die Angabe einer (oder mehrerer) Funktion(en) (DS-2) beinhalten. Ein nächster Schritt wäre, systematisch zu erkunden, wie diese DS-Angaben sprachlich in den Bedeutungsparaphrasen zum Ausdruck kommen können und auf dieser Grundlage eine systematische „Grammatik“ der Bedeutungsparaphrasen zu formulieren. Diese könnte dann dazu eingesetzt werden, die Bedeutungsparaphrasen zu vereinheitlichen und dazu, quasi-algorithmisch bestimmte semantische Merkmale aus den Bedeutungsparaphrasen zu ermitteln, die z. B. für onomasiologische Recherchezwecke genutzt werden könnten.

Um jedoch einen Überblick über sämtliche Bedeutungsparaphrasen im DDO zu bekommen, bedarf es rationellerer Verfahren als des menschlichen Lesens, Analysierens und Beurteilens einer jeden BP im Wörterbuch. Hier können korpuslinguistische Verfahren hilfreich sein.

4. Korpuslinguistische Verfahren

Im Hinblick auf eine Anpassung des DDO-Datenbestandes an die Anforderungen eines onomasiologischen Zugriffs wurden in Abschnitt 2.3 die folgenden beiden Fragen aufgeworfen:

1. Welche korpuslinguistischen Verfahren ermöglichen einen Einblick in den Aufbau der Bedeutungsparaphrasen?
2. Welche Optimierungen der semantischen Beschreibung sind notwendig, um einen onomasiologischen Zugriff zu ermöglichen?

Auf dem Hintergrund der Untersuchungen im vorhergehenden Abschnitt soll im Folgenden einer möglichen Beantwortung dieser Fragen nachgegangen werden.

4.1 Wörterbuch und Korpus

Wie bereits in Abschnitt 1 angeführt wurde, so verwendet die herkömmliche korpusgestützte allgemeinsprachliche Lexikographie ein Korpus als angenommene repräsentative Stichprobe des im Wörterbuch zu beschreibenden Sprachgebrauchs. Eine Überlegung wäre nun, das Quellkorpus auch als Grundlage für den Ausdruck der lexikalisch-semantischen Beschreibung zu verwenden. Bedeutungsparaphrasen könnten so kontrollierbar allgemeinsprachlich abgefasst werden und nur mit tatsächlich im Korpus auch belegbaren Wörtern und sprachlichen Strukturen. So ließe sich mithilfe verschiedener korpuslinguistischer Verfahren das Ausdrucksinventar für die Formulierung von Bedeutungsparaphrasen festlegen.

Beispiele für diese Tendenz sind zum Teil *COBUILD* und *lexiko*. Zwar mag eine solche Annäherung der sprachreflexiven Sprache des Lexikographen an die Objektsprache, die sie semantisch beschreiben soll, dem menschlichen Wörterbuchbenutzer entgegenkommen, gleichzeitig dürfte sie aber, da sie nicht bewusst für ihren Zweck, nämlich Bedeutungen zu erklären, festgelegt wurde, eine algorithmische Prozessierung solchermaßen abgefasster Bedeutungsparaphrasen erschweren. Möchte man herkömmliche, nicht-formale Bedeutungsparaphrasen einem onomasiologischen Zugriff zugänglich machen, so sollte man im Gegenteil das sprachreflexive Ausdrucksrepertoire für Bedeutungsparaphrasen bewusst und eindeutig festlegen.

Die Hyponymierelationen, die den Bedeutungsparaphrasen zugrunde liegen sollen, ließen sich ebenfalls mit verschiedenen Mitteln aus Korpora ableiten. Ein solches Unterfangen wäre durchaus von einem gewissen Erkenntnisinteresse, da es einen Vergleich mit den implizit vorhandenen GP-Hierarchien im Wörterbuch erlauben würde und sicher zur Optimierung dieser Hierarchien beitragen könnte. Dennoch wären so aufgestellte Hierarchien wahrscheinlich von noch mehr Inkonsistenzen geprägt als die aus den Bedeutungsparaphrasen ableitbaren, weshalb auch hier ein bewusst festgelegtes sprachreflexives, lexikographisches Ausdrucksinventar zweckmäßiger erscheint als die Anlehnung an ein allgemeinsprachliches Korpus.

4.2 Wörterbuch als Korpus

Um mithilfe korpuslinguistischer Verfahren einen Einblick in den Aufbau von Bedeutungsparaphrasen (vgl. Frage 1) zu bekommen, werden diese nunmehr

als ein eigenständiges Spezialkorpus betrachtet. Das BP-Korpus des DDO ist dabei so strukturiert, dass jede BP im Korpus mit drei Attributen versehen ist, nämlich der Angabe des zu dieser BP gehörenden Schlagwortes, einer Angabe seiner Wortart sowie des in der BP verwendeten GP. So lassen sich z. B. gezielt BP mit bestimmten GP-Ausdrücken oder BP bestimmter Wortarten aus dem Korpus für Untersuchungszwecke herausfiltern. Das BP-Korpus des DDO umfasst knapp 87800 Bedeutungsparaphrasen mit insgesamt etwa einer Million Tokens.

Wie im Abschnitt 3 dargestellt wurde, ergeben sich bei näherer Untersuchung des Aufbaus von Bedeutungsparaphrasen einige Regelmäßigkeiten, die eine algorithmische Prozessierung nahe legen, aber auch eine ganze Reihe von Unregelmäßigkeiten, die eine algorithmische Prozessierung zumindest erschweren, wenn nicht gar ganz in Frage stellen.

Es soll daher nunmehr untersucht werden, mit welchen korpuslinguistischen Mitteln sich die unregelmäßigen Fälle auf der Grundlage des BP-Korpus generell für das gesamte Wörterbuch isolieren lassen, sodass eine anschließende Anpassung dieser Abweichungen an die mehrheitlich determinierte „Norm“ ermöglicht wird.

Wie die Untersuchungen im Abschnitt 3 gezeigt haben, fallen die Abweichungen in folgende Gruppen:

1. GP-Angaben:

- (a) Hyponymie-Inkonsistenzen, Beispiel *kanin* ‚Kaninchen‘ mit dem GP *gnaver* ‚Nagetier‘ vs. *hare* ‚Hase‘ mit dem GP *pattedyr* ‚Säugetier‘
- (b) Hyponymie-Diskrepanzen zwischen einer Taxonomie und Wörterbüchern, Beispiel *muldyr* ‚Maultier‘ mit dem GP *dyr* ‚Tier‘ und nicht *hvirveldyr* ‚Wirbeltier‘ oder *pattedyr* ‚Säugetier‘
- (c) Unterschiedliche Qualitäten von Hyponymiebeziehungen, Beispiel *skadedyr* ‚Schädling‘ vs. *hvirveldyr* ‚Wirbeltier‘ beide mit dem GP *dyr* ‚Tier‘
- (d) Synonymische GP-Ausdrücke, Beispiel *fly* vs. *flyvemaskine* beide ‚Flugzeug‘
- (e) Polyseme GP-Ausdrücke, Beispiel *organ*
- (f) GP mit nur einem Hyponym, Beispiel *jerntang* ‚Feuerzange‘
- (g) Wörterbuchintern hyperonymlose GP (die keine Wurzelknoten in der Hyponymiehierarchie bilden), Beispiel *jerntang* ‚Feuerzange‘

DS-Angaben:

- (a) Uneinheitliche Merkmalangaben
- (b) Uneinheitliche Verfahren in der syntaktischen Anordnung von Merkmalangaben

Mithilfe der nachfolgend aufgelisteten Reihe quantitativer korpuslinguistischer Verfahren ist es zum Teil möglich, einige dieser Unregelmäßigkeiten in den Bedeutungsparaphrasen zu erfassen.

1. Similaritätsprofile über Bedeutungsparaphrasen: Mithilfe verschiedener statistischer Similaritätsmaße ließe sich die oberflächliche Übereinstim-

- mung von Bedeutungsparaphrasen z. B. hinsichtlich der Wahl und Anordnung der in ihnen verwendeten Tokens bestimmen. Hierbei ist anzunehmen, dass einander ähnliche Bedeutungsparaphrasen auch gemeinsame GP haben müssten. Werden nun einander ähnliche Bedeutungsparaphrasen mit unterschiedlichen GP gefunden, könnte dies womöglich auf eine Unregelmäßigkeit in der zugrunde liegenden semantischen Hierarchie hindeuten. Mithilfe einer solchen Methode ließen sich eventuell die unter (1-a) angeführten Unregelmäßigkeiten eingrenzen.
2. Frequenzprofile über das GP-Vokabular: Mithilfe solcher Frequenzprofile lassen sich vor allem niederfrequente GP ermitteln, die daraufhin durch solche auf höherem Niveau in der Hierarchie ausgetauscht werden können. Diese Methode eignet sich daher insbesondere zur Ermittlung der oben unter (1-f) angegebenen Unregelmäßigkeiten und zum Teil der unter (1-g) angeführten.
 3. GP-Lemma-Abgleich: Bei diesem Verfahren werden zwei Frequenzprofile oder lediglich zwei Type-Inventarlisten erstellt: eine über GP und eine über die Lemmata des Wörterbuchs. Diese Listen werden anschließend miteinander verglichen, sodass GP, für die es im Wörterbuch kein Lemma gibt, herausgefiltert werden. Mithilfe dieses Verfahrens dürften sich insbesondere Unregelmäßigkeiten vom Typ (1-g) ermitteln lassen.
 4. Frequenzprofile über das BP-Vokabular: Mithilfe dieses Verfahrens lassen sich insbesondere niederfrequente Wortformen ermitteln, die anschließend durch höherfrequente ersetzt werden könnten, was zu einer Vereinheitlichung des BP-Vokabulars (und der BP-Syntax) führen dürfte und somit zur Identifikation von Unregelmäßigkeiten der Typen (2-a) und (2-b) beitragen könnte.
 5. Vergleiche von *n*-Gramm-Profilen: Weiter lassen sich signifikante syntagmatische Muster in den Bedeutungsparaphrasen durch Vergleiche von *n*-Gramm-Profilen des Bedeutungsparaphrasenkörpus mit solchen des Quellkörpus ermitteln. Als statistischer Test empfiehlt sich hierbei *Log Likelihood* (G^2), vgl. Dunning 1994. Auch dieses Verfahren kann Aufschlüsse über das Inventar und den Aufbau von DS-Angaben geben und lässt sich für Unregelmäßigkeiten der Typen (2-a) und (2-b) verwenden.
 6. KWIC-Konkordanzen: Muster lassen sich anschließend mit einem Konkordanzwerkzeug für das gesamte Bedeutungsparaphrasenkörpus ermitteln, und es kann auf dieser Grundlage festgestellt werden, welche semantischen Spezifika durch welche Muster in den Bedeutungsparaphrasen zum Ausdruck gebracht werden. Auch dieses Verfahren dient insbesondere der Optimierung der DS-Angaben, Typen (2-a) und (2-b).
 7. Ermittlung hervortretender DS-Types: Mithilfe eines statistischen Verfahrens zum Ermitteln unerwartet häufig auftretender Phänomene in einer bestimmten Situation (z. B. *Log Likelihood* oder *Mutual Information*, vgl. Church u. Hanks (1989)) lassen sich augenfällige DS-Types in Bedeutungsparaphrasen mit einem gemeinsamen GP ermitteln. Hierdurch lassen

sich Rückschlüsse auf bestimmte semantische Merkmale ziehen; außerdem lassen sich Unregelmäßigkeiten der Typen (2-a) und (2-b) so identifizieren.

Die hier aufgelisteten Methoden eignen sich überwiegend zur Optimierung der DS-Angaben in den Bedeutungsparaphrasen des Wörterbuchs. Unregelmäßigkeiten in der Hyponymiestruktur lassen sich hingegen wesentlich schwieriger mithilfe korpuslinguistischer Verfahren ermitteln. So beruhen die Methoden (1)–(3) nicht so sehr auf der Rohinformation im BP-Korpus, sondern vielmehr auf seiner GP-Annotation. Gäbe es diese nicht, wäre die Optimierung der Hyponymiestruktur noch schwieriger.

Im Folgenden soll nun beschrieben werden, wie einige der Methoden für die Optimierung von Bedeutungsparaphrasen im Wörterbuch eingesetzt werden können (vgl. Frage 2 oben).

5. Optimierung

Unter *Optimierung* ist hier die gezielte Anpassung der lexikalisch-semantischen Beschreibung insgesamt an bestimmte Anwendungsszenarien zu verstehen. Im Rahmen eines herkömmlichen Wörterbuchs schließt dies u. a. die bewusste, gezielte Formulierung von Bedeutungsparaphrasen im Hinblick auf den anvisierten Benutzer ein. Im Rahmen einer sprachtechnologischen Ressource wie beispielsweise eines WordNets sollte die Anpassung hingegen mögliche sprachtechnologische Anwendungen im Blick haben. In einem solchen Kontext wäre eine logisch konsistente Formalisierung der lexikalisch-semantischen Beschreibung sinnvoll. Hier ist das Anwendungsszenario zunächst einmal ein digitales Wörterbuch, das auch eine onomasiologische Recherche erlauben soll.

Ziel der Optimierung ist eine Verbesserung der Bedeutungsparaphrasen, worunter teils eine strukturelle Vereinheitlichung, teils überhaupt größtmögliche Konsistenz in der semantischen Beschreibung zu verstehen ist. Hierbei sollten sowohl unnötige Redundanz als auch unnötige Spezifizierungen weitgehend vermieden werden. Ziel ist es nicht, aus den bestehenden Bedeutungsbeschreibungen im DDO *unmittelbar* eine sprachtechnologische Ressource zu derivieren – derlei Unterfangen, ein maschinenlesbares herkömmliches Wörterbuch zu einer lexikalisch-semantischen NLP-Ressource zu konvertieren, wurden in den vergangenen etwa fünfundzwanzig Jahren unter Anwendung unterschiedlichster Methoden immer wieder unternommen und konnten bislang eher nur bescheidene Erfolge verzeichnen. Ide u. Véronis (1994) stellten bereits vor Jahren fest, dass die Derivation lexikalisch-semantischer Ressourcen aus Wörterbüchern erhebliche Hürden zu überwinden hat. Seither hat sich das Hauptaugenmerk der Forschung auf die Verwendbarkeit von *Korpora* für das teilweise automatisierte Erstellen sowohl herkömmlicher als auch sprachtechnologischer lexikalisch-semantischer Ressourcen gerichtet, so z. B. das von Kilgarrieff u. Tugwell (2002) vorgestellte Verfahren, so genannte WordSketches aus Korpora abzuleiten.

Das Ziel der hier skizzierten Arbeit ist es also nicht, eine lexikalisch-semantiche NLP-Ressource unmittelbar aus dem DDO abzuleiten, sondern zunächst einmal die lexikalisch-semantiche Beschreibungen, allen voran die Bedeutungsparaphrasen, im Wörterbuch selbst zu vereinheitlichen, um dann von hier aus eine Grundlage für eine mögliche zukünftige Generalisierung der semantiche Beschreibung zu schaffen, sodass letztendlich ein und derselbe lexikalische Datenbestand sowohl für gedruckte als auch für elektronische Wörterbücher genutzt werden kann und außerdem gleichzeitig sprachtechnologischen Anforderungen genügt.

5.1 GP-Angaben

Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur lassen sich möglicherweise mithilfe des in Abschnitt 4.2 unter (2) dargestellten Verfahrens zum Teil identifizieren. Ein solches Verfahren ist augenblicklich bei der DSL in Entwicklung. Vorläufig erfolgt die Bereinigung von Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur allerdings ohne Zuhilfenahme eines aus der Korpuslinguistik entlehnten quantitativen Verfahrens. Statt dessen werden Teilhierarchien aus dem Wörterbuch graphisch dargestellt und daraufhin von Redakteuren auf ihre Konsistenz hin untersucht und es werden offensichtlich notwendige Korrekturen vorgenommen.

Etwas einfacher stellt sich die Situation bei der Ermittlung und Bereinigung niederfrequenter GP dar. Von insgesamt 13748 unterschiedlichen GP-Types werden 6776 nur ein einziges Mal verwendet; Tabelle 1 zeigt einige solcher GP-Types, die überdies weder im zugrunde liegenden Quellkorpus vorkommen noch selbst Stichwörter im Wörterbuch bilden, die somit außerhalb der algorithmisch ermittelbaren Hyponymiestruktur des Wörterbuchs liegen.

GP-Type	Äquivalent
pergamenthåndskrift	‚Pergamenthandschrift‘
teaterfigur	‚Theaterfigur‘
stegeredskab	‚Bratwerkzeug‘
munddel	‚Mundteil‘
konkurrenceprincip	‚Wettbewerbsprinzip‘
mindebygning	‚Denkmalgebäude‘
farveopløsning	‚Farblösung‘
kviksølvforbindelse	‚Quecksilberverbindung‘
symaskinedel	‚Nähmaschinenteil‘
lakridskugle	‚Lakritzkugel‘

Tab. 1: Einige GP-Types mit der Frequenz 1 im DDO

Bei den gezeigten GP-Types dreht es sich im Wesentlichen um okkasionell gebildete Komposita, die sich z. B. durch die kursivierten Teile in den Bedeutungsparaphrasen ersetzen ließen.

5.2 DS-Angaben

Da die DS-Angaben den größten Teil der Bedeutungsparaphrasen ausmachen, kommen korpuslinguistische Verfahren bei der Optimierung dieser Angaben besser zum Zug als bei den GP-Angaben. Im Folgenden werden nur die Möglichkeiten des in Abschnitt 4.2 unter (7) vorgestellten Verfahrens exemplarisch aufgezeigt.

Abbildung 3 zeigt die im DDO vorkommenden Lemmata, die in mindestens einer ihrer Bedeutungsparaphrasen das GP *gnaver* ‚Nagetier‘ haben. Untersucht man nunmehr sämtliche dieser Bedeutungsparaphrasen, so zeigt sich, dass bestimmte Merkmalsbezeichnungen in den DS-Angaben öfter wiederkehren – so z. B. ‚klein‘ bzw. ‚mausähnlich‘, ‚Haustier‘, ‚nicht einheimisch‘, ‚ohne Schwanz‘ – und zwar in den Bedeutungsparaphrasen derjenigen Lemmata, auf welche die vom entsprechenden Merkmal ausgehenden Pfeile in der Abbildung verweisen.

Abb. 3: Wiederkehrende Merkmale für *gnaver* ‚Nagetier‘

Sämtliche Hyponyme zu *gnaver* ‚Nagetier‘ ließen sich auf der Grundlage derartiger Merkmalsuntersuchungen mit systematisch vereinheitlichten Angaben so ermittelter Merkmale versehen. Eine solche konsequent bearbeitete digitale Version des Wörterbuchs würde eine relativ zuverlässige Beantwortung einfacher onomasiologischer Anfragen ermöglichen wie *Welche Wörter bezeichnen Nagetiere, die als Haustiere gehalten werden?*

Bei Hyponymen zu *gnaver* ‚Nagetier‘, wovon es lediglich gut zwanzig im DDO gibt, ist eine manuell vorgenommene Gruppierung verschiedener Merkmale aus den DS-Angaben relativ überschaubar. Wesentlich unübersichtlicher gestaltet sich diese Aufgabe mit Hyponymen sehr hochfrequenter GP, wie den in Tabelle 2 gezeigten. Hier wäre für Gruppierungszwecke eine statistische Hilfe vonnöten.

GP-Type	Äquivalent	Häufigkeit
person	‚Person‘	4382
del	‚Teil‘	799
sted	‚Ort‘, ‚Stelle‘	651
område	‚Gebiet‘	422
gruppe	‚Gruppe‘	421
plante	‚Pflanze‘	389
måde	‚Art/Weise‘	345
mængde	‚Menge‘	326
stof	‚Stoff‘	325
kvinde	‚Frau‘	318
farve	‚Farbe‘	314
genstand	‚Gegenstand‘	291

Tab. 2: Häufige GP-Types im DDO

So können hervortretende DS-Types in Bedeutungsparaphrasen mit gemeinsamem GP z. B. durch die Verwendung des informationstheoretischen Maßes *Mutual Information* (MI) ermittelt werden, das sonst typisch für die Ermittlung von Kookkurrenzen auf der Grundlage eines bestimmten Ausgangswortes verwendet wird, vgl. Church u. Hanks (1989). In der hier vorgestellten Anwendung sind diese Ausgangswörter GP und die potenziell kookkurrierenden Wörter sämtliche Wortformen der Bedeutungsparaphrasen.

Im Einzelnen funktioniert die MI-basierte Ermittlung von DS-Types so, dass zunächst ein Korpus *G* aus sämtlichen Bedeutungsparaphrasen im Wörterbuch gebildet wird. Daraufhin wird dasjenige Teilkorpus *T* von *G* ermittelt, das aus Bedeutungsparaphrasen mit einem bestimmten gemeinsamen GP besteht.

Für jedes Type in einem solchen Teilkorpus *T* wird seine relative Frequenz im Teilkorpus f_T und seine relative Frequenz im Gesamtkorpus f_G ermittelt. Der Quotient aus diesen beiden Frequenzzahlen ergibt für jedes Type in *T* einen (gerundeten) Score $s = \frac{f_T}{f_G}$. Je höher dieser Score, desto „enger“ die Bindung zwischen dem jeweiligen Type und dem GP, das für *T* bestimmend ist. Der niedrigste Wert, den *s* annehmen kann, ist 1, der dann eintritt, wenn die Häufigkeiten des Types in *T* und *G* identisch sind. Dies wäre dann so zu interpretieren, dass es keine besondere Bindung zwischen diesem Type und dem untersuchten GP gibt.

Ein Beispiel: Für das Token *doven* ‚faul‘ in einem Korpus *T*, das durch das GP *person* gebildet wurde, ergibt sich der Score:

$$s = \frac{f_T(doven)}{f_G(doven)} = \frac{281}{19} \doteq 15$$

Vereinfacht ausgedrückt kommt *doven* etwa 15 mal häufiger in Bedeutungsparaphrasen mit dem GP *person* vor als in allen Bedeutungsparaphrasen des Wörterbuchs zusammengekommen.

Sortiert man eine so ermittelte Liste von Bedeutungsparaphrasentypes nicht-steigend nach ihren Scores, erhält man die für das jeweilige GP signifikantesten Bedeutungsparaphrasentypes an der Spitze dieser Liste.

Tabelle 3 zeigt die Spitze einer solchen Liste für das GP *gnaver* ‚Nagetier‘; nämlich alle Types mit dem höchsten Score 2532. Der unmittelbare Eindruck von den aufgelisteten Types ist der, dass ihre Verbreitung in der Allgemeinsprache überwiegend wohl eher als gering einzuschätzen ist. In der Tat sind diese Types auch im BP-Korpus des DDO sehr niederfrequent: so kommen die Types *hårbeklædt* ‚haarbedeckt‘ und *nataktiv* ‚nachtaktiv‘ nur ein einziges Mal im Wörterbuch vor, nämlich in den Bedeutungsparaphrasen der Lemmata *studsmus* ‚Wühlmaus‘ und *hulepindsvin* ‚Stachelschwein‘. Dass niederfrequenten Types ein so hohes Gewicht in der Berechnung des Scores beigegeben wird, ist eine grundlegende Eigenart von MI: Wenn ein bestimmtes Type z. B. nur wenige Male in *G* vorkommt und gleichzeitig wenige Male im betrachteten Teilkorpus *T*; wenn *T* überdies nur wenige Tokens enthält wie im Falle *gnaver* ‚Nagetier‘, dann wird der Score für ein sehr niederfrequentes, vielleicht nur zufällig im Wörterbuch verwendetes Token sehr hoch, ohne dass aus diesem Grund etwas für eine sonderlich „enge“ Bindung zwischen dem Token und dem GP spräche. Diese Situation wird in Tabelle 3 widerspiegelt.

BP-Type	Äquivalent	Score
bæverlignende	‚biberähnlich‘	2532
fodsåler	‚Fußsohlen‘	2532
hårbeklædt	‚haarbedeckt‘	2532
nataktiv	‚nachtaktiv‘	2532
nutria	‚Nutria‘	2532
præriekområder	‚Präriegebiete‘	2532
Sibiriens	‚Sibiriens‘	2532
tundra	‚Tundra‘	2532
ungskov	‚Jungwald‘	2532

Tab. 3: BP-Types mit höchstem Score für *gnaver* ‚Nagetier‘

Unmittelbar scheint das MI-basierte Verfahren, hervortretende BP-Types für eine ganze Reihe von Kohyponymen zu ermitteln, seinen Zweck also zu verfehlen, denn niederfrequente Types bekommen scheinbar einen viel zu hohen

Score. Dies kann allerdings auch als Vorteil zur Optimierung von Bedeutungsparaphrasen genutzt werden. Denn es könnte durchaus etwas dafür sprechen, dass die so ermittelten Wortformen in den Bedeutungsparaphrasen überhaupt viel zu niederfrequent, zu speziell sind, als dass sie im Wörterbuch verwendet werden sollten.

So kommt keines der Types *bæverlignende* ‚biberähnlich‘, *hårbeklædt* ‚haarbedeckt‘, *prærieområder* ‚Präriegebiete‘ oder *ungskov* ‚Jungwald‘ im Querkorpus DDOC vor, was als ein weiteres Indiz für ihre mangelnde Zweckmäßigkeit in den Bedeutungsparaphrasen gewertet werden könnte; und es ließe sich daher erwägen, diese durch geläufigere Formen zu ersetzen wie *som ligner en baver* ‚der/die/das einem Biber ähnelt‘, *dækket af hår* ‚von Haaren bedeckt‘, *aktiv om natten* ‚aktiv nachts‘, *prærier* ‚Prärien‘ und *ungskov* ‚junger Wald‘. Ein hierdurch eventuell erforderliches höheres Maß an syntaktischer Komplexität sowie hieraus resultierende Auswirkungen auf die Rezipierbarkeit der Bedeutungsparaphrasen müssten Gegenstand einer gesonderten Untersuchung sein.

Die niederfrequenten Types in den Ergebnissen des MI-Verfahrens sowie auch Formen mit einem zu niedrigen Score lassen sich mithilfe quantitativer Filter entfernen.² Niederfrequente Types lassen sich etwa mit der heuristisch ermittelten Frequenzschwelle $\log d+1$ eliminieren, wo d die Anzahl der Bedeutungsparaphrasen mit dem untersuchten GP ist. Ist die absolute Frequenz eines kookkurrierenden Types höher als der Schwellenwert, wird das Type bei der Ermittlung der hervortretenden Types weiter berücksichtigt, andernfalls nicht. Entsprechend lassen sich Types mit einem zu niedrigen Score durch die Schwelle $2 \cdot (\log d+1)$ eliminieren: hierbei werden nur Types berücksichtigt, deren Score über diesem Schwellenwert liegt. Tabelle 4 zeigt sämtliche so ermittelte Types in den Bedeutungsparaphrasen mit dem GP *gnaver* ‚Nagetier‘. Ein Interpretationsansatz hierzu wäre, dass die durch die aufgelisteten

BP-Type	Äquivalent	Score
ører	‚Ohren‘	342
hale	‚Schwanz‘	194
pels	‚Fell‘	140
gråbrun	‚graubraun‘	127
korte	‚kurze‘	55
lever	‚lebt‘	53
lang	‚lang‘	52
forholdsvis	‚verhältnismäßig‘	24
kort	‚kurz‘	14
lille	‚klein‘	7

Tab. 4: Hervortretende BP-Types für *gnaver* ‚Nagetier‘ nach Filterung

² Die Verwendung eines statistischen Tests wie *Log Likelihood* anstatt MI wäre eine weitere Möglichkeit.

Types bezeichneten Merkmale zentral bei der Paraphrasierung von Bedeutungen mit dem GP *gnaver* ‚Nagetier‘ sind.

Es ließe sich nun für jede dieser Bedeutungsparaphrasen untersuchen, wie konsequent diese Merkmale erwähnt werden und in welchen syntaktischen Strukturen sie eingebettet sind: mangelnde Konsequenz ließe sich hierauf redaktionell bereinigen im Hinblick auf eine Vereinheitlichung des Aufbaus der Bedeutungsparaphrasen. So optimierte Bedeutungsparaphrasen ließen sich daraufhin besser algorithmisch behandeln. Eine solche Optimierung sei an einem weiteren Beispiel kurz aufgezeigt. Die Spitze der Liste der hervortretendsten Types in den Bedeutungsparaphrasen mit dem GP *kvinde* ‚Frau‘ ist in Tabelle 5 wiedergegeben.³

BP-Type	Äquivalent	Score
arrig	‚boshaft‘	310
prostitueret	‚prostituiert‘	241
stridbar	‚streitbar‘	233
gift	‚verheiratet‘	94
husholdning	‚Haushalt‘	93
pige	‚Mädchen‘	91
tiltrækkende	‚anziehend‘	78
parforhold	‚Zweierbeziehung‘	67
smuk	‚hübsch‘	53
gaden	‚die Straße‘	52

Tab. 5: Die zehn hervortretendsten BP-Types für *kvinde* ‚Frau‘

Auf der Grundlage einer solchen Tabelle lassen sich Suchanfragen an das Bedeutungsparaphrasenkörpus formulieren, z. B. nach sämtlichen Bedeutungsparaphrasen, die das GP *kvinde* ‚Frau‘ und das DS-Type *prostitueret* ‚prostituiert‘ enthalten. Die hieraus resultierende KWIC-Konkordanz ist in Abbildung 4 dargestellt.

Da jede BP im Korpus mit der Angabe des zu ihr gehörigen BP-Ausdrucks und des Lemmas annotiert ist, gestaltet sich eine solche Anfrage ziemlich einfach.⁴ In der wiedergegebenen Konkordanz ist die Annotation *Lemma* mit

³ Die zehn hervortretendsten BP-Types des GP *mand* ‚Mann‘ sind übrigens *forfører* ‚verführt‘, *homoseksuel* ‚homosexuell‘, *uopdragen* ‚unerzogen‘, *kone* ‚Ehefrau‘, *adelig* ‚adelig‘, *uforskammet* ‚unverschämt‘, *parforhold* ‚Zweierbeziehung‘, *dreng* ‚Junge‘, *gift* ‚verheiratet‘ und *tiltrækkende* ‚anziehend‘. Interessant wäre auf der Grundlage solcher quantitativ ermittelten Ergebnisse das in der Lexis verankerte Weltbild einmal näher zu untersuchen. In diesem Zusammenhang wäre ein Vergleich des Weltbildes in den redaktionell erstellten Bedeutungsparaphrasen mit einer entsprechenden quantitativen Untersuchung der dem Wörterbuch zugrunde liegenden Korpusdaten interessant im Hinblick auf die Frage, ob Wörterbuchredakteure das im Korpus vermittelte sprachliche Weltbild unmittelbar oder doch eher introspektiv in den Bedeutungsparaphrasen des Wörterbuchs vermitteln.

⁴ Als Anfragesystem wird die ursprünglich am Institut für Maschinelle Sprachverarbei-

angezeigt. Es fällt unmittelbar auf, dass eine Reihe von Bedeutungsparaphrasen nach folgendem Muster aufgebaut sind:

(8)

$[[_{DS} \text{ prostitueret }][_{GP} \text{ kvinde }][_{DS} \dots]]_{BP}$

≈ prostituierte Frau ...

[genus="kvinde" & norm="prostitueret"]			
lemma	left context	match	right context
kurtisane		Elegant prostitueret	kvinde med dannelse god smag og g
hore	Kvinde som bedriver hor	prostitueret	kvinde
demimonde	Kvinde der helt el delvist ernærede sig som	prostitueret	
massøse		Prostitueret	kvinde på en massageklinik
hetære		Prostitueret	kvinde i antikkens Grækenland ofte r
glædespige		Prostitueret	kvinde
gadetøs		Prostitueret	kvinde der trækker på gaden
gadepige		Prostitueret	kvinde der trækker på gaden
fruentimmer		Prostitueret	kvinde
callgirl		Prostitueret	kvinde der tilkaldes telefonisk ofte sc
bajadere		Prostitueret	kvinde

Abb. 4: BP-Konkordanz

Insgesamt gibt es im Wörterbuch elf Lemmata mit dieser Lesart, bei denen das unter (8) gezeigte BP-Muster verwendet wird.

Bei näherer Untersuchung der zu den weggefilterten DS-Types gehörenden Lemmata zeigt sich, dass unter ihnen weitere mit der Lesart ‚prostituierte Frau‘ auftreten z. B. *ludder* und *skøge* – beide überdies sogar monosem. Die zu diesen beiden Beispielen gehörenden Bedeutungsparaphrasen sind unter (9) wiedergegeben: deutlich zeigt sich, dass hier zwei Abweichungen vom Muster unter (8) vorliegen.

(9)

a. *ludder* ‚Nutte‘:

$[[_{GP} \text{ kvinde }][_{DS} \text{ som tilbyder samleje og andre seksuelle ydelser mod betaling }]]_{BP}$

≈ Frau, die Geschlechtsverkehr und andere sexuelle Leistungen gegen Bezahlung anbietet

b. *skøge* ‚Dirne‘:

$[[_{GP} \text{ kvinde }][_{DS} \text{ der ernærer sig ved prostitution }]]_{BP}$

≈ Frau, die sich durch Prostitution ernährt

Somit eignet sich dieses quantitative Verfahren zum Aufdecken von Unregelmäßigkeiten in der Struktur von an sich verwandten Bedeutungsparaphrasen.

tung der Universität Stuttgart entwickelte IMS Corpus Workbench (CWB) verwendet, vgl. Christ 1994. Eine Open-Source-Version hiervon wird unter <http://cwb.sourceforge.net/> zugänglich gemacht.

Eine auf dieser Grundlage systematisch durchgeführte Optimierung der Bedeutungsparaphrasen – möglicherweise unter Zuhilfenahme weiterer korpuslinguistischer Verfahren ließe sich somit für eine Vereinheitlichung der Bedeutungsparaphrasenstruktur im Wörterbuch verwenden. Solchermaßen optimierte oder standardisierte Bedeutungsparaphrasen erleichtern möglicherweise die Rezeption durch den menschlichen Benutzer, vor allem ermöglichen sie aber eine zuverlässigere algorithmische Auswertung als die nicht optimierten Bedeutungsparaphrasen, die in aller Regel in herkömmlichen Wörterbüchern anzutreffen sind. Von daher wäre ein so bearbeitetes Wörterbuch in einem digitalen Kontext wahrscheinlich funktionstüchtiger als die im Abschnitt 2.1 vorgestellten digitalen Versionen gedruckter Wörterbücher.

6. Ausblick

Optimierungen der Bedeutungsparaphrasen, wie sie hier ansatzweise beschrieben wurden, würden zwar das Wörterbuch als Grundlage onomasiologischer Anfragen zuverlässiger machen, dennoch würden weiterhin Unwägbarkeiten der freien Formulierung der Bedeutungsparaphrasen fortbestehen, die eine algorithmische Nutzung des Materials weiterhin relativ schwierig gestalten würden. Es scheint somit ein nur schwer zu überwindender Widerspruch zu bestehen zwischen den Bedürfnissen menschlicher Wörterbuchbenutzer, die rezipierbare Bedeutungsparaphrasen erwarten, und algorithmischen Systemen, die eine schlüssige, logische Struktur der semantischen Beschreibung im Wörterbuch erfordern. Diese Diskrepanz ließe sich zum Teil überwinden, wenn auf der Grundlage korpuslinguistischer Analysen der Bedeutungsparaphrasen zunächst einmal eine eher formal strukturierte lexikalisch-semantische Ressource entwickelt würde, die sich mit den semantischen Beschreibungen im bestehenden Wörterbuch verknüpfen ließe. Dieses Ziel verfolgt das zunächst bis Ende 2008 laufende Projekt *DanNet*, vgl. Pedersen u. a. (2006), das als Kooperation zwischen der DSL und dem *Center for Sprogteknologi, CST*, ein WordNet für das Dänische entwickelt, vergleichbar etwa dem Princeton WordNet, vgl. Fellbaum (1998), oder den verschiedenen EuroWordNets, vgl. Vossen (1999), jedoch ergänzt durch Angaben zur Qualiastruktur, vgl. Pustejovsky (1996). Die WordNet-Entwicklung erfolgt auf der Grundlage des DDO und unter Einsatz der in diesem Beitrag geschilderten Verfahren. Da dieses WordNet unmittelbar vom DDO abgeleitet wird, kann es als eine algorithmisch verwertbare Parallelversion zum Wörterbuch betrachtet werden, und ist als solche auch unmittelbar mit ihm verknüpft, indem die aus ihm abgeleiteten Synsets unmittelbar mit entsprechenden Bedeutungsparaphrasen im DDO verbunden sind. Auf dieser Grundlage dürften die angestrebten onomasiologisch orientierten Recherchen in der künftigen digitalen Version des DDO von einer besseren Qualität sein als solche, die sich ausschließlich auf eine optimierte semantische Beschreibung im Wörterbuch stützen würden.

Ein weiterer Vorteil dieses Verfahrens ist, dass zunächst nicht in die bestehende semantische Beschreibung im DDO eingegriffen wird. Gleichzeitig ermöglicht es eine gründliche Analyse und formale Umsetzung der in den Bedeutungsparaphrasen enthaltenen Informationen.

So wird die Bereinigung von Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur zunächst einmal nur im Rahmen des WordNets durchgeführt. Erst wenn hier eine akzeptable Hyponymiestruktur erreicht worden ist, wird auf dieser Grundlage dann in einem weiteren Schritt die Anpassung der GP in den Bedeutungsparaphrasen des Wörterbuchs durchgeführt. In dem Prozess der Bereinigung der Hyponymiestruktur werden die GP, die im DDO sowie in diesem Beitrag als GP bezeichnende Ausdrücke verstanden sind (vgl. Abschnitt 3.2.1) durch die von ihnen bezeichneten *Inhalte* ersetzt, was das Problem synonymer und polysemer GP-Ausdrücke löst. Auch qualitativ unterschiedliche Hyperonymierelationen werden entsprechend markiert, vgl. hierzu Pedersen u. Sørensen (2006).

Auf der Grundlage der DS-Analysen, z. B. der quantitativen Ermittlung hervortretender DS-Angaben, wird nicht nur eine Umsetzung der semantischen Information für DanNet vorgenommen, sondern als weiteres Ziel auch eine systematische Beschreibung der Mikrostruktur von Bedeutungsparaphrasen überhaupt angestrebt. Somit ließen sich unterschiedlich aufgebaute Bedeutungsparaphrasen bereits bei der Ausarbeitung eines Wörterbuchs weitgehend vermeiden. Eine solche systematische Strukturbeschreibung von Bedeutungsparaphrasen könnte in Kombination mit einer WordNet-Struktur wiederum die Grundlage für eine formale Beschreibung der lexikalischen Semantik hinter den eigentlichen Bedeutungsparaphrasen bilden – eine Beschreibung, die letztendlich anstelle der klassischen Bedeutungsparaphrasen treten könnte: diese würden dann je nach Bedarf aus den formalen Beschreibungen abgeleitet werden. Somit ließe sich auch die sprachreflexive Sprache der Bedeutungsparaphrasen zunächst einmal vermeiden. An ihrer Stelle würde dann intern eine eigentliche semantische Metasprache treten, die auf den jeweiligen Verwendungszweck des Wörterbuchs ausgerichtet in eine adäquate sprachreflexive Sprache umgesetzt werden könnte. So ließe sich auch das Problem, dass der sprachliche Aufbau und Ausdruck in den Bedeutungsparaphrasen eines Wörterbuchs möglicherweise nur ein zeitlich begrenztes Modephänomen darstellt, leichter umgehen.

Die in diesem Beitrag lediglich angerissenen Themen dürften einen Eindruck davon vermittelt haben, dass die Verwendung von Korpora und korpuslinguistischen Verfahren in der Lexikographie nicht nur einen Erkenntnisfortschritt für denjenigen Teil der lexikographischen Arbeit bedeuten, wo ein Korpus als Quelle für die lexikographische Beschreibung verwendet wird, sondern auch dort, wo sie als Werkzeug bei der lexikographischen Arbeit mit dem Wörterbuch selbst zum Einsatz kommen.

Ein digitales Wörterbuch, das wirklich die Möglichkeiten des Mediums Computer ausschöpfen will, sollte in seiner Konzeption eine multifunktionale

lexikalische Ressource sein, die sowohl in unterschiedlichen Buchversionen erscheinen kann als auch als digitales Nachschlagewerk mit unterschiedlichen Recherchemöglichkeiten – und die sich überdies auch für NLP-Zwecke verwenden ließe.

Literatur

- Asmussen, Jörg (2003): Zur geplanten Retrodigitalisierung des Ordbog over det danske Sprog. Konzeption, Vorgehensweise, Perspektiven. In: Burch, Fournier, Gärtner u. Rapp (Hg.) 2003: Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2000 (Abhandlungen der Akademie der Wissenschaften und der Literatur). Mainz.
- Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX'94. Budapest.
- Church, Kenneth u. Hanks, Patrick (1989): Word association norms, mutual information and lexicography. In: ACL Proceedings, 27th Annual Meeting. Vancouver.
- COBUILD: Sinclair, John u. a. (Hg.) (1987): Collins COBUILD English Language Dictionary. London.
- DDO: Hjorth, Kristensen, Lorentzen, Trap-Jensen, Asmussen u. a. (Hg.) (2005): Den Danske Ordbog 1–6. København.
- Dunning, Ted (1994): Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics 19(1), S. 61–74.
- DUW: Wermke, Matthias u. a. (Hg.) (2003): Duden Deutsches Universalwörterbuch. 5. überarbeitete Auflage. Mannheim.
- Fellbaum, Christiane (Hg.) (1998): WordNet: An Electronic Lexical Database. Cambridge, MA.
- Ide, N. u. Véronis, J. (1994): Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. In: Steffens, P. (Hg.): Machine Translation and the Lexicon. Springer-Verlag.
- Kennedy, Graeme (1998): An Introduction to Corpus Linguistics. London.
- Kilgariff, Adam u. Tugwell, David (2002): Sketching Words. In: Corréard, M.-H. (Hg.): Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins. Euralex.
- Klein, Wolfgang (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: Scharnhorst J.: Sprachkultur und Lexikographie. Berlin.
- LDOCE: Bullon, Stephen u. a. (Hg.) (2003): *Longman Dictionary of Contemporary English*. Pearson Education.
- Lemnitzer, Lothar u. Zinsmeister, Heike (2006): Korpuslinguistik. Tübingen.
- Lorentzen, Henrik (2004): The Danish Dictionary at large: presentation, problems and perspectives. In: Proceedings of the 11th EURALEX International Congress. Lorient.
- Lyons, John (1977): Semantics. Cambridge.
- McEnery, Tony u. Wilson, Andrew (2001): Corpus Linguistics. Edinburgh.
- MED: Rundell, Michael u. a. (Hg.) (2002): Macmillan English Dictionary for Advanced Learners. Oxford.
- Norling-Christensen, Ole u. Asmussen, Jörg (1998): The Corpus of The Danish Dictionary. In: Lexikos. Afrilex Series 8.
- ODS: Dahlerup, Verner u. a. (Hg.) (1956): Ordbog over det danske Sprog 1–28. København.
- Pedersen, Bolette S., Nimb, Sanni, Asmussen, Jörg u. a. (2006): DanNet – a WordNet for Danish. In: Proceedings of the Third International WordNet Conference. Jeju, Korea.

- Pedersen, Bolette S. u. Sørensen, Nicolai H. (2006): Towards Sounder Taxonomies in Word-Nets. In: Proceedings from the OntoLex Workshop in association with LREC 2006. Genova.
- Pustejovsky, James (1996): The Generative Lexicon. Reprint. Bradford Book.
- Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.
- Storjohann, Petra (2005): *elexico: A Corpus-Based Monolingual German Dictionary*. In: *Hermes, Journal of Linguistics* 34. Aarhus.
- Tognini Bonelli, Elena (2001): *Corpus Linguistics at Work*. Amsterdam/Philadelphia.
- Vossen, Pick (Hg.) (1999): *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*. Amsterdam.
- Widdows, Dominic (2003): *Geometry and Meaning. Center for the Study of Language and Information – Lecture Notes (CSLI-LN)*. Chicago.
- Wiegand, H. E. (1989a): Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Hausmann, Reichmann, Wiegand, Zgusta: *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin/New York.
- Wiegand, H. E. (1989b): Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Reichmann, Wiegand, Zgusta: *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin/New York.

GUNTER SENFT

„Ich weiß nicht, was soll es bedeuten ...“ –
Ethnolinguistische Winke zur Rolle von umfassenden
Metadaten bei der (und für die) Arbeit mit Corpora

Abstract

Arbeitet man als muttersprachlicher Sprecher des Deutschen mit Corpora gesprochener oder geschriebener deutscher Sprache, dann reflektiert man in aller Regel nur selten über die Vielzahl von kulturspezifischen Informationen, die in solchen Texten kodifiziert sind – vor allem, wenn es sich bei diesen Daten um Texte aus der Gegenwart handelt. In den meisten Fällen hat man nämlich keinerlei Probleme mit dem in den Daten präsupponierten und als allgemein bekannt erachteten Hintergrundwissen. Betrachtet man dagegen Daten in Corpora, die andere – vor allem nicht-indoeuropäische – Sprachen dokumentieren, dann wird einem schnell bewusst, wieviel an kulturspezifischem Wissen nötig ist, um diese Daten adäquat zu verstehen. In meinem Beitrag illustriere ich diese Beobachtung an einem Beispiel aus meinem Corpus des Kilivila, der austronesischen Sprache der Trobriand-Insulaner von Papua-Neuguinea. Anhand eines kurzen Ausschnitts einer insgesamt etwa 26 Minuten dauernden Dokumentation, worüber und wie sechs Trobriander miteinander tratschen und klatschen, zeige ich, was ein Hörer oder Leser eines solchen kurzen Daten-Ausschnitts wissen muss, um nicht nur dem Gespräch überhaupt folgen zu können, sondern auch um zu verstehen, was dabei abläuft und wieso ein auf den ersten Blick absolut alltägliches Gespräch plötzlich für einen Trobriander ungeheuer an Brisanz und Bedeutung gewinnt. Vor dem Hintergrund dieses Beispiels weise ich dann zum Schluss meines Beitrags darauf hin, wie unbedingt nötig und erforderlich es ist, in allen Corpora bei der Erschließung und Kommentierung von Datenmaterialien durch sogenannte Metadaten solche kulturspezifischen Informationen explizit zu machen.

1. Einleitung

Arbeitet man als muttersprachlicher Sprecher des Deutschen mit Corpora gesprochener oder geschriebener deutscher Sprache, dann reflektiert man in aller Regel nur selten über die Vielzahl von kulturspezifischen Informationen, die in solchen Texten kodifiziert sind – vor allem, wenn es sich bei diesen Daten um Texte aus der Gegenwart handelt. In den meisten Fällen hat man nämlich keinerlei Probleme mit dem in den Daten präsupponierten und als allgemein bekannt erachteten Hintergrundwissen. Mit der Idee, dass man die Dinge, die wir in unseren Kulturen als selbstverständlich und gegeben betrachten, auch aus anderer Perspektive sehen kann, hat Montesquieu im Jahre 1721 mit seinen „Persischen Briefen“ (*“Lettres persanes”*) eine eigenständige

literarische Tradition begründet. Die Grundidee von Montesquieus Buch und vieler Nachfolgepublikationen besteht darin, dass ein Autor Briefe, Tagebücher, Reisebeschreibungen oder Berichte von mehr oder weniger „exotischen“, d. h. nicht aus Europa kommenden Fremden publiziert, die verschiedene Aspekte der europäischen Kultur und Zivilisation aus ihrer eigenen – für uns exotischen oder ethnozentrischen Perspektive betrachten. Die Autoren dieser Publikationen behaupten, dass sie nur die Funktion von Übersetzern und Herausgebern der von ihnen veröffentlichten Materialien übernommen hätten – und holen damit natürlich die poetische Lizenz ein, die ihnen notwendig erscheint, um sich vor etwaigen Sanktionen der größtenteils kritischen Aufzeichnungen, die von diesen exotischen Fremden stammen, zu schützen. Ein in Deutschland vielbeachteter Beitrag zu dieser literarischen Tradition war die 1912/13 von Hans Paasche publizierte Briefesammlung mit dem Titel „Die Forschungsreise des Afrikaners Lukanga Mukara ins innerste Deutschland“. Das Plagiat dieser kritischen Briefesammlung, Erich Scheurmanns 1920 erstmals publizierte Werk mit dem Titel „Der Papalagi – Die Reden des Südseehäuptlings Tuiavii aus Tiavea“ gilt heute noch in manchen Kreisen als Kultbuch und hat es sogar geschafft, 1982 in den Lektürekanon für den Deutschunterricht an Gymnasien aufgenommen zu werden (vgl. Senft 1999).¹ Offenbar bereitet es uns ein gewisses Maß an Vergnügen, wenn wir literarisch gezwungen werden, uns aus anderer Perspektive zu betrachten.

Arbeitet man dagegen mit Daten in Corpora, die andere – vor allem nicht-indoeuropäische – Sprachen dokumentieren, dann wird einem schnell bewusst, wieviel an kulturspezifischem Wissen nötig ist, um diese Daten adäquat zu verstehen. Hier gewinnt man allenfalls Vergnügen daraus, dass man mit einer sorgfältigen und ausführlichen Beschreibung der Daten – mit sogenannten Metadaten – einen Text für nicht mit dieser jeweiligen Sprache und Kultur Vertraute so weit wie möglich verständlich erschließen kann. Im folgenden werde ich das kurz an einem Beispiel aus meinem Corpus des Kilivila, der austronesischen Sprache der Trobriand-Insulaner von Papua-Neuguinea illustrieren.

2. Zur Bedeutung von Metadaten für die Erschließung von Sprachdaten

Seit 24 Jahren untersuche ich die Sprache und Kultur der Trobriand-Insulaner in Papua-Neuguinea. Während meiner Feldforschungen lebe ich im Dorf Tauwema auf der Insel Kaile'una, die der Hauptinsel Kiriwina im Nordwesten vorgelagert ist. Im Verlauf meiner Feldforschung im Jahre 2003 habe

¹ Einen Überblick über diese literarische Tradition gibt Weißhaupt (1979). Ein jüngeres literarisches Beispiel, das dieser Tradition folgt, liefert Rosendorfer (1983). Einen authentischen Bericht eines Chinesen, der Frankreich in der Zeit von 1722 bis 1725 bereist hat, findet man in Spence (1988).

ich unter anderem ein Gespräch von sechs Männern gefilmt, die auf der Veranda meines Hauses saßen. Es wurde dabei schnell deutlich, dass sie sich über nichts Besonderes unterhielten, sondern einfach nur so miteinander tratschten. Bei den sechs Männern handelt es sich um meinen 25 Jahre alten Nachbarn Yabilosi, meinen 32 Jahre alten Freund Moagava und seinen 26 Jahre alten Bruder Keyeba – (diese drei Männer sind Mitglieder des Lukwasisiga Clans) – den 63 Jahre alten Yoya, den 34 Jahre alten Sose'ula und seinen 26 Jahre alten Neffen Mokeimwena (diese drei Männer gehören dem Malasi Clan an). Wenn man das Gespräch insgesamt betrachtet, so kann man feststellen, dass Yabilosi und Moagava sich am aktivsten an diesem Gespräch beteiligen. Beide gehören dem Lukwasisiga Clan an, dem an dritter Stelle rangierenden Clan in der trobriandischen Clanhierarchie (in der der Malasi Clan am bedeutendsten ist, gefolgt von dem Lukuba Clan, dem Lukwasisiga Clan und schließlich dem Lukulabuta Clan). Die Tatsache, dass Männer aus einem niederrangigen Clan mehr Gesprächsanteile haben, zeigt, dass beim Tratsch und Klatsch, den die Trobriander selbst metalinguistisch als „*kasilam*“ bezeichnen, soziale Barrieren und Schranken irrelevant sind.

Das Exzerpt aus diesem Gespräch, das ich hier vorstellen möchte, dokumentiert eine relativ kurze Interaktion von etwa zwei Minuten² nach etwa 20 Minuten vorangegangenen Klatsch und Tratsch. Während dieser ersten 20 Minuten sprachen die 6 Männer über die Arbeit im Garten, über Yamsknollen, über verschieden Arten von Fischen, über Betelnüsse, über eine „*sagali*“ genannte Trauerfeier (vgl. Senft 1985a; Weiner 1976) im Nachbardorf Koma, über Hausbau, über Rauchen und Tabak, über meine Arbeit, über Pläne für eine Party am Strand, über notwendige Renovierungsarbeiten am Haus, das als Vorschule genutzt wird, über das Regenwetter während der Trauerfeier in Koma, über Spannungen zwischen Einwohnern bestimmter Dörfer auf Kaile'una, über eine unerwartete Begegnung mit einem böartigen Totengeist (*kosi*), über Papageien, die die Gärten verwüsten, und über verschiedene Versuche, sich das Rauchen abzugewöhnen. Dann berichtet Keyaba das Folgende³:

- 1 Keyeba: Meine Freunde, gestern haben die Leute von Kuiava Tabakblätter gebracht, ihr werdet nicht glauben wie viele!
- 2 Moagava: Wirklich!

² Diese Interaktion beginnt auf dem digitalisierten Video-Band bei der Zeitanzeige von 20:29 Minuten und endet bei der Zeitanzeige von 22:20 Minuten.

³ Das Exzerpt besteht aus 24 ‚turns‘ die ich im Transkript numeriert habe. Ich gebe hier zunächst einmal nur die deutschen Glossen der Konversation im Kilivila an, um dem Leser einen ersten Eindruck davon zu geben, worum es hier überhaupt geht. Es versteht sich von selbst, dass die Analyse dieser Interaktion auf einer morpheminterlinearisierten Transkription basiert. Ich möchte darauf hinweisen, dass ich für die hier verfolgten Zwecke die Daten nicht nach den Konventionen der Konversationsanalyse präsentiere (vgl. z. B. Atkinson, Heritage 1984), und dass meine Kommentare und Anmerkungen ebenfalls nicht den dort geltenden Standards folgen. Ich gebe allerdings die Länge von Pausen (in Sekunden in Klammern) an.

- 3 Keyeba: Ein Kina das Blatt.
- 4 Moagava: (Sie sind) dick und groß, (2.0) ok (3.0) und trotzdem (sind sie) leicht.
- 5 Yabilosi: Ich (werde) Blumen im alten Garten (suchen gehen) wenn ich nicht ein Tabakblatt für mich kaufen kann!
- 6 Keyeba: Ein Kanu – hm, wie ist der Name – von den Bergen (von den D’Entrecasteaux Inseln) war (auch) angekommen. (2.0)
- 7 Yoya: (Von den) Bergen? (2.0)
- 8 Keyeba: Ja, (4.0), sie wollen Betelnüsse haben, sie wollen sie (bei einem Trauerritual) verteilen (und deshalb) wollten sie (sie) kaufen (10.0).
- 9 Sose’ula: Ich verstehe, hier gibt es Betelnüsse, man könnte sagen ein Trauerritual ohne Betelnüsse, (das ist schon schlimm) genug, (aber selbst) die Lebensmittel sind ausgegangen.
- 10 Mokeimwena: Ah nein, (weißt) Du, sie sind hungrig, (sie haben) nichts zu Essen.
- 11 Umstehender: Oh bei Gott, aber...
- 12 Mokeimwena: Sie sind gekommen...
- 13 Umstehender: ... sie haben in den Gärten gearbeitet!
- 14 Mokeimwena: ... Sie sind gekommen und haben zwei Dinge gesucht, (2.0) das eine waren Lebensmittel, (1.0) und das andere waren Betelnüsse.
- 15 Moagava: Wer?
- 16 Mokeimwena: Diese Leute von Kuiava sind gekommen. (4.0)
- 17 Yabilosi: Gib mir Streichhölzer!
- 18 Keyeba: Es, es fuhr ganz schnell dieses Dinghy, es kam und wollte nach Losuia fahren, es fuhr wirklich schnell, es kam an unser Dorf – es blieb stehen – kein Zoom (mehr), (1.0), sie kletterten auf die Betelpalmen in unserem Dorf und fuhren zurück (1.0), aber es kam und wollte zu Vapalaguyau auf Tuma fahren, (2.0), sie wollten (dort) auf die Betelpalmen klettern und es sollte eine Ladung von Betelnüssen aufnehmen und zurückfahren, (3.0), aber nein, es war alle das Zoom, und sie kamen zu unserem Dorf zurück. (7.0).
- 19 Mokeimwena: Dieser Regen da gestern, er hat diesen starken Regen gemacht während dieser Trauerfeier, (1s) gut, das war’s.
- 20 Moagava: Wirklich, sie haben sich schlecht benommen.
- 21 Mokeimwena: Das ist schlecht, heute brennt die Sonne (und) sie werden die Trauerfeier in unserem Dorf heute beenden.
- 22 Yoya: Sie werden (sie) beenden.
- 23 Sose’ula: Aber gestern haben sie beinahe gestritten (weil) es geregnet hat.
- 24 Keyeba: Mein Freund, heute werden wir bestimmt nicht nach Koma gehen, gleich arbeiten wir (und) dann wird es beendet sein.

Auf den ersten Blick erscheint das, was hier gesagt wird, ziemlich harmlos zu sein. Wenn man aber etwas genauer betrachtet, was hier passiert, dann wird schnell deutlich, dass dieser kurze Gesprächsausschnitt etwas absolut Unerhörtes dokumentiert – nämlich die allerschlimmste Beleidigung, die man auf den Trobriand-Inseln aussprechen kann. Aber bevor ich diesen Gesprächsausschnitt analysiere, gilt es zunächst, den Anfang der hier vorgestellten Konversation in die indigene trobriandische Typologie von Varietäten bzw. Sprach-Registern und die sie konstituierenden Genres einzuordnen.

Die Trobriander unterscheiden 4 wechselseitig verstehbare lokale Varietäten oder Dialekte des Kilivila (vgl. Senft 1986, S. 6–10); diese 4 Dialekte werden mit metalinguistischen Bezeichnungen voneinander unterschieden. Der Dialekt, den die Einwohner von Tauwema (und die Einwohner aller anderen Dörfer auf Kaile'una mit der Ausnahme der Leute von Kaduwaga) sprechen – und den ich beschrieben und gelernt habe, wird „*biga galawala*“ genannt⁴. Die hier vorgelegten Sprachdaten repräsentieren diese lokale Varietät des Kilivila.

Neben den 4 Dialekten des Kilivila unterscheiden die Trobriander aber auch noch spezielle Register, die ich „situations-intentionale Varietäten“ genannt habe (vgl. Senft 1986, S. 124–126; 1991). Mit diesem Terminus verweise ich darauf, dass im Kilivila bestimmte Register in bestimmten Situationen mit speziellen Intentionen gebraucht werden. Die muttersprachlichen Sprecher des Kilivila unterscheiden und bezeichnen metalinguistisch 8 solche Varietäten.

Darüberhinaus unterscheiden die Trobriander auch eine Reihe von Genres oder Textsorten, die für diese Varietäten konstitutiv sind. Diese Genres sind ebenfalls metalinguistisch bezeichnet. Eines dieser Genres ist „Tratsch“ oder „Klatsch“ – „*kasilam*“. Und „*kasilam*“ ist eine von mehreren Textsorten, die die sogenannte „*biga sopa*“, die „spaßige, lügenhafte, indirekte, uneigentliche, spielerische Sprachvarietät“ des Kilivila ko-konstituiert.

Diese für die Redeweise der Trobriander so charakteristische Varietät macht sich vor allem zwei der jede natürliche Sprache charakterisierenden Merkmale zu Nutze, nämlich „Vagheit“ und „Ambiguität“. Damit bietet sie ihren Sprechern ein Stilmittel, mit denen sie ihnen erlaubt, in bestimmten Sprechsituationen Konfrontation und zu aggressiv wirkende Direktheit zu vermeiden. Signalisiert nämlich ein Hörer, dass er sich durch einen Redebei-

⁴ Bei den anderen Dialekten handelt es sich um die „*biga galagoki*“, die „*biga besagala*“ und die „*biga galanani*“. Unterscheidungskriterium dieser Einteilung ist eine jeweils der Bezeichnung der Varietät (*biga* = Sprache) genannte Art von Shibboleth: Es handelt sich dabei um die unterschiedliche Art und Weise der Verneinung. Das Wort „*gala*“ bedeutet in allen Varietäten des Kilivila „nein, nicht“. Es ist für die Sprecher der Dialekte charakteristisch, dass sie – neben einer Reihe anderer phonologischer Unterschiede ihrer Varietäten untereinander – dem Negationswort „*gala*“ eine in dieser Verbindung semantisch bedeutungsleere Partikel wie „*besa*“ voranstellen bzw. wie „*goki*“, „*nani*“ und „*wala*“ anfügen (vgl. Senft 1985b, S. 128 f.).

trag getroffen fühlt, dann kann der Sprecher sich immer dadurch „aus der Affäre“ ziehen, dass er das Gesagte als „*sopa*“ verstanden wissen will, als etwas, was er ja „eigentlich“ nicht gemeint hat. Nach trobriandischer Etikette darf der Hörer sich dann nicht durch die explizit als „*sopa*“ bezeichneten Äußerungen beleidigt fühlen – sonst verliert er sein Gesicht! Die Trobriander nutzen diese Varietät im alltäglichen Gespräch, beim Klatsch und Tratsch, beim Flirten, in öffentlichen Debatten, in Mahnreden, in Liedern und beim Erzählen von Geschichten als rhetorisches Mittel zur Vermeidung von Konflikten und um die allgemeine Gesprächsatmosphäre zu entspannen. Mit der „*biga sopa*“ Varietät verfügen die Trobriander über ein Mittel, mit dem sie sich einen Raum, ein Forum zum sprachlichen Handeln auf Probe ohne Furcht vor Sanktionen schaffen (vgl. Senft 1987, S. 123). Darüber hinaus wird die „*biga sopa*“ auch dazu genutzt, um Leute entweder zu verspotten oder sie ironisch-satirisch zu kritisieren. Diese Kritik bezieht sich dann in aller Regel auf in der trobriandischen Gesellschaft als abweichend geltendes Verhalten und fordert die so Angesprochenen auf relativ freundliche Art und Weise auf, dieses Verhalten zu ändern. Schließlich gewährt die „*biga sopa*“ Varietät die Lizenz dafür, im verbalen Bereich fast alle Tabus zu brechen – mit Ausnahme der 6 übelsten und – für den Beleidiger im wahrsten Sinne des Wortes tödlichen – Beleidigungen, die im Kilvila möglich sind. Diese sechs Beleidigungen gelten immer noch als unentschuldigbar – und ein damit Beleidigter hat das Recht, den Beleidiger zu töten. Von daher liegen diese 6 Beleidigungen – im Gegensatz zu anderen Flüchen und Beschimpfungen, deren Gebrauch innerhalb der „*biga sopa*“ durchaus noch lizenziert ist – außerhalb des Bereiches, den dieses Register umfasst. Sie werden eindeutig der „*biga gaga*“ Varietät, der „bösen, situationsinadäquaten Sprache“ zugerechnet.

Nach diesen ersten Vorbemerkungen betrachte ich im folgenden das gerade angeführte Gesprächsexzerpt genauer – und zwar auf der Basis der dafür angemessenen morphem-interlinearisierten Transkription: Das Exzerpt beginnt mit einer im Kontext völlig neuen Referenz auf Dritte:

1 Keyeba:

<i>So-(gwe)</i>	<i>lova</i>	<i>minu</i>	<i>Kuiava</i>	<i>e-meye-si</i>
Freunde-(meine)	gestern	Leute.von	Kuiava.(Insel)	3.-bringen-Pl
Meine Freunde, gestern brachten die Leute von Kuiava				

<i>tombaiku</i>	<i>gala</i>	<i>ku-doki</i>	<i>ymalama!</i>
Tombaiku-Blätter	nicht	2.-denken	viele
Tombaiku-Blätter, Ihr glaubt nicht wieviele!			

2 Moagava:

Mokwita!

Wirklich!

3 Keyeba:

<i>One</i>	<i>Kina</i>	<i>ya-tala.</i>
Ein	Kina	KP.flexibel-eins
Einen Kina das Blatt.		

4 Moagava:

<i>Ya-popou</i>	<i>ya-vaka-veaka</i>	(2.0)	<i>bogwa</i>	(3.0)	<i>e</i>	<i>taga</i>
K.P.flexibel-dick	K.P.flexibel-Redup-groß		schon		und	aber

(Sie sind) dick und groß, (2.0), ok (3.0) und trotzdem

gagabila.

leicht

(sind sie) leicht.

5 Yabilosi:

<i>Yegu</i>	<i>pela</i>	<i>lala</i>	<i>o</i>	<i>ligaba</i>	<i>ki(da)mwa</i>	<i>gala</i>	<i>bogwa</i>
Ich	für	Blume	Loc	alter.Garten	wenn	nicht	schon

Ich (werde) nach Blumen im alten Garten (schauen gehen) wenn ich nicht

<i>ba-gimwali</i>	<i>ya-tala</i>	<i>agu</i>	<i>tombaiku!</i>
1.Fut-kaufen	CP.flexibel-ein	mein	Tombaiku-Blatt

ein Tombaiku-Blatt für mich kaufen kann!

Keyeba erzählt seinen Freunden, dass gestern Leute von der Nachbarinsel Kuiava gekommen waren und viele „*tombaiku*“-Blätter zum Verkauf mitgebracht hatten. Diese Tabak-ähnlichen Blätter wachsen auf den Inseln im Busch und wurden früher vor der Einführung von Tabak von den Trobriandern geraucht. Nachdem Tabak aber auf den Trobriand-Inseln von Händlern eingeführt war, wurden die „*tombaiku*“-Blätter jahrelang von den Trobriandern verschmäht, weil sie es vorzogen, Tabak (in den 80-er Jahren vor allem Tabak der Marke „Cowboy“) und später – seit den 90-er Jahren – gedrehte Zigaretten (der Marke „Mutrus“) oder andere Zigaretten (vor allem die der Marke „Benson & Hedges“) zu rauchen. In den letzten Jahren aber ist in den Läden auf der Insel Kiriwina der Preis für Tabak und Zigaretten dramatisch gestiegen, und die meisten Insulaner müssen sich – aus finanziellen Gründen – nun wieder mit den guten alten „*tombaiku*“-Blättern begnügen. Nachdem Moagava Keyebas Worte bestätigt hat, nennt der nun den Preis, zu dem die Blätter verkauft wurden. Daraufhin lobt Moagava die gute Qualität der Blätter. Yabilosi beteiligt sich nun auch an der Unterhaltung mit dem witzigen Kommentar, dass er in den Garten gehen wird und die Blätter irgendeiner Blume rauchen wird, wenn er nicht eines dieser „*tombaiku*“-Blätter kaufen kann. Keyeba spricht nun ein neues Thema an – er verweist auf ein anderes Kanu, das gestern ebenfalls am Strand von Tauwema angelegt hatte:

6 Keyeba:

<i>Ke-ta(la)</i>	<i>waga</i>	<i>mna</i>	<i>amyaga</i>	<i>va</i>	<i>koya</i>	<i>e-ma</i>	(2.0)
K.P.hölzern-ein	Kanu	hm	wie.heit's	Dir	Berg	3.-kommen	

Ein Kanu – hm, wie heißt's – von den Bergen ist es gekommen. (2.0)

7 Yoya:

Koya-koya (2.0).
 Berg-Redup
 (Von den) Bergen (wirklich). (2.0)

8 Keyeba:

E, bi-ne'i-si buva bi-sagali-si bi-gimwali-si.
 Ja 3.Fut-suchen-Pl Betelnuss 3.Fut-verteilen-Pl 3.Fut-kaufen-Pl
 Ja, sie suchen Betelnüsse, sie wollen sie verteilen (während einer Trauerfeier und deshalb) wollen sie (welche) kaufen. (9.0)

Keyeba erzählt, dass ein Kanu von den Bergen, das heißt von den Vulkaninseln der D'Entrecasteaux Gruppe etwa 100 km südlich von Kaile'una – nach Tauwema gekommen war. Yoya wiederholt den abkürzenden Ortsnamen noch einmal als Frage, um sicher zu gehen, ob Keyeba wirklich die D'Entrecasteaux-Inseln meint. Keyeba bestätigt das und berichtet weiter, dass die Männer im Kanu Betelnüsse (*Areca catechu*) kaufen wollten, um sie während eines „*sagali*“ genannten Verteilfestes im Rahmen eines Trauerrituals an die Teilnehmer dieser Feier zu verteilen. Die Insel Kaile'una ist im gesamten Massim-Gebiet (vgl. Malinowski 1922, S. 30) bekannt für ihre Betelpalm-Plantagen; von daher ist es nicht verwunderlich, dass Leute von weit her nach Kaile'una segeln, um dort von den Bewohnern der Insel Betelnüsse zu kaufen. In ganz Neuguinea (aber auch in Ägypten, Ostafrika, Indien, Südostasien und Ozeanien) sind Betelnüsse ein beliebtes, leicht berauschendes Genussmittel.⁵ Nach Keyebas Bericht folgt eine längere Pause, die Sose'ula mit der folgenden Bemerkung beendet:

9 Sose'ula:

a-gisi wala (i)-sima buva, ta-uvananai
 1.-sehe nur (3.)-sein Betelnuss Dual.inkl-sagen
 Ich hab's nur gesehen, hier gibt es Betelnüsse, man kann sagen

sagali gala buva desi wala kaula okwa.
 Verteilfest nein Betelnuss genug nur Lebensmittel zuende.sein
 ein Verteilfest ohne Betelnüsse, (das ist schon schlimm) genug, nur (noch schlimmer ist es, wenn) die Lebensmittel aufgebraucht sind.

Er bestätigt damit zunächst voller Stolz, dass wir genügend Betelnüsse auf unserer Insel haben. Dann konstatiert er, dass ein Verteilfest während eines Trauerrituals ohne Betelnüsse schon schlimm genug wäre, aber das es noch schlimmer sei, wenn die Lebensmittel – und hier vor allem die Yamsknollen⁶ – aufgebraucht wären. Einige Minuten vor dieser Aussage hatte Yoya schon behauptet, dass im Nachbardorf Koma während eines Verteilfestes am

⁵ Die Schalen der Nuss werden gekaut und ausgespuckt, der Kern wird dann zusammen mit Kalk und Blättern des Betelpfeffers (*Piper betle*) gekaut oder zuerst in einem Mörser zerstoßen und dann gekaut. Die euphorisierende Wirkung der Betelnuss beruht auf dem Alkaloid Arecolin ($C_8H_{13}NO_2$), das durch den zugesetzten Kalk spätestens im Mund zu Arecaidin ($C_7H_{11}NO_2$) umgewandelt wird. Arecaidin ist die eigentlich stimulierende Substanz (vgl. http://www.giftpflanzen.com/areca_catechu.html).

⁶ „*kaula*“ ist die Bezeichnung für Lebensmittel insgesamt, aber – und im Besonderen – auch die generische Bezeichnung für die verschiedenen Arten von Yams (z. B.: *Discorea alata*, *Discorea esculenta*). Damit ist im Lexikon des Kilivila die große Bedeutung, die diese Knolle für das Leben und die Kultur der Trobriander spielt, kodifiziert (vgl. Malinowski 1935).

Tage zuvor nur ein Schwein geschlachtet worden sei – das ist eine recht infame Verleumdung, denn die Feste, bei denen Lebensmittel verteilt werden, sind ausgesprochen kompetitiv: Michael Young (1971) hat dieses Verhalten, das man im gesamten Massim-Gebiet beobachten kann, treffend als „fighting with food“ (vgl. auch McDowell 1980) bezeichnet – und jede Art von Kritik im Hinblick auf die Lebensmittel, die für ein solches „*sagali*“ Verteilfest bereitgestellt werden, ist diffamierend und beleidigend für eine so kritisierte Dorfgemeinschaft. Bei diesem Gesprächsausschnitt muss aber auch noch besonders darauf hingewiesen werden dass Sose’ulas Kommentar am Schluss „*kaula okwa*“ – „die Lebensmittel sind aufgebraucht“ der übelsten und tödlichen Beleidigung der Trobriander ausgesprochen nahe kommt, nämlich der Behauptung, ein Mann oder eine Dorfgemeinschaft oder die Bewohner einer Insel hätten keine Lebensmittel mehr! Allerdings ist Sose’ulas Behauptung hier mit keiner spezifischen Personenreferenz verbunden. Die Trobrinder sind zuerst und vor allem Yams-Gartenbauern. Eine gute und reiche Yamsernte ist der größte Stolz eines jeden Trobrianders, denn eine solche Ernte dokumentiert, dass ein Mann hart und fleißig arbeitet und über gute Gartenmagie verfügt (vgl. Malinowski 1935, Vol II; Senft 1997a): der Ernteerfolg eines Mannes markiert seinen Status, der ihm innerhalb der Gemeinschaft zukommt und den er beanpruchen kann. Ein guter Ernteerfolg kann sogar die sonst sehr rigide Clanhierarchie innerhalb der Gemeinschaft überwinden. All das wird von dieser Beleidigung negiert – und eine solche Beleidigung kann durchaus zu einem Kampf oder gar zu Mord oder Krieg zwischen zwei Dorfgemeinschaften führen – egal, an wen sie gerichtet war. Sose’ula muss sich dessen bewusst sein – deshalb hat er seine Aussage wohl auch mit keiner spezifischen Personenreferenz verbunden und sich in seiner Wortwahl der eigentlichen Beleidigung nur angenähert – allerdings in einem recht hohen und durchaus gefährlichen Grad. Das ist natürlich eine spekulative Interpretation; was nun aber absolut verblüffend ist, ist das, was Sose’ulas Neffe Mokeimwena direkt im Anschluss an die Äußerung seines Onkels sagt:

10 Mokeimwena:

<i>A</i>	<i>gala</i>	<i>yokwa,</i>	<i>e-kamama-si</i>	<i>kaula</i>	<i>gala.</i>
ah	nein	du	3.-hungrig.sein-Pl	Lebensmittel	nicht

Ah nein (weiß) du, sie sind hungrig, (es gibt) keine Lebensmittel (mehr).

Damit wischt er die möglicherweise noch als elegant auffaßbare Anspielung seines Onkels zur Seite und behauptet recht plump, dass ‚sie‘ hungrig sind, weil keine Lebensmittel mehr da sind – und mit diesem zweiten Teil seiner Aussage – *kaula gala* – verbalisiert er die übelste Beleidigung der Trobriander im Wortlaut. Allerdings ist der Ausdruck der Personalreferenz in Form der die dritte Person Plural markierenden Verbalaffixe (*e-kamama-si*), den er dabei benutzt, noch immer nicht eindeutig. Nach Schegloff (1996) handelt es sich hier um eine „Folgereferenz in Initialposition“ („a subsequent sort of reference in an initial position“). Der Subjektspräfix (*e-*) und der den Plural markierende Suffix (*-si*) kann sich aber durchaus in anaphorischer Referenz hier

entweder auf die vorher erwähnten Leute von Kuiava oder auf die ebenfalls vorher erwähnten Leute „von den Bergen“ handeln. Die Situation eskaliert jetzt allerdings dadurch, dass sich ein Außenstehender, der bisher dem Tratsch der 6 Männer bloß zugehört hatte, einmischt und gegen diese Behauptung folgendermaßen protestiert:

11 Außenstehender (unbekannt, in der Video-Dokumentation nicht zu sehen)

A *Yaubada* *omatala* *taga* ...

Ah Gott vor(seinen.Augen) aber

Ah bei Gott aber ...

12 Mokeimwena (überlappend):

E-me-si ...

3.-kommen-Pl

Sie sind gekommen ...

13 derselbe Außenstehende (in der Videodokumentation nicht zu sehen)

... *e-bugubagula-si!*

... 3.-gärtnern-Pl

... sie haben in den Gärten gearbeitet!

14 Mokeimwena:

... *E-me-si* *e-nene'i-si* *kwe-yu* *vavagi*, (2.0)

3.-kommen-Pl 3.-suchen-Pl KP.generell.zwei Ding

... Sie sind gekommen und haben zwei Dinge gewollt: (2.0)

kwe-tala *kaula* (1.0) *kwe-tala* *buva*.
 KP.generell-ein Lebensmittel CP.generell-ein Betelnuss
 zum einen Lebensmittel (1.0) und zum anderen Betelnüsse.

Indem er Gott als Zeugen anruft – die Trobriander werden seit 1894 christlich missioniert, haben sich allerdings bis jetzt eine interessante Form von Synkretismus bewahrt (vgl. Malinowski 1916; Senft 1997b) –, protestiert der Außenstehende gegen Mokeimwenas Behauptung mit dem Hinweis, dass „sie in den Gärten gearbeitet haben“. Die anaphorische Personalreferenz ist hier immer noch nicht eindeutig, obwohl es eher unwahrscheinlich ist, dass jemand, der in Tauwema lebt, irgendetwas über die Gartenarbeiten auf den D'Entrecasteaux Inseln weiß. Aber Mokeimwena insistiert auf seiner Behauptung, indem er noch einmal explizit darauf hinweist, dass „sie“ gekommen sind und zwei Dinge wollten – und nach einer kurzen rhetorischen Pause nennt er dann zuerst „Lebensmittel“ und dann – wieder nach einer kurzen Pause – „Betelnüsse“. Diese Pausen geben seiner Aussage durchaus einen dramatischen Effekt. Die anaphorische Personalreferenz ist immer noch nicht eindeutig – aber es ist klar, dass die Situation jetzt eindeutig eskaliert ist. Offenbar hat auch Moagava das erkannt, denn direkt nachdem Mokeimwena seine Behauptung noch einmal und mit Nachdruck aufgestellt hat, fragt er ihn ganz direkt:

15 Moagava:

Avela?

Wer?

Und Mokeimwena – der sich jetzt gezwungen sieht, seine Personaldifferenz eindeutig zu machen – beantwortet diese Frage folgendermaßen:

16 Mokeimwena:

<i>M-to-si-na</i>	<i>mina</i>	<i>Kuiava</i>	<i>e-me-si. (4s)</i>
Dem-KP.männlich-Pl-Dem	Leute.von	Kuiava	3.-kommen-Pl

Diese Leute von Kuiava sind gekommen. (4s)

Er gibt diese Antwort allerdings mit leiser Stimme – wahrscheinlich hat er realisiert, was er gerade getan hat. Das mag auch den anderen Männern so gehen – jeder Trobriander weiß, dass Mokeimwena dafür von einem Mann aus Kuiava, der zufällig dem Gespräch zugehört hätte, auf der Stelle hätte getötet werden können. Wie dem auch sei, bevor die Männer ihr Gespräch wieder aufnehmen, verstreichen 4 Sekunden.

An dieser Stelle möchte ich noch einmal kurz zusammenfassen, was während der letzten 7 ‚turns‘ bei diesem Gespräch geschehen ist. Mokeimwena hat die für die Trobriander übelste Beleidigung ausgesprochen. Als dagegen jemand protestiert, weist er dessen Einwand zurück und betont noch einmal mit Nachdruck, dass die von ihm gerade beleidigten Leute wirklich gekommen sind, um Lebensmittel zu bekommen. Spätestens jetzt ist die Situation unrettbar eskaliert und Moagavas direkte Frage, wen Mokeimwena nun eigentlich meint, zwingt ihn schließlich auch noch dazu, die von ihm Beleidigten explizit zu nennen. Darüber, warum Moagava diese Situation mit seiner Frage zu ihrem bitteren Höhepunkt gebracht hat, lässt sich nur spekulieren. Es ist aber durchaus plausibel anzunehmen, dass Moagava Mokeimwena mit seiner Frage bloßstellen will, sicherlich auch, um ihn zu beschämen, vor allem aber, um ihm deutlich zu machen, dass er mit seinen letzten Äußerungen die Grenzen dessen, was durch die „*biga sopa*“ Varietät noch lizenziert wird, überschritten hat, und dass dieser Schritt absolut inakzeptabel ist. Mokeimwena hat seine Äußerungen hier nicht mehr im Rahmen der „*biga sopa*“ Varietät produziert; er hat damit den Rahmen der bisherigen verbalen Interaktion gebrochen und ist in den Bereich der „*biga gaga*“ Varietät abgedriftet – in die Varietät, die jeden stilistisch- und situations-inadäquaten Sprachgebrauch in allen kommunikativen Kontexten umfasst, und die außerdem noch durch das Genre „Beleidigungen und Flüche“ – „*matua*“ – konstituiert wird.⁷

Dieser Bruch der trobriandischen Etikette muss unbedingt wieder behoben werden, wenn die Männer ihr Gespräch weiterführen wollen. Sie sind sich alle ihrer besonderen interaktiven Situation bewusst – sie werden gefilmt, sie wis-

⁷ Die 6 schlimmsten Flüche im Kilivila lauten: „*Yokwa gala ka-m*“ – „Du hast keine Lebensmittel (keinen Yams)“, „*kweya m kwava / m mwala*“ – „Fick Deine Frau / Deinen Mann“, „*kweya lumta (bwadam) / akeya lumta*“ – „Fick Deine Schwester / Deinen jüngeren Bruder (Deinen älteren Bruder) / Ich fickte Deine Schwester“, „*kweya inam / tamam / akeya inam*“ – „Fick Deine Mutter / Deinen Vater / Ich fickte Deine Mutter“, „*migim lumta*“ – „Du siehst aus wie Deine Schwester“ und „*kukwam popu*“ – „Friß Scheiße“.

sen, dass das, was gerade passiert ist, video-dokumentiert ist und transkribiert, übersetzt und analysiert werden wird! Von daher müssen sie einen Weg finden, die Konversation zurück in den Rahmen der unbeschwerten und sicheren „*biga sopa*“ Varietät zu führen. Und das ist auch genau das, was nun passiert. Nach 4 Sekunden allgemeiner peinlicher Verlegenheit sind Yabilosi und Keyeba die ersten, die ihre Worte wiederfinden und das Gespräch folgendermaßen weiterführen:

17 Yabilosi:

Ku-meya *manusisi!*
2.-bringen Streichhölzer
Gib mir Streichhölzer!

18 Keyeba:

E-sa... *E-sakaula* *ma-ke-na* *dinga,*
3.- 3.-rennen Dem-KP.hölzern-Dem Dinghy
Es... Es fuhr schnell dieses Dinghy,

e-ma *bi-la-la* *Losuia,*
3.-kommen 3.Fut-gehen-Red Losuia,
es kam es wollte nach Losuia fahren

e-sakaula-ga *e-ma* *o* *da-valu-si* *e-okwa*
3.-rennen-Emph 3.-kommen Loc 2.incl.unser-Dorf-Pl 3.-zuende.sein
es fuhr wirklich schnell es kam zu unserem Dorf – es war zuende --

sopi *gala,* (1.0) *e-mwena-si* *buva* *o* *da-valu-si*
Zoom nicht 3.-klettern-Pl Betelpalme Loc 2.incl.unser-Dorf-Pl
kein Zoom (mehr), (1.0) sie kletterten auf die Betelpalmen in unserem Dorf

e-ke'ita-si, (1.0) *taga* *e-ma* *bi-loki-la*
3.-umkehren-Pl aber 3.-kommen 3.Fut-gehen.zu-Emph
(und) kehrten (dann) um, (1.0) aber sie kamen und wollten (zu)

Vapalaguyau *Tuma* (2.0) *bi-mwena-si* *buva*
Vapalaguyau Tuma.
Vapalaguyau (auf) Tuma, (2.0) sie wollten auf die Betelpalmen klettern

bi-una *bi-ke'ita,* (3.0) *gala-gola,*
3.Fut-laden 3.-umkehren nein-Emph
(und) es sollte (mit Betelnüssen) beladen (werden) und umkehren, (3.0) aber nein,

e-okwa *sopi* *gala,* *o* *da-valu-si* *e-ke'ita.* (7.0)
3.-zuende.sein Zoom nein Loc 2.incl.unser-Dorf-Pl 3.-umkehren
es war zuende, kein Zoom (mehr), (und dann) sind sie zu unserem Dorf zurück (7.0).

Yabilosi wendet sich an Moagava und bittet ihn um Streichhölzer – und Moagava wirft sie ihm in der typischen Manier der Trobriander zu (vgl. Malinowski 1922, S. 352).⁸ Keyeba, der zu gleichen Zeit wie Yabilosi etwas sagen wollte, wartet mit seinem Beitrag, bis Yabilosi ausgedet hat und erzählt, wie und

⁸ Malinowski (1922, S. 352) beschreibt solche Transaktionen während des „*Kula*“-Tausches folgendermaßen: “The etiquette of the transaction requires that the gift should be given in an off-hand, abrupt, almost angry manner, and received with equivalent non-chalance and disdain”.

warum die Leute von Kuiava – die gerade so beleidigt wurden – nach Tauwema kamen. Eigentlich wollten sie mit ihrem Dinghy nach Losuia, dem Ort mit dem Verwaltungsposten und den größten Läden auf der Insel Kiriwina fahren, um dort einzukaufen. Aber ihnen ging das ‚Zoom‘, das Benzin-Ölgemisch für den Außenbordmotor ihres Dinghys aus. Deshalb mussten sie auf halbem Weg nach Tauwema zurück, legten an, bekamen Betelnüsse und fuhren dann wieder zurück nach Kuiava. Sie fuhren also weder nach Losuia, wie ursprünglich geplant, noch zur Insel Tuma, auf der Vapalaguyau, einer der einflussreichen Männer von Tauwema, mit einer Reihe anderer Leute aus unserem Dorf vor einigen Jahren ein neues Dorf gegründet hat. Sie wollten eigentlich bei ihm Betelnüsse auf dem Rückweg von Losuia kaufen, aber wegen ihres knappen Zoom-Vorrats mussten sie ihren Plan ändern und die Betelnüsse in Tauwema kaufen. Dieser Bericht über den Besuch der Leute von Kuiava in Tauwema erwähnt nur noch Betelnüsse. Lebensmittel werden überhaupt nicht erwähnt – und so „überschreibt“ gleichsam Keyebas Bericht das, was Mokeimwena vor ein paar Sekunden gesagt hatte. Mokeimwenas Bruch der trobriandischen Etikette, sein Ausbruch aus dem Rahmen der „*biga sopa*“ Varietät mit dem Aussprechen der übelsten Beleidigung der Trobriander und damit sein Abdriften in die „*biga gaga*“ Varietät ist damit wieder sozial „repariert“. Keyaba ist es gelungen, die Situation im Goffman’schen Sinne neu zu rahmen (vgl. Goffman 1974) und die Konversation zurück ins „*kasilam*“ Genre und damit wieder in die „*biga sopa*“ Varietät zu transformieren. Auf Keyebas Beitrag folgt wieder eine relativ lange Pause – aber jetzt in entspannter Situation. Mokeimwena ergreift nun auch wieder das Wort und greift mit seinem Beitrag wieder das Thema „schlechtes Wetter beim „*sagali*“ Verteilfest in Koma am Tag zuvor“ auf:

19 Mokeimwena:

<i>Ma-na-kwa</i>	<i>lova</i>	<i>beya</i>	<i>kuna</i>	<i>e-vagi</i>	<i>beya</i>
Dem-Dem-KP.generell	gestern	hier	Regen	3.-machen	hier

Dieser Regen da gestern, er hat das gemacht,

<i>kunubwadela</i>	<i>m-kwe-na</i>	<i>sagali (1.0)</i> ,	<i>ka</i>	<i>kalabiga</i> .
starker Regen	Dem-KP.generell-Dem	Verteilfest	gut	das.ist's.

(den) starken Regen (während) dieses Verteilfests, (1.0) gut, das ist's.

20 Moagava:

<i>Taga</i>	<i>e-bubuna-si</i>	<i>gaga</i> .
aber.wirklich	3.-sich.benehmen-Pl	schlecht

Aber wirklich, sie haben sich schlecht benommen.

21 Mokeimwena:

<i>I-gaga</i>	<i>lagela</i>	<i>lilu</i>	<i>e-kile'i</i> ,	<i>bi-sagalimkolova-si</i>	<i>o</i>
3.-schlecht	heute	Sonne	3.-werfen	3.Fut-Verteilfest.beenden-PL	Loc

Das war schlecht, heute brennt die Sonne, sie hätten das Verteilfest in

<i>da-valu-si</i>	<i>lagela</i> .
2.inkl.unser-Dorf-Pl	heute

unserem Dorf heute beenden können.

22 Yoya:

Bi-gimkola-si.

3.Fut-beenden-Pl

Sie werden (es) beenden.

23 Sose'ula:

Taga lova bi-yova'i-si e-kuna-kuna.

aber gestern 3.Fut-kämpfen-Pl 3.-regnen-Redup

Aber gestern haben sie fast (miteinander) gekämpft (weil) es regnete.

24 Keyeba:

So-(gu) lagela ba-la-la Koma gala-go
 Freund-(mein) heute 1.Fut-gehen-Emph Koma nicht-Emph

Mein Freund, heute werden wir auf keinen Fall nach Koma gehen,

iga(u) ta-paisewa-si bi-vokuva.

später 1.inkl-arbeiten-Pl 3.Fut-zuende.sein

später werden wir arbeiten und (es) wird zuende sein.

Nach seinem Bruch trobriandischer Etikette bewegt sich Mokeimwena mit seiner Äußerung wieder im Rahmen der „*biga sopa*“ Varietät. Er tratscht wieder – und zwar in einer sehr typischen Weise: Der Genre-Begriff „*kasilam*“ kann auch als „hinter jemandes Rücken flüstern“ glossiert werden. Wieder versucht er, jemanden zu blamieren. Indem er das Thema des Gesprächs wieder auf den starken Regen beim „*sagali*“-Verteilfest in Koma bringt, klagt er einen ungenannten Wettermagier an und macht ihn verantwortlich für das schlechte Wetter. Diese Anklage impliziert aber auch, dass die Leute von Koma entweder keinen Wettermagier beauftragt haben, für schönes Wetter zu sorgen, oder dass sie zu geizig waren, für die Dienste eines solchen Wettermagiers zu bezahlen, oder dass die Wettermagier in Koma zu schwache wettermagische Formeln kennen, oder dass sie nicht in der Lage waren, ihre Formeln fehlerlos zu rezitieren. Die Trobriand Insulaner haben lange Zeit an die Macht der Magie geglaubt – dieser Glaube konnte erst vor relativ kurzer Zeit von den christlichen Missionaren erschüttert werden. Es gibt verschieden starke magische Formeln – stärkere Formeln können schwächere Formeln überwinden; darüber hinaus verlieren magische Formeln ihre Wirksamkeit, wenn sie nicht im überlieferten Wortlaut rezitiert werden oder wenn der Magier nicht die für seine Magie notwendigen Tabus – es handelt sich dabei meist um Nahrungstabus – beachtet hat (vgl. Malinowski 1935; Senft 1985c; 1997a). Mokeimwena setzt mit dieser Aussage so implizit eine ganze Reihe von Gerüchten frei. Aber diese Art des Gerüchte-Verbreitens ist durchaus Teil des „*biga sopa*“ Konzepts – und damit bewegt sich Mokeimwena auf sicherem Terrain mit seinen bössartigen impliziten Unterstellungen – sie sind durch die „*biga sopa*“ Varietät lizenziert. Moagava nimmt Mokeimwenas Anschuldigung auf und bestätigt, dass sich die Leute von Koma dadurch, dass sie nicht für besseres Wetter gesorgt haben, wirklich schlecht benommen haben. Dadurch bestärkt macht sich Mokeimwena weiter über die Leute von Koma lustig. Er weist darauf hin, dass jetzt in Tauwema schönes Wetter ist und merkt

an, dass die Leute von Koma ihr Verteilfest ja an diesem Tag in Tauwema beenden könnten. Yoyas Hinweis, dass sie ihr „*sagali*“ auf alle Fälle heute beenden werden, kann entweder als Kritik an Mokeimwena aufgefasst werden oder als eine weitere Anspielung darauf, dass die Leute von Koma einfach zu schlampig und sorglos sind, um so ein Fest adäquat vorzubereiten und durchzuführen. Beide Interpretationen sind möglich – und diese Mehrdeutigkeit ist es, die so charakteristisch für die „*biga sopa*“ Varietät ist. Soseula weist dann noch darauf hin, dass das schlechte Wetter auch noch dafür verantwortlich war, dass es zu ernsteren Spannungen zwischen den Leuten aus Koma und ihren Gästen gekommen war – Spannungen, die beinahe zu einem Kampf geführt hätten. Keyeba greift die Möglichkeit, dieses Thema weiter auszuführen, nicht auf. Er versucht, das Thema mit dem Hinweis abzuschließen, dass sie ja schon gestern beschlossen haben, nicht am Abschluss des Verteil-fests teilzunehmen, sondern sich an den Reparaturarbeiten am Haus zu beteiligen, dass den Kindern von Tauwema als Vorschule dient. Hier endet das Gesprächsexzerpt. Die gesamte Atmosphäre ist zu diesem Zeitpunkt wieder absolut entspannt – so wie zu Beginn der Interaktion, die ich hier vorgestellt habe – und alle 6 Männer tratschen und klatschen weiter und freuen sich an ihrem „*kasilam*“.

3. „... was soll es bedeuten...“:

Ein Corpus ist nur so gut wie seine Metadaten!

Die hier präsentierte Erschließung eines kurzen Gesprächsausschnitts hat gezeigt, wieviele Informationen notwendig sind, um diesem Gespräch nicht nur folgen, sondern um auch verstehen zu können, was bei diesem Gespräch eigentlich abläuft.

Zum einen hat die Erschließung des Texts auf der Basis von ethnolinguistischem Wissen im Hinblick auf die Gesprächsführung gezeigt, dass die Trobriander nicht nur metalinguistisch Dialekte unterscheiden, sondern auch sogenannte „situations-intentionale“ Varietäten, die durch verschiedene Genres, die ebenfalls metalinguistisch bezeichnet sind, konstituiert werden. Dabei wurde die Bedeutung und die Funktion der Varietäten „*biga sopa*“ und „*biga gaga*“ herausgearbeitet und gezeigt, wie ein Sprecher mit einer Äußerung, die eindeutig der „*biga gaga*“ zugeordnet ist, den Rahmen der „*biga sopa*“ sprengen kann, wie er damit trobriandische Etikette bricht, und wie die anderen Gesprächsteilnehmer durch Themenwechsel versuchen, diesen Bruch sozial zu reparieren, die prekäre Situation zu überwinden und das Gespräch wieder in den Rahmen der „*biga sopa*“ zurückzuführen. Außerdem wurde deutlich, dass bei Klatsch und Tratsch soziale Unterschiede im Rahmen der trobriandischen Clan-Hierarchie keine Rolle spielen.

Zum anderen hat die Erschließung des Texts auf der Basis von ethnolinguistischem Wissen im Hinblick auf die angesprochenen Themen gezeigt, welche Prozesse kulturellen Wandels im Zusammenhang mit Rauchverhalten

und Mobilität zu beobachten sind, welche Bedeutung Tabak und Betelnüsse vor allem aber Yamsknollen für die Trobriander haben – auch im Hinblick auf Verteilfeste im Rahmen von Trauer-Ritualen, welche Rolle der Magie immer noch zukommt, und dass Magie und christlicher Glaube durchaus noch in einer synkretischen Beziehung stehen.

Solche Erkenntnisse müssen Ihren Eingang in die das Textcorpus des Kili-vila begleitenden Metadaten finden, damit der Nutzer dieses Corpus optimal nutzen kann.⁹ Ich hoffe, dass ich mit diesem Beispiel aus dem Kilivila deutlich gemacht habe, dass jedes Corpus – welcher Sprache auch immer – nur so gut sein kann, wie die dieses Corpus begleitenden, erschließenden und erläuternden Metadaten. Wenn ich aber bedenke, wie weit diese Einsicht sich bisher in Metadaten-Beschreibungen bei Sprachdokumentations-Corpora niedergeschlagen hat, dann muss ich leider wieder im Rückgriff auf Heinrich Heine bekennen: „Mein Herz, mein Herz ist traurig ...“. Es gibt noch unglaublich Vieles zu tun. Packen wir es – adäquat – an!

Literatur

- Atkinson, J. Maxwell, Heritage, John (Hg.) (1984): *Structures of Social Action: Studies in Conversation Analysis*. Cambridge.
- Broeder, Daan, Brugman, Hennie, Senft, Gunter (2005): Documentation of languages and archiving of language data at the Max Planck Institute for Psycholinguistics in Nijmegen. In: *Linguistische Berichte*, 201, S. 89–103.
- Goffman, Erving (1974): *Frame Analysis. An Essay on the Organization of Experience*. New York.
- Malinowski, Bronislaw (1916): Baloma: The spirits of the dead in the Trobriand Islands. In: *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 46, S. 353–430.
- Malinowski, Bronislaw (1922): *Argonauts of the Western Pacific*. London.
- Malinowski, Bronislaw (1935): *Coral Gardens and their magic*. Vol. I: The description of gardening. Vol. II: The Language of Magic and Gardening. London.
- McDowell, Nancy (1980): It's not who you are but how you give that counts: the role of exchange in a Melanesian society. In: *American Ethnologist*, 7, S. 58–70.
- Montesquieu, Charles-Louis de Secondat, Baron de (1721): *Lettres persanes*. Cologne: Pierre Marteau.
- Paasche, Hans (1912/13): Die Forschungsreise des Afrikaners Lukanga Mukara ins innerste Deutschland. Der Vortrupp (= (1988) München, herausgegeben von Franziskus Hähnel und mit einem Nachwort von Iring Fetscher).
- Rosendorfer, Herbert (1983): *Briefe in die chinesische Vergangenheit*. München.
- Schegloff, Emanuel A. (1996): Some practices for referring to persons in talk-in-interaction: A partial sketch of a systematics. In: Fox, Barbara (Hg.): *Studies in Anaphora*. Amsterdam. S. 437–485.
- Scheurmann, Erich (1920): *Der Papalagi – Die Reden des Südseehäuptlings Tuiavii aus Tiavea*. Buchenbach: Felsenverlag (= 1977 Adliswil-Zürich: Tanner & Staehlin).

⁹ Wie wir das am MPI für Psycholinguistik geregelt haben, ist in Broeder et al. (2005) dokumentiert.

- Senft, Gunter (1985a): Trauer auf Trobriand – eine ethnologisch/linguistische Fallstudie. In: *Anthropos*, 80, S. 471–492.
- Senft, Gunter. (1985b): Kilivila – die Sprache der Trobriander. In: *Studium Linguistik*, 17/18, S. 127–138.
- Senft, Gunter (1985c): Weyeis Wettermagie – eine ethnolinguistische Untersuchung von fünf wettermagischen Formeln auf den Trobriand Inseln. In: *Zeitschrift für Ethnologie*, 110, Heft 1, S. 67–90, Heft 2, Erratum 2pp.
- Senft, Gunter (1986): Kilivila – The Language of the Trobriand Islanders. Berlin.
- Senft, Gunter (1987): Rituelle Kommunikation auf den Trobriand Inseln. In: *Zeitschrift für Literaturwissenschaft und Linguistik*, 65, S. 105–130.
- Senft, Gunter (1997a): Magical Conversation on the Trobriand Islands. In: *Anthropos*, 92, S. 369–391.
- Senft, Gunter (1997b): Magic, missionaries and religion. Some observations from the Trobriand Islands. In: Otto, Ton und Borsboom, Ad (Hg.): *Cultural dynamics of religious change in Oceania*. Leiden. S. 45–58.
- Senft, Gunter (1999): Weird Papalagi and a Fake Samoan Chief – A Footnote to the Noble Savage Myth. In: *Rongorongo Studies – A Forum for Polynesian Philology* 9(1): 23–32 & 9(2): 62–75.
- Spence, Jonathan D. (1988): *The question of Hu*. New York: Alfred A. Knopf.
- Weiner, Annette B. (1976): *Women of Value, Men of Renown: New Perspectives in Trobriand Exchange*. Austin.
- Weißhaupt, Winfried (1979): *Europa sieht sich mit fremdem Blick. Werke nach dem Schema der „Lettres persanes“ in der europäischen, insbesondere der deutschen Literatur des 18. Jahrhunderts*. (2 parts in 3 volumes.) Frankfurt am Main.
- Young, Michael (1971): *Fighting with Food. Leadership, Values and Social Control in a Massim Society*. Cambridge.

ALEXANDRA N. LENZ

Zur variationslinguistischen Analyse regionalsprachlicher Korpora¹

Abstract

Im Beitrag werden korpuslinguistische Aspekte und Fragestellungen aus variationslinguistischer Perspektive diskutiert, wobei der Fokus auf der Analyse regionalsprachlicher Daten und Korpora des Deutschen liegt. Unter *Regionalsprache* wird hier der areal gebundene, sprechsprachliche Gesamtbereich „unterhalb“ der normierten Standardsprache verstanden, der sich von den lokalen Basisdialekten über Varietäten bzw. Sprechlagen des mittleren Bereichs bis hin zum standardnächsten Pol der Regionalsprache (*Regionalakzent* oder *Regionalstandard*) erstreckt.

Am Beginn des Beitrags steht ein Überblick über Datenklassen, die als empirische Grundlage für regionalsprachliche Fragestellungen herangezogen werden können. Eine Präsentation regionalsprachlicher Korpora, die auf den vorgestellten Datenklassen aufbauen, schließt sich an, wobei hier lediglich einige ausgewählte, online zugängliche Korpora Berücksichtigung finden (können). In einem nächsten Schritt werden die vorgestellten Korpora zur konkreten variationslinguistischen Analyse von zwei regionalsprachlichen Phänomenen aus dem lautlichen („g-Spirantisierung“) bzw. grammatischen Bereich („Rezipientenpassiv“) herangezogen und auf ihre Ergiebigkeit geprüft.

1. Zu Spezifik und Problematik regionalsprachlicher Daten und Korpora

Im vorliegenden Beitrag werden korpuslinguistische Aspekte und Fragestellungen aus variationslinguistischer Perspektive diskutiert. Dabei wird der Fokus auf der Analyse regionalsprachlicher Daten und Korpora liegen.

Regionalsprache wird hier definiert als der areal gebundene, sprechsprachliche Gesamtbereich „unterhalb“ der normierten Standardsprache.² Dieser variative Gesamtbereich erstreckt sich von den lokalen Basisdialekten über Varietäten bzw. Sprechlagen des mittleren Bereichs bis hin zum standard-

¹ Für kritische Anmerkungen und hilfreiche Literaturhinweise danke ich herzlich Stephan Elspaß (Augsburg), Joachim Herrgen (Marburg), Arend Mihm (Duisburg) und Jürgen Erich Schmidt (Marburg).

² Hinter den Attributen „unten“ versus „hoch“ steht das Bild von der Zweidimensionalität der Regionalsprache. In diesem Bild entspricht die x-Achse der horizontal-arealen Ebene, während die y-Achse die vertikal-soziale Dimension von Sprache repräsentiert. Zum Begriff der Zweidimensionalität s. etwa Bellmann 1983 und Auer/Hinskens 1996.

nächsten Pol der Regionalsprache, dem sogenannten „Regionalakzent“ (Lenz 2003) oder „Regionalstandard“ (Spiekermann 2005).³

Aus der dem Regionalsprachenbegriff inhärenten Oralität und Variabilität leiten sich spezielle korpuslinguistische Probleme ab. Zu diesen gehören etwa das im Zusammenhang mit der Gewinnung oraler Daten häufig diskutierte „Beobachterparadoxon“, das nach William Labov (1972, S. 61 f.) darin besteht, dass die Erhebung natürlichen Sprachverhaltens eigentlich impliziert, Sprachverhalten gerade dann zu beobachten, wenn Sprecher nicht beobachtet werden.⁴ Mit wachsendem Methodenbewusstsein gehen in der Variationslinguistik ausgefeilte Erhebungsmethoden einher, mit denen eine Minimierung der Auswirkungen des Paradoxons und die Gewinnung möglichst „authentischer“ Sprachdaten angestrebt werden. Häufig bedingen diese Erhebungsmethoden eine Zunahme des personellen, zeitlichen, technischen und damit auch finanziellen Aufwands. Hinsichtlich der Erhebung natürlicher Sprachdaten kommt erschwerend hinzu, dass

“the low frequency of occurrence in normal speech of any specific sound, word, grammatical item, pragmatic pattern, or whatever else is being sought, means that a very large amount of natural language must be observed or collected in order to have a sufficient quantity for useful analysis.”

(Spolsky 2005, S. 1086)

Auch die Datenstrukturierung mittels Meta-Beschreibungen⁵ erweist sich als aufwändiges Moment (z. B. die areale oder soziale Einordnung von Sprachdaten, die ethnographische und soziodemographische Charakterisierung der Sprecher u. a.). Sind orale Sprachdaten erhoben und strukturiert, stellt die zeitintensive Datenaufbereitung in Form von (möglicherweise feinphonetischen) Transkriptionen und weiterführenden linguistischen Annotationen ein weiteres Problem dar, das sich aus der Spezifik gesprochensprachlicher „Primärdaten“⁶ ergibt. Bei der Entscheidung für ein konkretes Transkriptionssystem bzw. für eine bestimmte „Tiefe“ der Transkription spielen

³ Zu den komplexen Gesamtstrukturen der modernen deutschen Regionalsprachen ist bislang relativ wenig bekannt. Als gesichert gilt, dass im Vergleich der Großräume des deutschsprachigen Sprachraums von der Schweiz bis hin zu Norddeutschland mit deutlich differierenden Verhältnissen zu rechnen ist. Zur Varietätenstruktur, die etwa für den moselfränkischen Raum und darüber hinaus wohl für weite Teile des Westmitteldeutschen anzusetzen ist, s. Lenz 2003. Zu regional differierenden Strukturmodellen im Deutschen s. die Überblicksdarstellungen bei Schmidt 1998, König (¹⁵2005, S. 135 f.) und Spiekermann 2005.

⁴ Zum Beobachterparadoxon s. auch Wodak 1982 und Thun 2005.

⁵ Nach Skiba (2005, S. 1189) fasst eine Meta-Beschreibung „alle soziolinguistische und technisch relevanten Informationen, die zu einer *session* [abgeschlossenen Dateneinheit; A. N. L.] gehören, zusammen und kann deshalb zum Aufbau eines gut beschriebenen Sprachkorpus verwendet werden.“

⁶ Skiba (2005, S. 1187): „Audio- und Videoaufnahmen von sprachlichen Ereignissen bilden die primären Daten, und somit die Grundlage soziolinguistischer Untersuchungen.“

Aspekte der angestrebten Datenauswertung eine Rolle.⁷ Zu diesen gehören Fragestellungen wie: Welche linguistische Systemebene steht im Fokus der Analyse? Welche Varianten bzw. welche Grade der Variantendifferenzierung sind bei der Analyse des in Rede stehenden Phänomens zu berücksichtigen? Welche sprachlichen Varianten konstituieren eine regionalsprachliche Variable?

Der Beitrag, in dem diese und andere korpuslinguistische Probleme der Regionalsprachenforschung diskutiert werden, gliedert sich wie folgt: Am Beginn steht ein Überblick über Datenklassen, die als empirische Grundlage für regionalsprachliche Fragestellungen herangezogen werden können (Kap. 2). Eine Präsentation regionalsprachlicher Korpora, die auf den vorgestellten Datenklassen aufbauen, schließt sich an (Kap. 3), wobei hier lediglich einige ausgewählte, online zugängliche Korpora Berücksichtigung finden (können). In einem nächsten Schritt werden die vorgestellten Korpora zur konkreten Analyse von zwei Beispielphänomenen herangezogen und auf ihre Ergiebigkeit geprüft (Kap. 4). Der Beitrag schließt mit einer Zusammenfassung (Kap. 5).

2. Regionalsprachliche Datenklassen

Ein Blick in die Forschungslandschaft zeigt, dass zur Analyse regionalsprachlicher Phänomene bzw. zum Aufbau regionalsprachlicher Korpora unterschiedlichste Daten herangezogen werden. Neben „objektiven“ Sprachdaten rücken zunehmend „subjektive“ Spracheinstellungsdaten in den Untersuchungsfokus.⁸ Die Tatsache, dass Spracheinstellungsdaten in der Variationslinguistik nicht nur als Erklärungsfaktoren für Sprachverhalten herangezogen werden, sondern mitunter einen eigenständigen Untersuchungsgegenstand darstellen⁹, kommt in der Klassifikation regionalsprachlicher Daten in Abbildung (1) darin zum Ausdruck, dass auf einer ersten Klassifikationsebene Spracheinstellungsdaten neben Sprachdaten gestellt werden. Dabei wird das enge Verhältnis der beiden Datenkomplexe, die im Optimalfall miteinander erhoben und aufeinander beziehbar werden, durch den bidirektionalen Pfeil angedeutet.

⁷ Zur Diskussion verschiedener Transkriptionssysteme und -konventionen s. Richter 2005.

⁸ Nach Stickel (1999, S. 17) „spricht man in der deutschsprachigen Linguistik seit Anfang der 70er Jahre von ‚Spracheinstellungen‘ und meint damit wertende Dispositionen, die einzelne Menschen oder soziale Gruppen gegenüber sprachlichen Erscheinungen haben. Spracheinstellungen sind besonders Haltungen gegenüber Sprachen, Sprachvarietäten oder Sprachverhalten anderer Individuen oder Gruppen, oft mit wertender Berücksichtigung der jeweils eigenen Sprache. Wie andere Einstellungen gelten Spracheinstellungen als erlernt, relativ beständig, wenn auch veränderbar.“

⁹ Zur Diskussion s. Lenz (2003, S. 263–265), Tophinke/Ziegler (im Druck) und Häcki Buhofer 2002. Als Beispiel für eine regionalsprachliche Studie, deren Schwerpunkt auf Einstellungsdaten liegt, sei hier Huesmann 1998 genannt.

Der Fokus wird im Folgenden auf dem Komplex der Sprachdaten liegen. Diese werden auf einer nächsten Klassifizierungsstufe in medial mündliche und medial schriftliche Daten eingeteilt. Dabei bezieht sich – in Anlehnung an das Konzept von Koch/Oesterreicher (1985 und 1994) –¹⁰ das Medium auf die „Realisierung sprachlicher Äußerungen, wo ‚mündlich‘ = ‚phonisch‘ und ‚schriftlich‘ = ‚graphisch‘ ist“ (Koch/Oesterreicher 1994, S. 587). Dagegen trägt der Konzeptionsbegriff der Tatsache Rechnung, dass innerhalb ein und desselben Mediums verschiedene Grade an raum-zeitlicher, sozialer und emotionaler Nähe bzw. Distanz und damit verbunden verschiedene Grade der sprachlichen Variation auftreten können. Bezüglich des Mediums sind die Begriffe ‚mündlich‘/‚schriftlich‘ dichotomisch, bezüglich der Konzeption von Sprache und Sprechen stellen sie Endpunkte eines Kontinuums dar. Wie Koch/Oesterreicher (1994, S. 587) hervorheben, besteht zwar eine „prinzipielle Unabhängigkeit von Medium und Konzeption“, die jedoch „nicht im Widerspruch dazu [steht], daß einerseits zwischen dem phonischen Medium und konzeptionell mündlichen Äußerungsformen, andererseits zwischen dem graphischen Medium und konzeptionell schriftlichen Äußerungsformen eine ausgeprägte Affinität besteht.“

(1) Regionalsprachliche Datenklassen (ein Überblick)¹¹

Abb. 1

Die in Abbildung (1) aufgeführten literarischen bzw. stilisierten Daten medial schriftlichen Typs stellen einen (bislang) eher vernachlässigten Datentyp in der Regionalsprachenforschung dar.¹² Sie stehen in Konkurrenz mit konzeptionell mündlichen Daten.

¹⁰ Eine Weiterentwicklung des Ansatzes liefern Ágel/Hennig 2006.

¹¹ Zu einer Typologie von Sprachkorpora s. vergleichend Lemnitzer/Zinsmeister (2006, S. 102–107).

¹² Zur Spezifik und Problematik von „Dialekt als Sprache literarischer Werke“ s. Haas 1983. Als Beispiel für eine Sammlung von Texten aus der Dialektliteratur, die sich (auch)

tionell-mündlichen und mit elizierten Daten, zu denen deutlich mehr regionalsprachliche Analysen vorliegen. Neben den „traditionellen“ konzeptionell mündlichen Daten (Briefe, Tagebücher u. ä.)¹³ erfreuen sich die Daten der Neuen Medien (Internet, Handy) einer wachsenden Beliebtheit in der Variationslinguistik.¹⁴ Für regionalsprachliche Analysen werden dabei besonders häufig Chatdaten herangezogen. Bezogen auf das Medium handelt es sich bei diesen eindeutig um schriftliches Material, wohingegen die Konzeption eher auf Mündlichkeit verweist.¹⁵ Ein Vergleich von regionalsprachlichen Analysen zur Chatkommunikation in der Schweiz und in Deutschland deckt auf, dass die Aussagekraft des Materials deutlich von der soziolinguistischen Makrostruktur der Sprachgemeinschaft abhängt. Während etwa nach Siebenhaar (2005a) Dialekt in regionalen Chaträumen der Schweiz die unmarkierte Varietät darstellt, ergibt sich in Deutschland ein etwas anderes Bild.¹⁶ Hier ist es nur bedingt möglich, (potentiell) regionalsprachliche Merkmale bzw. ihre Sprecher einer Region eindeutig zuzuordnen. Die Aussagekraft regionalsprachlicher Chatdaten aus dem deutschen Raum beschränkt sich i. d. R. auf die areal-soziale Verbreitung zumindest großregionaler Phänomene, die Bewertung und Salienz regionalsprachlicher Varianten, auf die sozialsymbolischen Funktionen von Dialektalismen und Regionalismen u. a.

Die nach wie vor größte Menge medial schriftlicher Daten, die für regionalsprachliche Analysen herangezogen werden, sind elizitierte Daten. Die Erhebung von medial schriftlichen Sprachdaten geht auf eine lange Tradition zurück, die ihren Anfang in der Dialektologie nimmt. Trotz wachsenden Methodenbewusstseins und – damit einhergehend – häufig diskutierter Probleme,

als Untersuchungsgegenstand für regionalsprachliche Analysen anbietet, sei hier Cornelissen/Schmitt (Hg.) 1997 genannt.

¹³ S. etwa die Studie von Ziegler 2006, die auf „Zettelnachrichten“ aus bayerischen Schulklassen basiert. Wie Zieglers Analysen zeigen, gehören „bestimmte Höflichkeitsformen und Akronyme ebenso wie der kontextspezifisch motivierte Gebrauch von Nonstandard- und Standardformen“ zum sprachlichen Repertoire der Schüler. „Der Dialekt stellt – wie zu erwarten – eine wesentliche kulturelle Ressource für die Konstruktion von Nähe dar. Allerdings gilt dies nicht für alle Dialektvarianten gleichermaßen, denn die einzelnen Dialektvarianten haben einen unterschiedlichen Status im Ausdrucksrepertoire der Jugendlichen. Insbesondere deuten die Ergebnisse darauf hin, dass Regionalformen, die großräumig verteilt und sowohl für den mittel- wie oberdeutschen Sprachraum typisch sind, stärker favorisiert werden als Dialektvarianten, die sich durch eine gewisse bairische Exklusivität auszeichnen.“

¹⁴ Zur Frage „Ist das World Wide Web ein Korpus?“ s. Lemnitzer/Zinsmeister (2006, S. 43 f.).

¹⁵ Zur Internetkommunikation im Allgemeinen (Nutzungsmotivation u. a.) s. Siever [u. a.] (Hg.) 2005, Beißwenger (Hg.) 2001. Zu sprachlichen Merkmalen von Chatdaten s. Storrer 2000, Androutsopoulos/Ziegler 2003, Ziegler 2005, Dürscheid 1999 und 2005. Zu linguistischen Aspekten der Mail- und SMS-Kommunikation s. Günther/Wyss 1996, Dürscheid 2002 und Androutsopoulos/Schmidt 2002.

¹⁶ Zur Schweiz s. Aschwanden 2001, Siebenhaar 2003 und 2005b. Zum deutschen Sprachraum s. etwa Christen/Tophinke/Ziegler 2005, Androutsopoulos/Ziegler 2003, Ziegler 2005 und Vogel 2003.

die mit (insbesondere indirekt erhobenen) schriftlichen Befragungsdaten verbunden sind, ist das Verfahren nach wie vor sowohl in dialektologischen als auch regiolektal („umgangssprachlich“) orientierten Projekten sehr beliebt.¹⁷ Aktuelle regionalsprachliche Projekte, in denen schriftliche Befragungsdaten gesammelt werden, sind etwa der Syntaxatlas der Schweiz (s. Bucheli/Glaser 2002) und der „Atlas der deutschen Alltagssprache (ADA)“ (s. Elspaß 2005b). Letzterer dient auch als ein Beispiel für ein regionalsprachliches Forschungsprojekt, das die „Alltagssprache“ im deutschsprachigen Raum mittels Internetbefragungen in den Blick nimmt.

Unter den medial mündlichen Daten stellen sogenannte „Dokumentationsdaten“ einen besonders wertvollen Typ dar. Ihnen ist eigen, dass der Zweck ihrer Aufnahme kein linguistischer ist, sondern diese Daten aus bestimmten nicht-linguistischen Gründen dokumentiert werden. Die Auswirkungen des Beobachterparadoxons werden somit minimiert bzw. ausgeschaltet. Als ein Beispiel für regionalsprachliche Daten dieses Typs seien hier Gesprächsdaten von Polizisten aus Notrufannahmen genannt. Diese werden aktuell im Rahmen eines Kooperationsprojekts des Forschungsinstituts für deutsche Sprache – *Deutscher Sprachatlas* (DSA) in Marburg und des Instituts für Deutsche Sprache (IDS) in Mannheim für das Bundeskriminalamt in Deutschland gesammelt und ausgewertet. Eine erste Auswertung liegt mit Kehrein (2006) vor.

Hinsichtlich der Elizitierung von medial mündlichem Sprachmaterial bedient sich die Regionalsprachenforschung besonders häufig der mündlichen Befragung, die unterschiedliche Grade der Standardisierung und Strukturiertheit annehmen kann.¹⁸ So weist etwa ein informelles und freies Gespräch zwischen dem Explorator und seinen Informanten i. d. R. einen geringeren Standardisierungsgrad als ein gesteuertes Einzel- oder Gruppeninterview auf. Noch standardisierter und strukturierter sind mündlich zu bewältigende Übersetzungsaufgaben oder Lückentexte, wie sie im Rahmen von Sprachatlasprojekten oder von Kompetenzerhebungen eingesetzt werden.¹⁹ Auch was die Datenerhebung per Beobachtung angeht, finden verschiedene Varianten der Methode Anwendung. Diese differieren in der Stellung des Beobachters zum Geschehen, im Standardisierungsgrad der Beobachtung, in der Sichtbarkeit des Beobachters oder in der Art der Beobachtungssituation (s. Lüdtkke 2005). Die Grenze zwischen Beobachtungen bzw. Befragungen mit starkem Standardisierungsgrad und Experimenten ist fließend²⁰, wobei letztere –

¹⁷ Anstelle einer ausführlichen Diskussion über die Vor- und Nachteile der indirekten Erhebung sei hier lediglich auf weiterführende Literatur verwiesen: Atteslander 2005, Glaser 2000, Bucheli/Glaser 2002, Lenz 2004, Cornips/Poletto 2005, Elspaß 2005b.

¹⁸ Zur Methode der mündlichen Befragung s. Briggs 2005, Auer (1995, S. 431–433) und Thun 2005.

¹⁹ S. Bucheli/Glaser 2002, Cornips/Poletto 2005, Lenz (2003, S. 54–65) und Lenz 2004.

²⁰ Dabei ist nach Auwärter (2005, S. 1077) als definierendes Kriterium für ein Experiment „nicht die besonders sorgfältige, kontrollierte Datenerhebung, und erst recht nicht die

wie auch Testverfahren –²¹ relativ selten als Erhebungsmethode eingesetzt werden.

In der Soziolinguistik [wie auch in der Regionalsprachenforschung; A. N. L.] stellen Experimente schon wegen der genannten externen Validitätsprobleme keine uneingeschränkt empfehlbare Forschungsmethode dar. Aber bei Berücksichtigung der Tatsache, dass die alternativen Methoden: Interviews, Tests und Beobachtung, und erst recht die pure Intuition, zum Teil die gleichen Mängel aufweisen, vielleicht zwar, besonders im Falle der Beobachtungsverfahren, höhere ökologische Validität aufweisen, dafür aber mangelnde intersubjektive Überprüfbarkeit und geringere Generalisierbarkeit, gehören Experimente auch nicht zu den diskriminierten Forschungsmethoden.

(Auwärter 2005, S. 1083)

Neben Dokumentations- und elizitierten oralen Sprachdaten sind in Abbildung (1) auch Sprachdaten aus „Funk und Fern(sehen)“ aufgeführt, die hinsichtlich der genannten Aspekte der Natürlichkeit und Authentizität einen speziellen Datentyp darstellen. Da auf diesen Typus nicht weiter eingegangen werden kann, sei hier exemplarisch auf regionalsprachliche Analysen verwiesen, die auf Radio- bzw. Fernsehsprachdaten basieren: Werlen 2000, Christen 2002 und Burger (2005, S. 362–388).

3. Regionalsprachliche Korpora des Deutschen (in Auswahl)

Nach diesem Überblick über verschiedene Datenklassen, die für regionalsprachliche Fragestellungen herangezogen werden können, wird der Fokus auf regionalsprachliche Korpora gerichtet, die auf den vorgestellten Datenklassen aufbauen. Unter einem *regionalsprachlichen Korpus* wird hier eine abgeschlossene Menge an medial mündlichen oder medial schriftlichen Sprachdaten verstanden, die für die modernen Regionalsprachen des Deutschen bzw. für regionalsprachliche Spektrumsausschnitte (Varietäten) repräsentativ sind, durch Metadaten erschlossen sind und (im weiteren Sinne) computertech-nisch aufbereitet vorliegen.²²

Im Rahmen dieses Beitrags können nur einige ausgewählte regionalsprachliche Korpora ausführlicher diskutiert werden. Da sich die Mehrheit regionalsprachlicher Analysen – trotz wachsenden Interesses an insbesondere Internetkommunikationsdaten – nach wie vor auf elizierte Daten stützt, wird sich die folgende Diskussion auf regionalsprachliche Korpora dieses Datentyps konzentrieren.²³ Zu einer ersten Einordnung dient Abbildung (2), aus der die Datenbasis der im Zentrum stehenden Korpora ablesbar ist.

Laborsituation [anzusehen; A. N. L.], sondern ein Design, das wesentlich die ‚Manipulation‘, die systematische Variation von Faktoren vorsieht – oder im Grenzfall die Analyse der systematischen Variation dieser Faktoren durch ‚natürliche Vorkommnisse‘.“

²¹ Zu variationslinguistischen Tests s. Spolsky 2005.

²² Zu einem ähnlichen Korpusbegriff s. McEnery/Wilson (1996, S. 21). Zur Definition und Abgrenzung linguistischer Korpora s. auch Lemnitzer/Zinsmeister (2006, S. 40–44).

²³ Hinweise zu weiteren, hier unberücksichtigten Sprachkorpora des Deutschen liefern

(2) Regionalsprachliche Korpora des Deutschen (in Auswahl) und ihre Datenbasen

Abb. 2

Bei den in Abbildung (2) aufgeführten Korpora medial mündlicher Sprachdaten handelt es sich um Tonkorpora und zugehörige Transkripte, die in die „Datenbank Gesprochenes Deutsch (DGD)“ des Instituts für deutsche Sprache in Mannheim eingebunden sind bzw. noch werden und für wissenschaftliche Zwecke frei zugänglich sind.²⁴

Die Datenbank Gesprochenes Deutsch (DGD) ist die digitale Realisierungsform des Deutschen Spracharchivs (DSAv). Ihre Aufgabe ist, im Laufe der Zeit den Gesamtbestand des DSAv und weitere einzuwerbende Korpora digital zu archivieren und diese Materialien des Archivs in elektronischer Form für interne wie für externe Nutzer zu erschließen und variabel zugänglich zu machen. Die DGD erfüllt damit einerseits Archivfunktion, auf der anderen Seite ist sie ein Forschungsinstrument. (Fiehler/Wagener 2005, S. 139)

Schwitalla/Wegstein 2005, Storrer 2005, Schlobinski (1996, S. 72–74), Wagener/Bausch 1997, Glas/Ehlich 2000 und Lemnitzer/Zinsmeister 2006. S. auch die Literaturangaben in Fn. 25.

²⁴ Im Gegensatz dazu steht etwa das oben erwähnte Polizisten-Korpus aus Notrufannahmen, das aktuell in Mannheim und Marburg aufgebaut wird, aus datenschutzrechtlichen Gründen nicht öffentlich zur Verfügung.

Ausführlichere Informationen über die vier DGD-Korpora sind in Tabelle (3) zusammengestellt. Die Angaben sind v. a. den Webseiten der „Korpusstelle Gesprochenes Deutsch“ des Instituts für Deutsche Sprache in Mannheim entnommen (<http://www.dsav-oeff.ids-mannheim.de> [Stand: April 2006]).

(3) Regionalsprachliche Tonkorpora der „Datenbank Gesprochenes Deutsch (DGD)“ (in Auswahl) (<http://dsav-oeff.ids-mannheim.de>)²⁵

	„Zwirner-Korpus“	„TAVM-Korpus“ ²⁶	„DDR-Korpus“	„Pfeffer-Korpus“
Intendierte Varietät/Sprechlage	Dialekt (+ „Umgangssprache“)	Dialekt	Dialekt (+ „Umgangssprache“)	„städt. Umgangssprache“
Aufnahme-region	v. a. alte Bundesrepublik	ehemalige deutsche Ostgebiete	DDR	deutschsprachiger Raum
Ortsnetz	ca. 1000 Ortspunkte	ca. 900 Ortspunkte	440 Ortspunkte	57 Städte
Anzahl der Aufnahmen	5857	984	1625	398
Aufnahmedatum	v. a. 1955–1960	1962–1965	1960–1964	1961
Aufnahmedauer	jew. ca. 12 min (insges. ca. 1092 h)	jew. ca. 28 min (insges. ca. 462 h)	insges. ca. 393 h	jew. ca. 12 min (insges. ca. 79 h)
Inhalt	v. a. Dialoge, Erzählungen	Erzählungen, Wenkersätze, Wochentage	Erzählungen, Dialoge, Vergleichstexte	Dialoge, Erzählungen
Sprecher	je 3 autochthone Dialektsprecher (3 Generationen) + 3 Aussiedler	(heimat)dialekt-kompetente Vertriebene (> 60 Jahre)	je 3 autochthone Dialektsprecher (3 Generationen)	403 „Umgangssprache-sprecher“ (185 w + 218 m)
Ton/Transkr. (synchron.)	2032	261	–	398

²⁵ Zu den in (3) aufgeführten sowie zu weiteren Korpora der DGD s. Wagener/Bausch 1997 und Wagener 2002. Ausführliche Informationen zum Aufbau der „Datenbank Gesprochenes Deutsch (DGD)“ geben u. a. Fiehler/Wagener 2005 und Wagener 2005.

²⁶ In der DGD ist das TAVM-Korpus („Tonaufnahme der Vertriebenenmundarten“) unter der Bezeichnung „Deutsche Mundarten: ehemalige deutsche Ostgebiete“ und mit der Korpusigle „OS“ aufgeführt.

Die Sprachdaten der vier DGD-Korpora wurden überwiegend zwischen 1955 und 1965 erhoben. Das weitaus größte Korpus der DGD stellt mit fast 6000 Aufnahmen an ca. 1000 ländlichen Ortspunkten im Gebiet der alten Bundesrepublik das „Zwirner-Korpus“ dar.²⁷ Indem Eberhard Zwirner „als Mundartsprecher je eine alte Person aus der ländlichen Bevölkerung[,] dann je eine Person mittleren Alters aus kleinbürgerlichen Kreisen und schließlich je ein[en] Sprecher aus der schulentlassenen Jugend“ berücksichtigte, erreichte er, dass „nicht nur die Mundarten, sondern auch die ländliche und kleinstädtische Umgangssprache und gleichzeitig verschiedene Stadien der bereits im Gang befindlichen und weitgehend vom Alter abhängigen sprachlichen Ausgleichsvorgänge erfaßt“ wurden (Zwirner 1956, S. 12). Falls an einem Aufnahmeort auch Aussiedler aus den ehemaligen ostdeutschen Gebieten anzutreffen waren, nahm Zwirner auch von diesen drei Sprecher auf, und zwar aus der lokal am „häufigsten vertretene[n] landsmannschaftliche[n] Gruppe“ (Bethge/Zwirner 1958, S. 16). Die von Zwirner eingesetzte Erhebungsmethode war überwiegend die der (möglichst) wenig strukturierten Befragung. Die Aufnahmen umfassen vor allem Dialoge und initiierte Erzählmonologe. Da die Aussiedler aus den ehemaligen deutschen Ostgebieten im Rahmen des Zwirner-Projektes nur unsystematisch berücksichtigt werden konnten, diente die „Tonaufnahme der Vertriebenenmundarten (TAVM)“ der systematischen Füllung der arealen Lücken, wobei der Fokus des TAVM-Projektes auf der Erhebung der Basisdialekte lag (s. Bellmann 1964 und 1970). Neben den Dialekten im Osten des deutschen Sprachgebiets fanden auch deutsche Sprachinseln in Ost- und Südosteuropa Berücksichtigung.

Zu den bereits digitalisierten Korpora des Deutschen Spracharchivs in Mannheim gehört auch das sogenannte „DDR-Korpus“²⁸, das in seiner methodischen Anlage eng an das Zwirner-Korpus angelehnt ist, insofern als bei der Datenerhebung drei einheimische Sprecher aus verschiedenen Altersgruppen pro Aufnahmeort in der DDR berücksichtigt wurden. Im DDR-Projekt, das die Zwirner-Aufnahmen in den neuen Bundesländern komplementiert, wurden in mehreren Regionen auch bzw. seltener (beispielsweise in der Berliner Region) nur „umgangssprachliche“ Aufnahmen erstellt. Gemeinsam decken das Zwirner-, das TAVM- und das DDR-Korpus die deutschsprachigen Dialekte in der Mitte des letzten Jahrhunderts in einem relativ dichten Ortsnetz ab.

Ergänzend zu diesen drei Korpora sei hier das „Pfeffer-Korpus“ genannt, das zur Dokumentation „städtischer Umgangssprachen“ erstellt wurde (s. Pfeffer/Lohnes 1984). An insgesamt 57 Städten im deutschsprachigen Raum wurden im Jahr 1961 403 Sprecher aufgenommen, deren Sprechweise

²⁷ Ausführlichere Informationen zur Konzeption von Zwirners Großprojekt geben Zwirner 1956 und 1958 sowie Zwirner/Bethge 1958 und Wagener/Bausch (1997, S. 112–114).

²⁸ Zum zugrunde liegenden Projekt „Tonbandaufnahme der deutschen Mundarten in der Deutschen Demokratischen Republik“ s. Schädlich/Große 1961.

als „umgangssprachlich“ eingestuft wurde. Im Vergleich zu den Zwirner-, TAVM- und DDR-Aufnahmen, die mehrheitlich im dialektalen bis mittleren Bereich des variativen regionalsprachlichen Gesamtspektrums angeordnet werden können, ist das Pfeffer-Korpus – obwohl auch hier die „Varietätenbreite [...] relativ groß ist“ (Spiekermann 2005, S. 107) – insgesamt als standardnäher zu bewerten.

Der Bearbeitungsstand der einzelnen DGD-Korpora – d. h. ihrer dokumentarischen Daten²⁹, Tonaufnahmen und Transkriptionen – ist sehr unterschiedlich. Wie die „Bestandsübersicht“ der DGD aktuell (April 2006) angibt, liegen die Tonaufnahmen und Transkripte des Pfeffer-Korpus komplett digitalisiert und miteinander synchronisiert vor (<http://dsav-wiss.ids-mannheim.de/DSAv/KORPORAB.HTM>). Für das Zwirner-Korpus sind bereits etwa knapp die Hälfte aller Transkripte mit aligniertem Ton verfügbar, für das TAVM-Korpus ca. ein Drittel. Die Alignierung der Transkripte ermöglicht die gezielte Tonwiedergabe zum gewählten Transkriptausschnitt. Bei den verfügbaren Transkriptionen handelt es sich i. d. R. um literarische, weitgehend an der standardsprachlichen Orthographie orientierte Transkripte, die als reine Textdateien vorliegen, in denen mit den Suchbefehlen des Volltextretrievalinstruments COSMAS II recherchiert werden kann (vgl. Fiehler/Wagener 2005, S. 141). Die DDR-Aufnahmen und -transkripte sind bislang noch nicht online zugänglich. Alle vier Korpora sind durch ausführliche Metadaten erschlossen („Dokumentationen“), die u. a. soziodemographische Informationen zu den Sprechern, zu Inhalt und Länge der Tonaufnahmen, zur vertikalen Positionierung der realisierten Sprechlagen auf der Dialekt-Standard-Achse und zur dialektgeographischen Zuordnung der Aufnahmen enthalten.

Da der Fokus dieses Beitrags auf der variationslinguistischen Analyse regionalsprachlicher Korpora liegt, kann der Blick auf medial mündliche Sprachkorpora des Deutschen nur ausschnitthaft sein. Weitere, hier nicht genannte Sprachkorpora – bestehend aus Tonaufnahmen und/oder phonetischen Transkriptionen – sind bzw. werden u. a. in den folgenden Instituten und Archiven gesammelt, archiviert und aufbereitet.

(4) Institute/Archive medial mündlicher Sprachkorpora des Deutschen
(ohne Anspruch auf Vollständigkeit)

- Amt für rheinische Landeskunde (ARL) des Landschaftsverbandes Rheinland (LVR) (Bonn)
- Archiv für Gesprochenes Deutsch (AGD) (IDS Mannheim)
- Bayerisches Archiv für Sprachsignale (BAS) der Ludwig-Maximilians-Universität München

²⁹ „Eine Recherche in den dokumentarischen Daten gibt einen Überblick über die Bestände oder kann je nach der vom Nutzer formulierten Anforderung bestimmte Gruppen von Sprechern, bestimmte Gesprächstypen, Sprache aus verschiedenen Sprachlandschaften u. ä. suchen.“ (Fiehler/Wagener 2005, S. 140).

- Berliner Lautarchiv der Humboldt-Universität Berlin
- Deutsches Rundfunkarchiv (DRA) (Wiesbaden/Babelsberg)
- Hoerburger-Archiv an der Universität Regensburg
- Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universität Kiel
- Max Kade Institute for German American Studies (Madison)
- Phonogrammarchiv der Österreichischen Akademie der Wissenschaften (Wien)
- Phonogrammarchiv der Universität Zürich
- Schallarchiv der Philipps-Universität Marburg
- „Ulmer Textbank“ der Universität Ulm
- Unterfränkisches Dialektinstitut (UDI) der Julius-Maximilians-Universität Würzburg³⁰
- Wossidlo-Archiv (Institut für Volkskunde) der Universität Rostock

Das größte medial schriftliche regionalsprachliche Korpus des Deutschen ist das „Wenker-Korpus“. Es besteht aus ca. 50.000 zweiseitigen Fragebogen, die v. a. Ende des 19. Jahrhunderts von dem Marburger Dialektologen Georg Wenker (1852–1911) erhoben wurden. Wenker bediente sich des Verfahrens der indirekten Befragung, was angesichts der erhobenen Datenmasse sicher die einzig mögliche Erhebungsform darstellte. Im Rahmen seines sukzessive ausgeweiteten Atlasprojektes schickte er – beginnend im Jahr 1876 – an alle Schulorte des ehemaligen deutschen Reichs einen Fragebogen bestehend aus 38 bis 42 Sätzen mit der Bitte an die dortigen Volksschullehrer, „eine Uebersetzung der einliegenden hochdeutschen Sätzchen in die ortsübliche Mundart“ auf den jeweiligen Wenkerbogen einzutragen, wobei die Lehrer aufgefordert waren, die Übersetzung „durch einen oder einige geeignete Schüler anfertigen“ zu lassen.³¹

Die Wenkerbogen stellen die Grundlage von Wenkers „Sprachatlas des deutschen Reichs“ (1888–1923) dar. Auf den 576 zusammengesetzten Gesamtkarten (1643 Teilkarten) des Atlases sind jedoch nur ausgewählte Lemmata der Wenkerbogen kartiert. Auch sind weiterführende Informationen (Geburtsort des Lehrers, Angaben zu den für das Ausfüllen des Bogens ver-

³⁰ Am UDI entsteht aktuell die „Bayerische Dialektdatenbank (Baydat)“, die auf den Transkriptionen des Bayerischen Sprachatlases (BSA) basiert, dessen Daten 1984 ff. erhoben wurden. „Im Rahmen von Baydat soll das kostbare BSA-Material [...] zukunftsicher archiviert, nach modernen Kriterien aufbereitet, für EDV- und Online-Recherchen zur Verfügung gestellt werden und auf diese Weise allen Interessierten einen strukturierten Zugriff bieten.“ (<http://www.uni-wuerzburg.de/germanistik/spr/suf/baydat-udi/baydat.html>) [Stand: 07.06.2006]

³¹ Zitiert aus Wenkers Anschreiben an die Lehrer von 1879/1880 (Erhebungsbogen für Nord-, Mittel- und Süddeutschland). Mit Herrgen/Schmidt (in Bearb.) ist zu betonen: „Im Rahmen sprachdynamischer Analysen kann es [...] notwendig sein, zu kontrollieren, für welche Sprechergruppe die lokalen Sprachdaten überhaupt Geltung beanspruchen können.“

antwortlichen Personen u. a.) allein auf den Rückseiten der Bogen festgehalten. Trotz der Schwierigkeiten, die etwa die Kurrentschrift bei der Analyse der Daten verursacht, sind die Wenkerbogen als eine zentrale, die Wenkerkarten ergänzende Datenquelle für die Dialekte Ende des 19. Jahrhunderts anzusehen. Im von der DFG geförderten Projekt „Digitaler Wenker-Atlas (DiWA)“ sind deshalb nicht allein die Wenkerkarten erstmals publiziert worden, und zwar online, sondern auch die Wenkerbogen wurden digitalisiert und werden aktuell – in Form von Bilddateien – sukzessive online gestellt (<http://www.diwa.info>).³²

(5) „Wenker-Korpus“ (<http://www.diwa.info>)

Intendierte Varietät/Sprechlage	Dialekt („ortsübliche Mundart“)
Erhebungsregion	ehemaliges deutsches Reich (+ „Sprachinseln“)
Ortsnetz	ca. 50.000 Schulorte
Erhebungszeitraum	1876–1887 (+ spätere Nacherhebungen)
Umfang	ca. 50.000 Fragebogen
Inhalt	„Wenkersätze“
Informanten	Volksschullehrer (und Schüler)

Die Datenbankstruktur des DiWA stellt verschiedene Abfrage- und Registerfunktionen für alle bereits bzw. zukünftig integrierten Daten zur Verfügung. Zu diesen gehören neben den ca. 100.000 digitalisierten Fragebogenseiten (jeweils Vorder- und Rückseite) die Wenkerkarten, die Erhebungsorte und ihre (alten bzw. aktuellen) Namen, Karten moderner Regionalatlanten, weitere (historische, soziodemographische u. a.) Karten-„Interpretamente“, Tonaufnahmen, bibliographische und sonstige ergänzende Informationen. Die Wenkerbogen sind in gleicher Weise einsehbar wie die georeferenzierten Wenkerkarten. Beim Anklicken eines Ortspunktes erscheint der jeweilige Wenkerbogen in einem neuen Fenster, in dem beliebige Zoomausschnitte des Bogens definiert werden können.

4. Regionalsprachliche Korpora des Deutschen in der Analysepraxis

Im Folgenden werden drei der vorgestellten Korpora – zum einen das Zwirner- und das Pfeffer-Korpus, zum anderen das Wenker-Korpus – zur Analyse von zwei regionalsprachlichen Phänomenen aus dem lautlichen bzw. grammatischen Bereich herangezogen. Wie die beiden Beispielanalysen zeigen

³² Zur Konzeption und Technik des DiWA s. Herrgen/Lenz 2003, Herrgen/Lenz/Pennay 2003, Kehrein/Lameli/Nickel 2005.

werden, bedingt die deutlich differierende Anlage der Korpusarten – zum einen medial mündliche, zum anderen medial schriftliche Sprachdaten – und auch die deutlich differierende Zugänglichkeit – einmal Transkriptionen in Form von Textdateien, zum anderen Bilddateien mit reichem Zusatzmaterial – unterschiedlichste Analysemöglichkeiten und -methoden.

Beispiel 1: Variationslinguistische Analyse auf der Grundlage des Wenker-Korpus

Der Fokus der ersten Analyse wird auf spezifischen Problemen des Wenkermaterials liegen, die insbesondere mit der laiensprachlichen Verschriftung der Wenkersätze in Zusammenhang stehen. Wie bereits Wrede (1895) betont

„[darf man] die Übersetzungen und demgemäß das, was auf den fertigen Karten steht, [...] nicht als phonetisch genaue Dialektwiedergaben auffassen, sondern jede Karte erfordert, grade so wie jede alte Handschrift, eine besondere, häufig recht schwierige Interpretation, wie die darauf verzeichneten Sprachformen phonetisch zu übertragen sind.“

(Wrede 1895, S. 37)

Das zu diskutierende Phänomen betrifft die Reflexe des westgermanischen *g* im Moselfränkischen. In diesem Dialektraum stehen den standardsprachlichen plosivischen Realisierungen [g] bzw. [k] in- und auslautend regional-sprachliche Spiransvarianten gegenüber.³³ Beispielsweise trifft der standard-sprachliche Plosiv [k] (<g> im Silbenauslaut und präkonsonantisch) im Moselfränkischen auf die Spiranten [ʃ] bzw. [x], wie sie auslautend etwa in [be:ʃ] ‚Berg‘ oder in [gʰno:x] ‚genug‘ vorliegen.³⁴ Die Frage, die im Folgenden diskutiert werden soll, ist die nach der Dynamik der moselfränkischen Spiranten, das heißt nach Stabilität, Progression oder Abbau im Laufe der letzten 100 Jahre. Um diese Frage (zumindest ansatzweise) zu beantworten, werden die Wenkerdaten (Ende des 19. Jahrhunderts) mit anderen regional-sprachlichen Daten aus den vergangenen 120 Jahren kontrastiert.

Da die Graphien der Wenkerlehrer unverändert von den Fragebogen auf die Wenkerkarten übertragen worden sind, kommt ein Blick in die Wenkerkarten einem Blick in die Wenkerbogen (annähernd)³⁵ gleich, denn

³³ Zur standardsprachlichen Aussprachenorm von <g> s. Duden (⁴2000, S. 84–86). Zu den dialektalen Verhältnissen s. Schirmunski (1962, S. 309–316).

³⁴ Neben der alveolaren Spirans [ʃ] findet sich in den rezenten westmitteldeutschen Regionalsprachen u. a. auch koronalisiertes [ç]. S. ausführlicher Herrgen 1986 und Lenz (2003, S. 148–158 u. 165–177).

³⁵ Mit Kehrein/Lameli/Nickel (2005) ist zu betonen: „Das Ansetzen von Isoglossen ist ein subjektives Verfahren, das auf der wissenschaftlichen Analyse des Kartenzeichners basiert und seinem linguistischen Urteilsvermögen unterliegt. Daher ist für linguistisch ungeschulte Rezipienten bei der Arbeit mit den im DiWA publizierten Wenker-Karten grundsätzlich mit Bedacht vorzugehen. Die kombinierte Flächen- und Punktsymbolkartierung evoziert durch die mit markanten Farben dargestellten Leitformen oft einen ersten plakativen Eindruck der Kartensituation, der aber gewissenhaft hinterfragt und

[w]as die technische Herstellung der Sprachatlaskarten betrifft, so sei nur daran erinnert, dass die Kartenbilder aus dialektischen Übersetzungsformularen gewonnen sind, [...] und dass die in diesen [...] Formularen gebotenen Sprachformen ganz mechanisch nach der Originalschrift, ohne theoretisierende Besserungen oder dergl., in die Karten eingetragen wurden.

(Wrede 1895, S. 34)

In (6) ist ein Ausschnitt der Wenkerkarte „(ge)nug (Satz 16)“ aus dem linksrheinischen Moselfränkischen abgebildet (eingebettet in die DiWA-Benutzeroberfläche), wobei die Städte Trier und Koblenz zur Orientierung hervorgehoben sind. Kartiert ist die zweite Silbe von *genug*. Obwohl es sich um einen moselfränkischen Kartenausschnitt handelt, sind nördlicheres „nog“ und südlicheres „nug“ (im moselfränkisch-rheinfränkischen Übergangsgebiet) als Leitformen für die durch – im Original farbige – Isoglossen abgegrenzten Areale kartiert. Innerhalb der beiden Areale weisen verschiedene Zusatzsymbole auf lokale Abweichungen der Lehrergraphien von den Leitformen „nog“ und „nug“ hin. In der – hier nicht abgebildeten – Legende wird ersichtlich, dass es sich bei senkrechten schwarzen Strichsymbolen um *ch*-Graphien der Lehrer handelt, die aber nur vereinzelt auftauchen. Dies widerspricht auf den ersten Blick der Erwartung, in den abgebildeten Regionen flächendeckend auf Spiranten zu treffen. Andere Wenkerkarten bzw. Wenkerbogen, die zur Analyse der dialektalen Reflexe von westgermanisch *g* herangezogen werden können, bestätigen jedoch, dass die *g*-Allographen in den Transliterationen der moselfränkischen Lehrer keine Ausnahme darstellen bzw. die *ch*-Graphien jeweils deutlich seltener auftreten.

Zur Interpretation dieser Befunde sind weitere regionalsprachliche Quellen heranzuziehen. Zu diesen gehören nicht zuletzt „Ortsgrammatiken“ vom Ende des 19. Jahrhunderts bzw. Anfang des 20. Jahrhunderts, die als besonders wertvolle, das Wenkermaterial ergänzende Datenquellen anzusehen sind. Für die dialektalen Reflexe des alten *g* im moselfränkischen Sehlern gibt z. B. Ludwig (1906) folgende Beschreibung:

[...] inlautend erscheint mhd. *g* 1. als stimmhafte Spirans vor stimmh. Laut a) nach Hartgaumenvokalen sowie nach *l*, *r* palatal = *j*, [...]; b) nach Flachzungen- und Weichgaumenvokalen velar = *ɣ*, [...]; 2. als stimmloser Reibelaut vor stimmh. Konson., je nach der Natur des vorhergehenden Vokals [...], bald als palatales *ç*, [...] bald als velares *x*, [...]; c) auslautendes, altes germ. *g* (> mhd. *c*), wie auch das nach Apokope eines Schluss-*e* an die Endung rückende *g* ist nach obiger Regel (vgl. 2) bald *ç*, [...] bald *x*, [...]; d) intervokalisches *g* ist bisweilen völliger Unterdrückung ausgesetzt [...].

(Ludwig 1906, S. 43 f.)

erst durch kleinräumige Betrachtung des Analysegebiets belegt werden muss. Es ist wichtig, die impliziten Interpretationsleistungen der Kartenzeichner nachzuvollziehen, die durch die Georeferenzierung des historischen Kartenmaterials offen gelegt werden können, [...].“

Neben den i. d. R. von phonetisch geschulten Dialektologen erstellten Ortsmonographien weisen auch regionalsprachliche Tonaufnahmen (spätestens) aus der zweiten Hälfte des 20. Jahrhunderts die Spiranten flächendeckend für das Moselfränkische nach. Mit den Zwirner-Aufnahmen der DGD sind hier Wenkersatzaufnahmen des Forschungsinstituts für deutsche Sprache – *Deutscher Sprachatlas* in Marburg zu nennen, die aktuell in DiWA integriert und dort ortspunktgenau über die Wenkerkarten abgerufen werden können (www.diwa.info). Ebenso belegen der „Mittelrheinische Sprachatlas (MRhSA)“, ein moderner Regionalatlas mit Dialektdateien aus den 1980ern³⁶, sowie aktuelle variationslinguistische Untersuchungen eine stabile Verteilung der Spiranten in den Dialekten und auch in den Regiolekten („Umgangssprachen“) des Moselfränkischen in der zweiten Hälfte des 20. Jahrhunderts (s. Lenz 2003, S. 148–158). Die gesammelten Quellen stützen die Thesen, dass die Spiransvarianten erstens auch zur Wenkerzeit in den Dialekten des Moselfränkischen verankert waren und zweitens bis heute ein abbauresistentes Dialekt- und Regiolektmerkmal des Moselfränkischen darstellen, das bis in standardnahe Sprechlagen hinein Verwendung findet.³⁷

Vor dem Hintergrund dieser Ergebnisse werden die *g*-Allographen der Wenkerlehrer im Moselfränkischen nicht als Belege für plosivische Lautvarianten zur Wenkerzeit interpretiert, sondern als Spiranten, die von den Lehrern mit dem Schriftzeichen <g> wiedergegeben wurden. Es ist davon auszugehen, dass die Lehrer (und Schüler) der Wenkerzeit auch in ihrer Leseaussprache das Schriftzeichen <g> spirantisiert ausgesprochen haben. Diese Interpretation einer spirantischen <g>-Aussprache im intendierten „Hochdeutsch“ westmitteldeutscher Sprecher ist nicht neu.³⁸ Beispielsweise geht Schirmunski (1962) in seiner „Deutschen Mundartkunde“ auf das Phänomen ein:

³⁶ Beim MRhSA handelt es sich um einen bidimensionalen Regionalatlas, insofern als neben einer alten immobilen Sprechergruppe auch jüngere Berufsspendler bezüglich ihrer Dialektkompetenz befragt wurden. S. ausführlicher Bellmann 1994. Auch im intergenerationellen Vergleich weisen die dialektalen Spiranten eine (relative) Stabilität im Moselfränkischen bzw. sogar eine Ausweitung ins Rheinfränkische auf. S. hierzu etwa die MRhSA-Karten 391/1–2, 392/1–2 und 393/1–2.

³⁷ Wie u. a. für die Region Wittlich (Eifel) nachgewiesen werden konnte, korreliert die Abbaustabilität der *g*-Spiranten mit ihrer Stellung und ihrem lautlichen Kontext. Die größte Stabilität weist die Spirans [x] im Silbenauslaut auf. Im Moselfränkischen ist sie in dieser Position zu den quintären und damit zu den abbauresistentesten Merkmalen der Regionalsprache zu rechnen. Vgl. Lenz (2003, S. 186–191).

³⁸ S. neben Schirmunski (1962, S. 315f.) auch besonders Mihm (2004, S. 190) und Mattheier (1979, S. 220). Ein Blick in die Forschungsliteratur belegt, dass auch in älteren Texten der Vorwenkerzeit (etwa im Diminutivsuffix *-gen*) häufig eine Allographie von <g> und <ch> als spirantische Repräsentanten anzunehmen ist. Vgl. Wegera 2000, Elspaß (2000, S. 254, Fnn. 13 und 14) sowie Elspaß (2005a, S. 340–345). An dieser Stelle danke ich noch einmal herzlich Stephan Elspaß (Augsburg) und Arend Mihm (Duisburg), die mich auf ergiebige Literatur zum Phänomen aufmerksam gemacht haben.

Spirantische Aussprache des intervokalischen und des auslautenden *g* ist sowohl in Nord- als auch besonders in Mitteldeutschland bis heute als lokale Ausspracheweise der Literatursprache erhalten geblieben und wird weitgehend auch weiterhin beibehalten. [...] Goethe reimte als gebürtiger Frankfurter (Main) entsprechend seiner Aussprache *Buch* : *genug*, [...] *zeigen* : *reichen* [...]. In Verbindung mit der Frage der Schaffung einer einheitlichen literatursprachlichen Aussprachenorm für das Deutsche entspann sich noch im letzten Viertel des 19. Jhs. in den philologischen und pädagogischen Zeitschriften eine äußerst heftige Polemik über die richtige Aussprache des *g*. Eine bedeutende Gruppe der Beteiligten, und zwar vorwiegend solche, die aus Mitteldeutschland stammten (darunter der bekannte Phonetiker W. Vietor) verteidigte die mitteldeutsche Norm (explosives *g*- im Wortanlaut, Reibelaut im In- und Auslaut) als wünschenswerten Kompromiß zwischen den niederdeutschen und den oberdeutschen Aussprachegepflogenheiten. Doch die Autorität der ‚Bühnenaussprache‘ erlangte nach langem Streit das oberdeutsche Verschluß-*g*; [...]. (Schirmunski 1962, S. 315 f.)

Angesichts dieser Befunde ist es nicht verwunderlich, dass auch Wrede (1895) die <*g*>-Aussprache als ein Beispiel anführt, um die „richtige Interpretation der Sprachatlaskarten“ zu illustrieren:³⁹

Regelmäßige *g*-Schreibung lässt durchaus nicht ohne weiteres auf Verschlusslaut schließen, beruht vielmehr oft darauf, dass die betreffende Mundart für das *g* in allen Stellungen nur einen Laut besitzt, der auch beim Schriftdeutschsprechen gilt (mag es nun Spirans oder Explosiva sein), und daher keinen Grund zu unterscheidenden Schreibungen hatte.

(Wrede 1895, S. 41)

Belege für die auch über das 19. Jahrhundert hinaus währende Konstanz dieser Laut/Schriftzeichen-Zuordnung (Schriftzeichen <*g*> = spirantische Aussprache) – auch über das Moselfränkische hinaus – finden wir (mindestens) bis zum Ende des 20. Jahrhunderts. So nehmen etwa Klein/Mattheier/Mickartz (1978, S. 73–76) die „Verwechslung von Rheinisch /j/, /r/, /sch/, /ch/ und Hochdeutsch /g/“ in die Reihe der Interferenzfehler rheinischer Schüler auf,⁴⁰ und auch das „Wittlicher Wörterbuch“ (21984) enthält laiensprachliche Einträge aus der Mitte des letzten Jahrhunderts (s. (7)), in denen eine Interpretation der spirantischen Aussprache von <*g*> außer Frage steht (s. besonders (7a)).⁴¹

³⁹ Für diesen Literaturhinweis danke ich herzlich Jürgen Erich Schmidt (Marburg).

⁴⁰ Auch Stellmacher (1981, S. 63) zitiert „zahlreiche hyperkorrekte Schreibungen“ im Bereich der /g/-/x, ç/-Alternation, die er als „Anzeichen für die große Unsicherheit in diesem Fehlerbereich“ niedersächsischer Schüler deutet: „*Wir haben noch fröhliche Tage. [...] Weihnagsgeschenk.*“ Zu vergleichbaren Fehlern westfälischer bzw. hessischer Schüler s. Niebaum (1977, S. 47–50) bzw. Hasselberg/Wegera (1976, S. 40–42).

⁴¹ Zu einer ausführlicheren Analyse der Wörterbucheinträge s. Lenz (2003, S. 197–199).

(7) Einträge aus dem „Wittlicher Wörterbuch“ (1984)

(7a)

	Basisdialekt	Wörterbucheinträge
<i>Oberste Straße</i>	[ɛ:βɪʃd̥ʒa:s]	<i>Ewigdgaahs</i>
<i>Wittlich</i>	[vɛdɪʃ]	<i>Wedlig</i>

(7b)

	Basisdialekt	Wörterbucheinträge
<i>Waage</i>	[vɔ:x]	<i>Woog</i>
<i>Tage</i>	[d̥ɛ:j]	<i>Dääg</i>

(7c)

	Basisdialekt	Wörterbucheinträge
<i>genug</i>	[g̊ʰno:x]	<i>genooch</i>
<i>Krieg</i>	[kre:j]	<i>Kreesch</i>

Zusammenfassend ist aus Beispiel 1 phänomenspezifisch abzuleiten: Auch in der Wenkerzeit sind die spirantischen Reflexe des westgermanischen *g* nicht nur ein Charakteristikum der moselfränkischen Dialekte, sondern des gesamten mündlichen Spektrums moselfränkischer Sprecher.⁴² Bezüglich des Wenkermaterials ist generell abzulesen, dass sich aus den Dialektübersetzungen der Lehrer auch Aussagen über das Verhältnis von Laut und Schrift Ende des 19. Jahrhunderts und damit auch über die eigentlich von Wenker intendierte Varietät Dialekt hinaus, nämlich über das landschaftlich geprägte Hochdeutsch der Zeit, ablesen lassen.⁴³ Im Blick auf das Gesamtmaterial stellen solche vermeintlichen „Fehler“ der Lehrer den Ausnahmefall dar. Wie aktuelle und mit modernster Technik durchgeführte Analysen des Wenkermaterials belegen, können die Lehrertransliterationen insgesamt als verlässliche und exakte Datenquelle für die Dialekte Ende des 19. Jahrhunderts bewertet werden.⁴⁴

⁴² Nach Elspaß (2000, S. 268) „kann für das 19. Jh. wohl kaum von umgangssprachlichen Varietäten gesprochen werden, die den Sprecherinnen und Sprechern als Wahlmöglichkeiten zwischen Dialekt und Standard zur Verfügung standen. Für sie war ihr – aus heutiger normativer Sicht unzulängliches – intendiertes Hochdeutsch die Form der für sie größtmöglichen Annäherung an den Standard.“

⁴³ S. hierzu Herrgen/Schmidt (in Bearb.): „Im Rahmen der Sprachdynamikforschung beginnen wir soeben erst zu verstehen, dass uns dieser offensichtliche methodische Schwachpunkt [der Laienschreibungen im Wenkermaterial; A. N. L.] nicht nur die Möglichkeit beschert hat, den Dialektwandel anhand empirischer Daten über ein Jahrhundert hinweg zu verfolgen, sondern uns darüber hinaus ungeahnte Einblicke in das sprechsprachliche Gesamtsystem des Deutschen im ausgehenden 19. Jahrhundert ermöglicht.“

⁴⁴ S. etwa die Analysen von Herrgen (2006), Rabanus 2005 und Schmidt 2005.

Beispiel 2: Variationslinguistische Analyse auf der Grundlage von Zwirner- und Pfeffer-Korpus

Die zweite Beispielanalyse will einen Einblick in die variationslinguistische Arbeit mit den regionalsprachlichen Tonkorpora der „Datenbank Gesprochenes Deutsch (DGD)“ geben. Unter diesen sind bislang das Pfeffer-Korpus und das Zwirner-Korpus komplett bzw. zu größeren Teilen aufbereitet und zur online-Recherche freigegeben.⁴⁵ Als exemplarischer Untersuchungsgegenstand, der auf der Grundlage dieser verfügbaren Daten analysiert werden soll, wurde das „Rezipientenpassiv“ ausgewählt.⁴⁶ Dieses Phänomen steht neben anderen grammatischen und semantischen Besonderheiten, die das verbale Wortfeld des Gebens und Bekommens im Deutschen, aber auch in anderen, besonders westgermanischen Sprachen aufweist (s. Lenz 2007a).

(8) Aktiv, Vorgangs- und Rezipientenpassiv im Vergleich

Aktiv	<i>Sein Vater überreichte ihm das Geschenk.</i>
Vorgangspassiv	<i>Das Geschenk wurde ihm (von seinem Vater) überreicht.</i>
Rezipientenpassiv	<i>Er bekam das Geschenk (von seinem Vater) überreicht.</i>

Wie die Bezeichnung ausdrückt, ist es (häufig) der Rezipient der Handlung, der im Rezipientenpassiv fokussiert wird, indem ihm Subjektstatus zukommt. Als Auxiliare des Rezipientenpassivs dienen *bekommen* und *kriegen* (seltener auch *erhalten*). Dabei fungiert *bekommen* (bzw. *erhalten*) als stilistisch gehobene und schriftsprachliche Variante, während *kriegen* als Variante des mündlichen Registers bewertet wird (vgl. Duden ⁷2005, S. 557). Was die dialektale Verbreitung des Rezipientenpassivs im Deutschen angeht, hat erstmals und jüngst Glaser (2005) detailliertere Ergebnisse formuliert. Ihre Auswertung der Forschungsliteratur bzw. der Daten aus dem Syntaxatlas der Schweiz ergibt,

dass das Mitteldeutsche zweifellos als Kernlandschaft des ‚bekommen‘-Passivs angesprochen werden kann, während es sich für das Niederdeutsche zwar ebenfalls sehr weiträumig nachweisen lässt, sich aber dennoch der Eindruck ergibt, dass es keinesfalls die gleiche Produktivität und Vielfältigkeit besitzt [...]. Die Untersuchung zum Schweizerdeutschen hat [...] ergeben, dass es allenfalls Ansätze zur Herausbildung eines ‚bekommen‘-Passivs gibt, dass dieses aber mit grosser Wahrscheinlichkeit keine dialektale Tradition hat.

(Glaser 2005, S. 50 u. 54)

⁴⁵ Das TAVM-Korpus bleibt in der folgenden Analyse unberücksichtigt, da erstens bislang nur ca. ein Drittel der Tonaufnahmen/Transkripte zur Recherche freigegeben sind und zweitens diese aufbereiteten Daten fast ausschließlich zum Schlesischen gehören. Hinzukommt, dass das TAVM-Korpus aufgrund der methodischen Anlage der Datenerhebung deutlich basisdialektaler einzustufen ist als das Zwirner-Korpus. Insofern verfälschte eine quantitative Analyse auf der Grundlage der vereinten Zwirner- und TAVM-Daten das Gesamtbild.

⁴⁶ Zum Phänomen im Deutschen s. Leirbukt 1997, Glaser 2005, Duden (⁷2005, S. 556–558) und Zifonun [u. a.] (1997, S. 1824–1829) und Lenz 2007b.

Diese Thesen bezüglich der Arealität des Rezipientenpassivs werden im Folgenden auf der Grundlage der in der „Datenbank Gesprochenes Deutsch (DGD)“ verfügbaren regionalsprachlichen Tonaufnahmen aus dem Zwirner- und dem Pfeffer-Korpus überprüft.

In (9) sind erste Ergebnisse der Korpusrecherchen einzusehen. Die Analyse aller *bekommen*-, *kriegen*- und *erhalten*-Belege, die in den 2430 alignierten Transkripten der beiden Korpora ermittelt werden konnten, ergibt insgesamt 432 Rezipientenpassiv-Belege, was einem prozentualen Anteil von 17,8 % entspricht. Das bedeutet, dass in jeder fünften bis sechsten verfügbaren Tonaufnahme ein Beleg für ein Rezipientenpassiv zu finden ist. Die Tatsache, dass die Aufnahmen aus Zwirner- und Pfeffer-Korpus jeweils ca. 12 Minuten Aufnahmelänge umfassen und damit eine vergleichbare Dauer aufweisen, stützt die Aussagekraft der berechneten Prozentzahlen. (Genauere Angaben bezüglich des Wort- bzw. Satzumfanges der Teilkorpora, die der Recherche zugrunde lagen, können hier nicht gemacht werden, da diese Recherchefunktion in der DGD noch nicht möglich ist.⁴⁷)

(9) Rezipientenpassive im Zwirner- und Pfeffer-Korpus

	Anzahl der verfügbaren Tonaufnahmen	Rezipientenpassive	
		absolute Häufigkeiten	relative Häufigkeiten
Zwirner	2032	360	17,7 %
Pfeffer	398	72	18,1 %
GESAMT	2430	432	17,8 %

Als Rezipientenpassive wurden – in Anlehnung an Leirbukt (1997) – nur solche Konstruktionen bestehend aus „*bekommen/erhalten/kriegen* + Part. II“ gewertet, die eine Retransformierung des Passivsatz-Subjekts in einen Aktivsatz-Dativ (Dativobjekt oder auch frei(er)er Dativ) ermöglichen. Folglich wurden die in (10) aufgeführten Konstruktionen nicht als Rezipientenpassive gezählt. Während das Partizip *gehackt* in (10a) (aus einer Aufnahme des TAVM-Korpus) als Objektsprädikativ in Verbindung mit dem Vollverb *kriegen* im Sinne von ‚erhalten, empfangen‘ zu interpretieren ist und somit keine Verbalperiphrase vorliegt, exemplifiziert (10b) eine *kriegen*-Periphrase zum

⁴⁷ In den Korpusinformationen der DGD finden sich aktuell lediglich Angaben über den digitalisierten Gesamtumfang eines Korpus, unabhängig von seiner Verfügbarkeit und damit auch unabhängig von seiner Recherchierbarkeit im Internet (<http://dsav-wiss.ids-mannheim.de/DSAv/KORPORA/>). Angaben über Teilbestände, die einer Recherche zugrunde liegen (z. B. Wortumfang der alignierten Transkripte eines Teilkorpus zu einer ausgewählten „Mundartgruppe“) können noch nicht erfragt werden. Da das Pfeffer-Korpus komplett aufbereitet und online gestellt ist, liegt das gesamte digitalisierte Korpus auch der Berechnung in (9) zugrunde. Das Pfeffer-Korpus umfasst insgesamt 645.540 Wörter (tokens), darunter 42.075 unterschiedliche Wortformen (types).

Ausdruck des Bewerkstelligen. Bei dieser Modalkonstruktion handelt es sich nach Glaser (2005, S. 60) dialektologisch gesehen wohl „um eine im wesentlichen niederdeutsche (und niederfränkische)“ Konstruktion.

Zumindest im Westen reicht die Konstruktion auch in den hochdeutschen Raum hinein, worauf Belege aus dem Luxemburgischen [...] und dem Ripuarischen deuten. [...] Im Unterschied zum ‚bekommen‘-Passiv, das auch schriftlich gut und v. a. auch räumlich breit belegt ist, ist das modale Aktiv mit *kriegen/bekommen – erhalten* kommt hier nicht in Frage – aber offenbar bislang weitgehend an die gesprochene Sprache gebunden geblieben und hat auch dabei (noch) keine überregionale Verbreitung gefunden. Es ist zu vermuten, dass es aus den niederdeutschen Basisdialekten, wo es sich sogar wesentlich besser als das ‚bekommen‘-Passiv belegen lässt, in den (nördlich geprägten) regionalen Substandard eingedrungen ist.

(Glaser 2005, S. 59)

Die dritte, in (10c) exemplifizierte Konstruktion, die nicht in die Berechnung der Rezipientenpassivbelege eingegangen ist, weist ein Subjekt auf, das mit einem Aktivsatz-Akkusativ korrespondiert, bzw. umfasst Sätze, bei denen nicht zwangsweise von einer Ellipse eines Akkusativobjektes ausgegangen werden kann.⁴⁸

(10) „bekommenkriegenlerhalten + Part. II“-Konstruktionen ohne Passivstatus (im definierten Sinne)

(Beispiele aus den DGD-Korpora)

(10a) Objektsprädikativ mit Zustandsinterpretation

OS 494: [...] *dann kriegten sie auch ein gekochtes Hühnerei fein gehackt*
 ‚dann bekamen sie auch ein fein gehacktes gekochtes Hühnerei‘

(10b) Objektsprädikativ mit kausativ-resultativer Interpretation („Modalkonstruktion“)

ZW8L7: [...] *da haben wir so Mühe mit, dass wir das dann getan kriegen*
 ‚wir haben Mühe damit, es zu schaffen‘

(10c) Aktivsatz-Akkusativ

ZWY08: [...] *und deshalb hab ich auch oft geschimpft gekriegt*
 ‚und deshalb bin ich auch oft geschimpft worden‘

Der in (9) ermittelte korpusübergreifende Gesamtwert von 17,8 % weicht nur minimal von den ermittelten Frequenzen des Rezipientenpassivs in den beiden Teilkorpora ab (17,7 % bzw. 18,1 %). Obwohl das Zwirner-Korpus insgesamt dialektalere Aufnahmen als das „umgangssprachlichere“ Pfeffer-Korpus umfasst, zeigen die beiden Korpora hinsichtlich des Rezipientenpassivs

⁴⁸ Von der Ellipse eines Akkusativobjektes wurde dagegen in Konstruktionen mit *bezahlen* oder *verhauen* u. a. ausgegangen. Eine ausführlichere Diskussion des in (10c) exemplifizierten Konstruktionstyps kann hier nicht geleistet werden. S. dazu Leirbukt (1997, S. 13–15). Zur Klassifikation der Konstruktionen in (10) siehe auch Zifonun [u. a.] (1997, S. 1827).

keine signifikanten quantitativen Unterschiede. Differenzen zwischen den beiden Korpora werden jedoch offensichtlich, wenn nach den konkreten Auxiliaren gefragt wird, die in den Rezipientenpassivbelegen vorkommen. Wie aus Tabelle (11) abzulesen ist, wird das Rezipientenpassiv im Pfeffer-Korpus zu 79,2 % mit dem Auxiliar *bekommen* realisiert, wohingegen im Zwirner-Korpus *kriegen* mit 90,3 % eindeutig als Auxiliar überwiegt. *Erhalten* ist mit nur einem Beleg im Pfeffer-Korpus vertreten. Wenn auch Zwirner- und Pfeffer-Korpus hinsichtlich der Quantität realisierter Rezipientenpassive kaum differieren, sind sie somit in der Qualität der Belege klar voneinander abzugrenzen. Die dialektalere Anlage der Zwirner-Aufnahmen kommt bezüglich des Rezipientenpassivs in der Dominanz des Passivauxiliars *kriegen* zum Ausdruck, während in den standardsprachnäheren Pfeffer-Aufnahmen das Auxiliar *bekommen* dominiert.

(11) Auxiliare des Rezipientenpassivs im Zwirner- und Pfeffer-Korpus
(absolute und relative Häufigkeiten)

	<i>kriegen</i>		<i>bekommen</i>		<i>erhalten</i>	
Zwirner	325	90,3 %	35	9,7 %	–	–
Pfeffer	14	19,4 %	57	79,2 %	1	1,4 %
GESAMT	339	78,5 %	92	21,3 %	1	0,2 %

Aus den Frequenzwerten in (9) könnte die These abgeleitet werden, dass sich das Rezipientenpassiv – zumindest hinsichtlich seiner Vorkommenshäufigkeit – relativ gleichmäßig über die Dialekte (annähernd repräsentiert im Zwirner-Korpus) und Regiolekte (annähernd repräsentiert im Pfeffer-Korpus) des Deutschen verteilt. Um diese These zu verifizieren, ist ein Blick auf die areale Verteilung der Belege notwendig, die in (12) – zunächst für Zwirner- und Pfeffer-Korpus gemeinsam – einzusehen ist. Das Balkendiagramm zeigt die relativen Häufigkeiten von Rezipientenpassiven in ihrer Verteilung auf die „Mundartgruppen“, wie sie in der DGD-Abfrage aufgeführt sind. Unberücksichtigt blieben die Mundartgruppen Friesisch, Mecklenburgisch-Vorpommersch, Märkisch, Schleswigsch, Westthüringisch und Südbairisch, da die Anzahl ihrer verfügbaren Aufnahmen deutlich unter dem Durchschnitt von ca. 108 Aufnahmen pro Mundartgruppe liegen, sowie die Aufnahmen aus „außereuropäischen Sprachinseln“.

(12) Areale Verteilung der Rezipientenpassive aus Zwirner- und Pfeffer-Korpus

Das Rezipientenpassiv ist in allen hier differenzierten Regionen des deutschsprachigen Raums vertreten, allerdings in deutlich unterschiedlichen Frequenzen. Unter den DGD-Mundartgruppen zeigen das Niederfränkische (44,8%), das Mittelfränkische (35,3%) und das Rheinfränkische (38,8%) die höchsten Frequenzwerte, wobei zu bedenken ist, dass auch die Großgruppe der hessischen Aufnahmen in die DGD-Kategorie „Rheinfränkisch“ eingeordnet ist.

Auf einer höheren Klassifizierungsebene sind die Mundartgruppen den drei großen Sprachräumen Nieder-, Mittel- und Oberdeutsch zugeordnet. Wie aus der Zusammenschau von Balkenhöhe und Sprachraum ablesbar wird, weist das Mitteldeutsche auch insgesamt die höchsten Frequenzen an Rezipientenpassiven auf, nämlich insgesamt 33,6%, gefolgt vom Niederdeutschen mit insgesamt 18,5%. Die geringste Anzahl an Rezipientenpassiven ist im oberdeutschen Raum zu finden (8,4%). Innerhalb der nieder- und oberdeutschen Mundarten sind es jeweils die an das Westmitteldeutsche angrenzenden Mundarten, die die höchsten Frequenzen an Rezipientenpassive innerhalb ihres Großraumes aufweisen, nämlich das Niederfränkische im Niederdeutschen bzw. das Süd- und Ostfränkische im Oberdeutschen.

In (13) und (14) sind die Frequenzen des Rezipientenpassivs für Zwirner- und Pfeffer-Korpus getrennt aufgeführt. Es fanden jeweils nur solche Mundartgruppen Berücksichtigung, die mit mehr als 10 bzw. 20 recherchierbaren Aufnahmen im Pfeffer- bzw. Zwirner-Korpus vertreten sind.

(13) Areale Verteilung der Rezipientenpassive im Zwirner-Korpus

Aus dem Vergleich der dialektaleren Zwirner- mit den standardnäheren Pfeffer-Aufnahmen lassen sich folgende Thesen ableiten: Das Rezipientenpassiv ist ein genuin dialektales Phänomen des Westmitteldeutschen und – mit Blick auf Abbildung (13) – des nordwestlich angrenzenden Niederfränkischen.⁴⁹ In den Zwirner-Aufnahmen dieser Regionen tritt das Rezipientenpassiv in Frequenzen von über 40% auf, während die Vorkommenshäufigkeiten im Zwirner-Korpus für alle anderen Regionen bei unter 20% liegen. Ausgehend von diesem Ergebnis werden die drei benachbarten Regionen Niederfränkisch, Mittelfränkisch und Rheinfränkisch als dialektaler Kernraum des Rezipientenpassivs ausgewiesen. Von diesem Zentrum breitet sich das Phänomen sukzessive sowohl horizontal (in der Fläche) als auch vertikal (auf der Dialekt-Standard-Achse) in die Regionalsprachen des Deutschen aus. Dabei ist die Progression des Rezipientenpassivs in den 50ern und 60ern des letzten Jahrhunderts, in denen die Zwirner- und Pfeffer-Aufnahmen entstanden sind, im Vergleich der Graphiken (13) und (14) besonders im Westniederdeutschen und im Oberdeutschen (bes. nordalemannisch und mittelbairisch) nachzuvollziehen.

⁴⁹ Da das Pfeffer-Korpus keine Aufnahmen aus dem Niederfränkischen enthält, musste diese Dialektregion in Abbildung (14) unberücksichtigt bleiben. Die hohe Frequenz an Rezipientenpassivbelegen, die für das Niederfränkische aus Abbildung (12) abzulesen ist, basiert somit allein auf Zwirner-Aufnahmen.

(14) Areale Verteilung der Rezipientenpassive im Pfeffer-Korpus

Unter den oberdeutschen Regionen ist es das Bairische, das insgesamt die niedrigsten Frequenzen an Rezipientenpassiven aufweist (s. (12)). Für diese Region können aus dem Vergleich von (13) und (14) Evidenzen für die von Glaser (2005) aufgestellte These abgeleitet werden:

[...] Unsicherheiten bestehen bezüglich der Verhältnisse im Bairischen, wo bislang nur für das ans Mitteldeutsche grenzende Egerländische der Gebrauch verschiedener Passivperiphrasen belegt ist, während im übrigen die Konstruktionsmöglichkeiten sehr restringiert zu sein scheinen.

(Glaser 2005, S. 54)

Ebenso wird Glasers These (2005, S. 54), dass das Rezipientenpassiv im Schweizerdeutschen „mit grosser Wahrscheinlichkeit keine dialektale Tradition hat“, durch die nördlich ans Höchstalemannisch anschließenden Mundartgruppen Hochalemannisch und Nordalemannisch gestützt. Denn in diesen beiden Regionen – wie auch im Bairischen – sind es die Pfeffer- und damit die dialektferneren Aufnahmen, die eine höhere Frequenz an Rezipientenpassiven aufweisen als die Zwirner-Aufnahmen.

Phänomenspezifisch ist aus der zweiten Beispielanalyse abzuleiten: Die Ergebnisse, die aus der Analyse von Zwirner- und Pfeffer-Korpus gewonnen wurden, stützen nicht nur die Thesen, die auf der Grundlage anderer Datenquellen formuliert wurden, sondern die Korpusdaten ermöglichen auch einen tieferen Einblick in die regionalsprachlichen Verhältnisse des Rezipientenpassivs. Generell ist bezüglich der variationslinguistischen Arbeit mit den DGD-Korpora festzuhalten: Auch wenn es sich (bislang) bei den recherchierbaren Transkripten um reine Textdateien ohne linguistische Annotationen handelt,

können das Zwirner- und Pfeffer-Korpus bereits in ihrem aktuellen Bearbeitungsstand als eine höchst ergiebige Analysequelle für die Regionalsprachen der zweiten Hälfte des 20. Jahrhunderts bewertet werden. Wie am Beispiel des Rezipientenpassivs exemplifiziert werden konnte, repräsentiert das Zwirner-Korpus einen „tieferen“, das heißt dialektaleren regionalsprachlichen Spektrumsbereich als das standardsprachnähere Pfeffer-Korpus.

5. Zusammenfassung

Im Zentrum des Beitrags standen korpuslinguistische Fragestellungen aus variationslinguistischer Perspektive. Dabei lag der Fokus auf der Analyse regionalsprachlicher Daten bzw. Korpora.

Im Rahmen einer Überblicksdarstellung wurden zunächst Datenklassen präsentiert, die zu regionalsprachlichen Analysen herangezogen werden (können). Unter den beiden Hauptklassen „(objektive) Sprachdaten“ und „(subjektive) Spracheinstellungsdaten“ wurden erstere als Diskussionsgrundlage für diesen Beitrag ausgewählt. Zur weiteren Klassifikation der Sprachdaten diene – in Anlehnung an das Konzept von Koch/Oesterreicher (1985), (1994) – das Medium der sprachlichen Äußerung als zentrales Klassifikationskriterium. Wie ein Blick in die Regionalsprachenforschung zeigt, überwiegen sowohl unter den medial mündlichen als auch unter den medial schriftlichen Daten elizitierte Sprachdaten, die mit unterschiedlichsten variationslinguistischen Methoden der Datenerhebung gewonnen werden.

Vor dem Hintergrund der Beliebtheit und Ergiebigkeit elizitierter Sprachdaten für regionalsprachliche Fragestellungen konzentrierte sich die anschließende Präsentation regionalsprachlicher Korpora des Deutschen auf exemplarische Sprachkorpora, die auf elizitierten Daten aufbauen. Angelehnt an den allgemeinen Korpusbegriff von McEnery/Wilson (1996) wurden einem regionalsprachlichen Korpus folgende Merkmale zugeschrieben:

- Es handelt sich um eine abgeschlossene Menge an medial mündlichen oder medial schriftlichen Sprachdaten.
- Das Korpus ist für die modernen Regionalsprachen des Deutschen bzw. für bestimmte regionalsprachliche Spektrumsausschnitte (Varietäten) repräsentativ.
- Das Korpus ist durch Meta-Beschreibungen erschlossen.
- Das Korpus ist (im weiteren Sinne) computertechnisch aufbereitet.

Als ein Beispiel regionalsprachlicher Korpora, die auf elizitierten medial schriftlichen Sprachdaten basieren, diene das Wenker-Korpus, das im „Digitalen Wenker-Atlas (DiWA)“ online verfügbar ist. Als Untersuchungsgegenstand, an dem die Spezifik und Ergiebigkeit dieses Korpus exemplifiziert wurde, fungierte die Plosiv-Spirans-Variation im Bereich des Schriftzeichens <g>. Als Beispiele medial mündlicher elizitierter Sprachkorpora wurden vier Tonkorpora der „Datenbank Gesprochenes Deutsch (DGD)“ vorgestellt.

Unter diesen wurden das Zwirner- und das Pfeffer-Korpus im Rahmen einer Analyse zum Rezipientenpassiv auf ihre Analysemöglichkeiten und ihre Ergiebigkeit untersucht.

In den Beispielanalysen wurde offensichtlich, dass sowohl das Wenker-Korpus als auch das Zwirner- und Pfeffer-Korpus hinsichtlich des analysepraktischen als auch bezüglich des zu leistenden Interpretationsaufwands sicherlich keine „einfachen“ Korpora darstellen. Zum Beispiel liegen den drei Korpora keine linguistischen Annotationen zugrunde, die Voraussetzung für unkomplizierte(re) Abfragemöglichkeiten darstellen würden. Im Falle des Wenker-Korpus kommt hinzu, dass aufgrund der Kurrentschrift eine automatische Erfassung der Wenkerbogen als Textdatei (bislang) noch nicht möglich ist. Trotz der skizzierten Probleme, die die Korpora aus korpuslinguistischer Sicht aufweisen, konnten die vorgestellten Analysen die hohe Ergiebigkeit und Validität der ausgewählten Korpora nachweisen. Vor dem Hintergrund der skizzierten Ergebnisse bleibt auf weitere Erschließung der Materialien in naher Zukunft zu hoffen.

Literatur

- Ágel, Vilmos/Hennig, Mathilde (2006): Theorie des Nähe- und Distanzsprechens. In: Ágel, Vilmos/Hennig, Mathilde (Hg.) (2006): *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650–2000*. Tübingen: Niemeyer. S. 3–31.
- Ammon, Ulrich [u. a.] (Hg.) (2005): *Sociolinguistics/Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*. 2., vollständig neu bearbeitete und erweiterte Aufl. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 3.2).
- Androutsopoulos, Jannis/Schmidt, Gurly (2002): SMS-Kommunikation: Ethnografische Gattungsanalyse am Beispiel einer Kleingruppe. In: *Zeitschrift für Angewandte Linguistik* 36, S. 49–80.
- Androutsopoulos, Jannis/Ziegler, Evelyn (2003): Sprachvariation und Internet: Regionalismen in einer Chat-Gemeinschaft. In: Androutsopoulos, Jannis/Ziegler, Evelyn (2003): „Standardfragen“. *Soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation*. Frankfurt/M. [u. a.]: Peter Lang (VarioLingua 18). S. 251–279.
- Aschwanden, Brigitte (2001): „Wär wot chätä?“ Zum Sprachverhalten deutschschweizerischer Chatter. *Networx* 24 [<http://www.mediensprache.net/de/networx/networx-24.pdf>].
- Atteslander, Peter (2005): Schriftliche Befragung/Written Investigations. In: Ammon, Ulrich [u. a.] (Hg.), S. 1063–1076.
- Auer, Peter (1995): Ethnographic Methods in the Analysis of Oral Communication. Some Suggestions for Linguists. In: Quasthoff, Uta M. (ed.): *Aspects of Oral Communication*. Berlin/New York: Walter de Gruyter. S. 419–440.
- Auer, Peter/Hinskens, Frans (1996): The Convergence and Divergence of Dialects in Europe. New and not so New Developments in an Old Area. In: Auer, Peter [u. a.] (Hg.): *Konvergenz und Divergenz von Dialekten in Europa*. Tübingen: Niemeyer (Sociolinguistica 10). S. 1–30.
- Auwärter, Manfred (2005): Experiment/Experiments. In: Ammon, Ulrich [u. a.] (Hg.), S. 1076–1084.

- Beißwenger, Michael (Hg.) (2001): Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld. Stuttgart: Ibidem.
- Bellmann, Günter (1964): Wege und Möglichkeiten der Schallaufnahme ostdeutscher Mundarten heute. Zur Tonbandaufnahme der Vertriebenenmundarten. In: Zeitschrift für Mundartforschung 31, S. 62–79.
- Bellmann, Günter (1970): Einleitung. In: Bellmann, Günter/Göschel, Joachim: Tonbandaufnahme ostdeutscher Mundarten 1962–1965. Gesamtkatalog. Marburg: Elwert (Deutsche Dialektgeographie 73). S. 7–30.
- Bellmann, Günter (1983): Probleme des Substandards im Deutschen. In: Mattheier, Klaus J. (Hg.): Aspekte der Dialekttheorie. Tübingen: Niemeyer (Reihe Germanistische Linguistik 46). S. 105–130.
- Bellmann, Günter (1994): Einführung in den Mittelrheinischen Sprachatlas (MRhSA) von Günter Bellmann. Tübingen: Niemeyer.
- Bellmann, Günter/Göschel, Joachim (1970): Tonbandaufnahme ostdeutscher Mundarten 1962–1965. Gesamtkatalog. Marburg: Elwert (Deutsche Dialektgeographie 73).
- Bellmann, Günter/Herrgen, Joachim/Schmidt, Jürgen Erich (1994–2002): Mittelrheinischer Sprachatlas (MRhSA). Bd. 1–5. Tübingen: Niemeyer.
- Besch, Werner [u. a.] (Hg.) (1983): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Erster Halbbd. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 1.1).
- Briggs, Charles L. (2005): Sociolinguistic Interviews/Soziolinguistisches Interview. In: Amon, Ulrich [u. a.] (Hg.), S. 1052–1062.
- Bucheli, Claudia/Glaser, Elvira (2002): The Syntactic Atlas of Swiss German Dialects: empirical and Methodological Problems. In: Barbiere, Sjef [et al.] (eds.): Syntactic Microvariation. [<http://www.meertens.knaw.nl/books/synmic/>]
- Burger, Harald (³2005): Mediensprache. Eine Einführung in Sprache und Kommunikationsformen der Massenmedien. Berlin: de Gruyter.
- Christen, Helen (2002): „Swiss Date“ – Aspekte dialektalen Sprechens im Schweizer Privatfernsehen. In: Wiesinger, Peter (Hg.): Akten des X. Internationalen Germanistenkongresses Wien 2000 „Zeitenwende – Die Germanistik auf dem Weg vom 20. ins 21. Jahrhundert“. Bd. 3: Aufgaben einer zukünftigen Sprachgeschichtsforschung – Gesprochene Sprache in regionaler und sozialer Differenzierung – Sprache in der Öffentlichkeit. Bern [u. a.]: Peter Lang. S. 217–225.
- Christen, Helen/Tophinke, Doris/Ziegler, Evelyn (2005): Chat und regionale Identität. In: Krämer-Neubert, Sabine/Wolf, Norbert Richard (Hg.): Bayerische Dialektologie. Akten der Internationalen Dialektologischen Konferenz, 26.–28. Februar 2002. Heidelberg: Winter (Schriften zum Bayerischen Sprachatlas 8). S. 425–439.
- Cornelissen, Georg/Schmitt, Eva-Maria (Hg.) (1997): Dialekte und Dialektliteratur in der Euregio Rhein-Maas-Nord. Unter Mitarbeit von Pierre Bakkes. Mit CD. Akustische Dokumentation Hans-Peter Beyenburg. Köln: Rheinland-Verlag.
- Cornips, Leonie/Poletto, Cecilia (2005): On standardising syntactic elicitation techniques. Part I. In: *Lingua* 115 (7), S. 939–957.
- DiWA = Schmidt/Herrgen 2001ff.
- Duden (⁴2000): Aussprachewörterbuch. Wörterbuch der deutschen Standardaussprache. 4., neu bearbeitete und aktualisierte Auflage. Bearbeitet von Max Mangold in Zusammenarbeit mit der Dudenredaktion. Duden Band 6. Mannheim [u. a.]: Dudenverlag.
- Duden (⁷2005): Die Grammatik. Unentbehrlich für richtiges Deutsch. 7., völlig neu erarbeitete und erweiterte Auflage. Herausgegeben von der Dudenredaktion. Duden Band 4. Mannheim [u. a.]: Dudenverlag.

- Dürscheid, Christa (1999): Zwischen Mündlichkeit und Schriftlichkeit: die Kommunikation im Internet. In: *Papiere zur Linguistik* 60/1, S. 17–30.
- Dürscheid, Christa (2002): E-Mail und SMS – ein Vergleich. In: Ziegler, Arne/Dürscheid, Christa (Hg.): *Kommunikationsform E-Mail*. Tübingen: Stauffenberg. S. 93–114.
- Dürscheid, Christa (2005): Medien, Kommunikationsformen, kommunikative Gattungen. In: *Linguistik online* 22/1.
- Eggers, Eckhard/Schmidt, Jürgen Erich/Stellmacher, Dieter (Hg.) (2005): *Moderne Dialekte – Neue Dialektologie*. Akten des 1. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD) am Forschungsinstitut für deutsche Sprache „Deutscher Sprachatlas“ der Philipps-Universität Marburg vom 5.–8. März 2003. Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik*. Beihefte 130).
- Elspaß, Stephan (2000): Rheinische Sprachgeschichte von 1700–1900. In: Macha, Jürgen [u. a.] (Hg.): *Rheinisch-westfälische Sprachgeschichte*. Köln [u. a.]: Böhlau (*Niederdeutsche Studien* 46). S. 247–276.
- Elspaß, Stephan (2005a): Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert. Tübingen: Niemeyer (Reihe Germanistische Linguistik 263).
- Elspaß, Stephan (2005b): Zum Wandel im Gebrauch regionalsprachlicher Lexik. Ergebnisse einer Neuerhebung. In: *Zeitschrift für Dialektologie und Linguistik* 72, S. 1–51.
- Fiehler, Reinhard/Wagener, Peter (2005): Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*. Ausgabe 6, 136–147. [<http://www.gespraechsforschung-ozs.de>]
- Glas, Reinhold/Ehlich, Konrad (2000): *Deutsche Transkripte 1950 bis 1995*. Ein Repertorium. Hamburg (Arbeiten zur Mehrsprachigkeit. Folge A. 63).
- Glaser, Elvira (2000): Erhebungsmethoden dialektaler Syntax. In: Stellmacher, Dieter (Hg.): *Dialektologie zwischen Tradition und Neuansätzen*. Beiträge der internationalen Dialektologentagung, Göttingen, 19.–21. Oktober 1998. Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik*. Beihefte 109). S. 258–276.
- Glaser, Elvira (2005): *Krieg und kriegern*: zur Arealität der BEKOMMEN-Periphrasen. In: Kleinberger Günther, Ulla [u. a.] (Hg.): „Krieg und Frieden“ – Auseinandersetzung und Versöhnung in Diskursen. Tübingen: Francke. S. 43–64.
- Günther, Ulla/Wyss, Eva Lia (1996): E-Mail-Briefe – Eine neue Textsorte zwischen Mündlichkeit und Schriftlichkeit. In: Hess-Lüttich, Ernest W. B. [u. a.] (Hg.): *Textstrukturen im Medienwandel*. Frankfurt/M. [u. a.]: Peter Lang. S. 61–86.
- Häcki Buhofer, Annelies (2002): Der Zusammenhang von Dialektbewusstsein und Charakteristik der gesprochenen Sprache. In: Wiesinger, Peter (Hg.): Akten des X. Internationalen Germanistenkongresses Wien 2000 „Zeitenwende – Die Germanistik auf dem Weg vom 20. ins 21. Jahrhundert“. Bd. 3: Aufgaben einer zukünftigen Sprachgeschichtsforschung – Gesprochene Sprache in regionaler und sozialer Differenzierung – Sprache in der Öffentlichkeit. Bern [u. a.]: Peter Lang. S. 211–216.
- Haas, Walter (1983): Dialekt als Sprache literarischer Werke. In: Besch, Werner [u. a.] (Hg.), S. 1637–1651.
- Hasselberg, Joachim/Wegera, Klaus-Peter (1976): *Hessisch*. Düsseldorf: Schwann (Dialekt/Hochdeutsch – kontrastiv. Heft 1).
- Herrgen, Joachim (1986): Koronalisierung und Hyperkorrektur. Das palatale Allophon des /CH/-Phonems und seine Variation im Westmitteldeutschen. Stuttgart: Steiner (*Mainzer Studien zur Sprach- und Volksforschung* 9).

- Herrgen, Joachim (2006): Sprachgeographie und Optimalitätstheorie. Am Beispiel der t-Tilgung in Auslaut-Clustern des Deutschen. In: Zeitschrift für Dialektologie und Linguistik 72/2005, S. 277–317.
- Herrgen, Joachim/Lenz, Alexandra N. (2003): Digitale Dialektologie. Online-Publikation des Wenker-Atlas im Internet. In: Marburger UniJournal 14, S. 43–48.
- Herrgen, Joachim/Lenz, Alexandra N./Pennay, Mark (2003): The Research Institute for German Language – *Deutscher Sprachatlas* (Marburg). In: Dialektologia et Geolinguistica 11, S. 94–109.
- Herrgen, Joachim/Schmidt, Jürgen Erich (in Vorb.): Sprachdynamik. Eine Einführung in die moderne Regionalsprachenforschung. Berlin: E. Schmidt (Grundlagen der Germanistik).
- Hommer, Emil (1910): Studien zur Dialektgeographie des Westerwaldes. Diss. Univ. Marburg.
- Huesmann, Anette (1998): Zwischen Dialekt und Standard. Empirische Untersuchung zur Soziolinguistik des Varietätenspektrums im Deutschen. Tübingen: Niemeyer (Reihe Germanistische Linguistik 199).
- Kehrein, Roland (2006): Regional accent in the German language area. How dialectally do German police answer emergency calls? In: Hinskens, Frans (Hg.): Language variation. European perspectives. Amsterdam/Philadelphia: Benjamins. S. 83–96.
- Kehrein, Roland/Lameli, Alfred/Nickel, Jost (2005): Möglichkeiten der computergestützten Regionalsprachenforschung am Beispiel des Digitalen Wenker-Atlas (DiWA). In: Braungart, Georg [u. a.] (Hg.): Jahrbuch für Computerphilologie – online 7.
- Klein, Eva/Mattheier, Klaus J./Mickartz, Heinz (1978): Rheinisch. Düsseldorf: Schwann (Dialekt/Hochsprache – kontrastiv. Heft 6).
- Koch, Peter/Oesterreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36, S. 15–43.
- Koch, Peter/Oesterreicher, Wulf (1994): Schriftlichkeit und Sprache. In: Günther, Hartmut/Ludwig, Otto (Hg.): Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung. 1. Halbbd. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 10.1). S. 587–604.
- König, Werner (¹⁵2005): dtv-Atlas Deutsche Sprache. Mit 155 Abbildungsseiten in Farbe. Grafiker Hans-Joachim Paul. München: dtv.
- Labov, William (1972): Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press.
- Leirbukt, Oddleif (1997): Untersuchungen zum *bekommen*-Passiv im heutigen Deutsch. Tübingen: Niemeyer (Reihe Germanistische Linguistik 177).
- Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen: Narr (narr studienbücher).
- Lenz, Alexandra N. (2003): Struktur und Dynamik des Substandards. Eine Studie zum Westmitteldeutschen (Wittlich/Eifel). Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik. Beihefte 125).
- Lenz, Alexandra N. (2004): Zur Interpretation des Intendierten Ortsdialekts. In: Lenz, Alexandra N./Radtke, Edgar/Zwinkl, Simone (Hg.): Variation im Raum. Variation and Space. Frankfurt/Main: Peter Lang (VarioLingua 20). S. 113–131.
- Lenz, Alexandra N. (2005): Hyperdialektalismen und Hyperkorrekturen – Indizien für Varietätengrenzen. In: Lenz, Alexandra N./Mattheier, Klaus J. (Hg.): Varietäten – Theorie und Empirie. Frankfurt/M. [u. a.]: Peter Lang (VarioLingua 23). S. 75–95.
- Lenz, Alexandra N. (2007a): Zur Grammatikalisierung von *geben* im Deutschen und Lëtzebuergesch. In: Zeitschrift für Germanistische Linguistik.

- Lenz, Alexandra N. (2007b): Ergebnisse eines Sprachproduktionsexperiments zum Rezipientenpassiv. In: Patocka, Franz/Seiler, Guido (Hg.): *Dialektale Morphologie, dialektale Syntax*. Wien: Edition Praesens. [im Druck].
- Ludwig, Johannes (1906): *Lautlehre der moselfränkischen Mundart von Sehlern* (Reg.-Bez. Trier). Diss. Freiburg i. Br.
- Lüdtke, Hartmut (2005): Beobachtung/Observation. In: Ammon, Ulrich [u. a.] (Hg.), S. 1033–1051.
- Macha, Jürgen (1991): *Der flexible Sprecher. Untersuchungen zu Sprache und Sprachbewußtsein rheinischer Handwerksmeister*. Köln [u. a.]: Böhlau.
- Mattheier, Klaus J. (1979): *Sprachvariation und Sprachwandel. Untersuchungen zur Struktur und Entwicklung von Interferenzprozessen zwischen Dialekt und Hochsprache in einer ländlichen Sprachgemeinschaft des Rheinlandes*. Habilitationsschrift. Universität Bonn.
- McEnery, Tony/Wilson, Andrew (1996): *Corpus Linguistics*. Edinburgh: University Press.
- Mihm, Arend (2004): Zur Geschichte der Auslautverhärtung und ihrer Erforschung. In: *Sprachwissenschaft* 29, S. 133–206.
- Niebaum, Hermann (1977): *Westfälisch*. Düsseldorf: Schwann (Dialekt/Hochdeutsch – kontrastiv. Heft 5).
- MRhSA = Bellmann/Herrgen/Schmidt (1994–2002).
- Pfeffer, Jay Alan/Lohnes, Walter F. W. (Hg.) (1984): *Grunddeutsch. Texte zur gesprochenen deutschen Gegenwartssprache. Überregionale Umgangssprache aus der Bundesrepublik Deutschland, der Deutschen Demokratischen Republik, Österreich und der Schweiz. Textkorpora 1. Einführungs- und Registerband*. Tübingen: Niemeyer (Phonai 28).
- Rabanus, Stefan (2005): Dialektwandel im 20. Jahrhundert: Verbalplural in Südwestdeutschland. In: Eggers/Schmidt/Stellmacher (Hg.), S. 267–290.
- Richter, Helmut (2005): Transkriptionssysteme/Transcription Systems. In: Ammon, Ulrich [u. a.] (Hg.), S. 1126–1134.
- Schädlich, Hans-Joachim/Große, Rudolf (1961): Tonbandaufnahme der deutschen Mundarten in der Deutschen Demokratischen Republik. In: *Forschungen und Fortschritte* 35, S. 358–363.
- Schirmunski, Viktor M. (1962): *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Berlin (Veröffentlichungen des Instituts für deutsche Sprache und Literatur 25).
- Schlobinski, Peter (1996): *Empirische Sprachwissenschaft*. Opladen: Westdeutscher Verlag (WV studium 174).
- Schmidt, Jürgen Erich (1998): Moderne Dialektologie und regionale Sprachgeschichte. In: Besch, Werner/Solms, Hans Joachim (Hg.): *Regionale Sprachgeschichte*. Berlin [u. a.] (Zeitschrift für deutsche Philologie. Sonderheft. 117). S. 163–179.
- Schmidt, Jürgen Erich (2005): Sprachdynamik. In: Eggers/Schmidt/Stellmacher (Hg.), S. 15–44.
- Schmidt, Jürgen Erich/Herrgen, Joachim (2001 ff.): *Digitaler Wenker-Atlas (DiWA)*. Bearb. von Alfred Lameli, Alexandra N. Lenz, Jost Nickel und Roland Kehrlein, Karl-Heinz Müller, Stefan Rabanus. Erste vollständige Ausgabe von Georg Wenkers „Sprachatlas des Deutschen Reichs“. 1888–1923 handgezeichnet von Emil Maurmann, Georg Wenker und Ferdinand Wrede. Marburg: Forschungsinstitut für deutsche Sprache – „Deutscher Sprachatlas“. [<http://www.diwa.info>]
- Schwitalla, Johannes/Wegstein, Werner (Hg.) (2005): *Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003*. Tübingen: Niemeyer.
- Siebenhaar, Beat (2003): Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. In: *Linguistik online* 15, S. 125–139.

- Siebenhaar, Beat (2005a): Die dialektale Verankerung regionaler Chats in der deutschsprachigen Schweiz. In: Eggers/Schmidt/Stellmacher (Hg.), S. 691–717.
- Siebenhaar, Beat (2005b): Varietätenwahl und Code-Switching in Deutschschweizer Chatkanälen. Quantitative und qualitative Analysen. *Networx* 43 [http://www.medien-sprache.net/de/networx/networx_43.pdf].
- Siever, Torsten [u. a.] (Hg.) (2005): Websprache.net. Sprache und Kommunikation im Internet. Berlin [u. a.]: de Gruyter (Linguistik. Impulse & Tendenzen 10).
- Skiba, Romuald (2005): Computeranalyse/Computer Analysis. In: Ammon, Ulrich [u. a.] (Hg.), S. 1187–1197.
- Spiekermann, Helmut (2005): Regionale Standardisierung, nationale Destandardisierung. In: Eichinger, Ludwig M./Kallmeyer, Werner (Hg.): Standardvariation – Wie viel Variation trägt die deutsche Sprache? Berlin/New York: Walter de Gruyter (IDS-Jahrbuch 2004). S. 100–125.
- Spolsky, Bernard (2005): Tests/Test. In: Ammon, Ulrich [u. a.] (Hg.), S. 1085–1095.
- Stellmacher, Dieter (1981): Niedersächsisch. Düsseldorf: Schwann (Dialekt/Hochsprache – kontrastiv. Heft 8).
- Stickel, Gerhard (1999): Zur Sprachbefindlichkeit der Deutschen: Erste Ergebnisse einer Repräsentativumfrage. In: Stickel, Gerhard (Hg.): Sprache – Sprachwissenschaft – Öffentlichkeit. Berlin/New York: de Gruyter (Jahrbuch 1998 des Instituts für deutsche Sprache). S. 16–44.
- Storrer, Angelika (2000): Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet. In: Voß, G. Günter [u. a.] (Hg.): Neue Medien im Alltag: Begriffsbestimmungen eines interdisziplinären Forschungsfeldes. Opfaden: Leske + Budrich. S. 151–176.
- Storrer, Angelika (2005): Online-Corpora zur linguistischen Analyse der deutschen Gegenwartssprache. In: Zeitschrift für Germanistische Linguistik 33, S. 145–150.
- Thun, Harald (2005): Variation im Gespräch zwischen Informant und Explorator. In: Lenz, Alexandra N./Mattheier, Klaus J. (Hg.): Varietäten – Theorie und Empirie. Frankfurt/M. [u. a.]: Peter Lang (VarioLingua 23). S. 97–126.
- Tophinke, Doris/Ziegler, Evelyn (2006): „Aber bitte im Kontext“: Neue Perspektiven in der dialektologischen Einstellungsforschung. In: Gessinger, Joachim/Voeste, Anja (Hg.): Dialekt im Wandel. Perspektiven einer neuen Dialektologie. Duisburg: Gilles und Franke. (OBST 71). S. 203–222.
- Vogel, Petra M. (2003): Passiv in deutschsprachigen Chats. Eine Korpusanalyse. In: Linguistik online 15, 3/03, 141–160.
- Wagener, Peter (2002): Gesprochenes Deutsch online. Zur Modernisierung des Deutschen Spracharchivs. In: Zeitschrift für Dialektologie und Linguistik 69, S. 314–335.
- Wagener, Peter (2005): Korpuslinguistik im Internet. Neue Wege dialektologischer Forschung am Beispiel des virtuellen Deutschen Spracharchivs. In: Eggers/Schmidt/Stellmacher (Hg.), S. 363–373.
- Wagener, Peter/Bausch, Karl-Heinz (1997): Tonaufnahmen des gesprochenen Deutsch. Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven. Tübingen: Niemeyer.
- Wegera, Klaus-Peter (2000): Gen, oder wie Herr Gottsched will, chen. Zur Geschichte eines Diminutivsuffixes. In: Habermann, Mechthild [u. a.] (Hg.): Wortschatz und Orthographie in Geschichte und Gegenwart. Festschrift für Horst Haider Munske zum 65. Geburtstag. Tübingen: Niemeyer. S. 43–58.
- Werlen, Iwar (2000): Variation im gesprochenen Hochdeutschen in der deutschen Schweiz – am Beispiel der Nachrichten von Radio DRS 1 und Radio DRS 3. In: Häcki Buho-

- fer, Annelies (Hg.): Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte. Festschrift für Heinrich Löffler zum 60. Geburtstag. Tübingen, Basel: Francke (Basler Studien zur deutschen Sprache und Literatur 80). S. 311–327.
- Wittlicher Wörterbuch (²1984): Zusammengestellt von Herrn Georg Fischer. Herausgegeben von Buch- und Offsetdruck, Nels. Wittlich.
- Wodak, Ruth (1982): Erhebung von Sprachdaten in natürlicher und simuliert-natürlicher Sprechsituation. In: Besch, Werner [u. a.] (Hg.), S. 539–544.
- Wrede, Ferdinand (1895): Über richtige Interpretation der Sprachatlaskarten. Vortrag gehalten in der germanistischen Section der 43. Versammlung deutscher Philologen und Schulmänner zu Köln am 26. September 1895. In: Wenker, Georg/Wrede, Ferdinand: Der Sprachatlas des Deutschen Reichs. Dichtung und Wahrheit. Marburg: Elwert. S. 31–52.
- Ziegler, Evelyn (2005): Die Bedeutung von Interaktionsstatus und Interaktionsmodus für die Dialekt-Standard-Variation in der Chatkommunikation. In: Eggers/Schmidt/Stellmacher (Hg.), S. 719–745.
- Ziegler, Evelyn (2006): Identitätskonstruktion und Beziehungsarbeit in bayerischen Schülerzetteln. In: Dürscheid, Christa/ Spitzmüller, Jürgen (Hg.): Perspektiven der Jugendsprachforschung/Trends and Developments in Youth Language Research. Frankfurt/M. [u. a.]: Peter Lang. S. 161–179.
- Zifonun, Gisela [u. a.] (1997): Grammatik der deutschen Sprache. 3 Bde. Berlin/New York: de Gruyter (Schriften des Instituts für Deutsche Sprache 7.1–7.3).
- Zwirner, Eberhard (1956): Lautdenkmal der deutschen Sprache. In: Zeitschrift für Phonetik und Allgemeine Sprachwissenschaft 9, S. 3–13.
- Zwirner, Eberhard (1958): Vorwort. In: Zwirner/Bethge, S. 5–9.
- Zwirner, Eberhard/Bethge, Wolfgang (1958): Erläuterungen zu den Texten. Mit 2 Tafeln. Göttingen: Vandenhoeck & Ruprecht (Lautbibliothek der deutschen Mundarten 1).

WERNER KALLMEYER

Möglichkeiten der maschinellen Unterstützung bei der Arbeit mit Interaktionskorpora

Abstract

In diesem Beitrag sollen anhand von Materialien aus Gesprächskorpora des IDS Schwierigkeiten und Möglichkeiten der maschinellen Recherche vorgeführt werden. Grundlage dafür sind Gesprächstranskripte, die in digitaler Form vorliegen und in einem System mit Rechercheprozeduren zugreifbar sind. Mit diesem Ziel wird auf Rechercheverfahren zurückgegriffen, die in den 1990er Jahren in einem Projekt SERGES im IDS als Anwendung der COSMAS-Technologie auf Gesprächskorpora entwickelt wurden. Die hier gegebenen Recherchemöglichkeiten werden an einem Auswahlkorpus von Gesprächstranskripten mit einem Gesamtumfang von 87.629 laufenden Wörtern versuchsweise angewendet und in ihren Beschränkungen und ihrer Fruchtbarkeit für explorative Untersuchungen betrachtet. Damit soll ein Beitrag zur Klärung der Frage geleistet werden, welche Recherchemöglichkeiten aus einer gesprächsanalytischen Perspektive vorstellbar und erwünscht sind und insofern bei der weiteren korpustechnologischen Entwicklung berücksichtigt werden sollten.

1. Zu den Voraussetzungen maschineller Recherchen in Interaktionskorpora

Interaktionskorpora enthalten: Ton- bzw. Video-Dokumente von sprachlichen Interaktionen in ihrem natürlichen Vorkommenszusammenhang, Dokumentationen zur Situation und zur Einbettung der jeweiligen Interaktionszene in übergreifende Zusammenhänge, Transkriptionen der sprachlichen und manchmal auch der nicht-sprachlichen Kommunikationsvorgänge. Solche Korpora entstanden im Zusammenhang mit der Entwicklung der Konversationsanalyse seit den 60er Jahren, in Deutschland seit dem Beginn der 1970er Jahre. Es gab und gibt vielerorts kleinere Korpora von Interaktionsmaterialien als Ergebnis von Forschungsprojekten. Unter dem Einfluss der digitalen Aufnahme-, Präsentations- und Archivierungstechnik wird auch der weitere Aufbau von Interaktionskorpora inzwischen von der Erwartung geleitet, dass die Aufnahmen ebenso wie die Transkriptionen und sonstigen schriftlichen Materialien in digitaler Form vorliegen, über eine Verwaltungskomponente auffindbar und zugreifbar sind und dass nach Möglichkeit Aufnahmen und Transkriptionen über Zeitindices miteinander verknüpft sind (Alignment), so dass zu ausgewählten Transkriptionsstücken die Ton/Bild-Abschnitte präsentiert werden können.

Größere digitale Interaktionskorpora mit einer solchen Datenbankstruktur existieren gegenwärtig aber erst im Ansatz. In dieser Hinsicht liegen Interaktionskorpora weit hinter den Korpora geschriebener Sprache zurück. Im British National Corpus wurde als ein Kriterium für ein nationales Referenzkorpus festgelegt, dass 10 % mündliche Sprache enthalten sein müsste. Diese Anforderung wird von der Vorstellung geleitet, dass alle wesentlichen Merkmale einer Sprache von schriftlichen Texten abzulesen seien mit Ausnahme der phonetisch-phonologischen Aspekte. In diesem Sinne entsprechen die 10 % gesprochener Sprache einer Minimallösung. Aber auch diese 10 % werden nicht erreicht. Korpora gesprochener Sprache wachsen relativ langsam. Das gilt insbesondere für Interaktionskorpora.

Die beim Aufbau von Korpora gesprochener Sprache auftretenden Probleme von der Daten-Beschaffung über die Transkriptionsweise, die Dokumentation von nicht-sprachlichen Eigenschaften von Interaktionsereignissen bis zur Frage der mehrstufigen Annotation sind in den letzten Jahren im Zusammenhang mit bedeutenden Initiativen des Korpusaufbaus vielfältig reflektiert worden.¹ Diese Diskussion soll hier nicht noch einmal nachgezeichnet werden. Vielmehr konzentriert sich dieser Beitrag auf die antizipierende Entwicklung von Recherchemöglichkeiten in Interaktionskorpora.

Der große Herstellungsaufwand von Interaktionskorpora lohnt sich im Kern nur für Interaktionsanalysen, also Untersuchungen, die sprachliche Phänomene relativ zu Sprechern, Sprecherwechseln und sequenziellen Positionen innerhalb von Redebeiträgen und in Folgen von Redebeiträgen erfassen. Trotzdem ist es natürlich sinnvoll, die Interaktionskorpora so einzurichten, dass sie auch für andere Untersuchungsinteressen, die nicht unmittelbar auf die Interaktionsstruktur rekurrieren, nutzbar sind. Langfristig ist das auch für die Interessen der Interaktionsanalyse von großem Interesse. Am besten realisierbar sind zurzeit „einfache“ Dinge wie Wortsuche. Möglich ist damit die Untersuchung von „mündlicher Sprache“ anhand von Wortformen (und ggf. Wörtern) und Kollokationen. Je realistischer die schriftliche Wiedergabe der Lautform ist, umso schwieriger wird allerdings die Lemmatisierung. Diese ist bei einer von der Orthographie abweichenden Schreibweise, z. B. der literarischen Umschrift mit der Wiedergabe von Schnellsprech-Formen, Verschleifungen, dialektalen Prägungen usw. nicht unmittelbar möglich, sondern nur über eine Normalisierung der Verschriftung.

Lemmatisierung vorausgesetzt kann man anhand von Interaktionskorpora (zusammen mit Korpora monologischer Daten gesprochener Sprache) den Wortschatz des gesprochenen Deutsch dokumentieren und sein Vorkommen in Textsorten / Interaktionstypen und bezogen auf Sprecherkategorien (also z. B. Männer / Frauen, ggf. Altersklassen, soziale Schicht usw.) untersuchen.

¹ Vgl. u. a. Leech/Myers/Thomas 1955; Pusch/Raible (2002); Schwitalla/Wegstein (2005).

Diese Art der Auswertung hängt ab von den verfügbaren Metadaten (z. B. für eine Sortierung nach Interaktionstypen, Formalität und Informalität, Geschlecht und Alter der Sprecher und nach der Entstehungszeit). Das Problem ist hier eher die Verfügbarkeit der Daten in ausreichender Menge als die Recherchestrategien. Mögliche Fragen sind weiter die nach der Anzahl der laufenden Wörter in einer Interaktion, bei einzelnen Sprechern und ggf. auch noch innerhalb von Gesprächsphasen (bei einer entsprechenden Segmentierung des Gesprächsverlaufs).

Im Folgenden soll erläutert werden, was man mit einem Recherchesystem, das über Suchoperatoren für Wörter und Interaktionseigenschaften verfügt wie das experimentelle System COSMAS II für Gesprächsdaten, im Rahmen von qualitativ ausgerichteten Untersuchungen tun kann.² Es geht dabei darum, begrenzte Suchmöglichkeiten heuristisch zu nutzen, Grenzen zu reflektieren und Recherchestrategien weiter zu entwickeln und Entwicklungsbedarf zu präzisieren für die Recherche von komplexen Phänomenen, die gegenwärtig kaum erfassbar sind. Da das System bislang nur in einer Versuchsversion existiert und nur ein begrenztes Korpus angeschlossen ist, wurden für die Zwecke dieser Demonstration und Reflexion von Möglichkeiten und Grenzen die Suchvorgänge mit anderen, einfacheren Werkzeugen wie Wortsuche in WORD ansatzweise simuliert und sorgfältig von Hand kontrolliert. Es handelt sich also bei dem, was hier vorgestellt wird, nicht um eine Recherche in einem voll funktionstüchtigen System, sondern um die Simulation der Arbeit mit Recherche-Operatoren, für die anhand von einem Entwicklungskorpus bereits Erfahrungen gesammelt wurden, teilweise auch für die hier interessierenden Recherchen.³

2. Das Zielobjekt

Im Zentrum des Interesses steht ein grammatisch-pragmatisches Muster, in dem zwei Äußerungseinheiten miteinander verknüpft werden, zwischen denen eine Beziehung der Opposition bzw. partiellen Inkompatibilität besteht: JA-ABER. Aus früheren gesprächsanalytischen Studien zu derartigen Phänomenen hat sich als Hypothese ergeben, dass es sich um ein relativ komplexes, formal und funktional bestimmtes Muster handelt, das eine verfestigte gesprächs rhetorische Lösung für bestimmte Anforderungen des Formulierens im Gespräch darstellt. Die Grundaufgabe kann man angeben als: die Relevanzen anderer berücksichtigen und eigene Relevanzen verfolgen.⁴ Das Muster hat folgende Eigenschaften:

² Einige der im Experimentalsystem enthaltenen Suchmöglichkeiten sind auch der DGD implementiert, allerdings nicht alle; und zudem ist die Menge der Gesprächstranskripte in der DGD nach wie vor sehr begrenzt.

³ Das gilt z. B. für die Suche nach Verwendungen des Konnektors *aber*; vgl. Kallmeyer/Schütte 2005.

⁴ Die Beschäftigung mit solchen Konstruktionsmustern war Gegenstand eines Projekts

- Formuliert wird eine komplexe Äußerung mit zwei Teilen. Ein Teil allein ist hinsichtlich der Äußerungskonstitution unvollständig ohne den anderen Teil. Der Normalfall ist, dass beide Teile vom selben Sprecher produziert werden.
- Der erste Teil drückt eine Berücksichtigung und / oder Bestätigung einer Position aus und der zweite Teil eine Zurückweisung bzw. eine mit der ersten nicht kompatible andere Position.
- Der erste Teil bezieht sich auf eine fremde Position, insbesondere auf eine Position eines Interaktionspartners, aber auch auf Positionen Dritter oder auf allgemein mögliche und nicht personal zugeschriebene Positionen. Der zweite Teil bezieht sich auf eine Position des aktuellen Sprechers.
- Es besteht ein Relevanzgefälle zwischen dem ersten und dem zweiten Teil in dem Sinne, dass der erste Teil in der Relevanz rückgestuft wird und der zweite Teil hochgestuft.
- Für die Markierung des zweiten Teils werden adversative Konnektoren wie *aber*, *nur*, *bloß* verwendet, insgesamt eine relativ begrenzte Menge. Die Markierung des ersten Teils ist wesentlich variabler. Neben *zwar* gibt es vor allem *ja* und andere Formen der Zustimmung wie *sicher*, *natürlich*, *richtig*.
- Zu den Merkmalen der Verwendung eines zweiteiligen und in sich oppositiven Formativs gehört, dass der erste Teil einen zweiten, adversativen Teil projiziert. Das geschieht mit Verfahren der Zustimmungseinschränkung und Relevanzrückstufung. Die Projektion des oppositiven zweiten Teils ist in vielen Fällen sehr deutlich; sie gehört zum Musterwissen kompetenter Interaktionsteilnehmer und wird z. B. auch in kommunikativen Spielen verwendet.⁵

Da *ja* für den ersten und *aber* für den zweiten Teil mit Abstand die häufigsten Markierungen sind, wird als allgemeine Kennzeichnung der Funktionen der beiden Teile JA und ABER verwendet, unbeschadet der Tatsache, dass es Funktionsunterschiede in Abhängigkeit von den gewählten Konnektoren und sonstigen Formulierungsmustern gibt.

Der erste Teil des Komplexes kann unterschiedlich ausgebaut sein. Neben minimalen Fällen wie *ja aber eh * kaprizieren sie sich nicht auf die neue prüfungsordnung* gibt es auch Fälle, in denen die Ausdrücke für JA und ABER

„Initiative Reaktionen“ im SFB 245 „Sprechen und Situation“. Vgl. dazu die beiden SFB-Papiere Kallmeyer/Schmitt 1991a und b. Diese frühen Fassungen werden zur Zeit in grundlegend überarbeiteter Form für die Veröffentlichung vorbereitet.

⁵ In einem Comic aus der Serie „Hägar der Schreckliche“ tritt Hägar mit seinem Adlatus einer größeren Gruppe von sehr angriffslustig wirkenden Feinden entgegen und begrüßt sie: *Ah ... das Empfangskomitee! – Okay! Ihr seid uns vielleicht zahlenmäßig überlegen ... vielleicht seid ihr auch größer! – Vielleicht seid ihr stärker und zäher als wir! – Vielleicht könnt ihr uns sogar besiegen!* – An dieser Stelle schaltet sich der Adlatus ein: *Kommst du bald zu einem „Aber“?*, worauf Hägar irritiert und ängstlich zur Seite blickt und entgegnet: *Ich dachte, das wäre ich schon!*

einen großen Wortabstand haben. Im folgenden Beispiel beträgt er 30, aber der im Korpus festgestellte Höchstwert ist 59.⁶

- 86 RS: weiß jetzt nich ob es ne vorlesung is jetzt oder
 87 BR: |s=is en proseminar- das is en proseminar| was
 88 RS: |n seminar↓ oder n seminar↓ |
 89 BR: rodenberger macht↓
 90 RS: also sie würden eher an=ne ne
 91 BR: <äh na ja RÄUSPERT SICH |im
 92 RS: vorlesung- <ach so |ja
 93 BR: prinzip is es so daß die | daß |die äh |
 94 RS: dann hab doch hab ich auch noch>| |LACHT ETWAS|
 95 BR: daß man |einführung>| * äh * → und proseminar nicht im
 96 RS: |> (... ...) <|
 97 BR: gleichen semester machen soll↓ sondern erst die
 98 BR: einführung dann das proseminar↓ ←
 99 RS: ah ja↓ * ja dann lass
 100 BR: |aber | äh was rodenberger über kafka
 101 RS: ich das fa|llen↓|
 102 BR: macht- * dürfte auch einführenden charakter haben **

Das Beispiel zeigt u. a., dass JA-ABER-Konstruktionen in eine Sequenz eingebettet sind, zu der eine Anlass- bzw. Bezugsäußerung gehört (vgl. *also sie würden eher an=ne vorlesung-*, 90/92) und oft auch Folgeschritte der beteiligten Sprecher.

Die erwähnten Annahmen über die Mustereigenschaften sollen anhand einer größeren Materialmenge weiter belegt, ggf. weiter ausdifferenziert und auf ihre Nutzbarkeit für eine automatische Recherche hin untersucht werden. Um das Vorgehen plausibel und fruchtbar zu machen, soll die Materialmenge so groß sein, dass sie nicht ohne weiteres mehr im Kopf präsent gehalten werden kann. Zugleich muss das Material zumindest in größeren Teilen auch von Hand analysiert werden. Dazu empfiehlt sich ein Auswahlkorpus von wiederum begrenztem Umfang. Das hier benutzte Korpus enthält insgesamt 87.629 laufende Wörter. Die Erfahrung zeigt, dass dieser Umfang beide Bedingungen erfüllt und eine praktikable Größe als exploratives Versuchskorpus darstellt. (Vgl. die Korpusübersicht im Anhang).

⁶ Das Gesprächskorpus des IDS benutzt ein Partitursystem mit fortlaufenden Sprecherzeilen, ohne interpretative Segmentierungen (z. B. in Äußerungen). Die Transkriptionen haben einen mittleren Feinheitsgrad. Bei Überlappungen von Redebeiträgen und prosodischen Markierungen von Lautstärke und Tempo werden jeweils Anfang und Ende notiert; das hat u. a. Vorteile für die Recherchierbarkeit. (Vgl. die Transkriptionserläuterung im Anhang).

Das Korpus enthält 10 Beratungsgespräche unterschiedlicher Art (Studienberatungen, ärztliche Beratung, Finanzberatung, Sozialberatung), eine Referatsbesprechung zwischen Professor und Student, ein universitäres Prüfungsgespräch, 3 Fernsehdiskussionen und eine Arbeitsbesprechung. Die Zusammensetzung des Korpus mit einem relativ großen Kernbereich (Beratung) und einigen näher oder ferner liegenden Kontrastfällen entspricht einem explorativen Vorgehen zur Hypothesenentwicklung in qualitativen Untersuchungen. So wurde das vorliegende Korpus zusammengestellt, um Hypothesen über die Rolle von adversativen Konstruktionen in helfenden Interaktionen als Folge der Beratungsaufgabe zu entwickeln. Zu Beratern gehört die Nutzung von unterschiedlichen Perspektiven als Lösungsresource und damit die Aushandlung der Perspektivenunterschiede bzw. auch die Durchsetzung einer Perspektive. Insofern liegen immer auch Verfahren der Durchsetzung „zum guten Zweck“ nahe. Eine gesprächsanalytisch und auch für die Beratungspraxis interessante Frage ist, wie mit dem Durchsetzungspotential so umgegangen wird, dass es für Helfen funktional ist und nicht „Recht behalten“ dominant wird. (Vgl. u. a. Kallmeyer 2000).

Die folgende Betrachtung konzentriert sich auf den Zusammenhang von JA + ABER mit den oben skizzierten Formativ-Eigenschaften. Auf einige Erscheinungsformen, die Ähnlichkeiten haben, aber ggf. dem Zielprofil nicht ganz entsprechen, soll hier noch hingewiesen werden. Sie werden bei den ersten Recherche-Schritten mit erfasst, aber anschließend ausgeklammert. Die strukturellen Verwandtschaften und Unterschiede dieser Erscheinungsformen im Verhältnis zu den JA-ABER-Fällen sind gesondert zu untersuchen. Das sind insbesondere folgende Vorkommen von ABER-Konnektoren:

(a) ABER als Diskursgliederung

- 155 BR: also das-| ** aber äh * möglichst >wenig↓<
 156 RS: | gut↓ **
 157 BR: |>gut<|
 158 RS: **aber** ich glaub das wär=s dann |nich| ** ham sie

Der Ausdruck *gut* fungiert hier als Erledigungs- und Abschlussmarkierung, und die Äußerung mit *aber* formuliert die Erledigungsfeststellung noch einmal explizit.

(b) Initiales ABER ohne erkennbare Einräumung

Unter den 564 ABER-Vorkommen im Korpus finden sich 29 Kandidaten für initiales ABER. Diese Fälle sind interessant, weil sie in der Regel in Kontexten von andauernden Kontroversen erscheinen und insofern ein Dissensmerkmal darstellen. In dieser Hinsicht ist das entscheidende Kriterium für die Klassifikation als initiales ABER, dass die vorausgehende Äußerung desselben Sprechers nicht als JA-Teil einzustufen ist. Weiter zurück wird man in den meisten Fällen bei der Etablierung des Dissenses auch JA-ABER-Kon-

struktionen finden. Das ist auch beim folgenden Ausschnitt aus einer psychologischen Beratung der Fall:

- 180 BL: paniksätzen fütterst ** die hätte ich ganz gerne
 181 BL: >diese paniksätze-< *2,5*
 182 MA: ja eigentlich sind sie
 183 BL: |wo du| ni"cht schnaufen
 184 MA: aggressiv diese paniksätze |also |
 185 BL: kannst *2*
 186 MA: ja also sie lösen ja dann eine art wu"t
 187 BL: **aber** bei der wut ka"nnste doch
 188 MA: dann in mir aus
 189 BL: wieder schnaufen ** ich möchte gerne die"
 190 MA: ja↑ **
 191 BL: sätze hörn wo du ni"cht schnaufen kannst↓

Es gibt auch spezifische Verwendungen von ABER ohne Vorgänger-JA als Instrument der Gesprächssteuerung, z. B. bei Moderatoren von Diskussionen wie Sabine Christiansen. In der Fernsehdiskussion mit dem Kandidaten Stoiber formuliert sie die meisten ihrer Fragen ebenso wie thematisch steuernde Kommentare mit initialem ABER (insgesamt 9 mal).

(c) Es gibt eine gewisse Anzahl von Fällen, in denen ein Konnektor vom Typ ABER erscheint, aber nicht erkennbar ist, dass der vorausgehende Äußerungsteil eine Projektion eines oppositiven zweiten Teils enthält. Vielmehr wird in diesen Fällen in einer Äußerungserweiterung retrospektiv eine Oppositionsrelation eingeführt. Im folgenden Beispiel aus dem universitären Prüfungsgespräch hat der Student (BB) in einem langen Redebeitrag eine Textstelle interpretiert. Am Ende dieser Darstellung schließt er eine Äußerungserweiterung mit *allerdings* an, die in der vorausgehenden Äußerung durch nichts erwartbar gemacht wurde:

- 141 BB: ** und u:nd ** und hier wird eben der
 142 BB: zweikampf alexander menes geschildert- *
 143 AA: ja
 144 BB: bis achtundfünfzig- be/ bis
 145 BB: achtundfünfzig * den der alexander
 146 BB: **allerdings** verliert↓ ** den schwertkampf
 147 AA: aha↑ ** ←ist das→ typisch für
 148 BB: verliert er↓
 149 AA: das alexanderbild↑

Interessanterweise wird dieser „unmotivierte“, d. h. nicht projizierte oppositive Zusatz Gegenstand einer interaktiven Bearbeitung. Dabei spielen zwei Aspekte eine Rolle. Zum einen fällt auf, dass die Einschränkung bzw. Relativierung in einem abhängigen Satz (Relativsatz) erscheint. Ohne sonstige Hochstufungen, z. B. prosodischer Art oder in Form einer Fokussierungsformel, ist das für JA-ABER wichtige Merkmal der Relevanzhochstufung im zweiten Teil nicht gegeben. Immerhin bearbeitet BB diesen Aspekt nachträglich dadurch, dass er den Einschränkungssachverhalt noch einmal als Hauptsatz und damit als eigenständigen Sachverhalt formuliert. Dieses Verfahren ist vermutlich so zu interpretieren, dass der Sprecher ein thematisches Potential einführt und mit der Reformulierung in der Relevanz hochstuft. Wie die Reaktion von AA zeigt, sieht dieser Anlass, genau diesen Punkt mit seiner Nachfrage zu bearbeiten. Einerseits erfolgt durch *aha*↑ eine markierte Aufmerksamkeitszuwendung, also eine Relevanzhochstufung, und andererseits wird ein möglicherweise in Opposition stehender Sachverhaltsaspekt explizit eingeführt.

(d) Die Kombination von Zurückweisung + ABER, realisiert als *nein aber* mit der Funktion einer Voraussetzungskorrektur. Sie kommt im Korpus mehrfach vor (vgl. die Korpusübersicht im Anhang). Im folgenden Beispiel hat eine Studentin die Lösungsvorstellung, dass sie nach einer noch in Vorbereitung befindlichen neuen Prüfungsordnung Examen machen könnte, weil diese Prüfungsordnung für sie möglicherweise günstiger ist. BR erläutert die Unklarheit der Situation (*das is dann noch nich geläufig jetzt weil wir die neue prüfungsordnung noch nich haben*) und stellt die Nachteile der Lösung „neue Prüfungsordnung“ dar. Die Reaktion von RS enthält zwei Aussagen: *das macht ja nichts* und *das wollt ich jetzt grad wissen was da*. BR wiederum reagiert mit einer Zurückweisung bzw. Korrektur von Voraussetzungen, die RS in ihrer zweiten Aussage benutzt (die Möglichkeit, über die neue Prüfungsordnung Auskunft zu geben), eingeleitet mit *nein* und einer Richtigstellung der Voraussetzung:

- 39 BR: kö"nnen sie zwar theoretisch optieren für die neue
 40 BR: prüfungsordnung↑ aber die wird so viel andere
 41 RS: >mhm<
 42 BR: anforderungen enthalten↓ |die
 43 RS: ja das macht ja nichts |das
 44 BR: sie-| |ja |
 45 RS: | LACHT wollt ich jetzt grad wissen |was| da was
 46 BR: **nein** →das kann ich ihnen natürlich auch nicht
 47 RS: da-
 48 BR: sagen weil ich die neue prüfungsordnung nicht
 49 BR: kenne↓← *|aber | es sind zum beispiel dann also
 50 RS: |ah so↑|

- 51 BR: vier hauptseminarscheine erforderlich-
 52 RS: in der neuen
 53 BR: |insgesamt * statt bisher zwei"
 54 RS: abteilung | (...) |

Mit der Richtungstellung der Voraussetzung gibt BR seine Handlungsbedingungen an, im konkreten Fall seine mangelnden Voraussetzungen für eine vollgültige Antwort. Mit dem ABER-Teil unternimmt er aber dennoch einen Antwortversuch. Dabei reagiert er auf die Annahme von RS, dass die neuen Anforderungen mit ihren Lösungsvorstellungen kompatibel sein könnten. Die Behandlung eigener Voraussetzungen in Einräumungsteilen kommt häufiger vor, und solche Fälle werden auch berücksichtigt. Nur die Fälle mit Voraussetzungskorrekturen oder -zurückweisungen mit der Negationspartikel *nein* sollen hier vorerst ausgeklammert werden.

3. Wortsuche nach Konnektoren

3.1 Der ABER-Teil

In einem ersten Schritt wurden die Konnektoren für ABER im Korpus gesucht. Ihre Anzahl ist begrenzt.

Im vorliegenden Fall wurden in die Recherche sowohl adversative als auch konzessive Konnektoren aufgenommen, die sich dadurch auszeichnen, dass sie eine Relation der Inkompatibilität zwischen den zwei verknüpften Aussagen herstellen.⁷ Dabei kommen Konjunktionen wie *aber* und *obwohl* und auch Adverbien wie *nur*, *bloß*, *allerdings* in Betracht. Außerdem werden auch Konnektoren wie *auf der einen / anderen seite* bzw. *zum einen / anderen* berücksichtigt, welche die Unterschiedlichkeit von zwei verknüpften Aussagen markieren, weil sie tendenziell mit einer Relevanzhochstufung des zweiten Teils verbunden sind und diese häufig auch noch durch *aber* oder einen anderen Konnektor ausgedrückt wird; oft wird auch nur ein Teil des paarigen Konnektors verwendet. Für alle Ausdrücke wurden Aussprachevarianten berücksichtigt (wie *abba aba awwa awa* für *aber* oder *andre* für *andere* usw.) sowie alle Formen der in den Transkriptionen benutzen prosodischen Markierungen für Akzentuierung, Dehnung, größere Lautstärke und langsames Tempo.⁸ Die Ausdrücke *eine / andere seite* in den möglichen Varianten *die andere seite* bzw. *auf der anderen seite* sind mit *ander** bzw. *andre* seite* mit Wortabstand 1 problemlos recherchierbar.⁹

Auch *doch*, *jedoch* wurden am Material geprüft. Allerdings gab es keine Funde, in denen in einer zweiteiligen Äußerungskonstruktion die Opposition

⁷ Vgl. Zifonun/Hoffmann/Strecker et al. (1997), Kapitel H2.

⁸ Also: *a"ber* und *a:ber*, *<aber [xxx]>*, *←aber [xxx]→*; [xxx] bezieht sich hier darauf, dass die zweite korrespondierende Markierung sowohl direkt nach dem Konnektor als auch später erscheinen kann.

⁹ Vgl. die COSMAS-Operatoren *WORT(x)*, *BETONUNG*, *DEHNUNG*, *GLEICH(X, Y)*.

nur durch diese Ausdrücke markiert wurde. Mit *aber* kommt vor allem *doch* vor (*aber es ist doch so ...*), für *jedoch* gibt es auch in dieser Kombination keinen Treffer.

Der paarige adversative Konnektor *nicht sondern* wurde recherchiert, aber nicht in die weitere Untersuchung einbezogen, weil hier der erste Teil eine Zurückweisung und ggf. Richtigstellung beinhaltet und damit ein anderes grammatisch-pragmatisches Formativ konstituiert wird.

Konzessive Konnektoren wurden als ABER-Äquivalente berücksichtigt, wenn sie im zweiten Teil des Äußerungspaares erscheinen. Der Konnektor *obwohl* erscheint auch im ersten Teil des Formativs zur Markierung der Einräumung bzw. Berücksichtigung (vgl. unten zu JA). Besonders deutlich wird die Äquivalenz zu *aber*, wenn diese Subjunkturen als syntaktisch herausgelöst vorangestellt werden und der folgende Satz Hauptsatzstellung hat.¹⁰

Die Subjunkturen im zweiten Teil markieren ggf. ein anderes Relevanzverhältnis als ABER, indem hier der zweite Teil in der Relevanz rückgestuft wird. Das kann allerdings auch bei der Verwendung von *aber* der Fall sein. Solche Fälle wurden jedoch berücksichtigt, sofern die anderen Mustereigenschaften wie JA-Markierungen im ersten Teil gegeben sind.

Die Gesamtmenge der adversativen Konnektoren und als äquivalent betrachteter Verwendungen anderer Konnektoren beträgt im Korpus 546. Die Recherche ergibt große Unterschiede im Anteil der Vorkommen pro Textmenge zwischen den vier Konnektoren: Auf *aber* entfallen 92,7% aller 564 Verwendungsfälle der vier Konnektoren, auf *nur* 6,8%, auf *bloß* 0,5 und auf *obwohl* ebenfalls 0,5% (vgl. dazu die Tabelle im Anhang).

3.2 Der JA-Teil

Die Ausdrücke für die Markierung des ersten Formativ-Teils wurden von Hand in einem zweiten Schritt gesammelt, ausgehend von den Fundstellen für ABER. Die Ausdrücke sind viel variabler als die ABER-Markierungen und beziehen sich jeweils auf unterschiedliche Funktionsaspekte. Eine – noch unvollständige Liste – umfasst folgende Ausdrücke:

- Konnektoren: *zwar*, *obwohl*
- Ausdrücke der Zustimmung: *ja* mit Varianten (u. a. *ja:*, *jo*, *jau* usw.), *okay*, *naja*; die Adverbien *sicher*, *gewiss*, *natürlich*, *klar*, *richtig*, *eben*, *schon* usw., andere satzförmige Ausdrücke der Zustimmung wie *dem stimme ich zu*.
- Die Ausdrücke der Zustimmung können zugleich Relevanzrückstufungen transportieren. Das ist z. B. bei den Ausdrücken der Selbstverständlichkeit bzw. Fraglosigkeit wie *gewiss*, *natürlich*, *klar* der Fall.
- Hinsichtlich der Projektion von ABER sind weitere Ausdrücke für die Begrenztheit der Zustimmung relevant, z. B. bei *teilweise*. (*ja das stimmt zumindest teilweise aber*).

¹⁰ Vgl. Fiehler/Barden/Elstermann/Kraft (2004).

- Weiter werden Zustimmungseinschränkungen formuliert durch Modalisierungen, z. B. Ausdrücke der Möglichkeit wie *das kann so sein* bzw. *möglich ist das*, der Unsicherheit wie *vielleicht*, und der Perspektivierung, z. B. des perspektivierten Rechtgebens *so gesehen hast du recht* oder der Gültigkeit in einer bestimmten Hinsicht, z. B. *im allgemeinen gilt das, im prinzip ist es so*. Diese Ausdrücke können für die Konstruktion einer Opposition benutzt werden, z. B. *im prinzip* vs. „im konkreten Fall“ bzw. „im Einzelfall“, wobei der Gegenbegriff häufig nicht explizit erscheint, aber das Konzept inkorporiert wird. Das in Kap. 2 angeführte Beispiel aus „Theaterwissenschaftler“ zeigt eine solche Formulierungsweite: *naja im prinzip is es so dass die dass die äh dass man einführung *äh * und proseminar nicht im gleichen semester mach soll↓ sondern erst die einführung und dann das proseminar ** aber äh was rodenberger über kafka macht- * dürfte auch einführenden charakter haben* usw., d. h. im konkreten Fall kann anders entschieden werden, als es der allgemeinen Regel entsprechen würde.

Man kann die erwähnten Ausdrücke der Modalisierung und Perspektivierung als Oppositionsanker bezeichnen. Sie bilden keine geschlossene Klasse und erscheinen in vielen Varianten. Allerdings gibt es geläufige Verwendungen, die sich, ausgehend von bereits identifizierten Einräumungsteilen, auch maschinell suchen lassen.

Eine weitere Variante von JA ist *nein*, d. h. die Zustimmung zu einer negativen Aussage: *nein natürlich nicht aber darum geht es ja auch nicht*. Solche Vorkommen wurden für die Anwendung der Recherche-Verfahren vorläufig ausgeklammert.

4. Weitere Ansätze für die Erfassung des JA-ABER-Zusammenhanges

4.1 Der zweite Teil des Formativs

ABER-Konnektoren in initialer Position, d. h. am Beginn des zweiten Formativ-Teils, erscheinen relativ häufig mit voraufgehender oder folgender Pause und auffälliger Akzentuierung oder Dehnung, Tempoverringerung, Erhöhung der Lautstärke und/oder Grenztonmarkierung durch steigende oder fallende Intonation. Derartige Markierungen zeigen sowohl die Segmentierung als auch die Relevanzhochstufung an. Ob und wie häufig dies der Fall ist, lässt sich mit den COSMAS-Operationen für prosodische Markierungen von Wörtern und Wortfolgen sowie Pausen vor und nach Wörtern feststellen. Man bekommt so Treffer wie z. B. den folgenden, in dem *nur* durch Akzentuierung und langsames Tempo zusätzlich markiert ist:

2291 VF: |vorher gesagt haben↓| ich hab

2292 AS: |au"ch gut dass das | au"ch möglich ist↓

2293 VF: ja nur gesagt |i"ch möchte| ich

2294 AS: |←nu"r→ | ich

- 2295 VF: |auf die fra"ge↑| |LACHANSATZ|
 2295 AS: |mu"ss sie drauf| au"fmerksam |machen | ** es

Syntaktisch integriertes *aber*, *nur*, *bloß* oder *allerdings* erscheint innerhalb des ersten Syntagmas.¹¹ Um auch in diesen Fällen die unmittelbare Nähe des Konnektors zur Umschlagstelle zum ABER-Teil feststellen zu können, muss man Hilfsstrategien mit einbeziehen. Die Beobachtungen am Korpus haben ergeben, dass die erste Einheit von ABER mit integriertem Konnektor fast durchweg das Format einer einfachen und kurzen Konstruktion hat:

- 299 BR: a"ndre frage wo sie ihren lati"numschein machen
 300 BR: * das is völlich ihnen überlassen↓
 301 RS: **mhm ***
 302 BR: |also ich mein ob sie dann das| mü"ssen sie nicht in
 303 RS: |**ja** das geh"t **aber** schwe:r- |
 304 BR: dreibug machen
 305 RS: **ja** das is **aber** dann nun ganz
 306 RS: schwierig wenn ich hier studiere und dann * ich kann

Hier verwendet RS ein Konstruktionsmuster, das für einen unmittelbar bezogenen Kommentar charakteristisch ist mit einer Proform vom Typ *das* bzw. *da* und Kompositionen damit wie *dabei*, *darum* usw. Dieses Formulierungsmuster ist kurz und in besonderer Weise dafür geeignet, den syntaktisch integrierten Konnektor nah an der Umschlagstelle zu platzieren.

Vergleichbare Beobachtungen kann man auch an anderen Beispielen machen, in denen die Markierung des JA-Teils weit entfernt und manchmal auch undeutlich ist:

- 209 BR: **ja * ja *** es is also keine *
 210 RS: alles eigentlich↓
 211 BR: schwierigkeit s=is **nur** so wenn sie jetzt schon im
 212 BR: sechsten semester sind also * ich meine es is jetzt

Hier wird zunächst mit *ja * ja* und einer negativen Zustimmung/Einräumung Teil 1 gebildet. Der ABER-Teil beginnt dann mit einer Fokussierungsformel mit eingebettetem *nur*. Vergleichbar wären *es ist aber so* oder *es ist bloß so*. Auch dieses Formulierungsformat stellt die Nähe des Konnektors zur Umschlagstelle her. Im Unterschied zur Kommentar-Formel, die sich durch einen deutlichen und deiktisch-unmittelbaren Rückbezug auszeichnet, dominiert bei der Fokussierungsformel die kataphorische Funktion (in der Regel mit Hauptakzent auf *so* und Intonationsbruch). Die Suchvorgabe für solche Fälle

¹¹ Vgl. die Positionsklassen der Konnektoren in Pasch/Brauß/Breindl (2004), S. 502 ff.

Noch schwächer hinsichtlich der Bestätigungsfunktion als *ja* sind Rückmeldungen wie *hm* oder *mhm*. Eine praktikable Entscheidung scheint hier zu sein, dass einfache Rückmelder ohne besondere Markierung nur als Kandidaten einbezogen werden, wenn sie in kurzem Abstand dem ABER-Teil voraufgehen. Dieser Abstand könnte z.B. auf vier Wörter Diskurs (d. h. von einem anderen, gerade aktiven Sprecher) festgelegt werden. Ausscheiden könnte man als leise gesprochen transkribierte Rückmeldungen (z. B. >*mh*<).

Bei anderen Markierungen des ersten Teils durch markante und ggf. markierte Formen (also nicht durch nicht hervorgehobenes *ja*) ist ein Wortabstand von 10–15 in der Regel für die Recherche kein Problem, d. h. die zurückgelieferten Fälle sind mit großer Wahrscheinlichkeit echte Treffer. In dieser Hinsicht ist das folgende Beispiel ganz eindeutig:

- 521 BR: semester bequem machen ↓
 522 RS: **eben** ↓ und deshalb da fehlt"
 523 BR: |ja |
 524 RS: mir ja nich mehr so |viel| * **obwohl** in traun
 525 BR: |RÄUSPERT SICH |
 526 RS: |fehlt mir dann | wieder mehr" die haben *

Der Konnektor *obwohl* erscheint hier in syntaktisch herausgelöster Frontstellung. Das ist für das Recherchesystem allerdings nicht erkennbar. Aber die Zusammengehörigkeit von *eben* ↓ am Äußerungsbeginn und *obwohl* im Wortabstand 11 ist so wahrscheinlich, dass eine solche Stelle problemlos als Treffer behandelt werden kann.

Als Leithypothese kann man formulieren: Je größer der Abstand zwischen JA und ABER, desto größer ist der Bedarf an deutlichen Markierungen. Das bedeutet für die Recherchemöglichkeiten: Markierte Formen erlauben einen größeren Wortabstand. Im folgenden Beispiel aus „Luxemburgerin“ wird z. B. der Zusammenhang von JA und ABER durch Mehrfachmarkierung von JA gestützt:

- 480 BR: |würden sie | ein:
 481 RS: gehört hab das is ganz |deprimierend-|
 482 BR: deutsches staatsexamen dann auch anerkannt bekommen
 483 BR: in luxemburg ↓ * |für also- gymnasialbereich ↓|
 484 RS: |ja das" so" wieso- das | is
 485 RS: also das **auf jeden fall: ja** das wird" anerkannt **bloß**
 486 RS: die lu"xemburger examen werden ja in deutsch"land
 487 BR: ja: ja |ja| ja
 488 RS: nich anerkannt ↓ |es| wird schon anerkannt

Hier ist der JA-Teil vielfältig durch Zustimmungsausdrücke besetzt: *ja, so"wieso, auf jeden fall, ja*. Der Wortabstand ist relativ gering, aber wenn man Fälle mit Mehrfachmarkierung des ersten Teils sucht, bekommt man auch bei großem Wortabstand echte Treffer.

Ein typischer Treffer für Mehrfachmarkierung im JA-Teil ist auch das folgende Beispiel („Luxemburgerin“):

306 RS: schwierig wenn ich hier studiere und dann * ich kann

307 RS: ja nicht da und hier sein LACHT ich kann ja nich

308 RS: dauernd * mich hier * hin und her fahren

309 BR: |natürlich also äh | für des: für=s

310 RS: |mit=m auto des geht ja kaum ↓|

311 BR: ler"nen für die * äh für=s latinum * **das is klar-**

312 BR: **nur** ich meine wo sie die abschlussklausur

313 BR: schreiben ↓

Die beiden Ausdrücke *natürlich* und *das is klar* haben einen Wortabstand 12 und werden mit erfasst, wenn man ausgehend von *nur* mit einem Wortabstand von 15 (einschließlich der Pausenfüller *äh*) Manifestationen von JA sucht. Der Ausdruck *das is klar* wird durch die Wortsuche nach *klar* erfasst, kann darüber hinaus aber ggf. auch als Bestandteil eines Kommentarmusters identifiziert werden. Der Fall zeigt zudem eine interessante Eigenschaft von Einräumungsteilen: Das erste Adverb der Zustimmung (*natürlich*) erscheint am Äußerungsbeginn und *das ist klar* unmittelbar vor dem ABER-Konnektor. Dies entspricht einer Rahmung des JA-Teils. Solche Rahmungen können maschinell gefunden werden, wenn der Beginn der Einräumung identifizierbar ist wie im vorliegenden Fall (Äußerungsbeginn). Beginnt die Einräumung irgendwo im Verlauf einer längeren Äußerung, können Segmentierungsmuster hilfreich sein, aber die müssen erst noch systematisch erarbeitet werden.

Im folgenden, bereits eingangs zitierten Beispiel mit dem Wortabstand 30 zwischen *ja* und *aber* gibt es ebenfalls zusätzliche stützende Elemente im JA-Teil:

86 RS: weiß jetzt nich ob es ne vorlesung is jetzt oder

87 BR: |s=is en proseminar- das is en proseminar| was

88 RS: |n seminar ↓ oder n seminar ↓ |

89 BR: rodenberger macht ↓

90 RS: also sie würden eher an=ne ne

91 BR: <äh na ja RÄUSPERT SICH |im

92 RS: vorlesung- <ach so |ja

- 93 BR: **prinzip is es so dass** die | dass |die äh |
 94 RS: dann hab doch hab ich auch noch>| |LACHT ETWAS|
 95 BR: dass man |einführung>| * äh * → und proseminar nicht im
 96 RS: |>(… …)< |
 97 BR: gleichen semester machen soll ↓ sondern erst die
 98 BR: einführung dann das proseminar ↓←
 99 RS: ah ja ↓ * ja dann lass
 100 BR: |aber | äh was rodenberger über kafka
 101 RS: ich das fallen ↓
 102 BR: macht- * dürfte auch einführenden charakter haben **

Eine Stütze für die Erkennbarkeit des Zusammenhanges von BRs Äußerung ab 91 und seiner vorausgehenden Äußerung ist, dass BR den Beginn seiner Äußerung durch die expandierte *ja*-Form *na ja* markiert, der man aufgrund des *na* zudem eine Relativierungsprojektion zuschreiben kann; weiter durch den möglichen Oppositionsanker *im prinzip* sowie eine Ankündigung bzw. Fokussierungsformel *is es so dass*. Aus einer gesprächsanalytischen Perspektive würde man zudem die besondere Gestaltung des Beginns als ein Hervorhebungsverfahren mit projektiver Funktion ansehen: Anhebung der Lautstärke, Verzögerung *äh* + *na ja* + weitere Verzögerung durch Räuspern. Ob ein solches Markierungsmuster für die Recherche weitere Fälle erschließt, die sonst zweifelhaft geblieben wären, kann allerdings nur anhand eines größeren Korpus geprüft werden.

Weiter gibt es eine Reihe von Fällen, in denen die Einräumung unter Verwendung von Verben des Wahrnehmens, Denkens und Sagens formuliert wird. In diesen Zusammenhängen wird die Übereinstimmung häufig durch *auch* und Äquivalente markiert: *das denke ich auch*; *das glaub ich au"ch* ↓ * *ähm aber* (Sklavenreferat 406/407) und ähnliche Fälle.

Spätestens an dieser Stelle wird jedoch deutlich, dass die Recherche von relativ variablen Formulierungsmustern an ihre Grenzen stößt. Die Ausdrücke für Einräumung sind zahlreich und strukturell unterschiedlich. Neben den bisher genannten kommen noch viele in Betracht, die Wahrnehmung, Denken, Fühlen und Handeln des Sprechers betreffen, z. B. *das sehe ich auch so* usw. Unter den gegebenen Recherchebedingungen verspricht eher Erfolg, solche Einräumungsmuster ausgehend von bereits über Konnektoren und andere geläufige Ausdrücke identifizierte JA-ABER-Fällen zu eruieren. Für die erwähnten Fokussierungsformeln ist die Lage insofern besser, als einige von ihnen gut auffindbar sind.

Es gibt weitere Fokussierungsformeln, die für die Relevanzhochstufung an der Schnittstelle zwischen JA und ABER verwendet werden. Kandidaten sind *es geht um*, Formulierungen mit *folgendes* und auch Formulierungen mit *dazu* / *hier* / *an dieser stelle* + flektierte Form eines Modalverbs (*müssen* / *wollen* /

können / sollen, vor allem *sollte, sollten*) + Subjekt *ich / wir / man*, ggf. auch andere personale Referenz in der dritten Person + Verbum dicendi im Infinitiv (*sagen / erwähnen* usw.). So setzt im Beispiel auf S. 212 AS den ABER-Teil fort mit *ich mu"ss sie drauf au"fmerksam machen*. Neben den positiv formulierten Konstruktionen gibt es auch negative Äquivalente wie *wir sollten / man darf nicht verschweigen* usw.

Die Menge der Formulierungsmuster für Einräumung ist relativ groß, und dementsprechend selten kommen die einzelnen Muster vor. Um sie zu eruieren, braucht man ein wesentlich größeres Korpus als das hier benutzte. Wenn man mit großen Listen von möglichen Mustern arbeitet, wird das Recherche-Verfahren zudem schwerfällig. Der Aufwand lohnt sich allerdings bei sehr großen Korpora.

Ähnlich wie die Fokussierungsformeln können auch bestimmte Segmentierungsausdrücke für die Recherche hilfreich sein. In vielen Fällen steht die Einräumung nicht am Äußerungsbeginn. Wenn sie auch keine der gut recherchierbaren Ausdrücke für JA enthält, bleibt nur die Recherche von Segmentierungsformeln. Im folgenden Beispiel („Luxemburgerin“) beginnt der Sprecher zunächst eine ABER-Formulierung, bricht sie ab, platziert eine Einräumung und setzt dann mit ABER fort. Der Sprecher benutzt *ich meine* als typische Segmentierungsformel:

211 BR: schwierigkeit s=**is nur so** wenn sie jetzt schon im

212 BR: sechsten semester sind also * **ich meine** es ist jetzt

214 BR: nich im nächsten semester notwendig **aber** so langsam *

216 BR: müßten sie sich mal * |um | ihren prüfer *

217 RS: |latein|

Vgl.

218 BR: kümmern * neben dem latein↓

219 K LACHEND

Der Ausdruck *ich mein** (d.h. mit Deklinationsmorphemen oder ohne) ist gut recherchierbar. Er erscheint in den allermeisten Fällen als Eröffnungsmarkierung von Parenthesen und Expansionen (manchmal allerdings auch bei der Bearbeitung von Verständigungsproblemen, z. B. *nein ich meine*), und er erscheint relativ häufig am Beginn von Einräumungen (z. B. *ich mein ich weiß auch nicht wie=s im au"genblick aussieht nur also ich würd anrufen...*, „Wiedereingliederung“ 297; *nich das ich mein d=s natürlich glücklich dass es so wenige sind ds macht viel aus aber wir machen es uns ...*, „Sklavenreferat“ 1223-8). Im gesamten hier benutzten Korpus ist unter 25 Fällen von *ich mein** im Wortabstand bis 15 vor ABER nur eine Verständigungskorrektur (*nein ich meine ... aber*, „Wiedereingliederung“ 217), während in allen anderen Fällen *ich meine* den Einräumungsbeginn markiert. Insofern kann die Recherche von

Formulierungselemente bei einer formalen Analyse Zusammengehörigkeit signalisieren kann (*zum ei"nen is es* und *und da"nn is es*). Beim gegenwärtigen Kenntnisstand ist keine klare Entscheidung möglich. Es wird notwendig sein, anhand wesentlich größerer Korpora zu überprüfen, ob und ggf. mit welcher Wahrscheinlichkeit *dann* als zweiter Konnektor zu *zum einen* und ggf. anderen JA-Konnektoren vorkommt. Kombiniert mit *aber* ist das häufiger der Fall (*aber dann*), aber dabei gibt es auch für die Suche des ABER-Konnektors kein Problem.

Trotz aller Erfolge der bisher entworfenen kombinierten Recherche von Konnektoren und häufigen JA-Markierungen unter Zuhilfenahme von zusätzlichen Markierungen und Formulierungsmustern ist die Zahl der nicht einfach zu lösenden Fälle im Korpus immer noch beträchtlich. So bietet auch der folgende Fall einige typische Schwierigkeiten:

1575 SC: |können s|ie: * ja" oder nei"n sagen dass sie

1576 ES: da|rüber muss|

1577 SC: sagen * ich möchte es gerne aus=em wahlkampf

1578 SC: rau"shalten * ich möchte eine schne"lle einigung *

1579 SC: oder * wird=s innen wahlkampf gehen↓ **

[ES antwortet nicht direkt auf die Frage]

1585 SC: |also| ja" sie

1586 ES: wi"rklich nicht das wird immer wie|der |

1587 SC: möchten sich gerne * |zü"ig einigen |

1588 ES: |ja wenn wenn die| s p d *

1590 ES: auf unsere positionen * die wir schon früh" im

1591 ES: mai" des letzten jahres- ** c d u c s u↓ frau

1593 ES: merkel und ich haben das vorgestellt * die

1594 ES: eckpunkte von c d u c s/ wir waren ja früher dran

1596 ES: als herr schily- * wenn diese dinge * eingehalten

1597 ES: werden↑ * äh übernommen werden↑ * äh da"nn↑ äh: *

1598 ES: **seh ich überhaupt kein proble"m * nu"r-** * ich

1599 ES: möchte nicht irgendwelche au"sführungen von herrn

1600 ES: schily haben↑ * sondern ich möchte den

1601 ES: gese"tzestext haben↑ * ich möchte wi"ssen↑ was *

In diesem Fall ist der ABER-Konnektor markiert (Pause + Akzentuierung + Pause) und wird insofern bei der Wortsuche sofort erkannt. Im unmittelbaren

Kontext von *nu"r* erscheint die Formulierung *seh ich überhaupt kein proble"m*. Hierin steckt als Formulierungsmuster *kein problem*. Der Ausdruck *problem* alleine kommt vielfältig vor, teilweise auch als Fokussierungselement (*das proble"m ist*), aber auch in vielen anderen Kontexten. Der kombinierte Ausdruck *kein problem* ist seltener und wird spezifischer verwendet. Es gibt im Korpus eine Reihe von Fällen, in denen *kein problem* im Zusammenhang mit Oppositionsrelationen erscheint und dann immer im ersten Teil steht: *hab ich kei"n problem * ähm ** a:ber *3* >ä:hm< *4* für mi"ch* ... („Feldbusch-Schwarzer“ 212); *ich wollte ihnen kein problem ... einreden ich wi"ll nur sagen* ... („Feldbusch-Schwarzer“ 1369–74). Insofern kann *kein problem* unter die Formulierungsmuster für Einräumung aufgenommen werden. Im vorliegenden Fall wird die Stelle bei einer Suche nach ABER+ JA-Markierungen (anhand einer Liste relevanter Formen) gefunden.

Damit ist aber noch nicht der Beginn der Einräumung erkannt. Das Problem ist hier, dass das initiale *ja* unmarkiert und prosodisch integriert ist. Es kann als eine Markierung für Eröffnung und Bezugnahme angesehen werden, aber kaum als Zustimmungsausdruck. Hinzukommt der große Wortabstand von 59, wenn man die einzelnen Elemente der Abkürzungen für die politischen Parteien als Wörter zählt (so wie sie transkribiert sind). Interpretativ spielt bei der Analyse für das frühzeitige Erkennen einer Einräumung natürlich der vorausgehende Kontext eine große Rolle. SC hat Zustimmung manifest relevant gesetzt, indem sie die vorausgehenden Äußerungen von ES in Reaktion auf die frühere SC-Frage als *also ja" sie möchten sich gerne * zügig einigen* reformuliert. In diesem Kontext ist auch das unmarkierte *ja* von ES als Behandlung der Akzeptanz bedeutsam. Man erkennt sofort, dass hier mit *ja wenn* eine bedingte Zustimmung formuliert wird, die in den schon erkannten Einräumungsteil mündet: ** äh da"nn↑ äh: * seh ich überhaupt kein proble"m*. Das weiß aber das Recherchesystem nicht.

Eine auf diesen Fall zugeschnittene Suchstrategie könnte sein, ausgehend von der Einräumungsformel im unmittelbar vorausgehenden Kontext Konnektoren zu suchen, die das Potenzial haben, weitgespannte Formulierungszusammenhänge zu verklammern, sowie ggf. mögliche Partner für paarweise verwendete Konnektoren. Dabei stößt man dann auch auf *wenn-dann*. Wenn sich eine Anbindung über einen solchen Konnektor findet, könnte man die Kontextspanne bis zum Äußerungsbeginn öffnen. Um daraus ein praktikables Suchverfahren zu entwickeln, braucht man allerdings wiederum ein wesentlich größeres Korpus. Immerhin zeigen die bisherigen Erfahrungen mit der Recherche von Formulierungsmustern, dass solche Hilfsstrategien durchaus erfolgreich sein können.

4.3 Relative Häufigkeiten von Wörtern in begrenzten Kontexten

Um die bisherigen Möglichkeiten der maschinellen Recherche von komplexen Zusammenhängen für JA-ABER-Sequenzen einschließlich Anlassäußerung und

möglichen Anschlussäußerungen weiter abzusichern, sollte man auch andersartige Verfahren in Betracht ziehen. Die bisherigen Recherche-Verfahren gehen von festgelegten Wörtern bzw. Wortformen aus. Nun kann es auch hilfreich sein, nach Verteilungsmustern von vorher nicht bestimmten Wörtern zu suchen.

An einer Reihe von Gesprächsausschnitten mit JA-ABER-Konstruktionen fällt auf, dass in dem begrenzten Kontext dieser Konstruktion einzelne Ausdrücke mehrfach wiederholt werden, die im Gespräch insgesamt selten sind und ggf. nur in diesem lokalen Kontext vorkommen. Diese Wortwiederholungen sind Teil der engen inhaltlichen Bezugnahme zwischen Anlassäußerung und JA-ABER-Komplex. Besonders deutlich wird das in den Fällen, in denen alle Aktivitätsschritte in einer Interaktionssequenz jeweils das entsprechende Wort enthalten. Im folgenden Beispiel („Luxemburgerin“) gibt es zwei Kandidaten für eine solche Verdichtung in der Wortverwendung, *examen* und *anerkannt*, wobei *anerkannt* mit vier Vorkommen besonders hervortritt:

- 480 BR: |würden sie | ein:
 481 RS: gehört hab das is ganz |deprimierend-|
 482 BR: deutsches staatsexamen dann auch **anerkannt** bekommen
 483 BR: in luxemburg ↓ * |für also- gymnasialbereich ↓|
 484 RS: |ja das" so"wieso- das | is
 485 RS: also das auf jeden fall: ja das wird" **anerkannt** bloß
 486 RS: die lu"xemburger **examen** werden ja in deutsch"land
 487 BR: ja: ja |ja| ja
 488 RS: nich **anerkannt** ↓ |es| wird schon **anerkannt**
 489 RS: aber es is eben bei uns auch so wie hier in
 490 BR: |ja | ja:
 491 RS: deutschland un|heimli|cher lehrerüberfluß |jetzt

Der Ausdruck *anerkannt* kommt in diesem Gespräch relativ häufig vor (und erweist sich insofern als Schlüsselwort für ein zentrales Thema der Beratung), insgesamt 11mal (bei insgesamt 5128 Wörtern, was 0.194% entspricht), aber nicht in der lokalen Verdichtung wie in der obigen Stelle. Es gibt nur noch eine Stelle mit zwei Vorkommen in zwei aufeinander folgenden Äußerungen („Luxemburgerin“ 828+829).

Die ersten drei Vorkommen erscheinen in den drei Schritten der Sequenz Anlassäußerung (Frage von BR) – Einräumung – ABER. Das vierte Vorkommen liegt in der Einräumung eines nächsten JA-ABER-Komplexes. In diesem Fall fungiert die vorausgehende ABER-Äußerung als Anlass bzw. Bezugsäußerung für eine weitere JA-ABER-Konstruktion. Die Wiederholung des Ausdrucks *anerkannt* markiert die inhaltliche Bezugnahme, aber – anders als im vorausgehenden JA-ABER-Komplex – nicht den internen Zusammenhang zwischen JA und ABER.

Der Ausdruck *examen*, als Simplex oder Teil eines Kompositums, kommt im Gespräch 13mal vor und an anderen Stellen auch in höherer Verdichtung. Im Prinzip ist ein zweimaliges Vorkommen die minimale Bedingung für die mögliche Relevanz eines solchen Ausdrucks als Schlüsselwort, allerdings nur in Kontexten, in denen das Wort sonst nicht oder nie verdichtet vorkommt.

Für solche Recherchen kann man die fortlaufenden Indices aller Ausdrücke in COSMAS benutzen und Abstände (z. B. entsprechend einem Kontext von 80 Wörtern) in Indexzahlen und Mindesthäufigkeiten (z. B. gleich/größer 3) vorgeben, die an der Gesamtmenge des Vorkommens relativ zur Wortmenge des Gesprächs gemessen werden (also z. B. alle Ausdrücke, die nicht mehr als 0,2 % des Gesamtvorkommens ausmachen). Für die hier durchgespielte Analyse von JA-ABER-Äußerungen wird man zudem diese Recherche auf die Umgebungen von ABER-Ausdrücken beschränken. Entscheiden muss man auch, ob man nur Wörter oder auch Wortteile berücksichtigen will. Eine Recherche der geschilderten Art würde z. B. zum Gespräch „Christiansen-Stoiber“ zurückliefern, dass im Gespräch der Ausdruck *kompetenzteam* insgesamt dreimal vorkommt und dass alle Vorkommen in einem begrenzten Kontext liegen:

- 334 SC: wenn ich
 335 ES: diensttag=äh: * den oder die vorstellen↓ *
 336 SC: mir das **kompetenzteam** so a"nschae↑ dann denk ich
 337 SC: wo bleibt die kompetenz der frau"en↓ frau me"rkel
 338 SC: ist natürlich dabei- * aber das war=s dann auch
 339 SC: schon↓
 340 ES: ja das ist nicht ein **kompete"nzteam**↓ das
 341 ES: was wir jetzt=äh:: zunächst einmal festgelegt
 342 ES: haben↑ * äh das ist=äh * die führungsmannschaft
 343 ES: die den wahlkampf inhaltlich und organisatorisch *
 344 ES: äh * vorbereitet↑ * u:nd äh: **natürlich** ko"mmen
 345 ES: dann in einer späteren phase * kommt ein
 346 ES: **kompetenzteam**↑ * >äh:< mit einigen ausgewählten
 347 ES: persönlichkeiten die dann auch für einzelne *
 348 ES: sa"chfragen stehen↑ * **aber**- * das hat zeit
 349 ES: zunächst äh: geht es ja darum den wahlkampf zu
 350 ES: verzahnen↑ * bestimmte themen gemeinsam

Auch hier verteilen sich die Vorkommen des Schlüsselwortes auf drei Schritte der oppositiven Sequenz, die insgesamt aus einer Anlassäußerung von SC, einer Zurückweisung mit Richtigstellung und einer JA-ABER-Konstruktion von ES besteht. Der Ausdruck *kompetenzteam* erscheint in der Anlassäußerung (*wenn ich mir das kompetenzteam so anschau*), der Zurückweisung von ES (*ja das ist nicht ein kompetenzteam*↓ *das was wir jetzt=ä:h zunächst einmal festgelegt haben*) und seiner Einräumung (*u:nd äh: natürlich ko"mmen dann in einer späteren phase * kommt ein kompetenzteam*↑).

Wenn man bei der Recherche auch Wortbestandteile zulässt, bekommt man für „Christiansen-Stoiber“ u. a. folgendes Ergebnis: Für den Ausdruck *team* gibt es im Gespräch insgesamt 5 Vorkommen (was 0,041 % einer Wortmenge von 12.237 entspricht), die sich an zwei Stellen verdichten, die zudem dicht aufeinander folgen und zusammen eine Gesprächsstrecke von ca. 145 Wörtern ausmachen. ES verwendet in einer Äußerung zweimal den Ausdruck *team*, in den Verbindungen *teamkandidat* und *team stoiber*:

323 ES: wir haben uns da"zu entschieden dass wir * einen

324 ES: >äh< **teamkandidat** ein **team** stoiber machen mit

325 ES: einem: * exzellenten * mann an der spitze der **

Auf diese Äußerung reagiert SC mit der im vorigen Ausschnitt erscheinenden Frage von SC, auf die wiederum ES mit seiner oppositiven Äußerung reagiert. Alle fünf Vorkommen von *team* stehen also in einem interaktiven Zusammenhang.¹²

Solche kleinräumigen Verdichtungen zeigen eine enge inhaltliche Bezugnahme und, im Zusammenhang mit JA-ABER-Konstruktionen, Aushandlungsprozesse. Solche Aushandlungszusammenhänge müssen nicht so knapp zusammengedrängt sein wie in den vorliegenden Fällen. Aber zumindest für bestimmte Aushandlungsmuster ist charakteristisch, dass die Beteiligten unmittelbar und unter Verwendung von Kern- oder Schlüsselementen aufeinander Bezug nehmen. Dabei handelt es sich im übrigen um eine allgemeine, nicht auf JA-ABER beschränkte Erscheinung. Das soll noch mit einem Beispiel aus einem anderen Zusammenhang belegt werden. Es geht um einen Einwurf eines zweiten Sprechers in die laufende Äußerung eines ersten Sprechers:

324 AN: aber der kleine handwerker is da * von der mentalität eben

325 AN: noch anders und glaubt alles selber machen zu müssen-

326 DE: **künstler**↓

¹² Wenn man dann, wissend dass der Kampf um Wörter zum politischen Kerngeschäft gehört, den Wortverteilungen weiter nachgeht, stellt man fest, dass *mannschaft* als Synonym von *team*, von ES, und nur von ES, im Gespräch 5mal verwendet wird. Aber solche Erweiterungen der lexikalischen Analyse sind ein späterer Schritt.

326 AN: ja **künstler**↓ genau↓ * aber * wir haben * trotz dieser

327 AN: →wir müssen die **künstler** sein die denen das

328 AN: vermitteln↓ also wir sind genauso gut **künstler**↓←

329 DE: . ja

330 AN: wir haben **künstler** zu sein↓

331 DE: LACHT ja ja aber das=s ne

Das Wort *künstler* kommt im gesamten Gespräch sonst nicht vor. Aus der Betrachtung dieses Falles geht erneut hervor, dass die lokale Verdichtung mit den Strukturbedingungen und Verfahren der engen wechselseitigen Bezugnahme im Zusammenhang mit Aushandlungsprozessen zu tun hat. Das haben beide Kontexte gemeinsam.

5. Fazit

Die vorliegenden Ausführungen sind ein Plädoyer für die Weiterentwicklung von Recherche-Möglichkeiten für Interaktionskorpora. Dazu wurde ein experimentelles System auf der Grundlage von COSMAS II herangezogen, das in den 1990er Jahren im IDS entwickelt wurde. Die Ausführungen hier beziehen sich auf die in diesem System angelegten Möglichkeiten. Diese können im Prinzip natürlich auch in anderen Umgebungen funktionieren.

Das Zielobjekt ist ein komplexes Muster, JA-ABER, für das die Kombination von Zustimmung / Berücksichtigung / Einräumung und Widerspruch / oppositive Relevanzsetzung charakteristisch ist. Dieses Muster wurde gewählt, weil es einige Schwierigkeiten bietet, die für die Recherche von Gesprächseigenschaften typisch sind. Dazu gehören: eine große Variabilität der Ausdrucksweisen; interaktive Bezüge wie Sprecherwechsel, Folgen aufeinander reagierender Aktivitäten und Sequenzmuster; und schließlich relativ weit gespannte und in ihrer Ausdehnung nicht vorab begrenzbare Zusammenhänge. Als besonderes Problem erweist sich z. B. die Auffindung der Einräumungsteile und des strukturellen Zusammenhanges von JA und ABER.

Der Entwicklungsweg führte ausgehend von den Möglichkeiten der einfachen Suche von einzelnen Wortformen, die Erprobung von unterschiedlich großen Wortabständen zwischen den Zielausdrücken, die Zusammenstellung einer größeren Anzahl von Ausdrücken für eine multiple, gestaffelte Suche, über die Analyse von Formulierungsmustern, die für die Recherche geeignet erscheinen bis hin zu relativen Worthäufigkeiten in lokalen Kontexten.

Die angemessene Erfassung dieses Zusammenhanges von JA und ABER in komplexen Äußerungen erfordert eine Ausweitung der ursprünglich implementierten Rechercheverfahren. Das gilt für den Umgang mit Formulierungsmustern, die einerseits feste Strukturen, aber auch Variabilität aufweisen. Hierbei ist im Zweifelsfall eine größere Anzahl von Merkmalen abzu prüfen, was eine gestaffelte Suche erfordert. Und schließlich erfordert auch die Erfassung der relativen Häufigkeit von im Gesamtgespräch seltenen, nicht

vorab identifizierten Wörtern in begrenzten Kontexten eine Weiterentwicklung der ursprünglichen Möglichkeiten der COSMAS-II-Anwendung für Gesprächskorpora.

Der Versuch, die Möglichkeiten vorhandener und unter den gegebenen Voraussetzungen prinzipiell möglich erscheinender Recherche-Verfahren hat für sich genommen schon einen nicht zu unterschätzenden heuristischen Wert. Der von den Systemmöglichkeiten erzwungene sehr formale, viele interpretative Aspekte ausblendende und manchmal im Verhältnis zur gewohnten interpretativen Analysepraxis tangentielle oder auch „schräg“ wirkende Blick auf das Gesprächsmaterial schärft die Wahrnehmung der Eigenschaften des Materials in erstaunlicher Weise und stimuliert die Hypothesenbildung über strukturelle Eigenschaften des Zielobjekts.

Das Ziel der Entwicklung ist die Arbeit mit großen Korpora. Die damit verbundenen Vorteile werden teilweise auch schon bei der Arbeit mit einem noch begrenzten Auswahlkorpus deutlich. Es ergeben sich nicht nur große Fallmengen für die gesuchten Phänomene, sondern auch Einsichten in Häufigkeitsverteilungen. Die Arbeit mit Häufigkeitsverteilungen gehört nicht zur üblichen gesprächsanalytischen Praxis. Die gesprächsanalytische Orientierung auf die qualitative Untersuchung der den jeweiligen Fall konstituierenden Verfahren und Eigenschaften hat ein großes Erkenntnispotenzial, das auf keinen Fall zugunsten einer auf große Mengen von Material gestützten quantitativen Untersuchung aufgegeben werden soll. Aber die Arbeit mit größeren Korpora macht deutlich erkennbar, dass es zumindest heuristisch sehr fruchtbar ist, beide methodischen Perspektiven zu verbinden.¹³

Die Erweiterung der Musteranalyse um Häufigkeitsanalysen wird möglicherweise für die Entwicklung der qualitativen Analyse ein Schlüssel für weitere Entwicklungsschübe sein. Qualitative Analysen stützen sich in sehr vielen Fällen auf implizite Annahmen über Normalität und Auffälligkeit. Und diese Annahmen enthalten auch quantitative Größen. So z. B. wenn bestimmte Vorkommnisse am „Normalfall“ gemessen werden oder an dem, was „in der Regel“ passiert. Die Beurteilung von Auffälligkeit stützt sich einerseits auf die Behandlung des entsprechenden Aspekts in der Interaktion (z. B. wenn etwas als „erklärungsbedürftig“ markiert wird); aber spätestens die Analyse des Ausbleibens von Folgen (d. h. z. B. die Nichtbehandlung von möglicher Auffälligkeit) erfordert eine Bewertungsfolie („sonst“, „in der Regel“, oder auch nur „was auch hätte sein können“). Oft stützt sich die Analyse auf Vergleichsstellen im Kontext, die so interpretiert werden können, dass die fraglichen Normalformerwartungen an anderen Stellen beachtet werden bzw. aktivitätsleitend sind. Aber auch dabei liegen wiederum Annahmen zugrunde, nach denen diese Stellen signifikant sind. Für die Rekonstruktion des

¹³ Zur Debatte und die Verbindung von qualitativen und quantitativen Verfahren, auf die auch schon die Conversation Analysis reagiert hat (vgl. Schegloff 1993) und zu den Entwicklungsperspektiven vgl. u. a. Kallmeyer (2005).

interpretativen Charakters von Interaktion können Häufigkeitsuntersuchungen wichtige Anregungen geben. Solche Überlegungen haben am Beginn der Konversationsanalyse ja durchaus eine Rolle gespielt. Die Conversation Analysis war methodologisch gegen die übliche quantitative Sozialforschung gerichtet, und Harvey Sacks hat sich oft über die Sinnlosigkeit von Häufigkeitsberechnungen mit Stellen hinter dem Komma geäußert. Aber die relativ grobe Wahrnehmung von Verteilungen hat eine Rolle gespielt, z. B. wenn bei der Analyse der Sprecherwechselmechanismen festgestellt wird, dass etwas „in the vast majority of cases“ der Fall ist.¹⁴

Literatur

- Bodmer, Franck/Fach, Marcus L./Schmidt, Rudolf/Schütte, Wilfried (2002): Von der Tonaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für die Arbeit mit Gesprächskorpora. In: Pusch/Raible (Hg.), S. 209–243.
- Cook, Guy (1995): Theoretical issues: transcribing the untranscribable. In: Leech/Myers/Thomas (eds.), S. 35–53.
- Fiehler, Reinhard/Barden, Birgit/Elstermann, Mechthild/Kraft, Barbara (2004): Eigenschaften gesprochener Sprache. (Studien zur Deutschen Sprache. Forschungen des Instituts für Deutsche Sprache Bd. 30). Tübingen.
- Kallmeyer, Werner (1997): Vom Nutzen des technologischen Wandels in der Sprachwissenschaft: Gesprächsanalyse und automatische Sprachverarbeitung. In: Zeitschrift für Literaturwissenschaft und Linguistik 107, S. 124–149.
- Kallmeyer, Werner (2000): Beraten und Betreuen. Zur gesprächsanalytischen Untersuchung von helfenden Interaktionen. In: Zeitschrift für Qualitative Bildungs-, Beratungs- und Sozialforschung 2, S. 227–252.
- Kallmeyer, Werner (2005): Qualitative Methoden. In: Ammon, Ulrich/Dittmar, Norbert/Mattheier, Klaus J./Trudgill, Peter (Hg.), Sociolinguistics/Soziolinguistik. An international Handbook of the Science of Language and Society/Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft. Bd. 1, Artikel 100. 2., überarbeitete Auflage. Berlin/New York. S. 978–992.
- Kallmeyer, Werner/Schmitt, Reinhold (1991a): Zur Markierung von oppositiven Reaktionen in komplexen Äußerungen. Mannheim, SFB 245 „Sprechen und Situation“.
- Kallmeyer, Werner/Schmitt, Reinhold (1991b): Zur Formulierungsdynamik von Äußerungen mit Fokusopposition. Mannheim, SFB 245 „Sprechen und Situation“.
- Kallmeyer, Werner/Schütte, Wilfried (2005): Der Umgang mit Gesprächskorpora am IDS. Die Recherche in der COSMAS-II-Gesprächsdatenbank. In: Schwitalla, Johannes/Wegstein, Werner (Hg.), Korpuslinguistik deutsch: synchron – diachron – kontrastiv. S. 71–80.
- Leech, Geoffrey N./Myers, Greg/Thomas, Jenny (eds.) (1995): Spoken English on computer: transcription, mark-up, and application. New York.
- Mair, Christian (2002): Tracking Ongoing Grammatical Change and Recent Diversification in Present-day Standard English. The Complementary Role of Small and Large Corpora. In: Pusch/Raible (Hg.), S. 9–26.
- Pasch, Renate/Brauße, Ursula/Breindl, Eva u. a. (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der

¹⁴ Vgl. Sacks/Schegloff/Jefferson (1974) und Schegloff (1968).

- deutschen Satzverknüpfen (Konjunktionen, Satzadverbien und Partikeln). (Schriften des Instituts für Deutsche Sprache, 9). Berlin/New York.
- Pusch, Claus D./Raible, Wolfgang (Hg.) (2002): Romanistische Korpuslinguistik: Korpora und gesprochene Sprache. Romance Corpus Linguistics: Corpora and Spoken Language. Tübingen (ScriptOralia 126).
- Sacks, Harvey/Schegloff, Emanuel/Jefferson, Gail (1974): A simplest systematics for the organization of turn taking for conversation. In: *Language* 50, S. 696–735.
- Schegloff, Emanuel (1993): Reflexions on Quantification in the Study of Conversation. In: *Research on Language and social Interaction*, 26, S. 99–128.
- Schmitz, Ulrich (2000): Neue Medien als Arbeitsinstrument der Linguistik. In: Kallmeyer, Werner (Hg.), *Sprache und neue Medien* (Institut für deutsche Sprache, Jahrbuch 1999). Berlin/New York. S. 253–274.
- Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno et al. (1997): *Grammatik der deutschen Sprache*. (Schriften des Instituts für deutsche Sprache 7.17.3). Berlin/New York.

Anhang

A Korpusaufstellung

Die Aufstellung enthält die Wortmengen, die Vorkommen von ABER-Konnektoren, den prozentualen Anteil der ABER-Konnektoren zusammen an der Wortmenge und zu Vergleichszwecken Angaben zur Verwendung von NEIN in Voraussetzungskorrekturen und Zurückweisungen.

Telefonische Studentenberatung Dissertationskosten

Wörter	1428
Aber	3
Nur	0
Bloß	0
Obwohl	0
Summe	3
%	0,210
Nein	1

Medizinische Beratung Mongolismus

Wörter	2798
Aber	12
Nur	0
Bloß	0
Obwohl	0
Summe	12
%	0,429
Nein	0

Genetische Beratung

Wörter	780
Aber	6
Nur	0
Bloß	0
Obwohl	0
Summe	6
%	0,769
Nein	0

Schulpsychologische Beratung Erziehung

Wörter	12.976
Aber	57
Nur	2
Bloß	0
Obwohl	0
Summe	59
%	0,455
Nein	3

Studienberatung Theaterwissenschaftler

Wörter	995
Aber	7
Nur	0
Bloß	0
Obwohl	0
Summe	7
%	0,704
Nein	1

Studienberatung Luxemburgerin

Wörter	5128
Aber	50
Nur	3
Bloß	2
Obwohl	2
Summe	57
%	1,112
Nein	4

Studienberatung Student aus Mainz

Wörter	480
Aber	0
Nur	0
Bloß	0
Obwohl	0
%	0
Nein	1

Psychologische Beratung im Fernsehen Lämmle

Wörter	2763
Aber	19
Nur	1
Bloß	0
Obwohl	0
Summer	20
%	0,848
Nein	0

Nichtsesshaftenberatung Wiedereingliederung

Wörter	2357
Aber	14
Nur	3
Bloß	0
Obwohl	0
Summe	17
%	0,506
Nein	3

Finanzberatung Bausparen

Wörter	3612
Aber	24
Nur	5
Bloß	0
Obwohl	0
Summe	29
%	0,803
Nein	1

Referatbesprechung Professor und Student Sklavenreferat

Wörter	6580
Aber	40
Nur	2
Bloß	1
Obwohl	1
Summe	44
%	0,669
Nein	3

Prüfungsgespräch in der Universität

Wörter	6607
Aber	33
Nur	2
Bloß	0
Obwohl	0
Summe	35
%	0,530
Nein	1

Arbeitsbesprechung Cervantes

Wörter	8450
Aber	38
Nur	0
Bloß	0
Obwohl	0
Summe	38
%	0,450
Nein	1

Talkshow im Fernsehen Berlin 19 Aktionskünstler

Wörter	2771
Aber	12
Nur	0
Bloß	0
Obwohl	0
Summe	12
%	0,433
Nein	10

Talkshow im Fernsehen Feldbusch-Schwarzer

Wörter	17.667
Aber	129
Nur	15
Bloß	0
Obwohl	0
Summe	144
%	0,815
Nein	65

Politische Talkshow im Fernsehen Christiansen-Stoiber

Wörter	12.237
Aber	76
Nur	5
Bloß	0
Obwohl	0
Summe	81
%	0,662
Nein	27

B Suchoperatoren der COSMAS-Gesprächsrecherche

In der COSMAS-Gesprächsrecherche stehen folgende Operatoren zur Verfügung (vgl. auch Bodme Fach/Schmidt/Schütte 2002 und Kallmeyer/Schütte 2005):

- Der Wort-Operator *WORT*(*x*) sucht nach einem oder mehreren alternativen Suchbegriffen, die mit dem Platzhalteroperator * (Asterisk) versehen sein können. So findet *WORT(indirekt*)* Vorkommnisse von *indirekt*, *indirekte*, *indirekter*, *indirektes*, etc. Falls die Großschreibung in den Transkripten verwendet wird, kann durch entsprechende Optionen gesteuert werden, ob die Groß-/Kleinschreibung berücksichtigt werden soll. Durch die Trennung von Text und Annotationen in COSMAS II ist dieser Operator im Stande, Wörter zurückzuliefern, welche durch die Diskursgliederung oder Annotationen fragmentiert sind. Die Suchanfrage *WORT(indirekt)* findet demnach auch Textstellen wie *indirekt* mit einer Überlappungsgrenze, *indi*rekt* mit Mikropause oder *indirekt*↓ mit Intonationsmarkierung.
- Operatoren für die Suche nach Wörtern mit prosodischen Merkmalen: So ist *BETONUNG* ein Suchoperator, der Wörter findet, bei welchen auffallende Betonungen markiert wurden, z. B. *i*ndirekt* mit auffallend betonter Silbe *in*. Die Operatoren *DEHNUNG* und *DEHNUNG-LANG* finden Wörter mit einer auffallenden Dehnung. *INTONATION*(*x*) sucht nach Wörtern mit aufsteigender, absteigender oder schwebender Intonation; z. B. liefert *INTONATION(fallend)* Textstellen wie *indirekt*↓.

- Mit dem Operator *GLEICH*(*X*, *Y*) lassen sich Wortformen und prosodische Merkmale kombinieren, z. B. das Wort *nicht* sowohl mit steigender Intonation als auch mit einer besonderen Dehnung. Als Treffer liefert das Verfahren ein Textstück wie *willst du nicht↑*, übergeht jedoch Fälle wie *glaubst du nicht↑*.
- Operatoren für Pausen verschiedener Länge: *PAUSE* für Pausen beliebiger Dauer, *PAUSE-KURZ* für Mikropausen (in unseren Diskurstranskripten ca. $\frac{1}{2}$ Sek.) und *PAUSE-LANG* für Pausen längerer Dauer ab 1 Sek. Diese drei Pausen-Operatoren lassen sich grundsätzlich je nach Art und Genauigkeit der in den Transkripten vorkommenden Pausenkodierungen beliebig erweitern.
- Operatoren für nicht-lexikalisierte Äußerungen: Nicht-lexikalisierte Äußerungen bilden in unseren Transkripten eine offene Menge von kurzen Beschreibungen in der Sprecherzeile, z. B. *und als ich sah dass er lacht LACHT KURZ sachte ich*. Diese Beschreibungen sind mit Großbuchstaben festgehalten, damit der Leser sie als Annotation erkennen kann. In COSMAS II werden sie in den Annotationsstrukturen gespeichert und über eine Reihe von speziellen Suchoperatoren recherchierbar gemacht, z. B. über den *LACHT*-Operator, der Beschreibungen wie *LACHT*, *LACHT KURZ*, *LACHANSATZ*, etc. wiederfindet. Da diese Beschreibungen nicht ausschließlich geschlossene Listen von Ausdrücken verwenden, sondern auch aus freiem Text bestehen können, haben wir einen Operator *VOCAL*(*x*) eingeführt, der mit freiem Text als Eingabe arbeitet, z. B. *VOCAL* (*LACHANSATZ*), *VOCAL*(*GRINST*), *VOCAL*(*SEUFZT*).
- Operatoren für die Diskursgliederung liefern Wörter zurück, die sich an markanten Stellen der Diskursgliederung wie Äußerungsgrenzen oder Simultanpassagen befinden. So liefern *ÄÜßER-BEG* und *ÄÜßER-END* jeweils das erste bzw. letzte Wort einer Äußerung zurück und *SIM-BEG* bzw. *SIM-END* jeweils das erste bzw. letzte Wort einer Simultanpassage.
- Operatoren für Sprecherdaten wie Altersangabe oder Geschlecht, die mit den anderen Operatoren kombinierbar sind. *IN-SP-GESCHL*(*X* *Y*) filtert die Suchanfrage *X* auf Sprecher mit Geschlecht *Y*; *IN-SP-ALTER*(*X* *Y*) filtert die Suchanfrage *X* auf Sprecher mit Alter *Y*. Bislang enthalten nur wenige Gesprächstranskripte diese Informationen, COSMAS II ist aber darauf vorbereitet, diese Daten bei Suchanfragen berücksichtigen zu können.
- Ein sprecherbezogener Wortabstandsoperator *ABSTAND*(*X* *nw* *Y*) gewährleistet, dass die gefundenen *X*- und *Y*-Textstellen vom selben Sprecher stammen und höchstens den angegebenen Abstand (in Worteinheiten) aufweisen. Dieser Operator kann *X*- und *Y*-Stellen miteinander kombinieren, zwischen denen Einschübe von anderen Sprechern vorkommen. So liefert *ABSTAND*(*WORT*(*nicht*) 3*w* *WORT*(*weil*)) *nicht-weil*-Wortpaare eines Sprechers zurück, die höchstens 3 Wörter auseinander liegen; *ABSTAND* (*WORT*(*nicht*) 0*w* *ÄÜßER-END*) liefert *nicht*-Stellen am Ende einer Äußerung zurück; und: *ABSTAND*(*WORT*(*nicht*) +1*w* *INTONATION*(*steigend*)) liefert *nicht*-Stellen zurück, die unmittelbar vor einem beliebigen Wort (+1*w*) mit steigender Intonation gesprochen wurden.
- Der Abfolgeoperator *ABSTAND*(*X* *ns* *Y*) betrachtet das Transkript als Abfolge von Zeitsegmenten. Eine Simultanpassage gilt in dieser Betrachtungsweise als ein Zeitsegment. Wörter außerhalb von Simultanpassagen bilden jedes für sich ein Segment. Im Gegensatz zum Wortabstand stellt dieser Abstandsoperator sowohl für Wörter eines einzelnen Sprechers als auch verschiedener Sprecher einen Bezug her. Er ist insbesondere als Instrument zur Untersuchung von Phänomenen in und um Simultanpassagen wichtig. Mit dem Abfolgeoperator können nicht-lexikalisierte Äußerungen berücksichtigt werden, die der Wortabstandsoperator außer Acht lässt. So sucht die komplexe Anfrage *ABSTAND*(*IN-SP-GESCHL*(*WORT*(*weil*) *fem*) +1*s* *IN-SP-GESCHL*(*WORT*(*nein*

nee nō) masc)) nach Vorkommnissen von *weil* in Äußerungen von Sprecherinnen, denen unmittelbar (+1s) danach ein *nein* (oder *nee* oder *nō*) eines männlichen Teilnehmers folgt.

Transkriptionskonvention

ja aber		simultane Äußerungen stehen übereinander; Anfang und
nein nie mals		Ende werden auf den jeweiligen Textzeilen markiert
*		kurze Pause (bis max. 1/2 Sekunde)
**		etwas längere Pause (bis max. 1 Sekunde)
=		Verschleifung (Elision) eines oder mehrerer Laute
		<u>zwischen</u> Wörtern (z. B. <i>sa=mer</i> für <i>sagen wir</i>)
/		Wortabbruch
↑		steigende Intonation (z. B. <i>kommst du mit</i> ↑)
↓		fallende Intonation (z. B. <i>jetzt stimmt es</i> ↓)
–		schwebende Intonation (z. B. <i>ich sehe hier-</i>)
"		auffällige Betonung (z. B. <i>aber ge"rn</i>)
:		auffällige Dehnung (z. B. <i>ich war so: fertig</i>)
←immer ich→		langsamer (relativ zum Kontext)
→immerhin←		schneller (relativ zum Kontext)
>vielleicht<		leiser (relativ zum Kontext)
<manchmal>		lauter (relativ zum Kontext)
LACHT		Wiedergabe nichtmorphemisierter Äußerung auf der Sprecherzeile in Großbuchstaben
K	IRONISCH	Kommentar zur Äußerung (auf der Kommentarzeile)
K&	QUIETSCHEN	nicht-kommunikatives (akustisches) Ereignis in der Gesprächssituation (auf der global. Kommentarzeile)
Ü	<i>Mutter</i>	Übersetzungszeile

THORSTEN BRANTS

Statistisch basierte Sprachmodelle und maschinelle Übersetzung

Abstract

Statistische Methoden finden derzeit in der Sprachtechnologie vielfache Verwendung. Ein Grundgedanke dabei ist das Trainieren von Programmen auf großen Mengen von Daten. Für das Trainieren von statistischen Sprachmodellen gilt zur Zeit das Motto „Je mehr Daten desto besser“. In unserem System zur maschinellen Übersetzung sehen wir eine fast konstante qualitative Verbesserung (gemessen als BLEU-Score) mit jeder Verdoppelung der monolingualen Trainingsdatenmenge. Selbst bei Mengen von ca. 20 Milliarden Wörtern aus Nachrichtentexten und ca. 200 Milliarden Wörtern aus Webseiten ist kein Abflachen der Lernkurve in Sicht.

Dieser Artikel gibt kurze Einführungen in statistische maschinelle Übersetzung, die Evaluation von Übersetzungen mit dem BLEU-Score, und in statistische Sprachmodelle. Wir zeigen, welchen starken Einfluß die Größe der Trainingsdaten des Sprachmodells auf die Übersetzungsqualität hat. Danach wird die Speicherung großer Datenmengen, das Trainieren in einer parallelen Architektur und die effiziente Verwendung der bis zu 1 Terabyte großen Modelle in der maschinellen Übersetzung beschrieben.

Statistische maschinelle Übersetzung

Dieses Feld findet in den letzten Jahren mehr und mehr Interesse in der Forschung. Aufgrund der weltweit steigenden Verbreitung des Internets besteht auch ein steigender Bedarf an Übersetzungen von Texten, um Webseiten auch denjenigen Benutzern zugänglich zu machen, die die Originalsprache des Dokuments nicht verstehen. Leider läßt die Übersetzungsqualität der existierenden Systeme oft noch sehr viel zu wünschen übrig. Im Folgenden ein Beispiel für eine Übersetzung aus dem Arabischen ins Englische, bei dem die geringe Qualität ein Verstehen fast unmöglich macht:

The Bradi : The inspectors need to “a few months” for end important their

Paris 13 – 1 (aa so in in) – the general manager for agency announced international for energy atomic Mohammed the Bradi today Monday that inspectors of international nze’ the weapons need to “a few months” for end important their in Iraq.

Journalistic conference in end of meeting with French External Minister of Dominique de Villepin that the inspectors said during “a few their need to important months for end”.

Said that he not can to the accurate for conclusion schedule an appointment important the inspectors and definitive importance the appointment lessened from on January 27

indicative to that the report who Security Council at that the date of he will hand over to “temporary report”.

The Bradi that Security Council confirmed “ünderstands” that January 27 final term not.

Als positives Beispiel hier die Übersetzung, die unser Google Research System vom gleichen Originaltext produziert:

ElBaradei : Inspectors Need “a Few Months” to Complete Their Mission

Paris 13 – 1 (AFP) – The Director General of the International Atomic Energy Agency Mohamed ElBaradei today, Monday, that the international disarmament inspectors need “a few months” to complete their mission in Iraq.

He said during a press conference at the conclusion of a meeting with French Foreign Minister Dominique de Villepin that the inspectors “need a few months to complete their mission.”

He could not specify the exact date of the conclusion of the task of the inspectors and reduced the importance of the deadline of 27 January, pointing out that the report will be handed to the Security Council on that date is a “progress report”.

ElBaradei stressed that the Security Council “ünderstands” that the 27 January deadline is not final.

Auch ohne den Originaltext zu kennen, lässt sich unschwer erkennen, dass dies eine bessere Übersetzung ist. Bis auf einige kleinere Fehler, die das Verstehen nur wenig einschränken (z. B. Fehlen des Verbs „said“ im ersten Satz), hat der übersetzte Text in etwa die Qualität eines von einem Muttersprachler produzierten Textes.

Statistische maschinelle Übersetzungssysteme kombinieren in der Regel mindestens zwei statistische Modelle: das Übersetzungsmodell und das Sprachmodell (Brown et al. 1990). Das Übersetzungsmodell liefert Wahrscheinlichkeiten $p(f|e)$ dafür, dass einzelne Wörter oder relative kurze Sequenzen von Wörtern in der Quellsprache f Übersetzungen von Wörtern oder Sequenzen in der Zielsprache e sind.¹ Aufgrund dessen werden Satzbruchstücke als Kandidaten für Übersetzungen ausgewählt. Das Sprachmodell liefert Wahrscheinlichkeiten $p(e)$ für Sequenzen von Wörtern in der Zielsprache e . Diese sorgen dafür, dass die Kandidatenbruchstücke in die richtige Reihenfolge gebracht werden. Die meisten Systeme verwenden zusätzlich weitere Informationsquellen (*Features*) wie z. B. die Satzlänge, wie viele verschiedene Bruchstücke verwendet werden, wie stark die Reihenfolge in der Zielsprache von der Reihenfolge in der Ausgangssprache abweicht usw. Alle verwendeten statistischen Modelle und weiteren Features werden dann in einem log-linearen Model miteinander kombiniert:

$$\log(\text{score}) = \lambda_1 \log(p(f|e)) + \lambda_2 \log(p(e)) + \lambda_3 \dots$$

¹ Die Abkürzungen e und f standen ursprünglich für *english* als Zielsprache und *french* als Quellsprache, da ein Großteil der frühen Arbeiten zur statistischen maschinellen Übersetzung sich mit diesem Sprachpaar beschäftigte. Jetzt werden sie unabhängig vom Sprachpaar weiterverwendet.

Die Übersetzungswahrscheinlichkeiten $p(f|e)$ und die meisten anderen Features sowie deren Gewichtungparameter λ werden anhand von parallelen Trainingsdaten gelernt, d. h. anhand von Dokumenten, die Übersetzungen voneinander sind. Im phrasenbasierten Ansatz wird versucht, innerhalb von Sätzen möglichst lange Phrasen zu identifizieren, die häufig Übersetzungen voneinander sind (Och et al. 1999; Koehn et al. 2003). Dabei sind mit Phrasen in der Regel keine Phrasen im linguistischen Sinne gemeint, sondern beliebige Sequenzen von Wörtern, Zahlen und Satzzeichen, die häufig gemeinsam mit einer bestimmten Sequenz in der anderen Sprache auftreten.

Die Sprachmodell-Wahrscheinlichkeiten $p(e)$ werden anhand von monolingualen Daten gelernt. Hierfür stehen die größten Datenmengen bereit. Dies wird weiter unten eingehender betrachtet.

Qualitätsbestimmung durch BLEU-Scores

Traditionell wird die Qualität von Übersetzungen durch menschliche Evaluation festgestellt. Möglich sind z. B. das Bewerten von Übersetzungen auf einer mehrstufigen Skala für verschiedene Kategorien wie Adäquatheit oder Flüssigkeit (adequacy, fluency) oder der direkte Vergleich zweier verschiedener Übersetzungen, indem ein Annotierer bestimmt, ob Übersetzung A oder B die bessere ist. Menschliche Evaluation erlaubt sehr sichere Qualitätsurteile und ist damit auch heute noch die zuverlässigste Methode für Evaluationen. Sie hat jedoch den Nachteil, dass sie sehr zeit- und arbeitsaufwendig ist und damit nur schlecht für Entwicklungszyklen automatischer Systeme geeignet ist.

Papineni et al. (2002) stellten daher ein automatisches System mit dem Namen „BLEU“ vor (oder auch: BLEU-Score). Dieses System benötigt ebenfalls menschliche Übersetzer, jedoch lässt es von diesen nur genau einmal (vor der eigentlichen Evaluation) mehrere Übersetzungen des gleichen Texts anfertigen und vergleicht dann automatisch, wie nah die zu evaluierenden Übersetzungen den Referenzübersetzungen sind. Sie zeigten in ihrer Studie, dass dieser automatische Vergleich sehr stark mit menschlichen Urteilen korreliert, wenn eine große Zahl von Testsätzen verwendet wird, und dass in vielen Fällen sowohl menschliche als auch automatische Evaluation zur gleichen Rangreihenfolge der zu vergleichenden Übersetzungen führt.

Der BLEU-Score gab der Forschung zur maschinellen Übersetzung neuen Antrieb. Evaluationen, die bisher Wochen oder gar Monate dauerten, finden nun innerhalb weniger Sekunden statt. Die Idee hinter dem BLEU-Score wird in dem folgenden Beispiel erläutert. Wir nehmen an, dass es drei menschliche englische Referenzübersetzungen eines Textes gibt:

Referenz 1: Japan remained the biggest trading partner

Referenz 2: Japan is still the largest trade partner

Referenz 3: Japan still remains the number one trade partner

Jetzt soll bestimmt werden, welches von drei verschiedenen Systemen die beste Übersetzung erstellt. Der BLEU-Score zieht Wortsequenzen (N-gramme) der Längen 1 bis 4 in Betracht², die mit den Referenzübersetzungen übereinstimmen. Alle Übereinstimmungen sind im folgenden hervorgehoben:

System 1: **Japan** will continue to be **partner** big

System 2: **Japan is still the biggest trading partner**

System 3: Ukraine won 2:1 against Poland

System 1 generierte lediglich zwei übereinstimmende Unigramme: „Japan“ und „partner“. Die Übersetzung von System 2 ist komplett durch N-gramme in den Referenzen abgedeckt: durch „Japan is still the“ aus Referenz 1, und durch „the biggest trading partner“ aus Referenz 2. System 3 generierte keines der Referenzworte. Je mehr Text durch übereinstimmende N-gramme abgedeckt ist und je länger die N-gramme sind, desto besser die Qualität. Demnach lieferte System 2 die beste Übersetzung. Dies deckt sich sehr gut mit menschlichen Urteilen.

Der BLEU-Score verwendet:

- *Die modifizierte N-gramm-Präzision*: Für die Standard N-gramm-Präzision wird errechnet, welcher Anteil der maschinell erzeugten N-gramme der Länge k auch in den Referenzübersetzungen erscheint. Die modifizierte N-gramm-Präzision lässt nur die einmalige Benutzung jedes Teils der Referenzübersetzung zu.
- *Brevity Penalty*: Dieser Parameter bestraft die Erstellung von zu kurzen Übersetzungen.
- *BLEU-4-Score*: Das geometrische Mittel der N-gramm-Präzision der Längen N=1 bis 4, multipliziert mit der Brevity Penalty.

Bei großen Mengen von Sätzen besteht normalerweise eine hohe Korrelation zwischen dem BLEU-Score und menschlichen Urteilen zur Qualität der Übersetzung. Die BLEU-Score-Werte wachsen mit der Anzahl der Referenzübersetzungen. Beispielsweise erreichen die besten statistischen Systeme bei der Übersetzung von Nachrichtentexten vom Arabischen ins Englische BLEU-Scores von mehr als 50%, wenn 4 Referenzübersetzungen verwendet werden. Zum Vergleich: Menschliche Übersetzungen erreichen Werte zwischen 50 und 60% (leicht über statistischen Systemen), und die existierenden kommerziellen (nicht-statistischen) System erreichen lediglich Werte von 10 bis 35%. Die BLEU-Scores für Übersetzungen vom Chinesischen ins Englische liegen insgesamt niedriger (Werte in den 30ern), weisen aber ähnliche Muster auf.

Das amerikanische National Institute of Standards and Technology (NIST) führt jährlich eine Evaluation von automatischen Übersetzungssystemen.

² Dies ist der sogenannte BLEU-4-Score, der am häufigsten verwendet wird. Andere Maximalängen sind möglich geben jedoch in der Regel sehr ähnliche Ergebnisse.

men durch. Im Jahre 2005 erreichte das Google Research System BLEU-Scores von 51,37% für Arabisch-Englisch und 35,31% für Chinesisch-Englisch. Wir waren mit diesen Werten sehr zufrieden. Die Ergebnisse der anderen Teilnehmer sind auf den Webseiten von NIST zu finden.³

Statistische Sprachmodelle

Ein Standard-Ansatz für statistische Sprachmodelle ist die Verwendung von N-grammen, deren Wahrscheinlichkeiten aufgrund ihrer Auftretenshäufigkeiten in größeren Textmengen geschätzt werden. Die erste große Anwendung für diese Modelle war die automatische Spracherkennung (Jelinek/Mercer 1980). N-gramme sind Sequenzen von N Tokens, wobei ein Token aus einem Wort, einer Zahl oder Interpunktion besteht. Bei Verwendung von N-grammen wird die Wahrscheinlichkeit des nächsten Wortes bestimmt unter der Bedingung, dass die vorherigen $N-1$ Wörter bekannt sind. Die einzelnen Wahrscheinlichkeiten werden gemäß der Kettenregel multipliziert. Lautet z. B. die zu verarbeitende Wortfolge „NASA officials say they hope ...“, dann ergibt sich dessen Wahrscheinlichkeit in einem Trigramm-Modell ($N=3$) aus

$$\begin{aligned} p(\text{NASA officials say they hope ...}) \\ = p(\text{NASA}) p(\text{officials} \mid \text{NASA}) p(\text{say} \mid \text{NASA officials}) \\ p(\text{they} \mid \text{officials say}) p(\text{hope} \mid \text{say they}) p(\dots \mid \text{they hope}) \end{aligned}$$

oder allgemein für einen Satz mit den Wörtern $w_1 \dots w_k$:

$$p(w_1, \dots, w_k) = \prod_{i=1}^k p(w_i \mid w_{i-N+1}, \dots, w_{i-1})$$

N-gramm-Modelle sind auch unter der Bezeichnung Markov-Modell bekannt (Markov 1913). Die Wörter $w_{i-N+1} \dots w_{i-1}$ bilden den vorherigen Kontext (auch Geschichte, History), w_i ist das nächste Wort. Die Wahrscheinlichkeiten auf der rechten Seite der Gleichung werden anhand von relativen Häufigkeiten geschätzt,

$$p(w_i \mid w_{i-N+1}, \dots, w_{i-1}) = \frac{f(w_{i-N+1}, \dots, w_i)}{f(w_{i-N+1}, \dots, w_{i-1})}$$

wobei f die Häufigkeit im zugrunde liegenden Korpus repräsentiert. Im konkreten Fall ist dies zum Beispiel für die Wahrscheinlichkeit, dass das nächste Wort „say“ ist, wenn der vorherige Kontext „NASA officials“ war:

$$p(\text{say} \mid \text{NASA officials}) = \frac{f(\text{NASA officials say})}{f(\text{NASA officials})} = \frac{752}{12819} = 0.05866$$

³ (http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html)

Glättung

Die direkte Verwendung von relativen Häufigkeiten führt in der Regel zu Problemen mit Null-Wahrscheinlichkeiten, falls eines der beteiligten N-gramme nicht im Korpus auftritt. Diese sorgen dafür, dass die Wahrscheinlichkeit eines gesamten Satzes ebenfalls Null wird, da die einzelnen Wahrscheinlichkeiten multipliziert werden. Damit würde allen Sätzen, die eine solche Null-Wahrscheinlichkeit enthalten, die gleiche Wahrscheinlichkeit von Null zugewiesen, und wir könnten keine Aussage mehr darüber machen, welcher Satz der bessere oder schlechtere ist. Um dies zu vermeiden, wird Glättung (Smoothing) eingesetzt. Chen und Goodman (1998) geben eine umfassende Übersicht zu diesem Thema. Standardverfahren sind das Backoff und die gewichtete Interpolation. Im Backoff werden kürzere N-gramme verwendet, falls der vorherige Kontext seltener als ein vorbestimmter Schwellwert auftritt:

$$p(w_i | w_{i-n+1}, \dots, w_{i-1}) = \begin{cases} \hat{p}(w_i | w_{i-n+1}, \dots, w_{i-1}) & \text{falls } f(w_{i-n+1}, \dots, w_i) > \theta \\ \lambda p(w_i | w_{i-n+2}, \dots, w_{i-1}) & \text{sonst} \end{cases}$$

In der gewichteten Interpolation werden ebenfalls längere und kürzere N-gramme gemeinsam verwendet, allerdings gewichtet unter Verwendung eines Faktors λ :

$$p(w_i | w_{i-N+1}, \dots, w_{i-1}) = \lambda \hat{p}(w_i | w_{i-N+1}, \dots, w_{i-1}) + (1 - \lambda) p(w_i | w_{i-N+2}, \dots, w_{i-1})$$

Dies ist eine rekursive Definition, d. h. die Wahrscheinlichkeit des kürzeren N-gramms im rechten Teil der Gleichung wird wiederum durch eine gewichtete Interpolation mit dem noch kürzeren N-gramm bestimmt.

Korpora

Für das Trainieren eines Sprachmodells benötigen wir große Mengen an Daten in der Zielsprache. Wir benutzten unter anderem die Gigaword Corpora in Englisch, Chinesisch und Arabisch, die vom Linguistic Data Consortium (LDC) herausgegeben werden. Sie enthalten jeweils einige 100 Millionen bis einige Milliarden Wörter pro Sprache. Als sich im Laufe unserer Arbeit an dem System der Bedarf an mehr und mehr Daten ergab, begannen wir, Google-Daten aus dem Internet zu benutzen. Dies sind zum einen Nachrichtentexte aus dem „Google News“ Service. Im aktuellen System verwenden wir davon ca. 25 Milliarden Wörter englischer Texte und 17 Milliarden Wörter chinesischer Texte. Für das Deutsche stehen ca. 4 Milliarden Wörter aus Nachrichtentexten bereit. Weitere Sprachen wie z. B. Französisch, Italienisch, Spanisch, Arabisch und andere sind ebenfalls vorhanden. Nachrichtentexte haben den Vorteil, dass öffentlich zugängliche Daten dieser Domäne täglich wachsen und ständig neue Ereignisse durch sie abgedeckt werden. Weiterhin verwenden wir Texte aus allgemeinen Web-Seiten, die in noch größeren Mengen zur Verfügung stehen. So verarbeiteten wir z. B. 200 Milliarden Wörter aus englischen Seiten für unser aktuelles Übersetzungssystem. Daten für an-

dere Sprachen stehen ebenfalls zur Verfügung, z. B. ca. 88 Milliarden Wörter für das Chinesische und ca. 30 Milliarden Wörter für das Deutsche. Dieses sind nur kleine Bruchstücke der insgesamt bei Google indizierten Web-Daten, und wir arbeiten daran, noch größere Mengen für unser Übersetzungssystem zu verwenden.

Zum Trainieren des Sprachmodells in unserem System zur maschinellen Übersetzung werden unannotierte Daten verwendet, die nur einer geringfügigen Vorverarbeitung unterzogen werden (z. B. Tokenisierung und Satzende-Erkennung). Dies ist eine wichtige Voraussetzung. Der Einsatz manueller Schritte würde die Verwendung solch großer Datenmengen unmöglich machen. Weitere Gebiete, die ebenfalls von großen Mengen unannotierter Daten profitieren können, sind unter anderen die automatische Spracherkennung, Texterkennung und die Rechtschreibkorrektur.

Einfluss der Sprachmodellierung auf die Übersetzungsqualität

„Je mehr Daten desto besser“ ist das Motto, das hier überprüft werden soll. Dafür trainierten wir 5-gramm-Sprachmodelle auf verschiedenen Mengen von Nachrichtentexten und integrierten diese in unser Übersetzungssystem. Alle auftretenden N-gramme wurden verwendet, d. h. die Mindesthäufigkeit für N-gramme im Modell war 1. Für die Glättung wurde ein Backoff-Verfahren angewandt. Die kleinste Datenmenge ist 75 Millionen Wörter. Danach wird die Datenmenge in jedem Schritt in etwa verdoppelt, insgesamt achtmal, bis 18 Millionen Wörter erreicht sind. Getestet wird auf den Daten der NIST MT Evaluation von 2003 für Übersetzungen aus dem Arabischen ins Englische. Die Testdaten umfassen 100 Texte mit insgesamt 663 Sätzen.⁴ Die Ergebnisse sind in Abbildung 1 dargestellt.

Wir sehen eine fast konstante Verbesserung von ca. 0.5% im BLEU-Score bei jeder Verdoppelung der Datenmenge. Dies ist eine sehr deutliche Verbesserung, und viele in der Literatur beschriebenen Neuerungen liefern kleinere Verbesserungen des BLEU-Scores. Auch bei 18 Milliarden Wörtern scheint die Kurve nicht abzuflachen.⁵

Schließlich fügten wir noch ca. 200 Milliarden Wörter aus Webseiten zum Sprachmodell hinzu. Die Testdaten bestehen im wesentlichen aus Nachrichtentexten. Die Verwendung einer anderen Domäne lässt geringere Verbesserungen erwarten. Dies bestätigte sich auch in unseren Untersuchungen. Insgesamt ergibt der Sprung von 18 Milliarden Wörtern Nachrichtentexten auf 200 Milliarden Wörtern aus Webtexten eine Verbesserung von 0.9%. Dies

⁴ Diese Daten sind vom Linguistic Data Consortium unter der Katalognummer LDC2005T05 erhältlich.

⁵ 18 Millionen Wörter ist die größte Menge an Nachrichtentexten die zum Zeitpunkt der Experimente zur Verfügung stand. Inzwischen liegen mehr Daten vor und es wird interessant zu sehen, wie die Kurve weiter verläuft.

Abb. 1: BLEU-Scores für NIST 2003 MT Evaluationsdaten Arabisch-Englisch bei Verwendung verschiedener Mengen von Nachrichtentexten für das Sprachmodell (M = Millionen, G = Milliarden Wörter).

entspricht immerhin noch einer Steigerung von ca. 0.25% bei jeder Verdopplung der Datenmenge. Webtexte sind zwar weniger hilfreich für unsere Testdaten, liegen jedoch nur genügend davon vor, dann ist trotzdem eine deutliche Steigerung der Übersetzungsqualität sichtbar.

Ähnliche Untersuchungen mit Mengen paralleler Daten ergaben, dass jede Verdopplung der Datenmenge den BLEU-Score um ca. 2.5% verbessert. Der Einfluss paralleler Daten ist somit deutlich höher als der monolingualer Daten. Jedoch liegen letztere in viel größeren Mengen vor. Die vom Linguistic Data Consortium vertriebenen parallelen Korpora haben 100–200 Millionen Wörter für die größten Sprachpaare. Die in dieser Untersuchung verwendeten monolingualen Datenmengen sind um den Faktor 1000 größer.

Das auf 200 Milliarden Wörtern trainierte Sprachmodell ist riesig. Es enthält 45,6 Milliarden N-gramme, und die zugehörigen Tabellen benötigen ca. 1.1 Terabyte Speicherplatz. 66,5% der N-gramme sind Singletons, d. h., treten nur einmal auf. Jedoch schadet das Entfernen der Singletons der Übersetzungsqualität leicht. Im folgenden betrachten wir das Trainieren und die Anwendung dieser Modelle.

Verteiltes Trainieren von Sprachmodellen

Die Rohtexte werden verteilt in mehreren hundert Dateien im Google File System (GFS) gespeichert (Ghemawat et al. 2003). Dies ist ein Dateisystem, das auf Standardhardware läuft und jeden Datenblock auf drei verschiedenen Maschinen repliziert. Ausgefallene Maschinen werden automatisch erkannt, und die Software erstellt neue Kopien von Datenblöcken, so dass immer drei Kopien vorhanden sind. Das zugrunde liegende Netzwerk besteht aus mehr als 1000 Computern mit insgesamt mehr als 300 Terabyte Plattenspeicherplatz. Dieses dient als Basis für die parallele Verarbeitung der Daten.

Als Software-Infrastruktur dient MapReduce (Dean/Ghemawat 2004). Dieses parallelisiert die Verarbeitung automatisch. Die Daten werden unterteilt in Blöcke (die *Shards*), auf denen die Verarbeitung in zwei Phasen abläuft. Phase 1 („Map“) bildet die Eingabedaten auf temporäre Schlüssel-Werte-Paare ab. Jeder Schlüssel wird dann aufgrund einer Hashfunktion einem Shard für Phase 2 („Reduce“) zugewiesen. Diese Phase kombiniert alle temporären Paare mit dem gleichen Schlüssel und schreibt dann die Ausgaben.

Im ersten Schritt unserer Verarbeitung erstellen wir das Vokabular. In diesem konkreten Fall eines MapReduce dienen die Wörter als Schlüssel und deren Häufigkeiten als Werte. Die temporären Paare enthalten die Wörter und deren Häufigkeiten für einen bestimmten Teil der Eingabedaten (z. B. für eine der mehreren hundert Dateien). Alle Paare für dasselbe Wort werden zum selben „Reduce“-Shard geschickt, in dem dann die Häufigkeiten addiert werden. Schließlich wird das Wort zusammen mit der Summe der Häufigkeiten in die Ausgabedatei des entsprechenden Shards geschrieben. Wörter, die zu selten auftreten, z.B. seltener als einmal in 5 Milliarden Wörtern, werden entfernt. An ihre Stelle tritt das „unbekannte Word“ (<UNK>). Damit müssen später weniger N-gramme gespeichert werden, da z.B. die Häufigkeiten $f(A\ B\ SELTENES_WORT_1\ C)$ und $f(A\ B\ SELTENES_WORT_2\ C)$ denselben Eintrag belegen, nämlich $f(A\ B\ <UNK>\ C)$. Für alle anderen Wörter werden im nächsten Schritt direkt die N-gramme gesammelt. Als Beispiel für den Einfluss der Minimalhäufigkeiten auf die Datengröße geben wir hier Größen an, die sich bei der Bearbeitung von 200 Milliarden Wörtern aus Webdaten ergaben: bei einer minimalen Worthäufigkeit von 1 werden 400 Millionen verschiedene „Wörter“ gefunden, was zu 1.1 Terabyte 5-gramm-Daten führt, während bei einer minimalen Häufigkeit von 40 lediglich 11 Millionen verschiedene „Wörter“ verbleiben, was zu 800 Gigabyte 5-gramm-Daten führt.

Die Vokabularerstellung ist in Abbildung 2 schematisch dargestellt mit zwei Reduce-Shards. Tatsächlich verwenden wir 100 und mehr Shards für diesen Schritt. Abbildung 3 zeigt ein sehr vereinfachtes Beispiel mit drei Dokumenten. Die tiefgestellten Indizes vor den N-grammen zeigen den Wert der Hash-

Abb. 2: Schematische Darstellung der Vokabularerstellung als MapReduce.

Abb. 3: Beispiel für die Vokabularerstellung mit drei Eingabedokumenten.

funktion, die bestimmt, zu welchem Reduce-Shard das N-gramm geschickt wird.

Im zweiten Schritt werden die eigentlichen N-gramm-Tabellen zusammengestellt. Hier kommt ebenfalls MapReduce zum Einsatz. Jetzt sind die N-gramme die Schlüssel, die Häufigkeiten sind wiederum die Werte. Die Anzahl der Map- und Reduce-Shards ist jetzt viel größer um die viel größere Menge der N-gramme zu verarbeiten. Typischerweise verwenden wir hier 1000 Shards. Ein sehr einfaches Beispiel für die Erstellung von Bigrammen aus drei Dokumenten ist in Abbildung 4 dargestellt. Für dieses Beispiel wurden den Dokumenten zwei Wörter hinzugefügt, die jeweils nur einmal auftreten, um auch die Verwendung von $\langle \text{UNK} \rangle$ zu illustrieren.

Im nächsten Schritt wird die relative Häufigkeit r eines jeden N-gramms bestimmt. Diese ergibt sich aus der Häufigkeit f des N-gramms und der des linken Kontexts:

$$r(E \mid ABCD) = \frac{f(ABCDE)}{f(ABCD)}$$

Ein Problem dabei ist, dass die beiden Häufigkeiten f nicht notwendigerweise im selben Shard gespeichert sind, auf diese aber gleichzeitig zugegriffen werden muss. Ein wahlfreier Zugriff auf die auf Festplatte gespeicherten Tabellen wäre sehr langsam. Eine Lösung dafür ist, dass die gewählte Hash-Funktion N-gramme aufgrund ihres ersten Wortes verteilt. Im Beispiel oben wären damit alle N-gramme, die mit dem Wort „A“ beginnen, im selben Shard. Relative Häufigkeiten können so in nur einem Lauf über den betreffenden Shard berechnet werden. Ein Nachteil dieses Verfahrens ist, dass

Abb. 4: Beispiel für die N-gramm-Erstellung mit drei Eingabedokumenten.

einige Wörter (z. B. Funktionswörter) sehr viel häufiger sind als andere. Shards, die diese sehr häufigen Wörter speichern, werden in der Regel viel größer als andere Shards. Dies möchten wir vermeiden, damit die Verarbeitungszeiten für alle Shards in etwa gleich sind. Die ungleichmäßige Verteilung kann in den meisten Fällen dadurch gelöst werden, dass die Hash-Funktion die ersten *zwei* Wörter berücksichtigt und alle Unigramme in allen Shards repliziert werden (damit relative Häufigkeiten von Bigrammen ebenfalls innerhalb eines Shards berechnet werden können). Dies wird *Präfix-Sharding* genannt. Die Vervielfältigung der Unigramme hat nur einen sehr kleinen Einfluss auf den benötigten Speicherplatz, da ihre Anzahl viel kleiner ist als die der N-gramme höherer Ordnung.

Der letzte Schritt für die Erstellung des Sprachmodells ist das Smoothing. Dafür muss auf kürzere N-gramme zugegriffen werden, die durch Entfernen des jeweils ersten Wortes im linken Kontext entstehen. So muss z.B. zur Berechnung der Wahrscheinlichkeit eines 5-gramms $p(E|ABCD)$ im Smoothing auf diese fünf relativen Häufigkeiten zugegriffen: $r(E|ABCD)$, $r(E|BCD)$, $r(E|CD)$, $r(E|D)$ und $r(E)$. Diese sind in den N-gramm-Tabellen unter *ABCDE*, *BCDE*, *CDE*, *DE* und *E* zu finden. Da jedoch im vorherigen Schritt die Hash-Funktion die N-gramme anhand der beiden ersten Wörter verteilt hat, sind diese jetzt nicht notwendigerweise im selben Shard zu finden. Das macht das Smoothing sehr zeitaufwendig. Die Lösung dafür ist, eine neue Hash-Funktion zu wählen, die die N-gramme anhand der beiden *letzten* Wörter verteilt. Die N-gramme werden damit nach der Berechnung der relativen Häufigkeiten und vor dem Smoothing umverteilt. Diese Umverteilung ist bei

großen Datenmengen effizienter als der wiederholte wahlfreie Zugriff auf viele Shards.

Die vier Schritte für das Erzeugen eines Sprachmodells (Vokabularerstellung, Erstellung der N-gramm-Tabellen, Berechnung der relativen Häufigkeiten und Smoothing) benötigen bei 200 Milliarden Wörtern im Eingabetext ca. einen Tag bei Verarbeitung auf 1000 Maschinen.

Verwendung verteilter Sprachmodelle

Die Verwendung sehr großer verteilter Sprachmodelle stellt eine weitere Herausforderung dar. Wahlfreier Zugriff auf Dateien auf Festplatten verbietet sich meistens aus Effizienzgründen. Wir laden deshalb jeden Shard in den Hauptspeicher einer Maschine und greifen dann auf diese über das Netzwerk zu. Dabei ist der größte Engpass die Wartezeit im Netzwerk. Ein Zugriff im Speicher einer Maschine dauert im Mikrosekundenbereich während ein Zugriff über das Netzwerk im Millisekundenbereich liegt. Die Übersetzung eines Satzes benötigt Zugriff auf ca. 1 Million N-gramme. Setzen wir einen Zugriff über das Netzwerk sehr optimistisch mit einer Millisekunde an, dann würde bei naiver Verwendung die Übersetzung des Satzes 1000 Sekunden dauern (also mehr als 16 Minuten).

Ein Zugriff pro N-gramm ist damit viel zu ineffizient. Eine Lösung besteht darin, nicht jede Anfrage für ein N-gramm direkt zu senden, sondern diese zunächst in eine Warteschlange zu stellen, um dann später gleichzeitig auf viele N-gramme zuzugreifen. Die Verarbeitung könnte z. B. in 20 Schritten ablaufen, in jedem Schritt werden ca. 50.000 N-gramme abgefragt. Dies dauert zur Zeit für alle 20 Schritte zusammen weniger als 1 Sekunde.

Unser Übersetzungssystem wechselt jetzt zwischen dem Einreihen in die Warteschlange (den „Dummy Lookups“) und der eigentlichen Verwendung der N-gramme (den „True Lookups“) hin und her. Das Vorgehen ist im Suchgraphen in Abbildung 5 dargestellt. Zum Zeitpunkt t werden die 3 besten Hypothesen betrachtet: AB, A, C. Jede dieser Hypothesen hat zwei mögliche Fortsetzungen, es ergeben sich sechs neue Hypothesen: AB.D, AB.C, A.BC, A.D, C.A, C.AB. Zwei davon (AB.C und A.BC) ergeben dieselbe Wort-

Abb. 5: Beispiel für eine Graphensuche mit verteiltem Sprachmodell und abwechselnden Dummy Lookups und True Lookups.

sequenz. Im Dummy Lookup werden also die N-gramme für 5 neue Hypothesen in die Warteschlange eingereiht. Danach wartet das System auf die Antwort vom Server, der die Wahrscheinlichkeiten zurückliefert. Mit deren Hilfe wird nun jede Hypothese bewertet und der Graph beschnitten. Nur die besten drei verbleiben für den nächsten Schritt (ABC, CA, CB im Beispiel). Danach beginnt der Prozess von vorn. Wiederum werden Hypothesen generiert, die Warteschlange aufgebaut, usw.

Abschließende Bemerkungen

Die Verwendung großer verteilter Sprachmodelle hat einen starken Einfluss auf die Übersetzungsqualität automatischer Systeme. Ein 5-gramm-Modell trainiert auf 150 Millionen Wörtern passt in den Speicher einer Maschine und liefert mit Google's Übersetzungssystem einen BLEU-Score von 49.13 % für Arabisch-Englische Übersetzungen. Ein 5-gramm-Modell trainiert auf 200 Milliarden Wörtern belegt ca. 1 Terabyte bei Speicherung aller auftretenden N-gramme. Dies verhilft zu einem BLEU-Score von 53.10 % auf den gleichen Daten. Dies ist eine enorme Steigerung. Die verteilte Verwendung des größten Sprachmodells mit Dummy Lookups und True Lookups führt dabei kaum zu Geschwindigkeitseinbußen, die Übersetzung eines Satzes geschieht in ca. einer Sekunde.

Eine der Ursachen für die hohe Qualität des großen Sprachmodells ist, dass viele der N-gramme in zu übersetzenden Texten vom System bereits gesehen worden sind. Dies wird in der folgenden Tabelle veranschaulicht.⁶

Training	unbekannt	5-gramme	4-gramme	3-gramme
150 Mio	0.44 %	14.75 %	33.83 %	64.46 %
200 Mia	0.09 %	47.52 %	72.89 %	92.20 %

Die Tabelle zeigt Werte für zwei verschiedene Mengen von Trainingsdaten: 150 Millionen Wörter und 200 Milliarden Wörter. Die zweite Spalte gibt an, wie viele der Wörter in den Referenzübersetzungen unbekannt sind, d. h., nicht in den Trainingsdaten aufgetaucht sind. Dies sind deutlich weniger für das große Sprachmodell. Die weiteren Spalten geben an, wieviele der N-gramme bekannt sind. Die Unterschiede sind sehr groß. So „kennt“ das große Sprachmodell ungefähr dreimal so viele 5-gramme wie das kleine Sprachmodell (47.52 % gegenüber 14.75 %), kann also deutlich bessere Schätzungen für deren Wahrscheinlichkeiten angeben.

⁶ Alle Werte wurden auf den Arabisch-Englischen Übersetzungsdaten des Linguistic Data Consortium mit der Katalognummer LDC2005T05 gemessen.

Sehr große N-gramm-Tabellen sind für die Forschung in einer Reihe von Gebieten nützlich. Zusätzlich zur maschinellen Übersetzung sind dies z. B. die automatische Spracherkennung, die Zeichenerkennung, Rechtschreibkorrektur und Informationsextraktion. Allerdings ist die Erstellung solcher Tabellen sehr zeit- und ressourcenaufwändig. Google plant deshalb, 5-gramm-Tabellen, erstellt aus ca. 1 Billion (10^{12}) Wörtern englischer Web-Daten, für die Forschung zur Verfügung zu stellen. Damit soll allen Forschungsgruppen erlaubt werden, gemeinsam mit großen Datenmengen zu arbeiten, unabhängig von der Größe der Gruppe und ihrer Computer-Infrastruktur. Die Veröffentlichung wird ca. 6 DVDs umfassen und vom Linguistic Data Consortium vertrieben werden.

Danksagung

Vielen Dank für die gemeinsame Arbeit an der Erstellung der großen Sprachmodelle, für zahlreiche Diskussionen zum Thema, sowie für wertvolle Unterstützung bei der Erstellung dieses Artikels an Sabine Brants, Jeff Dean, Peter Dienes, Alex Franz, John Hawkins, Franz Och, Jay Ponte, Shankar Kumar, Ashok Popat, Ioannis Tsochantaridis und Peng Xu.

Literatur

- Brown, P./Cocke, J./Della Pietra, S./Della Pietra, V./Jelinek, F./Lafferty, J./Mercer, R./Roosin, P. (1993): A Statistical Approach to Machine Translation. In: *Computational Linguistics*, 16(2), S. 263–313.
- Chen, S. F./Goodman, J. T. (1998): *An Empirical Study of Smoothing Techniques for Language Modeling*. Technischer Report TR-10-98, Computer Science Group. Harvard University.
- Dean, J./Ghemawat, S. (2004): MapReduce: Simplified Data Processing on Large Clusters. Symposium on Operating System Design and Implementation. San Francisco, CA. USA.
- Ghemawat, S./Gobioff, H./Leung, S.-T. (2003): The Google File System. Symposium on Operating System Principles. Lake George, NY. USA.
- Jelinek, Frederick/Mercer, Robert L. (1980): Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*. S. 381–397. Amsterdam, The Netherlands.
- Koehn, P./Och, F. J./Marcu, D. (2003): Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference*. Edmonton, Canada.
- Markov, A. (1913): Russian [An example of statistical investigation in the text of Eugene Onyegin illustrating coupling of ‘tests’ in chains]. In: *Proceedings of the Academy of Sciences*. St. Petersburg, 7(VI), S. 153–162.
- Och, F. J./Tillmann, C./Ney, H. (1999): Improved alignment models for statistical machine translation. In: *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. S. 20–28.
- Papineni, K./Roukos, S./Ward, T./Zhu, W.-J. (2002): BLEU: a method for automatic evaluation of machine translation. In: *40th Annual Meeting of the Association for Computational Linguistics*. S. 311–318. Philadelphia, PA. USA.

MICHAEL STRUBE/MARGOT MIESKES/CHRISTOPH MÜLLER

Gesprächsprotokolle auf Knopfdruck: Die automatische Zusammenfassung von gesprochenen Dialogen

Abstract

Dieser Beitrag beschreibt computerlinguistische Arbeiten zur automatischen Zusammenfassung gesprochener Dialoge. Der Beitrag geht sowohl auf die notwendige Vorverarbeitung als auch auf die eigentliche Zusammenfassung durch automatische Erkennung von Themengrenzen und Extraktion relevanter Äußerungen ein. Ein weiterer Schwerpunkt liegt in der Beschreibung von Arbeiten zur automatischen Anaphernresolution in gesprochener Sprache. Der Beitrag betont vor allem die Rolle und Bedeutung von annotierten Korpora für die computerlinguistische Forschung und Entwicklung.

1. Einleitung

Täglich finden unzählige Besprechungen, Verhandlungen und Diskussionen statt. Oft schreibt einer der Teilnehmer fleißig Notizen, um später ein Protokoll anzufertigen – eine meist ungeliebte Aufgabe. Im von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt DIANA-Summ (*Dialog ANAphors and Summarization*) soll ein Computersystem entwickelt werden, das gesprochene Dialoge automatisch zusammenfasst. Langfristiges Ziel des Projektes ist es, die automatische Erzeugung von Gesprächsprotokollen zu ermöglichen. Das Verfahren, das dabei zur Anwendung kommt, basiert darauf, für jedes im Gespräch behandelte Thema (*topic*) relevante Äußerungen zu extrahieren und zu einer Zusammenfassung zu kombinieren.

Spontane gesprochene Sprache unterscheidet sich in mehrfacher Hinsicht von geplanter, geschriebener Sprache. Zum einen enthält sie viele abgebrochene Äußerungen, Wiederholungen, Selbstkorrekturen und Pausen, wobei letztere häufig mit Platzhaltern wie 'Ähh' gefüllt sind. Sind mehrere Sprecher beteiligt, kommt als weitere Komplikation hinzu, dass diese auch gleichzeitig sprechen können. Dies alles macht es erforderlich, dass vor der Erstellung einer Zusammenfassung eine Reihe von Vorverarbeitungsschritten durchgeführt wird, die diese Phänomene erkennen und ggf. entfernen. Eine weitere Besonderheit gesprochener Sprache ist der erheblich höhere Anteil an Pronomen wie (im Deutschen) *es* und *das* oder (im Englischen) *it* und *this* bzw. *that*. Eine der Hypothesen des Projektes ist, dass die Auflösung dieser Pronomen

(d. h. die Bestimmung ihrer jeweiligen Antezedenten) sowohl die Identifikation von relevanten Äußerungen erleichtern als auch die Verständlichkeit der so erstellten Zusammenfassungen erhöhen sollte.

Die folgenden Kapitel geben einen Überblick über das Projekt DIANA-Summ (Stand April 2006). Nach einer Beschreibung der verwendeten Datengrundlage (Kap. 2) wird zunächst auf methodische Aspekte des Projektes eingegangen (Kap. 3). Dabei steht insbesondere die Rolle des Korpus als empirische Grundlage für Systementwicklung und Evaluierung im Mittelpunkt. Außerdem wird das Konzept der Korpusannotation eingeführt. Datengewinnung durch Korpusannotation sowie die Sicherung der Qualität der so gewonnenen Daten werden ebenfalls erläutert. Im Anschluss daran werden die einzelnen Schritte der Vorverarbeitungskette (Kap. 4) beschrieben. Die drei letzten Kapitel gehen dann detaillierter auf die Teilaufgaben Anaphernresolution (Kap. 5), Erkennung von Themengrenzen (Kap. 6) und Zusammenfassung (Kap. 7) ein.

2. Datengrundlage

Im Rahmen des DIANA-Summ-Projektes soll keine eigene Sammlung von Rohdaten durchgeführt werden, da derartige Datenerhebungen mit erheblichem technischen und organisatorischen Aufwand verbunden, für die eigentliche Zielsetzung des Projektes aber nicht von besonderem Interesse sind. Als Datengrundlage wird deshalb das englischsprachige *ICSI Meeting Corpus* (Janin et al., 2003) verwendet. Bei diesem Korpus handelt es sich um Aufnahmen von 75 Diskussionen (gesamt ca. 72 Stunden). Jede Diskussion hat durchschnittlich 6 Teilnehmer, wobei nicht alle Teilnehmer Englische Muttersprachler sind. Themen umfassen u. a. Computerlinguistik, Sprachtheorie, Spracherkennung und Netzwerktechnologie. Zusätzlich zu den eigentlichen Sprecheräußerungen enthält das Korpus auch explizite Markierungen von abgebrochenen Äußerungen bzw. Wörtern (sog. *disfluencies*, dt. Häitationen), sowie Meta-Informationen über nicht-sprachliche Ereignisse (Atemgeräusche, Rauschen etc.). In struktureller Hinsicht besteht das Korpus aus einer Reihe von sog. Segmenten. Jedes Segment ist mit genau einem Sprecher assoziiert, wobei längere Äußerungen desselben Sprechers auch auf mehrere Segmente verteilt sein können. Vergleiche das folgende Beispiel (Nummerierung nicht im Original).

1. FE016 uh, echo cancellation to get basically back down to the individual
2. FE016 speaker utterances would be probably all that we would need to be able to do good recognition on the – on the close-talking mikes.
3. ME011 Um, why don't you, if you have a hard copy, why don't you email it
4. MN014 So these –
5. FE016 So, that's about the summary –
6. ME011 to the list.
7. FE016 But this is – Morgan has this paper. I mean he – he – it – it's that paper. Yeah, yeah.

8. MN014 Yeah, yeah.
9. ME013 Yeah, so it's the same thing? It's the same thing I mailed to every –
10. ME011 Oh it's in the paper.
11. ME011 O_K.
12. ME013 everybody that w- where it was, yeah.
13. FE016 So, we basically, um,
14. ME011 O_K then, it's already been mailed.

Es fällt sofort der Anteil an überlappender Sprache auf, der sich in den Daten in Form von diskontinuierlichen Segmenten (z. B. 3–6 bzw. 9–12) manifestiert. Auch ist erkennbar, dass Segmentgrenzen nicht unbedingt mit linguistisch motivierten Phrasen- oder Satzgrenzen zusammenfallen müssen (z. B. 1–2). Unterbrechungen bzw. abgebrochene Äußerungen sind durch Bindestriche angezeigt (z. B. 7 und 9).

Das *ICSI Meeting Corpus* liegt vollständig manuell transkribiert und im XML-Format vor. Da automatische Spracherkenner spontane Diskussionen mit mehreren Teilnehmern noch nicht in befriedigender Qualität verarbeiten können, wird für DIANA-Summ eine idealisierte Eingabe in Form der manuell erstellten Transkriptionen angenommen.

3. Methoden

In der Korpuslinguistik werden Textkorpora schon lange als maschinenlesbare Sammlungen natürlicher Sprache verwendet. Sie dienen dort als empirische Grundlage für Häufigkeits-, Kollokations- oder andere Analysen. Im Gegensatz dazu geht die Verwendung von Korpora in der Computerlinguistik über die reine Deskription hinaus, denn ihre Aufgabe ist die Entwicklung von Computersystemen zur automatischen Text- oder Sprachverarbeitung. Im Projekt DIANA-Summ sind dies in erster Linie die automatische extraktive Zusammenfassung und die Anaphernresolution. Ein weit verbreiteter methodischer Ansatz für die Entwicklung derartiger Systeme, der auch im Rahmen von DIANA-Summ Anwendung findet, ist der Einsatz von sog. überwachtem maschinellen Lernen (*supervised machine learning* (ML)) (Mitchell, 1997). Bei dieser Methode stellen annotierte Korpora eine unverzichtbare empirische Grundlage dar. Die Korpora werden mit Hilfe spezieller ML-Programme analysiert. Die Programme erzeugen statistische Modelle der in den Korpora vorgefundenen Phänomene, welche dann auf neue Korpora angewendet werden können. Die ML-Programme können allerdings nur solche Phänomene des analysierten Korpus in ihrem Modell berücksichtigen, welche explizit in diesem enthalten sind. Um z. B. ein Modell für die automatische Erkennung von Häitationen zu erzeugen (zu *trainieren*), muss das analysierte Trainingskorpus für jedes Wort Informationen darüber enthalten, ob es Teil einer Häitation ist. Das gleiche gilt für das Trainieren von Modellen zur Anaphernresolution: Jeder anaphorische Ausdruck im Trainingskorpus muss mit seinem korrekten Antezedenten verbunden sein, damit ein Modell trainiert werden

kann, das für jede Anapher den *wahrscheinlichsten* Antezedenten identifizieren kann.

Der Vorgang und das Ergebnis der manuellen Hinzufügung dieser Art von Information zu einem existierenden Textkorpus wird *Annotation* genannt. Außer zum Training von statistischen Modellen für die automatische Text- oder Sprachverarbeitung werden annotierte Korpora auch zu deren Evaluierung benötigt. Dabei wird die Leistungsfähigkeit eines Systems bestimmt, indem seine (potentiell falsche) Ausgabe mit der zuvor manuell erstellten, korrekten Annotation im Korpus verglichen wird. Bei der Evaluierung, welche ein wichtiger Schritt in der Entwicklung computerlinguistischer Systeme ist, kommen verschiedene quantitative Maße zum Einsatz, darunter v. a. Präzision (*precision*) und Vollständigkeit (*recall*). Präzision wird dabei berechnet aus „Anzahl der korrekten Antworten des Systems“ geteilt durch „Anzahl aller Antworten des Systems“. Die Vollständigkeit berechnet sich aus „Anzahl der korrekten Antworten des Systems“ geteilt durch „Anzahl der möglichen korrekten Antworten“. Das F-Maß (*f-measure*) setzt Präzision und Vollständigkeit ins Verhältnis zueinander.

Linguistische Annotation ist eine mühsame und zeitraubende Aufgabe. Da sie größtenteils manuell durchgeführt wird, ist sie auch mit erheblichen Kosten verbunden. Die Annotierer verwenden in der Regel spezialisierte Annotations-Tools (wie z. B. MMAX2 (<http://mmax.eml-research.de>)). Je nach der Schwierigkeit der Aufgabe (d. h. des zu annotierenden Phänomens) kann die manuelle Annotation erhebliche Interpretation von seiten der Annotierer erfordern, so dass verschiedene Annotierer selten völlig identische Annotationen produzieren. Der Grad der Übereinstimmung zwischen zwei (oder mehr) Annotierern für dieselbe Aufgabe (*inter-annotator agreement*) ist daher ein wichtiger Indikator für die Schwierigkeit bzw. die grundsätzliche Lösbarkeit der Annotationsaufgabe. Die Übereinstimmung wird in der Regel in Form der Kappa-Statistik berechnet (Carletta, 1996). Das Besondere an diesem Maß ist, dass es berücksichtigt, dass ein bestimmter Grad an Übereinstimmung auch durch Zufall zustande kommen kann. In der Regel bewegt sich Kappa zwischen -1 und 1 . Ein niedriger Kappa-Wert für eine bestimmte Annotationsaufgabe kann ein Hinweis darauf sein, dass das zu annotierende Phänomen an sich unscharf, ambig oder vage ist und keine eindeutige Lösung besitzt. Im Allgemeinen wird ein Kappa-Wert von $0,67$ als gerade noch akzeptabel angesehen. Liegt der Wert darunter, dann sind die durch die Annotation gewonnenen Daten nur von sehr eingeschränkter Nützlichkeit. Angesichts der zentralen Bedeutung der manuellen Annotation im Projekt DIANA-Summ ist es daher notwendig, die Datenqualität sowohl für die Vorverarbeitung (Kap. 4) als auch für die eigentlichen Aufgaben (Kap. 5, 6 und 7) durch Kontrolle des Kappa-Wertes zu sichern.

4. Vorverarbeitungsschritte

Um die transkribierten Dialoge zusammenfassen zu können, sind mehrere Vorverarbeitungsschritte nötig. Einige davon werden wir im Folgenden näher betrachten. Eine schematische Darstellung der Vorverarbeitung findet sich in Abbildung 1, wobei zu bemerken ist, dass wir nicht alle Vorverarbeitungsschritte selbst durchführen werden. Die so genannten Dialogakte etwa, die jede Aussage mit Information bezüglich der Intention versehen (Bitte, Frage, Aussage etc), wurden uns von einer anderen Forschungsgruppe zur Verfügung gestellt.

Abb. 1

4.1 Part-of-Speech Tagging

Part-of-Speech (POS) Tagging bezeichnet den Vorgang der Anreicherung/Annotation der Daten mit Wortarten. Da es ausgereifte Programme gibt, die die Annotation mit Wortarten automatisch vornehmen, setzten wir solche ein, die verschiedene Ansätze abdecken: den *Brill Tagger* (TBL; Brill 1994), *TnT* (Brants, 2000), zwei *Tagger* aus der *Stanford NLP Library* (*Left3* und *Bidirect*) (Toutanova et al., 2003, Toutanova und Manning, 2000). Diese Systeme sind

auf den *Wall Street Journal* Daten der *Penn Treebank* (Marcus et al., 1993) trainiert, also auf Zeitungstexten, die sich stark von gesprochener Sprache unterscheiden. Durch einen Mehrheitsentscheid (*Maj*) über allen vier Systemen hofften wir, die ursprüngliche Fehlerrate zu senken, die sowohl durch die Anwendung auf Dialoge als auch durch die unterschiedlichen Herangehensweisen der verschiedenen *Tagger* entstehen.

Dieser Mehrheitsentscheid wurde Annotierern als Grundlage für eine manuelle Korrektur der automatischen Annotation übergeben. Dies diente zum einen zur Evaluierung der ursprünglichen Annotation, zum anderen aufgrund der subjektiv relativ schlechten Ergebnisse der automatischen *Tagger* für deren Neutrainieren. In zwölf der 75 *Meetings* wurden die *POS-Tags* manuell von allen drei Annotierern korrigiert. In 25 weiteren wurden die *POS-Tags* von einzelnen Annotierern korrigiert. Die Übereinstimmung zwischen den Annotierern wurde mit Hilfe der Kappa-Statistik bestimmt und ist mit $k = 0,96$ sehr hoch. Dies zeigt, dass die Aufgabe der manuellen Korrektur schon vorgefertigter *POS-Tags* sehr zuverlässig möglich ist. Zudem ist dieses Vorgehen deutlich schneller, als eine rein manuelle Annotation.

Aus den manuell korrigierten Daten wurden zwei Testkorpora gewählt. Test1 besteht aus einem Teil der zwölf *Meetings*, die von allen drei Annotierern korrigiert wurden, Test2 besteht aus *Meetings*, die nur von einzelnen Annotierern korrigiert wurden.

Nach der manuellen Annotation waren wir in der Lage, die ursprüngliche Annotation auszuwerten. Die Ergebnisse sind in Tabelle 1 dargestellt.

Data	TBL	TnT	Left3	Bidirect	Maj
Test1	11,3	11,2	11,1	11,1	10,5
Test2	13,8	14,3	13,6	13,2	13,4

Tab. 1

Da diese Ergebnisse weit von den Ergebnissen entfernt sind, die in der Literatur für die *Tagger* angegeben werden (2,5 % bis 3 % Fehlerrate) wurden die manuell korrigierten *POS-Tags* dazu genutzt, die vier automatischen Systeme neu zu trainieren. Dazu wurden die vorhandenen Daten in 6 verschiedene Teilmengen unterteilt, die immer mehr Daten enthielten. Die Ergebnisse des Neutrainierens sind in Tabelle 2 dargestellt, ebenso wie die jeweilige Datenmenge.

Das Neutrainieren bewirkt eine deutliche Verbesserung. Alle vier *Tagger* erreichen Fehlerraten von 3,5 %–4 %, wobei insbesondere bei *TBL* vom ersten zum sechsten Schritt eine deutliche Verbesserung zu sehen ist. Im Durchschnitt liegt die Verbesserung über die Teilmengen bei ca. 1 %–1,5 %, wobei bei allen *Taggern* der Gewinn durch die Erhöhung der Trainingsmenge nach dem vierten oder fünften Schritt nur noch gering ausfällt.

Set Wörter	Set1 124K	Set2 162K	Set3 197K	Set4 221K	Set5 253K	Set6 283K
TnT						
Test1	3,4	3,4	3,3	3,3	3,4	3,4
Test2	5,4	5,1	4,9	4,6	4,5	4,5
TBL						
Test1	3,9	3,5	3,5	3,5	3,6	3,5
Test2	8,4	5,5	5,0	4,7	4,4	4,4
Left3						
Test1	3,2	3,0	3,0	3,2	3,2	3,2
Test2	5,2	4,7	4,5	4,3	4,2	4,1
Bidirect						
Test1	3,2	3,0	3,0	3,2	3,2	3,2
Test2	5,2	4,7	4,5	4,3	4,2	4,1

Tab. 2

Auch mit den neutrainierten Systemen wurde ein Mehrheitsentscheid durchgeführt, allerdings wurde der zweite Stanford-Tagger (*Bidirect*) aufgrund der Ähnlichkeit mit *Left3* nicht weiter berücksichtigt. Das Ergebnis des Mehrheitsentscheids ist in Tabelle 3 illustriert:

	Setup1	Setup2	Setup3	Setup4	Setup5	Setup6
Test1	3,0	2,9	2,9	3,0	3,0	2,9
Test2	5,1	4,7	4,4	4,1	4,0	3,9

Tab. 3

Auch hier ist eine deutliche Verbesserung im Vergleich zu den ursprünglichen Annotationen feststellbar. Allerdings ist die Verbesserung nur bis zum vierten Schritt deutlich und schwächt sich danach ab. Man kann daraus schließen, daß der erhöhte Aufwand durch das manuelle Annotieren ab dem vierten bis fünften Schritt nicht mehr zu einer deutlichen Verbesserung der Ergebnisse führt.

Der einfache Mehrheitsentscheid bietet also eine computationell unaufwendige Methode, um die Ergebnisse weiter zu verbessern. Damit liegt die Fehlerrate beim *Taggen* von transkribierter, gesprochener Sprache ähnlich niedrig wie die von geschriebener Sprache. Eine detailliertere Ausführung zu diesem Thema findet sich in Mieskes und Strube (2006).

4.2 Häitationen und Unterbrechungen

Häitationen, Unterbrechungen und Selbstkorrekturen sind ein Phänomen der gesprochenen Sprache. In geschriebenen Texten kommen sie nicht vor und nur äußerst selten bei gelesenen Texten wie beispielsweise Nachrichtentexten. Dabei unterscheidet man vier Hauptformen: Pausen (nichtlexikalische (NLFP) und lexikalische (LFP)), Wiederholungen (*repet*), Fehlstarts (abgebrochene Wörter (*abw*) und abgebrochene Äußerungen (*abutt*)) und Selbstkorrekturen (*repai*). Hierbei werden nur sogenannte „gefüllte Pausen“ berücksichtigt, da die leeren Pausen (Stille) nur schwer von anderen Pausen (Sprecherwechsel etc.) zu unterscheiden sind. Beispiele für die genannten Häitationen sind:

- NLFP: FE016 **uh**, echo cancellation to get basically back down to the individual
- LFP: FE016 ... **I mean** he – he – it – it's that paper.
- repet: FE016 ... I mean **he – he** – it – it's that paper.
- repai: FE016 **But this is** – Morgan has this paper.
- abw: ME013 everybody that **w-** where it was, yeah.
- abutt: FE016 **So, that's about the summary** –

Auch dieses Phänomen wurde in den Daten manuell annotiert. Werden nur Elemente berücksichtigt, die von beiden Annotierern markiert wurden, ist die Übereinstimmung $k = 0,95$. Bei den einzelnen Kategorien sind die Unterschiede sehr deutlich, wie Tabelle 4 zeigt, wobei die zweite Spalte (inkl. Elemente) alle markierten Elemente berücksichtigt und die dritte Spalte (gem. Elemente) nur jene, die von beiden Annotierern markiert wurden.

Art	inkl. Elemente	gem. Elemente
NLFP	0,52	100,0
LFP	0,47	0,98
Repet	0,63	0,98
Repai	0,24	0,78
Abw	0,54	0,96
Abut	0,28	0,85

Tab. 4

Wie man aus Tabelle 4 ersehen kann, sind besonders Reparaturen, Wiederholungen, aber auch lexikalische Pausen problematisch. Aus den Daten, die von beiden Annotierern markiert wurden, kann man jedoch erkennen, dass es eine grosse Übereinstimmung gibt. Dies ist gut genug, um eine automatische Erkennung von Häitationen zu entwickeln. Eine erste Mustererkennung, die

auf einfachen Buchstaben- und Wortabgleichungen basiert, erlaubt es, die nichtlexikalischen Pausen und wortwörtliche Wiederholungen zu bestimmen. Für die komplexeren Phänomene wurden Methoden des maschinellen Lernens eingesetzt. Die besten Ergebnisse mit mehr *Features*, als nur den drei vorangehenden *POS-Tags* sind in Tabelle 5 aufgeführt.

	total	richtig	falsch	Präz.	Vollst.	F-Maß
NLFP	8110	7404	2506	74,71	91,92	82,18
LFP	8489	6881	1737	79,84	81,06	80,45
abutt	4568	2198	1541	58,70	47,94	52,78
abw	2531	2275	989	69,70	89,89	78,52
repai	6620	3949	2494	61,29	59,65	60,46
repet	5156	3677	1347	73,19	71,19	72,18

Tab. 5

4.3 Lemmatisierung

Auf Basis der *POS*-Annotation lässt sich ein weiterer Vorverarbeitungsschritt automatisieren, der für die Zusammenfassung relevant sein wird, die sogenannte Lemmatisierung. Hierbei werden die Grundformen der Wörter ermittelt. Beispielsweise sind *is* und *was* beide Formen des Verbs *be*. *houses* ist die Pluralform von *house*. *be* und *house* sind jeweils die Lemmata von *is* und *was*, resp. *houses*. Für die Bestimmung der Lemmata verwendeten wir den *Tree-Tagger* (Schmid, 1997).

5. Anaphernerkennung und -resolution

In der Anaphernresolution im Projekt DIANA-Summ wird ein besonderes Augenmerk auf das Personalpronomen *it* und die Demonstrativpronomina *this* und *that* gelegt. Die Präferenz für diese Pronomen liegt neben ihrer Häufigkeit auch in der Anwendung begründet, in welche die Anaphernresolution eingebettet und anhand welcher sie zu evaluieren ist. Im Kontext einer extraktiven Zusammenfassung ist es nahe liegend, einzelne Äußerungen durch Resolution der in ihnen enthaltenen Pronomen und durch Ersetzung derselben durch ihre lexikalischen Antezedenten semantisch explizit und damit von ihrem Kontext unabhängig zu machen. Dies hat positive Auswirkungen sowohl auf die Extraktion als auch auf die Lesbarkeit der erzeugten Zusammenfassung.

Nicht alle Vorkommen von *it*, *this* und *that* sind indes resolvierbare Anaphern. Vor der eigentlichen Anaphernresolution muss daher eine Desambiguierung bzw. Vorauswahl erfolgen. Für *this* und *that* z. B. müssen pronominale Verwendungen von solchen unterschieden werden, in denen die Aus-

drücke als Determinierer einer NP fungieren. *That* kann außerdem auch Relativpronomen oder Komplementierer (entsprechend Dt. *dass*) sein. Die automatische Unterscheidung all dieser Fälle wird durch die in Kapitel 4 beschriebene Wortartenklassifizierung geleistet. Die Identifikation der nicht-referentiellen und somit nicht resolvierbaren Fälle des Pronomens *it* ist hingegen ungleich schwieriger. Der folgende Abschnitt gibt lediglich einen Überblick. Eine detailliertere Darstellung (inklusive Beispielen) findet sich in Müller (2006).

5.1 Identifikation von nicht-referentiellem *it*

5.1.1 Typen von *it*

In Müller (2006) werden fünf Typen von *it* unterschieden, welche in morpho-syntaktischer Hinsicht alle als Pronomen zu klassifizieren sind. Allerdings sind nur zwei der fünf Typen referentielle Pronomen, und lediglich ein Typ stellt eine resolvierbare Anapher dar.

normal:	Anaphorisches Pronomen mit identifizierbarem Antezedenten (referentiell, resolvierbar)
vage:	Anaphorisches Pronomen ohne identifizierbaren Antezedenten (referentiell, nicht resolvierbar)
verworfen:	Pronomen in abgebrochener Äußerung (nicht-referentiell, nicht resolvierbar)
Extraposition:	Pronomen als syntaktischer Platzhalter für nach rechts verschobenes Satzglied (nicht-referentiell, nicht resolvierbar)
Prop-it	(Quirk Pronomen als semantisch leere Besetzung der obligatorischen et al, 1991): Subjekt-Position (nicht-referentiell, nicht resolvierbar)

Als Vorstufe der Resolution des Pronomens *it* ist es daher notwendig, nicht-referentielle Fälle automatisch zu identifizieren und auszufiltern.

5.1.2 Manuelle Annotation von *it*

Zwei Annotierer erhielten die Aufgabe, unabhängig voneinander alle Vorkommen von *it* in einen Sub-Korpus von fünf Dialogen als einen der fünf Typen oder als ‚Sonstiges‘ zu klassifizieren. Die Annotation wurde von zwei Annotierern redundant ausgeführt, um die Übereinstimmung in Form von Kappa ermitteln zu können. Die Resultate dieser ersten Annotation finden sich in Tabelle 6.

Die Kappa-Werte für alle Kategorien mit Ausnahme von ‚verworfen‘ liegen teilweise deutlich unter dem geforderten Wert von 0,67 (vgl. Kap. 3). Ein erstes wichtiges Resultat von Müller (2006) ist daher, dass die differenzierte (d. h. fünffache) Klassifikation der verschiedenen Typen von *it* nicht unproblematisch zu sein scheint, da sie häufig nur mit einer Genauigkeit erfolgt, die an Zufall grenzt. Durch Zusammenlegung der Kategorien ‚normal‘ und

‚vage‘ zu ‚referentiell‘ auf der einen und ‚verworfen‘, ‚Extraposition‘ und ‚Prop-it‘ zu ‚nicht-referentiell‘ auf der anderen Seite kann die Übereinstimmung jedoch deutlich verbessert werden. Die Kappa-Werte für die resultierenden Super-Kategorien ‚referentiell‘ bzw. ‚nicht-referentiell‘ betragen 0,72 bzw. 0,74, so dass die Verwendbarkeit der erhobenen Daten für die weniger differenzierte zweifache Klassifikation gewährleistet ist.

	normal	vage	verwrf.	Extrap.	Prop-it	Sonst.	Ges.
Bed017	0,65	0,33	0,94	0,27	0,54	0,42	0,62
Bmr001	0,69	0,21	0,92	0,48	0,33	– 0,01	0,63
Bns003	0,59	0,18	0,75	0,55	0,21	0,32	0,55
Bro004	0,65	– 0,05	0,86	0,75	0,59	– 0,01	0,65
Bro005	0,57	– 0,03	0,84	0,58	0,36	0,23	0,58
Gesamt	0,64	0,11	0,86	0,56	0,43	0,2	0,61

Tab. 6

Als letzten Schritt der Annotation hatten die Annotierer die Aufgabe, aus ihren teilweise divergierenden Annotationen durch Diskussion von strittigen Fällen eine Referenz-Annotation zu erstellen, welche die Datengrundlage für die im folgenden beschriebene automatische Klassifikation darstellt.

5.1.3 Automatische Klassifikation von *it*

Ziel der automatischen Klassifikation ist es, mit Hilfe von Methoden des maschinellen Lernens (ML) einzelne Vorkommen von *it* als entweder ‚referentiell‘ oder ‚nicht-referentiell‘ zu erkennen. Jedes Vorkommen von *it* wird dafür in Form einer Liste von Merkmalen (*feature vector*) repräsentiert. Diese Merkmale beschreiben den Kontext (im weitesten Sinne), in welchem das *it* vorkommt. Dies sind z. B. syntaktische Konstruktion, regierendes Verb, Distanz (in Wörtern) zu bestimmten grammatischen Funktionswörtern, etc. Jede Merkmalsliste enthält außerdem ein Zielkonzept, d. h. sie ist als ‚referentiell‘ oder ‚nicht-referentiell‘ markiert.

Ein Teil der so gewonnenen Daten wird dann als Trainingsdaten für die ML-Programme genutzt, welche sie analysieren und ein statistisches Modell erzeugen. Dieses Modell kann dann verwendet werden, um neue, nicht im Training verwendete Daten automatisch zu klassifizieren. Die Korrektheit einer jeden Klassifikation kann überprüft werden, indem die automatisch vergebene Klasse mit der während der manuellen Annotation vergebenen verglichen wird. Das beste im Rahmen von Müller (2006) erzielte Ergebnis belief sich auf

Präzision = 80,0%, Vollständigkeit = 60,9% und F-Maß = 69,2%.

Dieses Ergebnis, insbesondere die hohe Präzision, erlaubt es, die hier entwickelte Komponente zur Identifikation von nicht-referentiellem *it* als Filter für die Anaphernresolution einzusetzen.

5.2 Resolution von anaphorischen Pronomen

Während die automatische Anaphernresolution für geschriebene Texte (v. a. Zeitungstexte) bereits eine etablierte Teildisziplin innerhalb der Computerlinguistik ist, steht diese Aufgabe im Bereich der (transkribierten) gesprochenen Sprache noch ziemlich am Anfang. Einer der Hauptgründe dafür ist sicherlich die weitaus größere Schwierigkeit der Aufgabe. Spontane gesprochene Umgangssprache weist einen weitaus höheren Anteil an vagen oder ambigen Referenzausdrücken auf als geschriebener Text. Die in Kapitel 5.1.2 erwähnte geringe Übereinstimmung zwischen Annotierern bei der Klassifikation von normalen und vagen Fällen von *it* kann als Beleg dafür gelten. Im Vergleich zu dieser Aufgabe ist allerdings die Identifikation des korrekten Antezedenten eines Pronomen nochmals erheblich schwieriger. Eine Kontrolle der Annotierer-Übereinstimmung ist demnach auch hier vonnöten.

Die im Rahmen von DIANA-Summ durchzuführende manuelle Koreferenzannotation ist zum aktuellen Zeitpunkt (April 2006) noch nicht abge-

Abb. 2

schlossen, so dass noch keine Zahlen für die Bewertung der Übereinstimmung vorliegen. Als erster Eindruck kann jedoch auch ein qualitativer Vergleich der von den verschiedenen Annotierern gebildeten Koreferenzklassen dienen. Koreferenzklassen kann man am besten in Form von Koreferenzketten visualisieren, indem man alle Elemente einer Klasse in der Reihenfolge ihres Erscheinens im Text verbindet. Diese Form der Visualisierung wird auch vom Annotationstool MMAX2 verwendet. Der *Screenshot* in Abbildung 2 zeigt die graphische Oberfläche des Tools mit einem selektierten Demonstrativpronomen (*this*) an der Cursorposition (Pfeil).

Die unterschiedlich gefärbten Linien (hier in verschiedenen Grautönen wiedergegeben) visualisieren die vier verschiedenen Koreferenzketten, denen das selektierte Pronomen von vier Annotierern unabhängig voneinander zugeordnet wurde. Während dies sicherlich ein Extrembeispiel darstellt, gibt es doch einen guten Eindruck von der Schwierigkeit der Koreferenzannotation in gesprochenen Dialogen.

6. Themengrenzen

Ein wichtiger Schritt für die Zusammenfassung ist die Abgrenzung der verschiedenen Themen, die im Verlauf der Gespräche auftauchen. Da die Gespräche sehr lang sind, behandeln sie häufig mehr als nur ein Thema. In manuellen Protokollen ist es ebenfalls üblich, das Protokoll in thematische Abschnitte zu unterteilen.

In der Vergangenheit wurden verschiedene Methoden zur automatischen Bestimmung der Themengrenzen entwickelt, denen häufig Wortfrequenzen oder statistische Modelle zu Grunde liegen. Neuere Methoden basieren auf lexikalischen Ketten, die zusammengehörende Wörter verbinden. Bisher werden Wörter als zusammengehörend betrachtet, wenn ihnen gleiche Lemmata (vgl. Kapitel 4.3) zu Grunde liegen (beispielsweise *house* und *houses*). Werden lexikalische Ketten über mehrere Sätze hinweg nicht fortgesetzt, so ist das ein Indiz für eine Themengrenze (Galley, 2003).

Auch für die Auswertung automatischer Themengrenzen sind manuelle Annotationen nötig. Da es sich nicht um eine Klassifikations- sondern um eine Segmentierungsaufgabe handelt, müssen die Evaluierungsmethoden berücksichtigen, dass die Grenzen nicht unbedingt exakt übereinstimmen müssen. Evaluierungsmethoden, die diesem Phänomen Rechnung tragen, sind: *Pk* (Beeferman et al., 1999) und *WindowDiff* (Pevzner und Hearst, 2002). Beide Maße geben an, wie groß die Wahrscheinlichkeit ist, eine Grenze falsch zu ziehen. Kleine Zahlen stehen für gute Ergebnisse.

Bei der manuellen Annotation gibt es mehrere Schritte. Zunächst werden nur vorläufige Grenzen gezogen. Diese werden in einem zweiten Schritt korrigiert und zum Teil wieder entfernt. In einem dritten Schritt werden die Grenzen unterschieden in Haupt- und Unterthemen bzw. „*Small Talk*“. *Small Talk* ist für die spätere Zusammenfassung irrelevant und sollte deshalb gefiltert

werden. Bei den hier vorliegenden *Meetings* gibt es viele, in denen es für die menschlichen Annotierer schwierig ist, eindeutige Themengrenzen zu ziehen. Die Ergebnisse liegen je nach Dialog für die manuellen Annotierer zwischen 20,0 und 7,0 (Angabe nach Pk-Maß).

Das erfolgreichste System (im Folgenden *g03* aus Galley et al. (2003)) arbeitet mit lexikalischen Ketten, die auf reiner Wortgleichheit basieren. Dieses System wurde zusammen mit üblichen Baselines getestet. Die Ergebnisse finden sich in Tabelle 7.

	random	even	all	g03
Pk	48,9	55,1	82,3	20,4
WD	53,5	61,3	99,9	24,2

Die drei *Baselines* sind *Random*, wobei eine bestimmte Anzahl an Grenzen (in diesem Fall 25) zufällig im Text gesetzt werden, *Even*, wobei diese 25 Grenzen gleichmäßig über den Text verteilt werden, und *All*, wobei eine Grenze nach jedem Segment eingefügt wird.

In einem nächsten Schritt werden Erweiterungen, in denen wir Wortähnlichkeiten mit Hilfe von *WordNet* testen, eingebaut. Das Ziel dieser Erweiterung ist es, die lexikalischen Ketten nicht nur auf Wortgleichheit aufzubauen, sondern auch auf Beziehungen zwischen Wörtern. Im ursprünglichen System wurden lexikalische Ketten aufgebaut, indem identische Wörter verbunden wurden. Beispielsweise werden im ursprünglichen System nur *house* und *houses* verbunden. Werden Wortähnlichkeiten verwendet, werden auch Wörter wie *houses* und *flat* berücksichtigt. Dies wollen wir für die automatische Themenabgrenzung nutzen, um so die vorhandenen Systeme zu verbessern.

7. Automatische Zusammenfassung

Im Verlauf der letzten 50 Jahre wurden verschiedene Methoden für verschiedene Text- und Zusammenfassungsarten entwickelt. Bei den Methoden gibt es zwei Möglichkeiten: Zusammenfassung durch Extraktion und Zusammenfassung durch Abstraktion. Bei der Extraktion werden Elemente des Textes herausgegriffen und wortwörtlich wiedergegeben. Bei der Abstraktion wird die Zusammenfassung aus Elementen, die nicht im Text vorkommen müssen, gebildet.

Bei den Zusammenfassungsarten gibt es drei prinzipielle Möglichkeiten: eine indikative Zusammenfassung, die die wichtigsten Ideen und Gedanken des Textes wiedergibt, allerdings ohne in die Details zu gehen. Eine themenorientierte Zusammenfassung orientiert sich an den Themen in den Texten, kann sich jedoch auch nur auf ein Thema aus dem Text konzentrieren. Generierte Zusammenfassungen geben auch Standpunkte des Autors wieder.

Textarten, die in der Vergangenheit im Hinblick auf Zusammenfassung untersucht wurden, sind Einzeltexte aus den Bereichen Wissenschaft, Nachrichten aber auch Patentschriften. Auch die Zusammenfassung mehrerer Texte aus einem Themenbereich wurde untersucht. In den letzten Jahren kamen auch Zusammenfassungen von gesprochener Sprache dazu, die sich primär an vorgelesenen Nachrichten bzw. Nachrichtendialogen orientierten.

Wir werden im Folgenden nur auf die Methoden der extraktiven Zusammenfassung eingehen. In frühen Arbeiten wurden satzbasierte Merkmale für die Extraktion verwendet. Diese Merkmale dienten dazu, die zusammenfassungsrelevanten Informationen eines Textes zu finden. Diese Merkmale waren unter anderem die Position des Merkmals im Satz, Schlüsselworte, aber auch häufig wiederholte Wörter und Phrasen. Seit einigen Jahren werden maschinelle Lernverfahren genutzt, um diese Informationen zu finden. Dazu wurde vermehrt Textanalyse eingesetzt, um die Schlüsselworte und -passagen zu finden. Zu dieser Analyse gehören Wortbeziehungen und -ähnlichkeiten, aber auch Diskursstruktur und Anaphernanalyse. Für einen ausführlichen Überblick über die bisherigen Ansätze siehe Radev et al. (2002), Mani und Maybury (1999) und Mani (2001).

Die Zusammenfassung von Diskussionen zwischen mehreren Teilnehmern ist ein sehr neues Forschungsgebiet. Auch wird die Zusammenfassung eher einem Protokoll ähneln als einem fortlaufenden Text, wie bei den textbasierten Ansätzen. Hierfür werden Teile der Diskussionen, die sich zwischen zwei Themengrenzen befinden, als ein Text betrachtet. In nächster Zukunft werden wir extraktive Zusammenfassungsmethoden auf diese Textteile anwenden. Im Bereich der Textzusammenfassung wurden häufig *Lead* und *Tail* als *Baselines* verwendet, die aufgrund der besonderen Struktur von Zeitungstexten auch recht gute Ergebnisse erbringen. Dies ist bei unseren Daten nicht der Fall.

Erfolgversprechendere Möglichkeiten der Zusammenfassung basieren auf Lernverfahren. Hierfür müssen die zusammenfassungsrelevanten Passagen manuell annotiert werden, damit diese dann gelernt werden können. Im Anschluss an die Themengrenzenbestimmung werden zwischen zwei Grenzen die zusammenfassungsrelevanten Segmente oder auch größere Teile der Aussagen einzelner Sprecher manuell annotiert. Diese sollen zum einen dazu dienen, erste Zusammenfassungen zu erstellen und ihre Lesbar- und Verständlichkeit zu prüfen, aber auch um mit Hilfe eines maschinellen Lernverfahrens automatisch zusammenfassungsrelevante Elemente zu bestimmen. Die manuellen Elemente werden dann dazu dienen, die automatischen Verfahren zu evaluieren. Eine zweite Variante der Zusammenfassung nutzt die schon erstellten lexikalischen Ketten. Diese liefern Hinweise auf zusammenfassungsrelevante Begriffe und Phrasen, die bei der Behandlung eines Themas immer wieder vorkommen.

Ein Problem der automatischen Zusammenfassung ist die Evaluierung. In den Bereichen, die bisher vorgestellt wurden, gibt es formale Maße für die

Güte der Ergebnisse. In der Evaluierung von Zusammenfassungen fehlt dieses einheitliche Maß. Es gibt zwei prinzipielle Möglichkeiten: intrinsische und extrinsische Auswertung. Die intrinsische Evaluierung bezieht sich allein auf die Abdeckung der Zusammenfassung als solche. Dabei wird häufig verglichen, inwiefern die automatisch gewählten Passagen mit manuell gewählten Passagen übereinstimmen. Bei der extrinsischen Auswertung wird die Güte der Zusammenfassung im Hinblick auf den Informationsgehalt geprüft. Dabei werden häufig Fragebögen erstellt, bei denen Fragen zum Gesamttext zu beantworten sind, wobei verglichen wird, ob die Fragen mit Hilfe der Zusammenfassung genauso gut beantwortbar sind wie mit Kenntnis des vollständigen Texts.

8. Schluss

Die vorangegangenen Kapitel haben gezeigt, dass die vollautomatische Anaphernresolution in gesprochenen Dialogen und die Zusammenfassung derartiger Dialoge sehr schwierige Aufgaben sind. Für die erste Phase des Projektes DIANA-Summ wurden daher gewisse Vereinfachungen vorgenommen, insbesondere die Verwendung manuell erstellter Transkriptionen anstelle von automatischer Spracherkennung. Für den praktischen Einsatz unter realistischen Bedingungen muss das System aber natürlich auch mit vollständig automatisch erkannter Sprache umgehen können. Es ist daher geplant, im zweiten Teil des Projekts einen automatischen Spracherkennung zu integrieren.

Das hier beschriebene System hat die Form einer Kaskade von einzelnen Komponenten. Fehler, die in vorgeordneten Systemkomponenten (z. B. vom *POS-Tagger* oder Lemmatisierer) gemacht werden, setzen sich durch die folgenden Komponenten hindurch fort, verstärken sich möglicherweise und beeinträchtigen damit das Endergebnis. Die Qualität der Anaphernresolution sowie der Zusammenfassung hängt daher stark von der Qualität der Vorverarbeitungsschritte ab. Die Vorverarbeitungskomponenten beruhen größtenteils auf Methoden des überwachten maschinellen Lernens, welche wiederum von annotierten Daten abhängig sind. Die Qualitätssicherung der zugrundeliegenden Daten hat daher einen hohen Stellenwert.

Danksagung

Das Projekt DIANA-Summ wird gefördert von der Deutschen Forschungsgemeinschaft (STR 545/2-1) und von der Klaus Tschira Stiftung.

Literaturverzeichnis

- Beeferman, Doug/Berger, Adam/Lafferty, John (1999): Statistical models for text segmentation. In: Machine Learning, 34, 177–210.
- Brants, Thorsten (2000): TnT – a statistical Part-of-Speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Wash., 29 April–4 May 2000, 224–231.

- Brill, Eric (1994): Some advances in transformation-based part-of-speech tagging. In: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, Wash., 1–4 August 1994, 722–727.
- Carletta, Jean (1996): Assessing agreement on classification tasks: The kappa statistic. In: Computational Linguistics, 22(2), 249–254.
- Galley, Michel/McKeown, Kathleen/Fosler-Lussier, Eric/Jing, Hongyan (2003): Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan, 7–12 July 2003, 562–569.
- Janin, Adam/Baron, Don/Edwards, Jane/Ellis, Dan/Gelbart, David/Morgan, Nelson/Peskin, Barbara/Pfau, Thilo/Shriberg, Elizabeth/Stolcke, Andreas/Wooters, Chuck (2003): The ICSI meeting corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong 6–10 April 2003, 364–367.
- Mani, Inderjeet/Maybury, Mark T. (Hg.) (1999): Advances in Automatic Text Summarization. Cambridge, MA.
- Mani, Inderjeet (2001): Automatic Summarization. Amsterdam/Philadelphia.
- Marcus, Mitchell P./Santorini, Beatrice/Marcinkiewicz, Mary Ann (1993): Building a large annotated corpus of English: The Penn Treebank. In: Computational Linguistics, 19(2), 313–330.
- Mieskes, Margot/Strube, Michael (2006): Part-of-Speech tagging of transcribed speech. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genua, Italy, 22–28 May, 2006, 935–938.
- Mitchell, Tom M. (1997): Machine Learning. New York.
- Müller, Christoph (2006): Automatic detection of nonreferential *It* in spoken multi-party dialog. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, 3–7 April, 2006, 49–56.
- Pevzner, Lev/Hearst, Marti (2002): A critique and improvement of an evaluation metric for text segmentation. In: Computational Linguistics, 28 (1), 19–36.
- Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1991). A Comprehensive Grammar of the English Language. London, UK.
- Radev, Dragomir R./Hovy, Eduard/McKeown, Kathleen (2002): Introduction to the special issue on text summarization. In: Computational Linguistics, 28(4), 399–408.
- Schmid, Helmut (1997): Probabilistic Part-of-Speech tagging using decision trees. In: Jones, Daniel/Somers, Harold (Hg.): New Methods in Language Processing. London, UK, 154–164.
- Toutanova, Kristina/Klein, Dan/Manning, Christopher D./Singer, Yoram (2003): Feature-rich Part-of-Speech tagging with a cyclic dependency network. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Alberta, Canada, 27 May–1 June, 2003, 252–259.
- Toutanova, Kristina/Manning, Christopher D. (2000): Enriching the knowledge sources used in a maximum entropy Part-of-Speech tagger. In: Proceedings of the Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong 7–8 October 2000, 63–70.