12. Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität

1. Korpusdefinitionen

Der Terminus *Korpus* wird — wie so viele linguistische Termini — höchst uneinheitlich verwendet. Schon bei der Schreibweise herrscht keine Einigkeit: Während in diesem Band die eindeutschende Version mit *K* gewählt wurde, findet man daneben auch *Corpus* mit *C* (so im Titel von Bergenholtz/ Schaeder (eds.) 1979) wie im Englischen und Französischen. Das Genus schwankt ebenfalls: Neutrum überwiegt (vgl. die Angaben in Duden-DUW, HWDG und anderen Wörterbüchern), zuweilen stößt man aber auch auf *der Korpus* in einer der uns hier interessierenden Bedeutungen. Mit dem Terminus werden in der Linguistik zudem recht verschiedenartige Objekte bezeichnet. Gemeinsam ist ihnen lediglich, daß es sich um Sammlungen von Sprachmaterial handelt. Die Unterschiede betreffen insbesondere zwei Punkte:

(1) Es besteht keine Einhelligkeit darüber, ob es sich bei einem Korpus um eine Sammlung von Texten handeln muß, oder ob auch Sammlungen isolierter Sätze oder gar einzelner Wörter so bezeichnet werden sollen. Zur Verdeutlichung empfiehlt sich daher der Terminus *Textkorpus,* den wir hier gleichbedeutend mit *Korpus* verwenden, um eine Sammlung von Texten oder Teiltexten zu bezeichnen. Unter *Text* soll dabei informell eine relativ selbständige, inhaltlich kohärente Folge natürlicher sprachlicher Äußerungen verstanden werden, unter *Teiltext* ein zusammenhängender, lückenloser Ausschnitt aus einem Text. Vom *Textkorpus* wird hier insbesondere die *Belegsammlung* nach Art der z. B. für lexikographische Zwecke üblichen Karteien unterschieden, bei denen aus den verwendeten Texten lediglich einzelne Zitate mehr oder minder willkürlich exzerpiert werden, so daß das Kriterium „zusammenhängend“ nicht erfüllt ist.

(2) Es besteht keine Einhelligkeit darüber, ob ein Korpus nur Texte enthalten darf, die natürlichen Kommunikationssituationen entstammen, oder ob die Bezeichnung auch auf Sammlungen elizitierter (experimentell gewonnener) oder selbstgebildeter Texte angewendet werden soll. Da in der Diskussion über die Datenbasis für die Sprachbeschreibung gerade der Gegensatz zwischen Korpus und Intuition eine zentrale Rolle spielt (vgl. z. B. Ulvestad 1979), halten wir es für unangebracht, ihn dadurch zu verwischen, daß auch Sammlungen selbstkonstruierter Beispiele als Korpora bezeichnet werden (so z. B. von Greenbaum 1984, 194: “Linguists use their knowledge of the language to create a corpus of samples of the language” und “Linguists use [...] a corpus created from introspection”). In kommunikativer Absicht entstandene Äußerungen sind aus vielfältigen Gründen nicht mit Beispielen gleichzusetzen, die ein sich für „kompetent“ haltender Sprecher im Rahmen der Diskussion linguistischer Theorien bildet (für eine solche Gleichstellung plädiert z. B. Altmann, H. 1981, 73—78). Auch elizitierte Daten sind nicht ohne weiteres vergleichbar. Im ungünstigsten Fall handelt es sich um recht unnatürliche Testsituationen (z. B. Ergänzungsaufgaben); im günstigsten werden spontane Reaktionen erreicht, ohne daß die Testperson sich ihrer Rolle bewußt ist (vgl. Labov 1975, 49). Die höchst unterschiedliche Qualität elizitierter Äußerungen läßt es sinnvoll erscheinen, für entsprechende Sammlungen mit Bausch (1975, 132 ff.) die Bezeichnung *manipuliertes Korpus* zu verwenden und von einem Korpus im eigentlichen Sinn nur bei authentischen (d. h. aus natürlichen und nicht manipulierten Kommunikationssituationen stammenden) Texten zu sprechen.

In Anlehnung an Bausch (1979, 78) und Bungarten (1979, 34) gehen wir von folgender Definition aus:

Ein Korpus ist eine begrenzte Sprachdatenmenge einer Sprache, die ausschließlich aus (geschriebenen und/oder gesprochenen) Texten oder Teiltexten besteht.

Im Rahmen der Computerlinguistik gilt das Interesse solchen Korpora, die in einer für die elektronische Datenverarbeitung ge-

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

eigneten Form vorliegen (im allgemeinen auf Magnetband). Neben der Textsammlung selbst werden üblicherweise auch Statistiken und Wortlisten verschiedener Art sowie Indizes und Konkordanzen zur Verfügung gestellt, was für viele Anwendungen überaus hilfreich ist. Solche Materialien sollten bereits bei der Konstruktion des Korpus eingeplant werden, zumal sich — was hier nicht näher ausgeführt werden kann — Entscheidungen über die Datenaufbereitung (s. 5.) sowohl auf die Möglichkeiten der maschinellen Erstellung von Indizes, Konkordanzen usw. als auch auf deren Benutzerfreundlichkeit auswirken.

2. Zur Leistung von Korpora

Korpora sind ein Hilfsmittel für die Sprachbeschreibung, dessen Wert nicht unumstritten ist. Der Erörterung von Konstruktionsprinzipien wollen wir deshalb eine Einschätzung dessen vorausschicken, was man überhaupt von Korpora erwarten darf. Mit Lutzeier (1981, 53) sind wir der Meinung, daß sie keine „Wundermittel" sind, mit deren Hilfe man z. B. Bedeutungen ohne (subjektive) Interpretation und unabhängig von wissenschaftstheoretischen Annahmen fast automatisch bestimmen könnte (s. hierzu van de Velde 1974, 1979). Anders als Lutzeier sehen wir in diesem Argument allerdings keinen entscheidenden Einwand gegen Textkorpora. Schwerwiegender sind sicherlich Bedenken wie die folgenden (vgl. zu den Vor- und Nachteilen der Korpuslinguistik auch Bergenholtz/Schaeder (eds.) 1979, darin bes. Bungarten 1979, 36—38, sowie Johansson 1982 und Svartvik 1982 a, 11).

(1) Da ein Korpus nur eine begrenzte Datenmenge umfaßt, kann es nicht alle Möglichkeiten einer Sprache dokumentieren — viele Formen oder Konstruktionen, die grammatisch sind, kommen zufällig nicht vor (vgl. z. B. Greenbaum 1977, 128). Wir konnten beispielsweise in einem 2,5 Mio. Textwörter umfassenden Korpus des Deutschen die Wendung *damit kannst du mich jagen* nicht belegen. Mit einem noch größeren Korpus (vor allem der gesprochenen Sprache) ließe sich diese „Lücke" wohl füllen; ob man aber etwa das von Fodor/Garret (1966, 137) vermißte Beispiel *my friend owns three-eighths of an elephant* jemals bei einem unvorbelasteten Sprecher/Schreiber finden wird, darf man bezweifeln. Je nach Fragestellung ist der Mangel an Belegen allerdings sehr unterschiedlich einzuschätzen:

— Wenn es um die Grammatikalität von Beispielen geht, kann man aus den vorgefundenen Belegen extrapolieren (s. (4)). Es genügt also, die Struktur des Satzes *my friend owns three-eighths of an elephant* und die darin vorkommenden lexikalischen Elemente zu finden, was kein allzu großes Korpus erfordern dürfte.

— Wenn es darauf ankommt, die Existenz selten auftretender Formen oder Konstruktionen nachzuweisen, mag es ergiebiger sein, willkürliche Belege zu sammeln, statt ein Korpus auszuwerten (vgl. Mackin 1983, vi zu idiomatischen Wendungen). Damit besteht aber für Aussagen über Häufigkeit oder Üblichkeit keine Grundlage mehr.

— Wenn sprachliche Zweifelsfälle geklärt werden sollen, gibt es zur Korpusauswertung keine überzeugende Alternative: Da kompetente Sprecher sich bei introspektiven Urteilen über ihre Kompetenz irren können und nicht in der Lage sind, sich die Regeln ihrer internalisierten Grammatik bewußt zu machen (s. Chomsky 1965, 8), ist auf Intuition kein Verlaß. In der Tat haben sich intuitive Einschätzungen immer wieder als höchst fragwürdig erwiesen (s. z. B. Ulvestad 1979). Auch die Befragung von Informanten kann (aus ähnlichen Gründen) kaum die Grundlage einer zuverlässigen Beschreibung abgeben (vgl. z. B. Bergenholtz 1980, 49—53). Willkürliche Zitatensammlungen führen hier erst recht nicht zu einer Lösung, weil sie keinerlei statistische Aussagen gestatten. Wir halten es daher für einen Irrweg, die Korpusmethode zugunsten solcher Verfahren aufzugeben, wenn die Belege im Korpus nicht ausreichen. Daß z. B. das Brown Corpus für eine Untersuchung der Negation von *need* und *dare* zu wenig Beispiele liefert, ist kein Grund, auf intuitive Urteile auszuweichen (so Greenbaum 1984, 193); es unterstreicht lediglich, daß Korpora dieser Größenordnung (1 Mio. Textwörter) viel zu klein sind (s. 9.). Manche Elemente oder Strukturen sind freilich so selten, daß sie auch in einem angemessen großen Korpus (von 50 oder 100 Mio. Textwörtern) nicht oder nur durch wenige Belege zu dokumentieren sind. Das dürfte im Deutschen z. B. für das Präteritum von *sieden* gelten (*sott* oder *siedete* ?). Solche Zweifelsfragen wird man dann unabhängig von der Methode als nicht lösbar betrachten müssen.

(2) Durch die Korpusauswertung allein

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

läßt sich nicht feststellen, inwieweit die vorgefundenen Belege stilistisch, sozial o. ä. markiert sind (vgl. Béjoint 1983, 72 f.). Das Problem, wie man den Sprachgebrauch mit Sozialdaten usw. korrelieren kann, ist allerdings von der Verwendung eines Korpus unabhängig. Dabei wird man zu intersubjektiv nachprüfbaren Ergebnissen sogar noch am ehesten kommen können, wenn man mit einem Korpus arbeitet, bei dessen Erstellung Faktoren wie Thema, Textsorte, regionale Herkunft usw. berücksichtigt wurden (vgl. 8.). Subjektive Einschätzungen sind hingegen auch in diesem Bereich nicht brauchbar, zumal das Repertoire eines einzelnen Sprechers ein viel begrenzteres Spektrum sprachlicher Mittel umfaßt als ein geeignet zusammengestelltes Korpus. Dementsprechend sind Fehlurteile an der Tagesordnung, wie z. B. die in deutschen Wörterbüchern übliche Einstufung von *eh* und *halt* (in *das weiß halt eh keiner*) als „süddeutsch/schweizerisch/österreichisch“ o. ä. (s. Bergenholtz/Mugdan 1986, 80).

Daß ein Korpus es nicht gestatten soll, zwischen normalem und ungewöhnlichem Gebrauch zu unterscheiden (so Béjoint 1983, 72) trifft schon gar nicht zu — im Gegenteil bietet es im Unterschied zur Belegsammlung gerade die Voraussetzung für statistische Untersuchungen, ohne die eine Trennung zwischen „üblich“ und „ungewöhnlich“, „oft“ und „selten“ usw. völlig unbegründet und beliebig bleiben muß (s. Bergenholtz/Mugdan 1984, 83 f.; Mugdan 1985, 196—199).

(3) Da ein Korpus auch ungrammatische Sätze enthalten kann, führt es angeblich nicht zu einer Grammatik, die ihrer Aufgabe gerecht wird, die grammatischen Sätze einer Sprache und nur diese zu generieren (vgl. Bierwisch 1963, 9; Fodor/Garret 1966, 137). In der Praxis sind aber eindeutige Fehler recht leicht zu erkennen, und ihr Anteil ist keineswegs so hoch, wie das gerne suggeriert wird — nicht einmal in der gesprochenen Sprache (vgl. Labov 1972, 203). „Druckfehler“ oder „Versprecher“, wie sie im Alltag heißen, fallen also kaum ins Gewicht. In allen wirklich interessanten Fällen urteilen jedoch die Sprecher einer Sprache über die Grammatikalität uneinheitlich (vgl. dazu Levelt 1972; Spencer 1973; Labov 1975 und zahlreiche publizierte Befragungen). Vor diesem Hintergrund ist auch das Argument hinfällig, Korpora seien deshalb nicht brauchbar, weil sie keine Auskunft über ungrammatische Äußerungen gäben (vgl. Rainer 1984, 292).

Bei schwankendem Sprachgebrauch muß, wie erwähnt, auch die Korpusauswertung nicht notwendigerweise zu klareren Ergebnissen hinsichtlich der Verteilung der konkurrierenden Varianten führen. Sie deckt aber zumindest die Existenz solcher Varianten auf, die — nicht zuletzt unter dem Einfluß normativer Grammatiken — oft gar nicht allgemein bekannt ist. Insofern übt die Korpusmethode einen heilsamen Zwang aus: Während introspektiv arbeitende Linguisten vielfach unliebsame Beispiele als „ungrammatisch“ ausschließen, um ein vorgefaßtes Regelsystem nicht in Frage zu stellen (vgl. auch Spencer 1973 und Labov 1975, 30 u. ö. zu Intuitionen von Linguisten), kann man bei der Korpusanalyse unerwartete Fälle nicht einfach ignorieren, sondern muß schon stichhaltige Begründungen vorweisen, um sie als „Irrtümer“ ausscheiden zu können.

(4) Die Beschreibung der in einem Korpus belegten Daten gilt vermeintlich nur für dieses begrenzte Material und nicht für die betreffende Sprache insgesamt (vgl. Greenbaum 1977, 128). Eine solche Beschränkung wird gerne der deskriptiven (strukturalistischen) Linguistik zugeschrieben (so von Bondzio 1980, 129). In Wahrheit haben jedoch Analysen endlicher Texte stets dazu gedient, Regularitäten zu ermitteln, aufgrund derer neue Äußerungen interpretiert oder produziert werden können. Das von Chomsky (1961, 237 f.) beschriebene Verfahren, durch Klassenbildung aus einzelnen Sätzen generative Regeln zu extrapolieren, ist kein anderes als das seit jeher verwendete. Wahr ist allerdings, daß z. B. die ermittelten Häufigkeiten einzelner Elemente zunächst nur für das Korpus selbst gelten und sich nicht ohne weiteres verallgemeinern lassen (s. dazu im Kontext von Häufigkeitswörterbüchern Alekseev 1984, 30, 38 f.).

3. Zielsetzungen

Textkorpora werden für eine Fülle verschiedener Zwecke verwendet, im linguistischen Bereich etwa zur:

— Bestimmung von Wortbedeutungen,

— Ermittlung der Vorkommenshäufigkeit konkurrierender Formen und Konstruktionen,

— Aufstellung syntaktischer Muster,

— Erstellung von Häufigkeitswörterbüchern und Grundwortschätzen,

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

— Untersuchung stilistischer Mittel.

Dabei können auch diachrone oder kontrastive Beschreibungen (verschiedener Sprachen oder verschiedener Varietäten einer Sprache) angestrebt werden.

Außerhalb der Linguistik werden Korpora ebenfalls verwendet, so in der Literaturwissenschaft (z. B. zur Erstellung von Konkordanzen oder Autorenwörterbüchern) und in der Psychologie (zur Auswertung von Therapiegesprächen) u. a. Dementsprechend gibt es ganz unterschiedliche Typen von Korpora (vgl. Art. 13).

Das jeweilige Forschungsinteresse ist für die Konstruktion des Korpus von entscheidender Bedeutung. Das betrifft insbesondere die Art der Texte, die sich nach Kriterien wie synchron/diachron, Gemeinsprache/Fachsprache, gesprochen/geschrieben sowie mittels Typologien von Textsorten oder literarischen Gattungen beschreiben läßt. Aber auch zahlreiche Details der Korpuskonstruktion hängen von der gewählten Zielsetzung ab — selbst bei Korpora, die einen für viele (linguistische) Anwendungen brauchbaren Querschnitt durch eine moderne Standardsprache bieten wollen (z. B. das Brown Corpus für das amerikanische und das Lancaster-Oslo-Bergen (LOB) Corpus für das britische Englisch, das LIMAS-Korpus für das Deutsche). Wir werden im folgenden primär auf solche Mehrzweckkorpora eingehen, denn generelle Hinweise zur Konstruktion von Textsammlungen jeglicher Art lassen sich kaum geben.

4. Technische Voraussetzungen

Bei der Zusammenstellung eines maschinenlesbaren Textkorpus spielen auch die technischen Gegebenheiten eine Rolle, und zwar

(1) die Möglichkeiten der Texteingabe und

(2) die Möglichkeiten der Textverarbeitung.

Zu (1): Korpora geschriebener Texte lassen sich heute — anders als noch vor einigen Jahren — sehr leicht erstellen. Zum einen liegen durch die Einführung neuer Technologien im Schriftsatz immer mehr Bücher, Zeitschriften und Zeitungen bereits in maschinenlesbarer Form vor, und zum anderen hat die Qualität von Klarschriftlesern ein solches Niveau erreicht, daß eine zügige Textaufnahme mit geringer Fehlerquote in technischer und finanzieller Hinsicht realistisch geworden ist.

Für die automatische Aufnahme gesprochener Texte gibt es hingegen noch keine geeigneten Verfahren (vgl. Sinclair 1982, 4 f.). Die Transkription der Texte und ihre Übertragung auf Datenträger ist äußerst zeitaufwendig (s. z. B. Bausch 1971; Müller 1971; Quirk/Svartvik 1979; Svartvik et al. 1982). Ob die von Francis (1979, 116 f.) angestellte Berechnung (ein Dollar pro gesprochenem Wort) auch heute noch zutrifft, kann hier nicht beurteilt werden. Zweifellos sind aber finanzielle Gründe entscheidend dafür verantwortlich, daß bislang erst einige kleinere Korpora gesprochener Sprache maschinenlesbar vorliegen (das Freiburger Korpus und das Textkorpus Grunddeutsch Sprechsprache für das Deutsche, das London-Lund Corpus für das Englische) und daß offenbar weitere derartige Korpora nicht in Arbeit sind.

Zu (2): Der maschinellen Verarbeitung von Textkorpora sind durch Speicherplatz und Rechengeschwindigkeit der verfügbaren Computer gewisse Grenzen gesetzt. Die Entwicklungen im Hardwarebereich haben aber in den letzten Jahren diese Grenzen erheblich verschoben, so daß auch größere Korpora mit mehreren Millionen Textwörtern z. B. in Hochschulrechenzentren verwendet werden können. Korpora von der Größe des Brown Corpus und des LIMAS-Korpus (1 Mio. Textwörter) lassen sich sogar auf einem Personal Computer mit geeigneter Peripherie bearbeiten. Allerdings sind die erforderlichen Operationen (Suchen, Sortieren und Kopieren von Texten usw.) immer noch ziemlich rechenzeitintensiv.

5. Datenaufbereitung

Bei der Aufnahme gesprochener wie geschriebener Texte auf Datenträger ist ein gewisses Maß an Bearbeitung der Primärdaten erforderlich. Es handelt sich hierbei um

(1) Umkodierung von Informationen,

(2) Reduktion von Informationen,

(3) Hinzufügung von Informationen.

Bei gesprochenen Texten ist außerdem eine vorherige Verschriftung nötig.

Zu (1): Bei maschinenlesbaren Korpora steht nur ein begrenzter Zeichensatz zur Verfügung, mit dem alle Informationen kodiert werden müssen. So waren auf älteren Rechnern nur 64 verschiedene Zeichen darstellbar, was keine Unterscheidung zwischen Groß- und Kleinbuchstaben erlaubte. Um diesen Unterschied rekonstruieren zu können, mußte man daher z. B. jedem Großbuchstaben das Sonderzeichen $ voranstellen. Heute ist auf allen gängigen Rechnern der 128 Zeichen umfassende ASCII-Code

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

verfügbar, so daß dieses Problem nicht mehr besteht. Damit sind aber z. B. unterschiedliche Schrifttypen und -größen nicht darstellbar; es muß also etwa der Wechsel zwischen recte und kursiv nach wie vor durch Sonderzeichen kodiert werden.

Zu (2): Es ist weithin üblich, bei der Aufnahme geschriebener Texte z. B. die ursprüngliche Aufteilung in Zeilen und Seiten nicht beizubehalten. Sonderzeichen, mathematische Formeln oder Passagen in anderen Alphabeten werden teilweise nicht umkodiert, sondern ausgelassen oder durch ein Platzhaltersymbol ersetzt. Bei Illustrationen u. ä. ist dieses Verfahren sogar nahezu unumgänglich.

Zu (3): Bei vielen Korpora werden die Originaltexte durch zusätzliche Informationen angereichert (vgl. z. B. Engelien 1979). Hierzu gehören beispielsweise:

— die Disambiguierung von Satzzeichen (Punkt am Satzende vs. Abkürzungspunkt u. ä.),

— die Kennzeichnung von Eigennamen,

— die Ergänzung der „Auslassungen" bei Konstruktionen vom Typ *Tages- und Nachtzeit* (zu *Tages/zeit/ und Nachtzeit* o. dergl.),

— die Markierung lexikalischer oder grammatischer Eigenschaften (s. u.).

Ferner werden üblicherweise Textkennungen und Zeilenzählungen angebracht.

Bei der Verschriftung gesprochener Texte stellen sich insbesondere zwei Probleme:

(1) die Wahl des Transkriptionssystems,

(2) die Kodierung von Pausen, Intonation, Sprecherwechsel, parasprachlichen Phänomenen (z. B. *äh, hm* u. dergl.), nonverbalem Verhalten usw.

Zu (1): Der geringste Informationsverlust gegenüber dem Originaltext könnte bei einer phonetischen Transkription erreicht werden. Für eine maschinenlesbare Version wäre jedoch eine Umkodierung (z. B. aus dem Internationalen Phonetischen Alphabet) in den verfügbaren Zeichensatz nötig. Zudem ist der Zeitaufwand selbst für eine breite phonetische Transkription enorm hoch. Die bislang vorliegenden Korpora verwenden daher die Standardorthographie der betreffenden Sprache und deuten allenfalls einige umgangssprachliche oder dialektale Eigentümlichkeiten an (z. B. *dann kann mer sagen* und nicht *dann kann man sagen* im Korpus Grunddeutsch Sprechsprache, s. Pfeffer/Lohnes (eds.) 1984, 31). Damit ist einerseits ein beträchtliches Maß an Interpretation verbunden (streng genommen wird der Originaltext in einen „Standard" übersetzt), andererseits sind solche Verschriftungen nicht selten mehrdeutig (z. B. steht *eh* oft sowohl für einen Pausenfüller — vermutlich /ə:/ oder /ɛ:/ — wie in *und, eh / also, ich war sternhagelvoll* als auch für /e:/ wie in *undann brauch se eh schon ma zwei Jahre,* so in Brons-Albert 1984, 18, 102).

Zu (2): Die Kodierung nichtsegmentaler Merkmale gesprochener Äußerungen ist von Korpus zu Korpus sehr unterschiedlich, sowohl hinsichtlich der berücksichtigten Informationen als auch hinsichtlich der verwendeten Symbole. Das Freiburger Korpus benutzt z. B. *xxxxxxaa, xxxxxxab* usw. zur Kennzeichnung der Sprecher, +p+ für eine Pause; das LOB Corpus markiert kurze Pausen mit . und längere mit — sowie steigende und fallende Intonation (was das Freiburger Korpus nicht berücksichtigt) mit Pfeilen, usw.

Es erscheint sinnvoll, bei Korpora, die für verschiedene Zwecke verwendbar sein sollen, möglichst viele Merkmale des Originaltexts zu bewahren (ggf. durch Umkodierung) und die Reduktion von Informationen zu vermeiden. Ärgerlich ist z. B. die gelegentlich praktizierte Eliminierung fremdsprachiger Zitate u. ä. Auch die Anreicherung mit Zusatzinformationen ist nicht unproblematisch. So kann sich z. B. die Behandlung von Fällen wie *halb- oder vollautomatisch* in unerfreulicher Weise auf die Wortstatistik auswirken (wenn etwa *halb-* wie *halb* oder wie *halbautomatisch* gezählt wird).

Die Kodierung grammatischer und lexikalischer Merkmale, für die sich die Bezeichnung 'tagging' eingebürgert hat, ist für die maschinelle Analyse der Texte sinnvoll und angebracht (vgl. hierzu Svartvik/Eeg-Olofsson 1982; Johansson/Jahr 1982; Svartvik et al. 1982; Leech/Garside/Atwell 1983; de Haan 1984 a). Ein Beispiel hierfür ist die manuelle oder partiell automatisierte Kodierung der Wortartzugehörigkeit (s. z. B. Leech/Garside/Atwell 1983, 18—20), die eine weitgehende Monosemierung leistet und damit eine automatische Lemmatisierung und eine syntaktische Analyse wesentlich erleichtert, wenn nicht überhaupt erst ermöglicht. Weiterhin werden in manchen Korpora besondere Phänomene wie Parenthesen, Tmesis, Ellipsen oder Idiome gekennzeichnet (s. z. B. Bausch 1971; Leech/Garside/Atwell 1983) oder Art und Umfang von Phrasen oder Sätzen (s. Svartvik/Eeg-Olofsson 1982, 104—107). Diese und andere Zusatzinforma-

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

tionen bieten für viele Anwendungen wesentliche Vorteile gegenüber einem unanalysierten Korpus. Es ist aber zu bedenken, daß die mitgelieferten Interpretationen und die dabei zugrunde gelegten Theorien sich in den meisten Fällen nicht mit denen anderer Forscher decken. Das Korpus sollte daher zumindest auch in einer unbearbeiteten Version zur Verfügung gestellt werden. Im übrigen sind Korpora mit vielen Kodierungen auch schlecht zu lesen, vor allem wenn die Kodes nicht unmittelbar einleuchtend sind (wie beim LIMAS-Korpus) oder wenn sie sehr viel Platz einnehmen (so beim Freiburger Korpus).

6. Repräsentativität

In Diskussionen um eine „Korpuslinguistik" ist „Repräsentativität" ein immer wiederkehrender Topos (vgl. hierzu Bungarten 1979, 42 f.; Rieger, B. 1979 b, 58—63). Dabei hat es bei Befürwortern wie Gegnern der Korpusanalyse einige Mißverständnisse gegeben.

Eine Stichprobe kann dann als repräsentativ gelten, wenn sie hinsichtlich bestimmter Eigenschaften mit der Grundgesamtheit übereinstimmt, aus der sie stammt. Offenkundig läßt sich das nur dann feststellen, wenn über die Grundgesamtheit ebenso viel bekannt ist wie über die Stichprobe — womit es sich erübrigt, eine Stichprobe zu erheben. Es gibt jedoch gewisse Verfahren der Stichprobenbildung, die mit hoher Wahrscheinlichkeit (wenn auch nicht mit Sicherheit) zu einer repräsentativen Stichprobe führen. Hier ist insbesondere das Prinzip der zufälligen Auswahl zu nennen, bei der jedes Element der Grundgesamtheit die gleiche Chance (Wahrscheinlichkeit) haben muß, in die Stichprobe aufgenommen zu werden (s. z. B. Rieger, B. 1979 b, 63 ff.). Das ist bei einer willkürlichen Auswahl nicht gegeben, z. B. wenn ein Forscher für ein Korpus moderner deutscher Romane nach Gutdünken einige Werke aussucht, die er gerade kennt oder schätzt. Für eine zufällige Stichprobe müßte man dagegen z. B. alle Elemente der Grundgesamtheit — hier also alle in einem bestimmten Zeitraum veröffentlichten deutschen Romane — durchnumerieren und (etwa per Zufallszahlengenerator) aus diesen Nummern die gewünschte Anzahl von Werken auslosen.

Oftmals ist jedoch ein solches Verfahren nicht praktikabel. Für viele Anwendungen wird daher die Stichprobe so gewählt, daß bestimmte Merkmale in der Grundgesamtheit und in der Stichprobe die gleiche Verteilung aufweisen. Die Stichprobe wird also in Teilmengen aufgeteilt, die bestimmten Bedingungen genügen müssen. Erfolgt innerhalb der Teilmengen die Auswahl willkürlich, so spricht man von einem Quotaverfahren, ist sie zufällig, so handelt es sich um eine geschichtete Stichprobe. Diese Methoden sind z. B. aus Meinungsumfragen vertraut: Es wird ein Personenkreis befragt, der hinsichtlich Alter, Geschlecht, Beruf, Einkommen usw. genauso zusammengesetzt ist wie die Gesamtbevölkerung. Voraussetzung ist dafür natürlich, daß die Verteilung dieser Merkmale in der Grundgesamtheit bekannt ist. Die entscheidende Annahme ist nun, daß auch bestimmte Merkmale, über die in der Grundgesamtheit nichts bekannt ist, in der Stichprobe genauso verteilt sind. Das setzt voraus, daß es einen Zusammenhang zwischen diesen und den für die Quotenbildung verwendeten Merkmalen gibt. So werden bei Meinungsumfragen die genannten Faktoren (und nicht etwa Körpergröße, Haarfarbe oder Vorname) kontrolliert, weil man davon ausgehen kann, daß gerade sie den politischen Standort der Befragten beeinflussen.

Wenn nun ein Textkorpus mit einiger Wahrscheinlichkeit für eine Sprache insgesamt repräsentativ sein soll, müßte vor allem die Grundgesamtheit bekannt sein, also die Menge aller Texte der betrachteten Sprache L. Das ist offenkundig bei keiner Sprache der Fall. (Selbst bei einer toten Sprache kennt man nur die überlieferten Texte, nicht aber alle, die jemals in ihr produziert wurden.) Eine zentrale Rolle spielt hier die Frage, ob zwei Texte zur selben Sprache oder zu verschiedenen gehören, also das notorische Problem der Abgrenzung zwischen Sprache und Dialekt. Auch wenn es ein Verfahren gäbe, mit dem sich feststellen ließe, ob der Text T zur Sprache L gehört oder nicht, wäre damit noch keine vollständige Aufzählung aller zu L gehörigen Texte möglich. Eine Zufallsauswahl von Texten aus L kann es daher nicht geben.

Das Quotaverfahren ist ebenfalls nicht anwendbar, weil die Verteilung der möglicherweise relevanten Merkmale (Textsorte, Thema usw.) in Grundgesamtheiten wie „der deutschen Gegenwartssprache" nicht bekannt ist — ganz abgesehen davon, daß sich derartige Merkmale nur schwer objektiv bestimmen lassen.

Da nun nicht einmal eine Wahrscheinlichkeit dafür berechnet werden kann, daß die

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

gewählte Zusammenstellung von Texten hinsichtlich gewisser Merkmale repräsentativ ist (eine Repräsentativität „an sich“ kann es ja ohnehin nicht geben), scheint es wenig sinnvoll, bei Textkorpora überhaupt von Repräsentativität zu sprechen — es sei denn, man bezieht sich auf eine klar definierte Grundgesamtheit (die dann eben nicht „die deutsche Gegenwartssprache“ o. ä. sein kann). Es bietet sich an, statt dessen die Bezeichnung „exemplarisch“ zu verwenden: „Ein Korpus ist exemplarisch, wenn seine Repräsentativität nicht nachgewiesen ist, andererseits weniger formale Argumente /.../ für eine sinnvolle Vertreterfunktion des Korpus plädieren“ (Bungarten 1979, 42 f.).

Wenn es prinzipiell nicht möglich ist, ein für die Sprache L mit gewisser Wahrscheinlichkeit repräsentatives Korpus zu konstruieren (s. auch Rieger, B. 1979 b), bieten sich zwei Auswege an. Zum einen kann man statt der Menge aller zu L gehörigen Texte eine wohldefinierte Teilmenge als Grundgesamtheit wählen (s. 7.); zum anderen kann man für das Quotaverfahren eine plausibel scheinende hypothetische Verteilung bestimmter Merkmale zugrunde legen (s. 8.). Schließlich spielt für die Einschätzung, ob ein Korpus als exemplarisch gelten kann, auch dessen Größe eine Rolle (s. 9.).

7. Festlegung der Grundgesamtheit

Anstatt bei der Konstruktion eines Textkorpus von „der deutschen Gegenwartssprache“ auszugehen, kann man eine klar definierte Menge von Texten als Grundgesamtheit wählen, z. B.

— die Menge aller 1984 in der Deutschen Bibliographie verzeichneten Veröffentlichungen;

— die Menge aller im Jahrgang 1967 des „Mannheimer Morgen“ erschienenen mit vollem Namen gezeichneten Artikel;

— die Menge aller vom 1. 1. bis zum 30. 6. 1980 im ersten Hörfunkprogramm von Radio Bremen gesendeten Nachrichten.

Wenn es der Umfang gestattet, kann diese Grundgesamtheit vollständig als Korpus verwendet werden (so beim Lunder Zeitungskorpus; s. Rosengren 1972). Andernfalls lassen sich die in 6. genannten Verfahren der Stichprobenbildung anwenden. So könnte man z. B. aus den in der Deutschen Bibliographie nachgewiesenen Publikationen mit Hilfe der fortlaufenden Numerierung eine Zufallsauswahl treffen oder (im Quotaverfahren) jedes Sachgebiet entsprechend der ihm zugeordneten Zahl von Veröffentlichungen berücksichtigen. Auf diese Weise erhält man ein Korpus, das mit gewisser Wahrscheinlichkeit für die gewählte Grundgesamtheit repräsentativ ist (vgl. auch Schaeder 1979, 237—239 zum Bonner Zeitungskorpus).

Es stellt sich nun die Frage, inwieweit eine solche Grundgesamtheit als exemplarisch für eine größere Textmenge gelten kann. Sind z. B. die 1967 im „Mannheimer Morgen“ erschienenen signierten Artikel exemplarisch für

— die deutsche Gegenwartssprache,

— die deutsche Schriftsprache in der Bundesrepublik der sechziger Jahre,

— die bundesdeutsche Tagespresse oder auch nur

— den gesamten Jahrgang 1967 des „Mannheimer Morgen„?

In Ermangelung präziser Kriterien für Exemplarität sind solche Fragen kaum zu beantworten. Es dürfte aber einleuchten, daß die Grundgesamtheit für „die deutsche Gegenwartssprache“ o. dergl. um so weniger exemplarisch ist, je restriktiver die Bedingungen für die Zugehörigkeit zu dieser Grundgesamtheit sind.

Andererseits sind gewisse Restriktionen die Voraussetzung dafür, daß die Grundgesamtheit überhaupt eindeutig bestimmt werden kann. So lassen sich die in einem Land erschienenen Bücher, Zeitschriften oder Zeitungen aufgrund der heute üblichen Registrierungspflicht für derartige Veröffentlichungen meist unschwer ermitteln; Flugblätter, Formulare, Broschüren, Gebrauchsanleitungen und viele andere Arten von Druckerzeugnissen werden jedoch nicht systematisch erfaßt, von maschinenschriftlichen oder handgeschriebenen Texten ganz zu schweigen. Bei gesprochener Sprache könnte man etwa Rundfunk- und Fernsehsendungen, Vorlesungen an Universitäten oder Parlamentsdebatten lückenlos nachweisen, aber bei Alltagsgesprächen ist es schier unmöglich, in irgendeiner Hinsicht Vollständigkeit zu erreichen. Wohldefinierte Grundgesamtheiten sind also nur bei bestimmten Arten von Texten möglich.

8. Auswahl der Stichprobe

Wenn man die in 7. beschriebene Eingrenzung der Grundgesamtheit nicht vornehmen will, um die damit verbundenen Beschrän-

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

kungen hinsichtlich der Art der Texte zu vermeiden, kann man eine Stichprobe bilden, in der bestimmte Merkmale eine plausibel erscheinende Verteilung aufweisen. Üblicherweise stützt man sich dabei (a) auf eine Typologie der Textsorten (wie beim London-Lund Corpus des gesprochenen Englisch; s. Quirk/Svartvik 1979) und/oder (b) eine Systematik von Sachgebieten und Themen (wie beim LIMAS-Korpus).

Bei der Festlegung der Quoten für die einzelnen Textsorten oder Sachgebiete gibt es zwei Vorgehensweisen:

(1) Es wird die Verteilung des betreffenden Merkmals in einer wohldefinierten Grundgesamtheit zugrunde gelegt. So ist man beim LIMAS-Korpus von der Sachgebietssystematik der Deutschen Bibliographie ausgegangen und hat jedes Gebiet entsprechend der Zahl der nachgewiesenen Titel berücksichtigt. Die Grundgesamtheit war jedoch nicht auf die in der Deutschen Bibliographie genannten Publikationen beschränkt; es wurden z. B. auch Zeitungstexte, Flugblätter u. v. a. aufgenommen. Eine Gewichtung der Sachgebiete nach der bloßen Zahl der Buchveröffentlichungen ist freilich angreifbar. Ein plausibleres Maß für die relative Wichtigkeit von Texten wäre ihre Verbreitung; es ist jedoch nicht praktikabel, wenn nicht zumindest die Auflagenhöhe bekannt gemacht wird (wie z. B. in der Sowjetunion) — die Zahl der Leser läßt sich schon gar nicht ermitteln.

(2) Die Gewichtung stützt sich auf intuitive Urteile. Sie müssen nicht notwendigerweise die Häufigkeit oder Verbreitung der Textkategorien betreffen. So sollte beim London-Lund Corpus der Anteil der Textsorten davon abhängen, welcher Textumfang erforderlich ist, um die Charakteristika der Textsorte deutlich werden zu lassen. Dabei wurde angenommen, daß etwa die grammatischen und stilistischen Eigenheiten der juristischen Fachsprache schon in relativ kurzen Textproben erkennbar werden, während z. B. private Gespräche eine größere Variationsbreite aufweisen und daher mit einem größeren Anteil vertreten sein sollten (s. Quirk/Svartvik 1979, 206).

Wenn die Anteile der verschiedenen Texttypen festgelegt sind, müssen die einzelnen Texte ausgewählt werden. Hier stellen sich erneut die bereits erörterten Probleme: Eine im statistischen Sinne zufällige Auswahl (die zu einer geschichteten Stichprobe führen würde) wird oft nicht möglich sein, so daß willkürlich irgendwelche Texte des betreffenden Typs herangezogen werden, die den Bearbeitern zugänglich sind (womit eine — weniger überzeugende — Quotenstichprobe entsteht).

Schließlich stellt sich die Frage, ob die gewählten Texte vollständig oder auszugsweise ins Korpus aufgenommen werden sollen. Die vollständige Aufnahme ist vor allem dann problematisch, wenn die zu berücksichtigenden Texte von sehr unterschiedlicher Länge sind oder wenn die Zahl der Texte relativ gering ist und die Eigentümlichkeiten des einzelnen Textes somit die Eigenschaften des Gesamtkorpus erheblich beeinflussen. Zweifellos ist es auch angemessener, bei Quotenstichproben oder geschichteten Stichproben die Anteile verschiedener Texttypen nicht nach der Zahl der Texte, sondern nach deren Umfang zu bestimmen. Als Maß hierfür dient üblicherweise die Anzahl der Wörter.

Sollen die Texte nicht notwendigerweise vollständig aufgenommen werden, kann man entweder eine Obergrenze für den Umfang festlegen (und einen längeren Text entsprechend kappen) oder alle Texte gleich lang machen. So wurde erstmals beim Brown Corpus das mittlerweile mehrfach übernommene Prinzip befolgt, alle Texte auf die gleiche Länge von 2 000 Wörtern zu bringen. Bei zu langen Texten ist dann ein Ausschnitt zu wählen, wobei man zwischen Proben von Anfang, Mitte und Ende der Texte abwechseln sollte; bei zu kurzen Texten müssen mehrere nominell als einer gezählt werden.

9. Größe der Stichprobe

Mit einem Umfang von 1 Mio. Textwörtern hat das Brown Corpus ein wiederholt kopiertes Vorbild gesetzt. Ein solches Korpus ist jedoch für die meisten Fragestellungen viel zu klein. Auch bei 4 Mio. Textwörtern kann man keineswegs (wie Engelen 1984, 12) von einem großen Korpus sprechen. So wird man für lexikographische Zwecke wenigstens 50 Mio. Textwörter auswerten müssen, um zu vertretbaren Aussagen zu kommen (vgl. Henne/Weinrich 1976, 347 f.). Die Gründe hierfür liegen in statistischen Gegebenheiten, die hier am Beispiel des Lunder Zeitungskorpus illustriert werden sollen (s. Rosengren 1972).

Der größere Teil dieses Korpus besteht aus 2 476 571 Textwörtern aus „Die Welt“. Die Zahl der verschiedenen Wortformen beträgt 166 484, von denen 93 614 (also 56%)

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM

nur ein einziges Mal vorkommen. Andererseits entfallen auf die 36 häufigsten Wortformen schon ein Drittel aller laufenden Wörter. Die Zahl der verschiedenen Lexeme dürfte bei rund 50 000 liegen, wenn das in einem Korpus von 0,5 Mio. Textwörtern ermittelte Verhältnis von durchschnittlich 3,3 Wortformen pro Lexem (s. Siliakus 1979, 157) übertragbar ist. Zum Vergleich seien hier Daten zitiert, die in Korpora von je 200 000 Textwörtern (Fachsprache Elektronik) für vier Sprachen ermittelt wurden (s. Alekseev 1984, 78):

	versch. Wortformen	versch. Lexeme	Formen/ Lexeme
Russisch	21 648	6 826	3,18
Englisch	10 582	7 160	1,48
Spanisch	13 507	7 564	1,79
Rumänisch	14 292	5 708	2,50

Im „Welt"-Korpus von 2,5 Mio. Wörtern weisen lediglich 2 362 Wortformen eine absolute Häufigkeit von 100 oder mehr auf. Hundert Belege sind jedoch nicht viel, wenn es z. B. darum geht, Bedeutungs- und Gebrauchsunterschiede nachzuweisen (s. Bergenholtz 1980 zum Wortfeld „Angst"), grammatische Zweifelsfälle zu klären (vgl. Bergenholtz/Mugdan 1984) usw. Nur bei relativ unproblematischen Fällen kann man mit einem solchen Minimum auskommen; viele schwierigere Fragen (z. B. bei der Kasusrektion von Präpositionen) lassen sich jedoch nur mit mehreren Tausend Belegen hinreichend sicher beantworten und erfordern Korpora in der erwähnten Größenordnung von 50 Mio. Textwörtern.

10. Ausblick

Wenn man weder wie Itkonen (1976, 65) die Auswertung von Textkorpora für eine „überflüssige Zeremonie" hält noch wie Rainer (1984, 292) den damit verbundenen Zeitaufwand scheut — der übrigens bei einer sinnvollen Informantenbefragung nicht geringer wäre —, so muß man bedauern, daß noch immer zu wenig geeignete maschinenlesbare Korpora zur Verfügung stehen. Textsammlungen wie das Brown Corpus oder das LIMAS-Korpus, die trotz einiger Schwächen als exemplarisch gelten dürfen, sind zu klein, während die größeren Korpora sich auf spezifische Textsorten beschränken oder in Auswahl und Umfang der Texte etwas unausgewogen sind.

Um bessere Voraussetzungen für maschinelle Korpusanalysen zu schaffen, müßten zunächst die bislang erstellten Korpora für wissenschaftliche Zwecke zu vertretbaren Konditionen verfügbar gemacht werden. (Merkwürdigerweise ist derzeit die Verwendung einiger Korpora durch hohe Miet- oder Kaufpreise erschwert, obwohl sie mit öffentlichen Mitteln zusammengestellt wurden.) Hier sind Clearingstellen gefragt, die die vorliegenden Korpora dokumentieren und sammeln (ansatzweise geschieht das bereits beim Norwegian Computing Centre for the Humanities in Bergen für das Englische und beim Institut für deutsche Sprache in Mannheim für das Deutsche).

Allerdings läßt sich aus den vorhandenen Korpora des Deutschen kein hinreichend großes und dabei noch exemplarisches Gesamtkorpus zusammenstellen (s. Mugdan 1985, 205 f.). Es müssen daher neue Korpora aufgebaut werden, wobei einerseits ein großer Bedarf für Sammlungen gesprochener Texte besteht und andererseits eine regelmäßige Aktualisierung der schriftsprachlichen Korpora wünschenswert wäre. Es steht zu hoffen, daß sich sowohl bei Linguisten als auch bei den für die Forschungsförderung zuständigen Instanzen die Erkenntnis durchsetzt, daß die Erstellung eines großen Korpus der deutschen Gegenwartssprache eine vordringliche Aufgabe der heutigen Sprachwissenschaft ist.

11. Literatur (in Auswahl)

Aarts/Meijs 1984 · Alekseev 1984 · Altmann, H. 1981 · Bausch 1971 · Bausch 1975 · Bausch 1979 · Béjoint 1983 · Bergenholtz 1980 · Bergenholtz/Mugdan 1984 · Bergenholtz/Mugdan 1986 · Bergenholtz/Schaeder (eds.) 1979 · Bierwisch 1963 · Bondzio 1980 · Brons-Albert 1984 · Bungarten 1979 · Chomsky 1961 · Chomsky 1965 · De Haan 1984 a · Duden-DUW · Engelen 1984 · Engelien 1979 · Fodor/Garret 1966 · Francis 1979 · Greenbaum 1977 · Greenbaum 1984 · Henne/Weinrich 1976 · HWDG · Itkonen 1976 · Johansson 1982 · Johansson/Jahr 1982 · Labov 1972 · Labov 1975 · Leech/Garside/Atwell 1983 · Levelt 1972 · Lutzeier 1981 · Mackin 1983 · Mugdan 1985 · Müller 1971 · Pfeffer/Lohnes (eds.) 1984 · Quirk/Svartvik 1979 · Rainer 1984 · Rieger, B. 1979 b · Rosengren 1972 · Schaeder 1979 · Siliakus 1979 · Sinclair 1982 · Spencer 1973 · Svartvik 1982 a · Svartvik/Eeg-Olofsson 1982 · Svartvik et al. 1982 · Ulvestad 1979 · Van de Velde 1974 · Van de Velde 1979.

Henning Bergenholtz, Aarhus (Dänemark)/
Joachim Mugdan, Münster
(Bundesrepublik Deutschland)

Brought to you by | ZHAW Zuercher Hochschule für Angewandte Wissenschaften
Authenticated
Download Date | 5/14/18 12:30 PM