KORPUSLINGUISTIK - RÜCKKEHR ZUM STRUKTURALISMUS ODER ERNEUERUNG DER COMPUTERLINGUISTIK?

Wolf Paprotté
Westfälische Wilhelms- Universität Münster

1 Bestandsaufnahme

Die vierte internationale Konferenz über theoretische und methodologische Probleme der maschinellen Übersetzung, vom 25. - 27. Juni 1992 in Montreal, hatte "Empiricist versus Rationalist Methods in MT" zum Thema. Dies unterstreicht die tiefgreifende methodologische U morientierung, die sich seit etlichen Jahren in allen Bereichen des Natural Language Processing und der Computerlinguistik vollzieht. Während sich die sogenannten Rationalisten auf die Repräsentationsmechanismen der verschiedenen Schulen der formalen Linguistik (GB, GPSG, HPSG, CUG, LFG usw.) stützen und Sprache regelbasiert zu erklären versuchen, orientieren sich die sogenannten Empiristen an großen Korpora und stochastischen Sprachmodellen oder konnektionistischen Verfahren, so ließen sich äußerst simplifiziert die Positionen beschreiben. Auf den ersten Blick reißt die Themenstellung einen unüberbrückbaren Graben zwischen "Empiristen" und "Rationalisten" auf und suggeriert, die um sich greifende Korpusorientierung in der Linguistik habe mit einer Rückkehr zum Strukturalismus Bloomfieldscher oder Harrisscher Prägung zu tun. Beides trifft nicht zu.

Noam Chomsky und die von ihm geprägten Entwicklungsstadien der formalen Linguistik haben den Deskriptivismus und Distributionalismus mit dem darunter liegenden Behaviorismus und mit ihm seine typischen Denkansätze unwiederbringlich desavouiert und verdrängt!; was geblieben ist, sind einige methodologische Prinzipien des Empirismus und die Orientierung an authentischen Daten. Auch die kognitiven Wissenschaften und die KI trugen dazu bei, daß Forschungen über Gehirn, Denkprozesse, Lernen und Sprache frühere simple Modelle ersetzten. Deshalb gilt es noch immer als unfein und nicht mainstream linguistics, sich deskriptiv und empirisch zu orientieren. Jedoch wird momentan in der vordersten Linie der Linguistik eine empirische Ausrichtung auf Korpora rehabilitiert und eine seit den Strukturalisten ununterbrochene Strömung korpusbasierten Arbeitens öffentlich anerkannt. "Those who work with computer corpora are suddenly finding themselves in an expanding universe." (Leech 1991: 25) Dies geschieht unter Einsatz neuer probabilistischer Methoden, an Korpora mit einer Größenordnung von 100 Millionen von Textwörtern und mehr, in der Verbindung mit ande-

1 Es ist aufschlußreich, sich in diesem Zusammenhang noch einmal mit Chomsky's Syntactic Structures (1957) und seinem "A Review of Skinner's Verbal Behavior" aus dem Jahre 1959 zu beschäftigen und festzustellen, wie gering der Erkenntnisfortschritt in der Linguistik seit damals geblieben ist.

ren Wissenschaftsgebieten wie der KI, aber ohne Infragestellung des in der Linguistik und KI seit 1957 erreichten Grades an Formalisierung.

Die Neuorientierung auf die maschinelle Verarbeitung authentischer Sprachmassendaten, häufig auch unter weitgehendem Verzicht auf regelbasiertes Vorgehen, erscheint dabei als:

=> Ergebnis der technologischen Entwicklung von Hardware und Software und gestiegener Anforderungen an maschinelle Werkzeuge für die Verarbeitung der Informationsflut in Wissenschaft und Gesellschaft und der mono- und multilingualen elektronischen Dokumentberge in den Büros; und als

=> Ergebnis der erfolgreichen Forschung im Bereich der akustischen Signalverarbeitung und Spracherkennung.

Die heutige Forschungssituation scheint also von der Einsicht in die Notwendigkeit geprägt, authentische Sprachdaten zum Ausgangspunkt linguistischer Forschung und Theoriebildung zu machen. Die von muttersprachlich "kompetenten" Linguisten introspektiv gewonnenen Beispielssätze 2 der Art: (1) *"Jeder Bauer, der einen Esel hat, schlägt ihn auch."* taugen zwar für heuristische Zwecke, können aber die in Korpora belegten Probleme für eine technologische Beherrschung von Sprachmassendaten nicht im Ansatz verdeutlichen. Inzwischen sind die Hardware Voraussetzungen für die automatische Verarbeitung großer Datenmengen selbst in Universitäten leicht zu schaffen; die Daten sind praktisch frei verfügbar ("information glut"; "terrabytes online") und Anwenderbedürfnisse in Forschung und Industrie weisen den Weg zu einer in der Linguistik bisher nicht geleisteten Theoriebildung mittels neuer Methoden.

Es sei hier angemerkt, daß mit den Korpora neuerdings auch das Lexikon als Beschreibungsebene die Aufmerksamkeit der Forschergemeinschaft findet, obwohl es lange Jahre als einer theoretischen Betrachtung unwürdig erachtet wurde. Die lexikologische Orientierung auf das Wort in der Vielfalt seiner grammatischen Bezüge und Eigenschaften ist ebenfalls auf umfangreiche Textsammlungen eher denn als auf Satzbelege angewiesen.

Die Unterschiede zum traditionellen Strukturalismus, ob amerikanischer oder europäischer Prägung, liegen demnach im Einsatz neuer (quantitativer) Methoden, in

2 Ich schlage den Band 26 der Reihe Syntax and Semantics; Syntax and the Lexicon, hrsg. von Tim Stowell und Eric Wehrli (1992) auf und finde dabei folgende Beispielsätze für die kasuistischlinguistische Argumentation:

> On that hill appears to be located a cathedral.
> Stuffing himself night and day eventually killed John.
> John caused Bill to die.
> John caused Bill to die on Sunday by stabbing him on Saturday.
> In the swamp was found a child. In this village was located for many years after the war a church which the Germans had bombed.
> The toys amused the children.

Diese Sätze beeindrucken durch Künstlichkeit und Mangel an Kontext. Man vergleiche damit die folgenden Beispiele aus dem Münsteraner Korpus:

> Im Gegensatz zu den 'Helden der Arbeit', die es reichlich gab, waren die 'Helden der DDR' hand verlesen.
> So werden in der Troisdorfer Sortieranlage die Müllsäcke - Papier wird extra behandelt - aufs Band gekippt, dort mit Hilfe von Förder- und Siebtechnik vorsortiert und schließlich von den Arbeitern handverlesen.
> Dann blickt der Fußballehrer, der sein Leben lang in Benningen und im acht Kilometer davon entfernten Affalterbach seßhaft war, durch ein Teleskop in die Galaxien.
> Die Nachricht, ein Ostberliner Linguistenkollektiv mit Joachim Schildt an der Spitze habe seine schöpferische Tätigkeit bei der Akademie der Wissenschaften der DDR planmäßig beendet und sich als Masseninitiative dem Institut für einheimische Sprache (IdS) in Mannheim angeschlossen, schreckt allenfalls die Toten.

den Voraussetzungen und Bedingungen ihrer technologischen Umsetzung und Anwendung, in der Größendimension der zu berücksichtigenden Daten, und darin, neue Methoden mit bekannten Formalismen für die Repräsentation linguistischer Information zu verbinden. Anders gesagt, der Trend zu Korpora und neuen stochastischen Verfahren ist kein restaurativer Prozeß mit Blick auf den Strukturalismus sondern eine empirische Neuorientierung der Linguistik und Computerlinguistik, als deren Ergebnis zunächst hybride Grammatiken mit statistischen und Regelkomponenten zu erwarten sind.

2 Die historische Perspektive

Die Grundpositionen des amerikanischen Deskriptivismus, wie in seinem Buch Language formuliert, legte Bloomfield, der sich selbst auf Arbeiten des Psychologen A.P. Weiss stützte und Entwicklungen der Junggrammatiker und Wundts aufgriff, bereits 1926 in seinem Artikel "A Set of Postulates for the Science of Language" fest.

Form und Bedeutungseigenschaften einer Sprechhandlung (act of speech) oder Äußerung (utterance) werden dort als beobachtbare Lautäußerungen und Stimulus - "Reaktionsbündel begriffen und über die Begriffe "gleich" oder "verschieden" zu "Formen" oder "Bedeutungen" zusammengefasst. Interessanterweise definiert Bloomfield eine Sprache als "totality of utterances that can be made in a speech community" und stellt fest, daß Linguisten Vorhersagen als Erklärung abgeben; der Topos der Unabschließbarkeit der Sprache treibt diese Gedanken schließlich noch etwas weiter. Bereits dort macht Bloomfield dieses Problem einer rein deskriptiven Linguistik deutlich und lässt implizit neben dem Korpus Informantenbefragung und "the investigator's own language" zu (cf. Joos 1957: p. 26 f.) [3]

Sieht man sich diese theoretische Position in der Praxis der Bloomfieldschen deskriptiven und vergleichenden Studien zur AIgonquian Sprachfamilie an, so findet man folgende Datengrundlagen und Methoden der Datengewinnung:

1. mittels direkten Kontakts und durch Beobachtung von Informanten, bzw. durch Elizitation von Äußerungen vom Linguisten gewonnene primäre Korpusdaten als Grundlage für die Beschreibung;

2. philologische Interpretation von "Sprachdenkmälern" (hier z.B. Aufzeichnungen von Missionaren, Händlern und Landvermessern) ;

3. primäre Daten als Eigenaufzeichnungen von geschulten muttersprachlichen Informanten;

4. Texte, die von Ethnologen und Anthropologen gesammelt worden waren.

Man kann (1)-(4) als Teilkorpora eines Korpus auffassen, bemerkt aber sogleich, daß sie wegen der Verschiedenheit ihrer Erstellung in Güte und Zuverlässigkeit stark variieren, insbesondere deshalb, weil Bloomfield unter synchroner Perspektive eine phonologische Beschreibung, unter diachronischer Perspektive die Gewinnung von Gesetzmäßigkeiten des Lautwandels anstrebte. Für beides waren Korpora dieser Art nicht zuverlässig genug. Der Datenumfang war naturgemäß gering, so daß das zur Verfügung stehende Gesamtkorpus

3 "4. Def. The totality of utterances that can be made in a speech-community is the language of that speech-community. We are obliged to predict; hence the words "can be made". We say that under certain stimuli a Frenchman (or Zulu, etc.) will say so-and-so and other Frenchmen or (Zulus, etc.) will react appropriately to his speech. Where good informants are available, or for the investigator's own language, the prediction is easy;
elsewhere it constitutes the greatest difficulty of descriptive linguistics." (Bloomfield 1926, in Joos 1957: 26 f.)

für die gesamte Sprachfamilie (Fox, Cree, Menomini, Ojibwa) kaum mehr als 100 000 Textwörter umfaßt haben mag.

Es ist wichtig hervorzuheben, daß diese Korpusmaterialien häufig aus Wortlisten bestanden, die entweder als Minimalpaare phonologische Kontraste belegten oder als primitive, zweisprachige Wörterbücher gedient hatten. Es kann also nicht davon die Rede sein, daß diese Materialien eine wohlstrukturierte Auswahl aus der Sprache oder "repräsentativ" für die zu beschreibende Sprache waren. Für die manuelle Auswertung und das auf Phonologie bzw. Morphophonemik gerichtete Erkenntnisinteresse waren diese Korpora jedoch ausreichend, obwohl mit Sicherheit grammatische Kerneigenschaften und große Teile des Lexikons nicht adäquat belegt waren. Materialsammlungen diesen Typs, kennzeichnend für die strukturalistische Arbeitsweise, kann man wohl als Korpora der ersten Generation bezeichnen.

Erst Harris (1951) formulierte eine pointierte Orientierung auf das Korpus: "Investigation in descriptive linguistics consists of recording utterances in a single dialect and analyzing the recorded material. The stock of recorded utterances constitutes the corpus of data, and the analysis which is made of it is a compact description of the distribution of elements within it. The corpus does not, of course, have to be closed before analysis begins" (1951:12). Seine wichtigen Neuerungen betrafen vor allem eine Menge distributionalistischer Verfahren der Segmentierung und Klassifikation, die gleichermaßen auf allen linguistischen Beschreibungsebenen angewendet werden sollten und mit wenigen statistischen Grundannahmen einher gingen4.

Für die linguistische Analyse waren vor allem minimal unterschiedliche, bzw. teilweise gleiche Umgebungen (environments) in Korpusbelegen von Interesse, so daß die Erstellung, zumindest aber Erweiterung eines Korpus, noch immer wesentlich auf Informantenbefragung beruhte (p.368). Wie bei Bloomfield ist ein Korpus auch bei Harris eine noch im wesentlichen durch den Begriff der utterance bestimmte Sammlung linguistischen Materials, da Phonologie und Morphologie im Zentrum der Analyse standen, obschon die syntaktische Analyse, wie dann von Chomsky betrieben, bereits programmatisch formuliert ist.

Bereits für morphologische Analysen stellte Harris Probleme der Größe eines Korpus und implizit der Möglichkeit seiner manuellen Auswertung fest:

"In many languages, several hundred hours of work with an informant would yield a body of material containing all the different environments (over short stretches of speech) of the phonemic segments. If the operations of 3-11 are carried out for one such corpus of the language, and then again for another such corpus of that language, no difference in relevant data would appear. It would require a corpus many times this size to give us almost all the morphemic segments of the language, ... That is, only a very large corpus would permit of the extraction of so many morphemes that no matter how much more material we collect in that language, we would hardly ever find any new morphemic segment. ...

However, if we could state all the morphemes, each with its exact distribution, for a corpus consisting of all the utterances in the language over a period, showing that a given morpheme has not occurred in a given environment in any utterance of that language, we could still not be able to predict with high probability that that morpheme might not appear in the given environment,

4 Vgl. Harris (1951), p. 372 und p. 13. Zum einen betont er, daß das Korpus adäquat sein müsse, zum anderen hebt er hervor, daß ein Korpus nur dann als deskriptive Stichprobe einer Sprache angesehen werden könne, wenn zwei Analysen, gestützt auf unterschiedliche Stichproben-Korpora, zu denselben Ergebnissen geführt haben.
Harris weist an dieser Stelle (p. 13) auch auf die Abhängigkeit von Korpusgröße und Untersuchungsziel hin, z.B. genügt ein kleines Korpus für die phonologische Analyse.

for the first time in the history of the language, in some new utterances soon to be said". (Rarris 1951: 254f).

Rarris spricht hier den in der heutigen Korpus Diskussion kaum berücksichtigten Gedanken an, daß Korpora nur eine zeitlich begrenzte Gültigkeit für die synchrone Beschreibung haben und daß darüber hinaus sprachliche Innovation dafür sorgt, daß eine Sprache eine nie abschließend extensional beschreibbare Entität ist. Nicht nur in dieser partiellen Vorwegnahme des Kompetenzbegriffs nimmt Rarris Gedanken der generativen Transformationsgrammatik vorweg. Es fällt z.B. auch auf, daß aus strukturalistischer Zeit keine Sammlung von Sprachmaterialien in einer Form existiert und überliefert ist, wie man sie heute etwa als Brown Korpus kennt. Die Unterschiede, mit anderen Worten, zwischen der Methode der heuristischen Beispiels- und Gegenbeispielsfindung konstruierter Sätze und den Korpora der Strukturalisten sind rein quantitativ: die in der gTG benutzten Inventare solcher Sätze sind vergleichbar einem auf das absolute Minimum reduzierten Korpus, welches nach Rarris ohnehin nur so umfassend zu sein brauchte, daß außer Innovationen nichts wesentlich Neues an sprachlichen Strukturen in weiteren Korpora entdeckt werden konnte. Dieselbe Überlegung steckt hinter dem Argumentieren mit Beispielssätzen. Dieser Gedanke liest sich bei Rarris folgendermaßen:

"For the linguist, analyzing a limited corpus consisting of just so many bits of talking which he has heard, the element X is thus associated with an extensionally defined class consisting of so many features in so many of the speech occurrences in his corpus. However, when the linguist offers his results as a system representing the language as a whole, he is predicting that the elements set up for his corpus will satisfy all other bits of talking in that language. The element X then becomes associated with an intensionally defined class consisting of such features of any utterance as differ from other features, or relate to other features in such and such a way." (1951:17)

Empirischer Rationalismus, könnte man sagen: die methodologische Differenz zum Chomskyschen Mentalismus bis ca. Mitte der 60er Jahre bleibt gering.

Trotz Chomskys vernichtender Kritik am Behaviorismus (Chomsky 1959) und trotz der Veröffentlichung seiner revolutionären Studie Syntactic Structures begannen Francis und Kucera gegen den vorherrschenden Trend im Jahre 1959 die Arbeit am Brown University Standard Corpus of Present-day American English, welches 1964 für die universitäre Forschung verfügbar wurde. Das Brown Korpus hatte insgesamt 1 Million Textwörter (word tokens). Es enthielt 500 Texte mit je 2000 Textwörtern, die sich auf 15 Textkatego-rien oder Genres verteilten.
Alle Texte stammten aus Publikationen des Jahres 1961 (Francis 1974). Ebenfalls im Jahre 1959 planten Randolph Quirk und Jan Svartvik ihr "Survey of English Usage" Korpus, welches gesprochene und geschriebene Texte des britischen Englisch enthalten sollte. Durch Informantenbefragung sollten im Korpus nicht belegte Strukturen oder für eine umfassende Beschreibung des Englischen nicht ausreichend dokumentierte Merkmale elizitiert werden (Quirk/ Svartvik 1979: 206). Im Gegensatz zum Brown Korpus war der Survey zunächst nicht als maschinenlesbares Korpus gedacht, zielte jedoch bereits auf ein Teilkorpus der gesprochenen Sprache. Das London - Lund Korpus mit ca. 435 000 Textwörtern, stellt die andere Hälfte des Survey of English Usage für das gesprochene Englisch dar und ist inzwischen maschinenlesbar vorhanden.

Im Jahr 1970 begann an der University of Lancaster die Arbeit an einem Korpus des britischen Englisch mit derselben Struktur und zusammengestellt nach denselben Auswahlprinzipien wie das Brown Korpus. In Zusammenarbeit mit den Universitäten Oslo und Bergen wurde die Textsammlung 1978 abgeschlossen.

Das Brown und das LOB Korpus sowie der Survey sind typische Vertreter von Korpora der zweiten Generation: sie sind für die automatische Auswertung im Rechner ausgelegt; beide haben eine Größenordnung von ca. 1 Million Textwörtern. Dies spiegelt sowohl die Probleme der Datenerfassung per Tastatur wie Begrenzungen der maschinellen Verarbeitung wider, die in dieser Zeit beim damaligen Stand der Hardware Entwicklung existierten. Gegenüber den Materialsammlungen der ersten Generation des Deskriptivismus fällt vor allem auf, daß Einsichten in die Variabilität einer Sprache bezogen auf Textsorte und Thematik sich in dem Versuch der Begründer der Korpuslinguistik zeigen, nach Genre, Textsorte und subjektiv bewerteter Häufigkeit diversifizierte Textsammlungen zu erstellen. Die Stichprobenziehung erfolgte demnach nicht nach strengen Maßstäben der statistischen Methodenlehre. In ihrer Größe unterschieden sich Materialsammlungen der ersten und zweiten Generation ungefähr um den Faktor 10.

Die dritte, heutige Generation von Korpora ist vor allem durch einen enormen quantitativen Zuwachs ca. um den Faktor 100 und eine enorme Verbesserung der Verarbeitungsmöglichkeiten im Rechner gekennzeichnet. Die Sammlung englischer Texte, die unter Leitung von John Sinclair in Birmingham als Grundlage für das beeindruckende monolinguale Cobuild Wörterbuch diente, umfasste mehr als 20 Millionen Textwörter und strebt wie bereits das Brown Korpus eine strukturierte Auswahl aus gedrucktem Prosamaterial an (Renouf 1987). IBM verfügt, wie man hört, am Thomas J. Watson Forschungszentrum über ein Korpus von 60 Millionen Textwörtern neben anderen Korpora (vgl. Garside et al. 1987:6). Die ACL/ DCI hat nach einer großangelegten Initiative ein Korpus von ca. 80 Millionen Textwörtern nach dem Kriterium der Verfügbarkeit gesammelt und beabsichtigt, diese Textsammlung zu erweitern und nach Textsorten und Themenbereichen ausgewogener zu gestalten.

Obwohl dies nicht die Stelle ist, mehr als nur die Umrisse der gegenwärtigen Korpusaktivitäten zu skizzieren, seien auch einige deutsche Aktivitäten erwähnt. Neben dem kleinen und betagten LIMAS Korpus der Universität Bonn, einem deutschen Zeitungskorpus der zweiten Generation mit 1 Million Textwörtern, ist vor allem das IdS Korpus in Mannheim als Korpus der dritten Generation zu erwähnen, bei dem auch die automatische Analyse relativ weit fortgeschritten ist. In Münster existiert ein deutsches Korpus (der dritten Generation) von nunmehr 100 Millionen Textwörtern, bestehend aus fast zwei Jahrgängen der FAZ und der ZEIT und ergänzenden Materialien. Daneben wurden in Münster umfangreiche Materialien für das Spanische (40 Millionen); das Neugriechische (8 Millionen), das Französische (25 Millionen) und eine Reihe englischer Korpora gesammelt und maschinell aufbereitet. In Münster wird eine ausgewogenere Struktur der Textsammlungen angestrebt, im Bewußtsein der Tatsache, daß die statistische Modellierung des Begriffs eines "repräsentativen" Korpus schwieriger ist als zunächst vermutet (vgl. Rieger 1979).

Wie auch Leech (1991) bemerkt, ist die Größe eines Korpus nur ein Qualität stiftendes Merkmal unter anderen Merkmalen der Datenbasis. Die Diversifizierung nach Textsorten, Themenbereichen, Fachsprachen, möglicherweise nach Parametern, die die Soziolinguistik bereitstellt, vor allem die Sammlung gesprochener Sprache bleiben Desiderate der meisten Korpora der dritten Generation, trotz der inzwischen verbesserten Möglichkeiten, maschinenlesbare Daten aus elektronischen Publikationen oder in Netzen verfügbare Daten für den Korpusaufbau direkt einzusetzen.

3 Motive der Korpusorientierung und ihre Methoden

Der heute recht verbreitete Einsatz von Computerkorpora ist also in den letzten drei Dekaden aus den Arbeiten von Francis & Kucera und Quirk und Mitarbeitern entstanden und fast gleichzeitig an der Universität Lancaster mit dem LOB Korpus und an einer Reihe weiterer Zentren betrieben worden, zu denen Lund, Amsterdam, Nijmegen und Birmingham gehören. Die Universitäten Oslo und Bergen koordinierten die Informationsdistribution über Korpusforschung im Rahmen von ICAME. Die sich hierin manifestierende Tendenz zum Empirischen resultiert einerseits aus der Tatsache, daß sich angesichts des Flaschenhalses großer Lexika in sprachverarbeitenden Systemen und des nicht eingelösten Versprechens schneller Fortschritte in der Sprachverarbeitung, wie sie von der formalen Linguistik immer wieder gemacht wurden, sprachtechnologisches Engineering auf die notwendig schnell zu lösenden Probleme der Sprachmassendatenverarbeitung und des Informationsmanagements besinnt. Andererseits beruht sie auch darauf, daß es trotz der so überzeugend ausgebreiteten Argumentationen in formalen Modellen und Theorieentwürfen keinen nennenswerten explanatorischen Zuwachs, etwa vom ursprünglichen Theorieentwurf Chomskys 1957, über die Standardtheorie von 1965, die Extended Standard Theory zu G & B gegeben hat. Vielmehr beschleunigte sich in der letzten Dekade das Tempo, mit dem neue partielle Theorieansätze entwickelt wurden, so daß heute neben GB, LFG, GPSG und HPSG, Tree Adjoining Grammar, Word Grammar etc. noch viele weitere Modelle im Schwange sind. Eine scheinbar monolithische Linguistik mit festen empirisch untermauerten Überzeugungen und einem umfassenden Theorieentwurf steht in Frage. Hier deutet sich möglicherweise ein Paradigmenwechsel der Linguistik im Kuhnschen Sinne an, dessen allgemeines Kennzeichen ein ausgeprägter Lexikalismus mit gleichzeitiger Orientierung auf große, authentische Datenmengen ist.

Wohin also sollte und könnte sich der linguistische Zeitgeist wenden, wenn die Überzeugungskraft der mentalistischen Position abnimmt, die Rückkehr zum Strukturalismus unmöglich ist und zugleich die versprochenen Erkenntnisfortschritte weder in der Theorie noch in ihren Implementierungen in der Sprachtechnologie erbracht werden?

Es scheint heute, als bestimme die Leistungsfähigkeit eines im Rechner implementierten Grammatikmodells seine Validität. Scheitert der Rechner an authentischem Material, wird das Modell verworfen. Korpora sind von daher dazu bestimmt, Testfall für die Leistungsfähigkeit von Systemen zu sein und systematisch abfragbare Daten bereitzustellen. Diese natürlich vorkommenden Sprachdaten ersetzen die introspektive Evidenz von Beispielsätzen. N ur in dieser Hinsicht hat sich also die programmatische Orientierung des Strukturalismus auf ein Korpus realer Daten durchgesetzt. Leech (1991:9) weist im übrigen darauf hin, daß über Korpora und neue probabilistische Methoden robuste Systeme für die Verarbeitung natürlicher Sprache entwickelt werden können, die die bisher nur auf "small scale problems" ausgelegten regelbasierten Techniken ersetzen werden.

Da der Korpuslinguistik interdisziplinäre Bezüge inhärent sind, die sie mit allen Zweigen der Informationswissenschaften, der Künstlichen Intelligenz, der Signalverarbeitung etc. verbinden, forcieren auch diese Bezüge den Aufbau von Korpora und ihren Einsatz. Dabei stellt die grammatische Analyse des Korpus durch automatische Annotierung oder syntaktisches Parsen ein Beispiel für die Informationsextraktion von lexikalen und grammatischen Merkmalen dar. In dieser Hinsicht liegt die wohl wichtigste Motivation für die Arbeit an annotierten Korpora in ihrer Falsifikationsfunktion für Grammatikmodelle.

In einer zweiten Hinsicht jedoch sind annotierte Korpora, Verbindung von Rohdaten mit analytischen Marken, Wörterbuch und Grammatikersatz und haben zugleich eine Trainingsfunktion für probabilistische Sprachmodelle, die sich bereits in der Spracherkennung als fruchtbar erwiesen haben.

4 Die automatische Analyse großer Korpora

Unter Analyse von Korpora wird zunächst automatische Analyse mit unterschiedlich großen Anteilen menschlicher Intervention verstanden, z.B. bei der Vorgabe der Analyseparameter oder beim manuellen Posteditieren automatischer Analyseergebnisse. Dabei gibt es zwei (durchaus interdependente) Analyserichtungen: die eine zielt auf die Extraktion von Information, die letztlich in eine vom Korpus unabhängige Daten- bzw. Wissensbank eingeschrieben wird; die zweite fügt dem Rohtext Analyseergebnisse hinzu und führt zu annotierten Korpora, in denen Datum und analysiertes Datum zugleich vorkommen. In jedem Fall weicht die Methode von der traditionellen distributionalistischen Analyse ab, die sich an mengen theoretischen Konzepten orientiert wie

=> den Äquivalenzklassen frei variierender Elemente x, y mit gleicher Distribution: nach dem Harrisschen Wiederholungs- oder Paartest sind x, y miteinander austauschbar und variieren frei sofern sie von Informanten als "gleich" beurteilt werden;

=> der komplementären Distribution: die Elementen x, y stehen in komplementärer Distribution, wenn die Mengen der Umgebungen von x und y kein gemeinsames Element haben;

=> den distinkten Elementen x, y, die in der Relation der Opposition stehen und in ähnlichen (partiell gleichen) oder sogar gleichen Umgebungen vorkommen, wie z.B. bei Minimalpaaren oder kontrastierenden Paaren (vgl. Paprotté 1974).

Automatische Analysen von Korpora richten sich auf die üblichen linguistischen Beschreibungsebenen. Dabei werden analytische Techniken eingesetzt, die auf statistischen Methoden und den Axiomen der Wahrscheinlichkeitstheorie beruhen. Ein Vorteil dieses Ansatzes besteht darin, daß Regularitäten und Idiosynkrasien, grammatische und ungrammatische aber noch verständliche, in der Intention des Sprechers rekonstruierbare Produktionen robust verarbeitet werden können (Sampson (1987) in Garside et al. 1987:17fl} Mit ihm wird die dichotomische Unterscheidung von grammatisch und ungrammatisch durch ein robustes System aufgehoben, welches Häufigkeiten, Verteilungen, Varianz für Wörter, Phrasen, Sätze und Texte berechnet. "Within our approach, by contrast, the concept of a grammar which defines 'all and only' the forms of the language plays no part at all. Our algorithms deal only with relative frequencies; they recognize no absolute distinctions between "well-formed" and "ill-formed". (Sampson in Garside et al 1987:20) Bereits die einfachsten Häufigkeitsanalysen erlauben interessante Einsichten in den Wortschatz einer Sprache, den tatsächlichen Gebrauch von orthographischen, morphologischen, morphosyntaktischen Varianten, in Rektionsverhältnisse oder Wortartwechsel, z.B. einer subordinierenden Konjunktion in eine koordinierende.

Zugegebenermaßen sind solche Aussagen relativ uninteressant, weil sie nur wenig zur "Gesamtsicht" einer Sprache, zum Grammatikmodell beitragen, wie es in den gängigen Ansätzen vom Anspruch her vorgelegt wird. Zumindest aber erlauben solche Methoden Aussagen darüber, "was der Fall ist" und bleiben so in der Tat dem strukturalistischen Wissenschaftsideal verhaftet.

Der eigentliche Reiz der Arbeit mit Korpora liegt jedoch im Einsatz informationstheoretischer Methoden und Maße, die

quasi selbstorganisierend relevante Kategorisierungen finden und lernfähig sind (Smyth, Goodman 1992). Kollokationen oder phraseologische Einheiten lassen sich z. B. mit dem Maß der mutual information (oder des association ratio) bestimmen (vgl. Church/Hanks 1990), welches die Wahrscheinlichkeit zweier Wörter $w1, w2$, isoliert vorzukommen, mit der Wahrscheinlichkeit vergleicht, sie zusammen zu beobachten:

$$I_{Kovorkommen}(w1, w2) = log \frac{P_{Kovorkommen}(w1, w2)}{P_{isoliert}(w1, w2)}$$

Während das Maß an mutual information die Grade an "Assoziation" von Wörtern angibt, kann mit dem t–score ein Maß für die Unterschiedlichkeit von Wörtern gegeben werden; beim Einsatz beider Maße müssen die Ergebnisse nachgearbeitet werden. Kompliziertere Methoden liegen den informationtheoretischen Sprachmodellen zugrunde, mit denen z.B. die Wahrscheinlichkeit eines Wortes geschätzt wird, gegeben ein oder mehrere Vorgänger in der linearen Folge der Kette von Wörtern. Es werden also Bigramme, Trigramme, etc. berechnet.

An zwei Analysearten, dem syntaktischen Annotieren und dem probabilistschen Parsen erweist sich der neue informationstheoretische Ansatz als fruchtbarer und zuverlässiger als jedes bisher bekannte regelbasierte Verfahren. Beim probabilistischen Annotieren erhält jedes Textwort (word token) im Korpus eine Wortartmarkierung bzw. eine disjunkte markierte Menge von tags, deren relative Häufigkeiten bekannt sind, wenn morphosyntaktische Ambiguität vorliegt. Aus dem Kontext der Vorgänger und Nachfolger Mengen von tags ergeben sich mögliche Pfade mit unterschiedlichen Übergangswahrscheinlichkeiten. Das System erzeugt eine Matrix von jedem tag mit jedem anderen und weist dann die Wortartmarkierung zu, deren Pfad die höchste summierte Übergangswahrscheinlichkeit aufweist. Eine gegebene Folge von Wörtern $w1, w2, w3, w4$ sei mit eindeutigen tags $t1, t4$; oder ambigen tags $t21, t22, t31 t32$ versehen,

w1,	w2,	w3,	w4
t1,	t21,	t31,	t4
	t22,	t32	

es werden dann die Wahrscheinlichkeiten der Folgen

t1,	t21,	t31,	t4
t1,	t21,	t32,	t4
t1,	t22,	t32,	t4
t1,	t22,	t31,	t4

aus einem Trainingskorpus berechnet und die Wahrscheinlichkeit für jeden ambigen tag auf der Grundlage der beobachteten bedingten Übergangswahrscheinlichkeiten vorhergesagt (Garside in Garside et al 1987:39).

Ein solcher Disambiguierungsmechanismus benötigt also Angaben zur Übergangshäufigkeit zwischen Paaren von tags und bezieht diese aus einem bereits annotierten Teilkorpus. Für das Englische wird oft auf die annotierte Version des Brown Korpus zurückgegriffen. Es ist erstaunlich, daß ein so einfacher Mechanismus hohe Trefferquoten von korrekt zugewiesenen tags erreicht: Das in Lancaster eingesetzte System Claws weist 96%–97% korrekte tags zu, ohne daß grammatisches Regelwissen bei der Zuweisung der Wortartmarken eine Rolle spielte.

Die zweite probabilistische Technik hat mit syntaktischem Parsen und dem Auffinden der/einer korrekten etikettierten Phrasenstruktur zu tun. Als korrekter Baum gilt dabei derjenige aus der Menge aller logisch möglicher Strukturbäume, welcher eine einfache Funktion aller Werte maximiert, die sich aus den Zuordnungen von Folgen von Tochterknoten zu einem sie dominierenden Mutterknoten ergeben. Die individuellen Werte für solche Zuordnungen werden empirisch aus den beobachteten Häufigkeiten der Folgen von Tochterknoten

abgeleitet. Da zunächst alle Folgen von Kategorien möglich sind, können auch mit diesem System auch "ungrammatische" Äußerungen analysiert werden. Auch hier wird der unmittelbare Vorgänger und der unmittelbare Nachfolger, nicht aber die gesamte vorhergehende oder nachfolgende Struktur betrachtet; es handelt sich also um ein Markov Modell der ersten Ordnung, welches jedoch einfach durch Betrachtung zweier Vorgänger und zweier Nachfolger in ein Markovmodell der zweiten Ordnung überführt werden kann. Die automatische PS-Analyse mittels rein statistischer Methoden setzt wie schon das Annotieren ein Trainingskorpus voraus, welches im Regelfall durch umfangreiches manuelles Posteditieren erst erstellt werden muß. Es sei angemerkt, daß Atwell (1984) für die syntaktische Analyse auf eine kontextfreie Phrasenstruktur Grammatik mit statistischen Elementen, also auf ein hybrides Sprachmodell, zurückgreift und damit den Suchraum für mögliche Strukturbäume stark einschränkt.

Sampson (Garside et al. 1987:22) bemerkt, dieses System habe höheren Anspruch auf psychische Realität und komme dem tatsächlichen menschlichen Funktionieren näher als die Regelsysteme der kognitiven Linguistik. Zumindest aber ist es ein entscheidender Schritt fort vom Strukturalismus, der grammatische Regeln weder in seiner God's Truth noch in seiner Hokus Pokus Variante auf statistische Gesetzmäßigkeiten stützte, sondern sich meist auf operationalistisch abgesicherte Kategorienbildung berief (Hockett 1948).

Korpusbasierte Lexikographie ist ein weiterer Anwendungsfall neuer Methoden, der auf eine grundlegende deskriptive und empirische Orientierung aufbaut. Es ist nach den Arbeiten der Birminghamer Cobuild Gruppe um John Sinclair deutlich geworden, wie erfolgreich und fruchtbar maschinenlesbare Korpora- für die Gewinnung authentischer "keywords in context" als Belege und Beispiele für den Wort gebrauch eingesetzt werden können. Dieser Anwendungsfall kann nun in Verbindung mit den oben genannten Methoden weitgehend automatisiert werden. In einem annotierten Korpus werden unter Einsatz des association ratio / mutual information Maßes statistisch signifikante Kovorkommen von Wörtern (Wortformtypen) berechnet. Zusammen mit einem Lemmatisierungsprogramm bzw. einem morphologischen Parser werden alle morphosyntaktischen Wortformtypen einem Lemma zugeordnet und kategorisiert.

Semantische Analysen und Diskursstudien bestimmen wahrscheinlich die nächsten Anwendungsschritte in der Annotierung von Korpusdaten. Da semantische Beschreibung bisher vor allem darauf aus war, die Bedeutungsdefinitionen existierender Wörterbücher zu parsen (Vossen et al 1988), kommt der Entwicklung geeigneter Techniken für die Extraktion semantischer Information aus Korpora große Bedeutung zu.

5 Mittelfristige Ziele in der Korpuslinguistik

Von John Sinclair stammt der Begriff des Monitorkorpus. Es handelt sich dabei um ein großes Korpus der dritten Generation mit ca. 500 Millionen Textwörtern, in welches kontinuierlich neue Daten aufgenommen und aus welchem beständig Materialien wieder ausgeschieden werden. Ein solches Korpus ist topisch stratifiziert und enthält im übrigen entsprechend große Anteile gesprochener Daten. Ein Teilmenge davon müßte als Trainingskorpus annotiert werden, der Rest würde automatisch annotiert und geparst.

Die "tags" oder Marken, mit denen die einzelnen Wörter gekennzeichnet werden kann man sich als Attribut-Merkmals-Matrizen, wie aus der HPSG bekannt als typed feature structures, vorstellen. In ihnen wird komplexe lexikale Information so repräsentiert, daß mit dem annotierten Korpus eine lexikale Datenbank als

Wörterbuch des Sprachgebrauchs vorliegt. Mit geeigneten Werkzeugen kann problemlos quantitative und kategoriale Information extrahiert werden.

Werden nun, wiederum zunächst für eine Teilmenge des Korpus, pro Satz die PS-Struktur und möglicherweise sogar satzübergreifende anaphorische Bezüge dargestellt, liegt zunächst für Trainingszwecke eines probabilistischen Parsers eine "treebank" als Trainingskorpus vor. Der lernfähige Parser kann anschließend für das Parsen neuer Materialien eingesetzt werden. Ein solcherart annotiertes und syntaktisch analysiertes Korpus enthielte implizit eine realistische Grammatik einer Sprache und könnte die Grundlage für die Entwicklung realistischer Regelapparate werden.

Bibliographie

Aijmer, Karin; Altenberg, Bengt (eds) (1991), English Corpus Linguistics. Studies in Honour of Jan Svartvik. London & New York: Longman.

Aarts, J.; Meijs, W. (eds) (1984), Corpus linguistics: re cent developments in the use of computer corpora in English language research. Amsterdam: Rodopi.

Atwell, Eric S. (1988), "Transforming A Parsed Corpus into a Corpus Parser." In Kytö et al. 1988: pp. 61 -69.

Atwell, Eric S.; Leeeh, G. N.; Garside, R.G. (1984), "Analysis of the LOB Corpus: Progress and Prospects". In Aarts & Meijs (1984), 41 - 52.

Bergenholtz, Henning & Schaeder, Burkhard (eds.) (1979), Empirische Textwissenschaft. Aufbau und Auswertung von Text- Corpora. Königstein/T: Scriptor.

Bloomfield, Leonard (1926), "A Set of Postulates for the Science of Language" Language 2: 153 -164; repr. in Martin Joos (ed), Readings in Linguistics I, The Development of Descriptive Linguistics in America 1925 - 56. University of Chicago Press: Chicago , London 1957: 26 - 31.

Chomsky, Noam (1957), Syntactic Structures. The Hague: Mouton.

Chomsky, Noam (1959), "A Review of B.F. Skinner's Verbal Behavior". Language 35,1: 26 - 58.

Church, Kenneth W.; Ranks, Patrick (1989), "Word Association Norms, Mutual Information, and Lexicography." In Proceedings of the 27th Annual Meeting of the ACL, pp. 76 - 83.

Francis, Nelson W. (1974), "Problems of Assembling and Computerizing Large Corpora." In Bergenholtz / Schaeder eds.(1979: 110123).

Garside, Roger; Leeeh, Geoffrey & Sampson, Geoffrey (eds) (1987), The Computational Analysis of English. A Corpus-Based Approach. London, New York: Longman

Harris, Zellig S. (1951), Structural Linguistics. Chicago & London: University of Chicago Press.

Hockett, C.F. (1948), A Note on Structure. IJAL 14:269 - 271.

Merja Kytö, Matti Rissanen (1988), "The Helsinki Corpus of English Texts: Classifying and Codifying the Diachronie Part." In: Kytö et al. 1988: 170 - 178

Merja Kytö, Ossi Ihalainen, Matti Rissanen (eds.), Corpus Linguistics, Hard and Soft. Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi 1988.

Leeeh, Geoffrey (1991), "The State of the Art in Corpus Linguistics". In Aijmer, Altenberg (eds.), p. 8 - 29.

Paprotté, Wolf (1974), Zur Entwicklung und Kritik des Amerikanischen Strukturalismus Bloomfields und seiner Schule mit besonderer Berücksichtigung der distributionalistischen Phonologie. Diss. TU Berlin.

Quirk, Randolph & Svartvik, Jan (1979) "A Corpus of Modern English." In Bergenholtz / Schaeder eds.(1979: 204..,.. 218).

Renouf, Antoinette (1987), "Corpus Development." In J.M. Sindair (ed.) (1987), Looking Up. London & Glasgow: Collins: pp. 1 - 40.

Rieger, Burghard (1979), "Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung." In Bergenholtz / Schaeder eds.(1979: 52 - 70).

Smyth, Padhraic; Goodman, Rodney M.
(1992), "An Information Theoretic Approach to Rule Induction from Databases". IEEE Transactions on Knowledge and Data Engineering Vol. 4, 4 : 301 - 316.

P. Vossen, M. den Broeder, W. Meijs (1988),
"The 'Links'- Project: Building A Semantic Database For Linguistic Applications. In: Merja Kytö, Ossi Ihalainen, Matti Rissanen (eds.) (1988): pp. 279 - 293.