LDV - FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-Forum 9.2 (1992)
Forum der Gesellschaft für Linguistische Datenverarbeitung e.V.

Herausgeber
Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)
Anschrift: Prof. Dr. Burghard Rieger, Universität Trier, FB 11: LDV /CL, D-5500 Trier, Postfach 3825, Tel.: (0651)201-2272/2270; Fax (0651)201-3946; Email: rieger@ldv0l.Uni-Trier .de

Redaktion
Burghard Rieger, Roland Fraese, Amancio Kolompar

Wissenschaftlicher Beirat
Dr. Karin Haenelt, Hans Haugeneder, Prof. Dr. Peter Hellwig, Prof. Dr. Gerhard Knorz, Prof. Dr. Jürgen Krause, Prof. Dr. Winfried Lenders, Dr. Dietmar Rösner

Erscheinungsweise
Zwei Hefte im Jahr, halbjährlich zum 30. Juni und 30. Dezember

Bezugsbedingungen
Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von DM 40,- (incl. Versand),

Editorial

Mit dem vorliegenden Heft des LDV-Forum - leider wieder mit einiger Verzögerung erschienen - endet mein Engagement als Schriftleiter und Herausgeber unseres Periodikums, für das ich als 1.Vorsitzender der GLDV seit nun vier Jahren ohnehin verantwortlich war, und dessen Redaktion mit Herstellung, Druck und Versand ich vor drei Jahren aus den bekannten Gründen auch übernommen und organisiert hatte. Obwohl nicht ohne Freude an und mit Dank für Mitarbeit und Unterstützung, die andere mir während dieser Jahre zukommen ließen, bin ich doch froh, das *LDV-Forum* ab der nächsten Nummer in verantwortlicher redaktioneller Leitung von Georg K n 0 r z , Darmstadt, die Produktion und den Versand in Zukunft bei Hans Haller, Saarbrücken, zu wissen. Ihnen beiden ist zu wünschen, daß sie - nach unserer Trierer Übergangslösung für das Erscheinen des *LDV-Forum - nun* auf der Grundlage gesicherter Finanzen und langfristig geregelter Verantwortlichkeiten zur weiteren Konsolidierung der Zeitschrift und unserer GLDV nachhaltig beitragen können.

Zum vorgeschlagenen Themenschwerpunkt dieses Heftes *(LDV /CL* und *empirische Sprachdaten)* hatte ich mir allerdings ein größeres Echo erhofft. Dies umso eher als korpuslinguistische Ansätze des *Natural* Language *Processing* (NLP) - wie der Aufsatz von *Wolf Papprotté* unterstreicht -aufzunehmen des linguistisches Interesse treffen und begründete Hoffnung besteht, daß über eine Verbindung empirisch-quantitativer Analysen performativer Sprachdaten mit wissensbasiertregelorientierter Modellierungen ihrer Generierungsprozesse nicht bloß klassisch-strukturalistische Auffassungen erneuert, sondern eine Neuorientierung der künftigen computerlinguistischen Forschung möglich wird.

Mir scheint, daß hiermit sogar eine Revision der wissenschaftstheoretischen Basis der Linguistik eingeleitet werden könnte. Denn angesichts sehr großer sprachlicher Textkorpora (2: 107 Worte) erweisen sich traditionelle linguistische Kategorien und Modellierungsformen (wie regelbasierte Grammatikmodelle, symbolische Repräsentationen, monotone Logiken, diskretisierende Raum-Zeit-Modelle, etc.) erkennbar als zunehmend inadäquat. Vermehrte Randunschärfen, große Variationsbreiten und vielfältige Ambiguitäten bei der Analyse solcher Sprachdatenmengen könnten dazu führen, klassisch-kategoriale Konzepte der Linguistik zu überprüfen

und gegebenenfalls als *fuzzy* categories neu zu entwickeln mit der Chance, Wissen (über Sprache und über Welt) als *kognitive* Strukturierungsleistung *semiotischer* Prozesse zu verstehen, deren prozedurale Modellierungen überdies eine quasi-empirische Überprüfung durch Simulation im Rechner erlaubt.

Mein Angebot in der letzten Nummer des *LDV-Forum* 9.1 fand dagegen eine erfreulich große Resonanz: nicht nur für die zur Besprechung *von* uns angebotenen Titel fanden sich einige .Rezensenten, es wurden auch neue Titel vorgeschlagen, so daß in diesem Heft sechs Besprechungen erscheinen können. N eben den noch nicht vorliegenden aber zugesagten Rezensionen möchte ich daher für die nächsten Hefte hier nochmals einige Titel nennen, die der Redaktion als Rezensionsexemplare zugingen. Gleichzeitig weise ich erneut darauf hin, daß die Verlage zunehmend dazu übergehen, Rezensionsexemplare ihrer Neuerscheinungen nur dann zu versenden, wenn eine Zusage zur baldigen Besprechung und deren Veröffentlichung gegeben wird. Ich möchte Sie daher weiterhin einladen und Sie auffordern, uns Ihre Rezensionsanregungen und thematischen Vorschläge (möglichst mit der Bereitschaft zur Besprechung der vorgeschlagenen Veröffentlichungen) zukommen zu lassen, damit wir die zusätzlichen Titel mit der Publikationszusage des *LDV-Forum* bei den betreffenden Verlagen anfordern können.

=> *Schierholz, Stefan J.:* Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive (Linguistische Arbeiten 269). Tübingen (Max Niemeyer Verlag) 1991 (251 S.)

=> *Heintzeler,* Mirjam: Raumausdrücke im Konzeptlexikon: Die Darstellung der Komposition lokaler Verben und Präpositionen in einem konzeptuellen Lexikon. Konstanz (Hartung-Gorre Verlag) 1992 (271 S.)

=> *Schmitz, Ulrich:* Computerlinguistik. Eine Einführung. Wiesbaden (Westdeutscher Verlag) 1992 (238 S.)

=> *Dik, Simon C.:* Functional Grammar in Prolog. An Integrated Implementation for English, French and Dutch. (Natural Language Processing 2). Berlin/ New York (Walter de Gruyter) 1992 (264 S.)

=> *Laffling, John:* Towards High-Precision Machine Translation based on Contrastive Textology (Distributed Language Translation 7). Berlin/ New York (Walter de Gruyter) 1992 (178 S.)

=> *Leitner, Gerhard* (Hrsg.): New Directions in English Language Corpora. Methodology, Results, Software Developments (Topics in English Linguistics 9). Berlin/ New York (Walter de Gruyter) 1992 (368 S.)

=> *Domenig,* Marc/ *tenHacken, Pius:* Word Manager: A System for Morphological Dictionaries. Hildesheim (Georg Olms Verlag) 1992 (212 S.)

Kompetenten und sachkundigen Rezensenten unter unseren Lesern, die eines (oder mehrere) Bücher beurteilen können und möchten, senden wir gern die entsprechenden Exemplare zu, welche - nach Erscheinen der Besprechung -wie üblich ihr Eigentum werden.
Das nächste Heft 10.1 (1993) des *LDV-Forum* wird Mitte des Jahres erscheinen.

Trier, im Januar 1993 B.R.

Einzelexemplare zum Preis von DM 20,- (zuzügl. Versandkosten) bei der Redaktion bestellt werden.

Titelgestaltung
Monika Schmitz, Düsseldorf

Fachbeiträge
Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei Referentlnnen begutachtet. Manuskripte (dreifach) sollten daher möglichst frühzeitig eingereicht und bei Annahme zur Veröffentlichung in jedem Fall zusätzlich auch noch auf Diskette (5 ¼“ bzw. 3 ½“) oder elektronisch (Email: ld vforum@ldv01.Uni-Trier.de)
als ASCII oder 1\TEX-Datei übermittelt werden. Formatierungshilfen *(LDVforum.sty)* werden auf Wunsch zugesandt.

Rubriken
Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der Autoren wider. Einreichungen sind - wie bei Fachbeiträgen - an die Redaktion zu übermitteln.

Redaktionsschluß
Für alle Rubriken mit Ausnahme der als Fachbeiträge eingereichten Manuskripte:
für Heft 10.1/93: 30. April 1993; für Heft 10.2/93: 31. Oktober 1993

Herstellung
Druckerei Print-In, Schönbornstr. 11, D-5500 Trier

Auflage
550 Exemplare

Anzeigen
Preisliste und Informationen: Prof. Dr. Johann Haller, Institut für Angewandte Informationsforschung (IAI), Martin-Luther-Str. 14, D6600 Saarbrücken 3; Tel.: (0681) 39313; Fax (0681) 397482;
Email: hans@iai.uni-sb.de

Bankverbindung
GLDVforum (Prof. Rieger): Stadtsparkasse Trier (BLZ 585 500 80) KtoNr. 680.280

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.

"JAZYKO-ZNANIE" IN MOSKAU

Reinhard Köhler, Universität Trier

In Moskau gibt es mehrere computerlinguistisch einschlägige Gruppen, von denen z.B. die Gruppe "Smysl'⇔Tekst" (Melcuk, Apresjan) seit Jahren international hohes Ansehen genießt, während andere bei uns noch wenig bekannt sind, so z.B. das Institut für Russische Sprache (über das noch zu berichten sein wird). Zunehmende Bedeutung kommt auch dem *Jazyko-Znanie*[1] genannten Zentrum zu, dem der vorliegende Bericht gewidmet ist.

Seit 1980 arbeitet die Forschungsgruppe

Theoretische und angewandte Lexikologie,

ein Zweig des *Labors für angewandte Linguistik* der Philologischen Fakultät der Moskauer Universität, an theoretischen linguistischen Problemen und an der elektronischen Verarbeitung natürlicher Sprache.

Seit 1990 wird diese Arbeit in größerem Rahmen organisatorisch als *Internationales Forschungszentrum Jazyko-Znanie* unter der Leitung von A. A. Polikarpov weitergeführt. Diese Organisation wiederum ist einer der Träger von *Gumanitarnoe Znanie* (etwa: Menschliche Kognition).

Zwischen dieser Gruppe und dem Fach LDV in Trier bestehen inzwischen intensive Kontakte; so wurde im vergangenen September ein Vertrag über Kooperation und Austausch (vor allem im Bereich quantitative und synergetische Linguistik) zwischen der Philologischen Fakultät der Moskauer und dem Fachbereich II der Universität Trier geschlossen. Der Autor hatte 1991 und 1992 Gelegenheit[2] zu Besuchen und Gesprächen mit den dortigen Kollegen.

Jazyko-Znanie ist zur Zeit mit etwa zwölf Mitarbeitern (Linguisten, Programmierern und Wissenstechnikern) ausgestattet. Darüber hinaus koordiniert und integriert die Einrichtung fallweise die Arbeit externer Wissenschaftler (Linguisten, Mathematiker, Psychologen, Musikologen und KI-Forscher) und Institute, darunter die Fakultät für Mathematische Informatik und Kybernetik der Moskauer Universität, *das Institut für Russische Sprache* der (vormals sowjetischen) Akademie der Wissenschaften, die Institute für Linguistik der Ukrainischen Akademie der Wissenschaften, und der Universität Tartu und das Staatliche Konservatorium Tiflis.

Jazyko-Znanie gliedert sich in mehrere problemorientierte Arbeitsgruppen und unterhält zwei Labors - eins für Computerlinguistik uns eines für 'systemische Linguistik'. Publikationsorgan ist die Reihe "Kvantitativnaja lingvistika i avtomaticeskij analiz tekstov" (Quantitative Linguistik und automatische Textanalyse), Universitätsverlag Tartu, herausgegeben von J. A. Tuldava.

Hauptaktivitäten der Arbeitsgruppen sind zur Zeit:

=> Modellbildung für den Bereich Evolution der natürlichen Sprache(n) und für Kommunikationsprozesse auf der Grundlage kognitiver Mechanismen und

1 nicht übersetzbares Wortspiel, ungefähr 'Sprach- Wissen[schaft]'

2 Mit Unterstützung durch die Deutsche Forschungsgemeinschaft

Randbedingungen für die Lösung von Polysemieproblemen und die semantische Interpretation auf der lexikalischen, morphemischen und syntaktischen Ebene.

=> Theoretische Analyse grundsätzlicher Fragen zur Organisation des Vokabulars und Besonderheiten der Repräsentation in verschiedenen Arten von Wörterbüchern.

=> Entwicklung von Algorithmen und von Software für die automatische Analyse (morphologisch, syntaktisch und lexikalisch-semantisch) von Texten in russischer und englischer Sprache.

=> Vergleichende Studien von Russisch und Englisch auf allen linguistischen Analyseebenen;

=> Modellierung des bilingualen Sprecher/Hörers; Untersuchungen zu Regularitäten von Interferenzen bei der Fremdsprachenverwendung, Entwicklung einer Fehlertypologie.

=> Entwicklung eines Moduls für idiomatische Ausdrücke für Übersetzungssysteme.

=> Erstellung von Softwarewerkzeugen für die maschinengestützte Lexikographie.

=> Aufbau von Textkorpora und lexikalischen Datenbanken.

Zu den konkreten Projekten gehören u. a. maschinenoperable Wörterbücher der Antonyme und der Synonyme des Russischen und weitere Spezialwörterbücher, eines Englisch-Russischen und Russisch-Englischen Übersetzungswörterbuchs (je ca. 100000 Einträge) mit Zugriffs- und Pflegesoftware, der "Thesaurus der russischen Sprache", lexikalische Datenbanken für weitere Sprachen (u. a. Chinesisch und Ukrainisch) und Software für verschiedene Zwecke, z.B. eine Volltext-Datenbank, ein elektronischer Redaktionsassistent zur Überprüfung von Orthographie, Vokabular, Grammatik, Stil und Interpunktion, ein Statistik-Paket für die quantitative Textanalyse und Module für Lehr- und Lernsysteme.

Die Forschungseinrichtung beteiligt sich laufend an internationalen Projekten (z.B. dem Bochum-Trierer Projekt zur synergetischen Linguistik) und Konferenzen (so das erste internationale Kolloquium über synergetische Linguistik, das von Jazyko-Znanie im September 92 ausgerichtet wurde). Die zweite internationale Konferenz zur quantitativen Linguistik, QUALICO-94, wird von dieser Organisation, zusammen mit der Philologischen Fakultät der Moskauer Universität, organisiert werden.

ASV AN DER UNIVERSITÄT LEIPZIG - EIN RÜCKBLICK

Bernd Koenitz, Leipzig

Zu Beginn des Jahres 1981 wurde entgegen skeptischen Stimmen an der Sektion Theoretische und angewandte Sprachwissenschaft (TAS) der Leipziger Universität mit der Bildung eines Forschungskollektivs der Aufbau der Forschungsrichtung "Automatische Sprachverarbeitung" (ASV) in Angriff genommen. Die Leitung war dem Verfasser dieser Mitteilung übertragen worden, de facto teilte er sie sich mit Rudi Conrad, dem auch das Verdienst gebührt, Ende der 70er Jahre drängend die Frage auf die Tagesordnung gesetzt zu haben, ob es nicht der größten sprachwissenschaftlichen Hochschuleinrichtung der DDR wohl anstünde, sich dem zukunftsträchtigen Forschungsfeld ASV zuzuwenden.

Das Forschungskollektiv war zunächst mit sehr bescheidenen Kräften - bemüht, den Forschungsstand aufzuarbeiten, zugleich aber auch wissenschaftliche Nachwuchskräfte auf dieses Gebiet zu lenken. Längere Zeit bestand einige Unsicherheit bezüglich der Art von Themen und Projekten, die unter den gegebenen Bedingungen sinnvollerweise anzugehen wären. Einerseits lag es nahe, den Umstand zu nutzen, daß an der Sektion TAS in Lehre und Forschung mehrere Fremdsprachen vertreten waren. Andererseits erschien das jedoch als nicht real, denn unter Berücksichtigung der Anzahl der Stellen (bzw. "Vollbeschäftigteneinheiten"), mit denen man für die ASV rechnen konnte, wäre ein auf einzelne Fremdsprachen ausgerichtetes tragfähiges Projekt, das den Prämissen entsprochen hätte, denen Rudi Conrad und ich Priorität beimaßen, nicht zustande gekommen. Diese Prämissen bestanden darin, daß, soweit wir es zu verantworten hatten, an der Sektion TAS Computerlinguistik in erster Linie als "Linguistik für den Computer" betrieben werden und außerdem an theoretischen Problemstellungen anknüpfen sollte, wie wir sie auf dem Gebiet der reinen Linguistik bisher bearbeitet hatten.

In der zweiten Hälfte der achtziger Jahre war die konzeptionelle Seite der Bemühungen klar: Im Vordergrund sollten allgemeine theoretische Grundlagen und Probleme des natürlichsprachigen Mensch-Maschine-Dialogs stehen. Als auch aus Sicht der Theorie besonders interessant und zugleich in der bisherigen computerlinguistischen Forschungslandschaft sehr wenig vertreten sollte die Projektrichtung "Intelligente automatische Examinatorsysteme" (mit solchen theoretischen Fragestellungen wie den mit dem automatischen Generieren von Fragen aus einer Wissensbasis verbundenen) betrieben werden.

Unsere Bemühungen auf dem Gebiet der ASV trafen sich ab Mitte der 80er Jahre mit Diskussionen über die Notwendigkeit, an der Universität die Informatik zu institutionalisieren. Nach langen Verzögerungen wurde schließlich entschieden, zum September 1989 eine Sektion Informatik zu gründen, an der ein Jahr nach Gründung mit der Ausbildung von Diplom-Informatikern begonnen werden sollte. Nach anfänglichem (wohlbegründeten) Zögern stimmten wir dem drängenden Vorschlag des Rektors zu, an der zu gründenden Sektion einen Wissenschaftsbereich ASV aufzubauen. Von dem Übergang an die Sektion Informatik versprachen wir uns wesentlich bessere rechentechnische Bedingungen, als sie (mit einem 8-bit-PC und einem Dutzend Kleincomputern) an der Sektion TAS gegeben waren, und die Möglichkeit enger interdisziplinärer Zusammenarbeit sowie der Einbeziehung erfahrener Programmierer und Systemspezialisten in die Bearbeitung unserer Pro-

jekte. Nicht zuletzt sahen wir die Möglichkeit, endlich - erstmalig an einer DDR-Universität - eine in die Informatik eingebettete Ausbildung von ASV -Spezialisten zu realisieren, wie sie ein 1983 von dem damals gegründeten Problemrat für die automatische Verarbeitung sprachlicher Daten unter der Leitung von Jürgen Kunze (Leiter des Arbeitsbereiches ASV am Zentralinstitut für Sprachwissenschaft der Akademie der Wissenschaften der DDR) erarbeiteter "Vorschlag zur Schaffung einer Studienrichtung Automatische Sprachverarbeitung" nachdrücklich gefordert hatte.

Der Wissenschaftsbereich ASV wurde schließlich (im dritten Jahr seiner Existenz) von zwei ordentlichen Professoren, einem Oberassistenten, einem wissenschaftlichen Mitarbeiter, einer Assistentin und drei Forschungsstudenten gebildet. Ein Mitarbeiter kam aus dem Rechenzentrum und war ein erfahrener Programmierer und Systemspezialist, alle übrigen Mitglieder des Wissenschaftsbereiches waren von Haus aus Linguisten.

An der Sektion Informatik wurden zum Wintersemester 1990/91 40 und zum Wintersemester 1991/92 50 Studenten für den Studiengang Informatik immatrikuliert. Die Studierenden hatten nach der ersten Konzeption die Wahl zwischen zwei Vertiefungsrichtungen, von denen eine die ASV war. N ach einer Modifikation der Studienordnung, die für den zweiten Studentenjahrgang wirksam wurde, bestand darüber hinaus die Alternative, sich für ein "gewöhnliches" Nebenfach (z. B. Psychologie oder Physik) zu entscheiden. Der Stundenanteil für die Vertiefungsrichtungen ergab sich aus der Addition der Stunden für einen Studienschwerpunkt innerhalb des Hauptfaches und der Stunden, wie sie ein Nebenfach beanspruchen kann. Daraus ergab sich ein relativ hoher Stundenanteil für die Vertiefungsrichtung, und im Falle der ASV war damit vom Zeitfonds her eine gründliche, der Komplexität des Aufgabengebietes gerecht werdende Ausbildung gewährleistet. Unsere Konzeption für die Ausbildung von Spezialisten der ASV innerhalb der Informatik, die wir in den Jahren 1988 - 90 erarbeitet hatten, entsprach fast genau der an der Universität Koblenz-Landau entwickelten und bereits länger praktizierten, so daß wir uns im Jahre 1990 bei der in kurzer Frist zu bewerkstelligenden Ausarbeitung einer Studienordnung und einer Prüfungsordnung nach bundesdeutschem Muster die entsprechenden Koblenzer Dokumente zum Vorbild nehmen konnten. Es ist daher kein Zufall, daß sich seit 1991 zwischen den computerlinguistischen Bereichen beider Universitäten hoffnungsvolle Beziehungen kollegialer Zusammenarbeit entwickelten.

Die Vertiefungsrichtung ASV stieß auf reges Interesse der Studenten; beispielsweise wurde im Studienjahr 1991/92 die Vorlesung "Einführung in die Grundlagen der ASV" etwa von der Hälfte der Informatikstudenten besucht. Fakultative Kurse "ASV mit LISP" und "ASV mit Prolog" (die von uns eigentlich für Studenten der Sprachwissenschaft angeboten wurden) fanden den regen Zuspruch der Informatikstudenten der ersten Semester. Ersteinsgesamt positive - Erfahrungen wurden 1992 mit dem im 4. Semester angesiedelten Projektpraktikum sowie einem Problemseminar gesammelt.

In der Forschung wurden vor al1em zwei Projekte betrieben: erstens das bereits oben genannte Examinatorprojekt, konkretisiert als "Intelligentes Examinatorsystem zu Vorfahrtsregeln", und zweitens - als Gemeinschaftsprojekt mit einem slawistischen Bereich der Sektion TAS - das eines multilingualen automatischen Sprachmittlerhilfssysterns. Beide Projekte konnten mit ihrer Anlage Originalität beanspruchen und kamen gut voran.

Die Erwartungen, die wir an die Verlagerung des Bereiches ASV in die Informatik geknüpft hatten, hatten sich sehr weitgehend erfüllt, zumal mittlerweile am Institut und am Rechenzentrum eine sehr gute rechentechnische Ausstattung erreicht war. Im Jahre 1992 brach die hier skizzierte Ent-

wicklung infolge äußerer Eingriffe abrupt ab. Beide Professoren sowie zwei von drei Mitarbeitern mußten mit unterschiedlichen Begründungen die Universität bzw. die inzwischen in ein "Institut für Informatik im Fachbereich Mathematik/Informatik" umgewandelte Sektion verlassen. Den 1990 und 1991 immatrikulierten Studenten empfahl man, an der Sektion TAS Lehrveranstaltungen des Bereiches Theoretische Linguistik zu belegen. Im Stellenplan des Instituts für Informatik wurde eine C3Professur für ASV vorgesehen und im Sommer 1992 ausgeschrieben. Den Gründern des Wissenschaftsbereiches ASV war deutlich gemacht worden, daß sie sich für die ausgeschriebene Stelle nicht mit Erfolg bewerben könnten. Im übrigen dürften es die Vorgaben des Stellen planes nicht zulassen, die ursprüngliche Konzeption einer dem Gegenstand angemessenen gründlichen und umfassenden Ausbildung von Fachleuten für die automatische Verarbeitung der natürlichen Sprache beizubehalten. Dies ist im Einklang mit dem seit anderthalb Jahren am Institut für Informatik von maßgeblicher Seite geäußerten Bekenntnis zu einer "0815-Informatik".

Anmerkung: Ergänzende Informationen siehe in: Studienführer Linguistische Datenverarbeitung/ Computerlinguistik für die deutschen Universitäten, ermittelt und bearbeitet von Magdalene Lutz-Hensel, Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV), Saarbrücken 1991, S. 51 ff.; Blätter zur Berufskunde. Sprachwissenschaftler /in, Computerlinguist/in, Phonetiker/-in, Bundesanstalt für Arbeit, Neuauflage (in Vorbereitung).

TAGUNGSBERICHT ZUM 1. SHOE-WORKSHOP (EXTRACTION OF HIERACHICAL STRUCTURE FOR MACHINE LEARNING OF NATURAL LANGUAGE)

27-28.2.92, ITK, Tilburg

Ziel des Workshops war es, einen Kooperationsrahmen für die in diesem Bereich tätigen Wissenschaftlerinnen und Wissenschaftler zu schaffen und einen Überblick über den aktuellen Forschungsstand zu vermitteln. Der Teilnehmerkreis (23 Personen) stammte aus Belgien, der Bundesrepublik, Großbritanien und den Niederlanden.

A. van den Bosch und *W. Daelemans* (ITK, Tilburg) untersuchten die Fähigkeit neuronaler Netze (Elman- und Jordan-Netze), Silbengrenzen in orthographischen und phonologischen Repräsentationen niederländischer Wortformen zu bestimmen. Getestet wurden sowohl unterschiedliche Codierungen der Eingaben als auch verschiedene Netzwerk-Architekturen. Zusammenfassend konnte festgestellt werden, daß die verwendeten Netze die Aufgabe zwar bewältigen konnten, jedoch nicht signifikant besser als symbolische Ansätze.

G. *Durieux* (UIA, Antwerpen) entwickelte ein Analogie-basiertes Verfahren, um einfachen Holländischen Wortformen (Monomorphemen bzw. Lehnwörter, die sich wie solche verhalten) den Hauptakzent zuzuweisen. Es basiert auf der Identifikation von Ähnlichkeiten bzw. Abweichungen zwischen einer Wortform und anderen im Lexikon gespeicherten Wortformen. Vorteile gegenüber traditionellen regelbasierten Ansätzen bietet dieses Verfahren vor allem aufgrund seiner Fehlertoleranz und der der Fähigkeit partielle Übereinstimmungen zu nutzen.

T.M. Ellison (CCS, Edinburgh) beschrieb ein unüberwachtes Lernverfahren für Vokal-Harmonien. Zur Repräsentation wurden nicht-deterministische endliche Automaten mit zwei Zuständen verwendet, die nicht nur einfache Harmonien sondern auch Transparenz und Opazität darzustellen erlauben. Das vorgestellte System wurde erfolgreich für die Sprachen Türkisch, Kirgisisch, Ungarisch und Yoruba getestet. Es zeigte sich, daß das Verfahren auch ohne sprach-spezifisches Wissens korrekte Vokal-Harmonien lernen konnte.

S. Finch (CCS, Edinburgh) beschrieb einen statistischen Ansatz zum Lernen lexikalischer Kategorien auf der Basis ungetaggter Korpora. Dazu wird jede Wortform durch einen Vektor beschrieben, der die Distribution der lokalen Kontexte beschreibt, in denen die betreffende Wortform vorkommt. Mit Hilfe einer Clusteranalyse werden diese dann zu Gruppen zusammengefaßt. Erfolgreich getestet wurde das Verfahren anhand größerer englischer Korpora.

P. Flach (ITK, Tilburg) gab eine Einführung in die Grundlagen des Konzeptlernens und der Induktiven Logikprogrammierung. Er beschränkte sich auf induktive Verfahren des Konzeptlernens; d.h. auf Verfahren, die auf der Grundlage positi-

ver und negativer Beispiele Klassifikationsregeln bilden und stellte das *Version Space* Modell als ein allgemeines Modell für das Konzeptlernen vor. Ziel der Induktiven Logikprogrammierung ist es Verfahren zu entwickeln, die den beispielgesteuerten Erwerb von Logikprogrammen ermöglichen.

S. Gillis (UvA, Antwerpen) skizzierte die verschiedenen Positionen, die in der Psycholinguistik zu folgenden, für die erste Phase des Spracherwerbs zentralen Fragen eingenommen werden:

. Wie ist die beim Spracherwerbs von Kindern gezeigte Kreativität adäquat zu charakterisieren?

. Gibt es sprachspezifische Prädispositionen oder wird Spracherwerb durch allgemeine kognitive Mechanismen gesteuert?

. Sind die biologische Grundlage des Spracherwerbs artspezifisch ?

. Ist es korrekt Spracherwerb als einen in seiner Form und zeitlichen Realisierung universalen Prozeß zu beschreiben?

B. Manderick (AI-LAB, Brüssel) berichtete über die vielfältigen Anwendungsmöglichkeiten der durch adaptive biologische Systeme inspirierten Genetischen Algorithmen(GA). Mit Hilfe der beiden "genetischen Operatoren" der Mutation und Selektion bieten GAen häufig eine effiziente Lösung für viele Such- und Optimierungsprobleme (wie z.B. das Rundreiseproblem). Darüberhinaus wurden GAen in der Bildverarbeitung, der Mustererkennung, dem maschinellen Lernen sowie in den Wirtschafts- und Sozialwissenschaften erfolgreich eingesetzt.

S. Naumann und *J. Schrepp* (LDV /CL, Trier) beschrieben ein System zum Erwerb syntaktischen Wissens. Den Kern des Systems bildet ein inkrementeller Lernalgorithmus, der ausgehend von einem Satzkorpus eine Folge reversibler Syntaxen generiert. In jedem Schritt wird aus dem Korpus eine kleine Menge von Sätzen ausgewählt und von einem speziellen Parser analysiert, der den von der aktuellen Grammatik nicht erfaßten Sätzen partielle Strukturbeschreibungen zuordnet. Auf der Grundlage dieser Strukturbeschreibungen werden Hypothesen zur Erweiterung der Syntax formuliert, die schließlich zur Generierung der neuen Syntax führen.

D.M. W. Powers (ITK, Tilburg) berichtete von einer Reihe von Experimenten zum Erwerb lexikalischen und syntaktischen Wissens. Ziel dieser Experimente war das Erlernen von Wort klassen und syntaktischen Regeln. Verwendet wurden in allen Fällen statistische bzw. konnektionistische Verfahren. Bei den Wortklassenexperimenten zeigte sich, daß geschlossene Wortklassen vor den offenen Klassen gelernt wurden. Zu diesem Ergebnis wurde auf von einem weiteren Experiment (Klassifikation von Graphemen/Phonemen) bestätigt.

J. C. Scholtes (UvA, Amsterdam) stellte einen Ansatz zum daten-orientierten Parsen mit Hilfe eines Kohonen-Netzes (feature map) vor. Der Hauptunterschied zu anderen korpus-basierten Verfahren ist die Verwendung von statistischer als auch syntaktischer und semantischer Information während der Trainingsphase. Ein so angereichertes Korpus wird in einem Kohonen-Netz "gespeichert", das dann zum Parsen verwendet wird.

J. Schrepp (LDV /CL, Trier) gab einen kurzen Überblick über Ziele und Methoden der quantitativen und synergetischen Linguistik. Besonders hervorgehoben wurden die oft vernachlässigten Anwendungsmöglichkeiten statistischer Information in praktischen Systemen.

S.N. & J.S.

TAGUNGSBESUCH IN RUSSLAND

Enno Leopold, Universität Trier

Ende September 1992 machte die Arbeitsgruppe Quantitative Linguistik des Fachs Linguistische Datenverarbeitung/Computerlinguistik der Universität Trier zu einem Tagungsbesuch in Rußland auf den Weg.

Zweck der Reise war der Besuch der Konferenz über linguistische Synergetik (COLISYN), die vom 22.9.1992 bis zum 25.9.1992 in den Räumen der Moskauer Universität stattfand. Die Konferenz war, gemessen an dem relativ speziellen Thema, mit 40 Teilnehmern aus fünf Ländern gut besucht.

Von den 32 Vorträgen waren fünf den theoretischen Grundlagen synergetischer Sprachmodellierung gewidmet. Die meisten anderen Beiträge befaßten sich mit spezifischen Teilproblemen und stellten einzelne Resultate dar.

Interessant waren hier einige Vorträge zur zeitlichen Entwicklung der Polysemie sowie zum Zusammenhang zwischen Worthäufigkeit und Polysemie gehalten.

Yu.K. Krylov aus St. Petersburg stellte mit seinen beiden Beiträgen "Von der Lexikostatistik zu einer Theorie der sprachlichen Selbstorganisation" und "Gnoseologische und ontologische Grundlagen des Welle- Teilchen Dualismus in quantitativen Linguistik" einen Zusammenhang zwischen thermodynamischen Problemen und Problemen der quantitativen Linguistik her.

Eine besondere Überraschung präsentierte eine Forschergruppe aus Kiev (Ukraine), die in den letzten Jahren eine umfangreiche Erfassungsarbeit zu quantitativ-linguistischen Daten der russischen und ukrainischen Sprache geleistet haben. Resultat dieser Arbeit sind eine russische und eine ukrainische Wortdatenbank mit je ca. 150.000 Einträgen.

Zwischen den Vortragsblöcken bestand die Gelegenheit zu sehr fruchtbarem Gedankenaustausch mit den ausländischen Kollegen. Ein gemeinsamer Besuch des russisch orthodoxen Klosters Troize Sergieva Lavra bei Zagorsk sowie ein gemeinsames Abendessen rundeten das Programm der Tagung ab.

Rezensionen

Henry G. Burger:
The Wordtree. A Transitive Cladistic for Solving Physical & Sodal Problems. First Edition (The Wordtree Publisher) 1984-86, (379 S.)

Was ist von einem Buch zu halten, das nicht nur in großem Format (B,H,D: 220 x 285 x 40mm) und kleinem, vierspaltigen Druck (3pt serifenlose Schreibmaschinen-Type) daherkommt, sondern sich als *WORTBAUM* preist (für übrigens US-$ 155,- plus Versand) und behauptet, ein *"handbook of physical and social engineering TM"* zu sein, das als *Wörterbuch der Ursachen und Wirkungen. . . eine neue Art Problemlöser* und eine Studie" *in logicolinguistics"* (S.16) darstelle:

> The volume provides a materialistic blueprint of the world, a link between substance and process, and a [!] antonym dictionary. It is simultaneously a semantic theorizer, offering a non-circular simplifier and a non-circular complexifier. It reveals an evolutionary history and a word-taxonomy. (S.25, C53)

Das macht nicht nur neugierig, weil ja immerhin einiges (was eigentlicht nicht?) versprochen wird, sondern eben auch einige Mühe, wenn es darum geht, derlei üppige Formulierungen durch gezielten Testgebrauch wenigstens im Ansatz zu überprüfen.

Was der Autor, Henry G. Burger, Professor der Anthropologie und Erziehungswissenschaft, University of Missouri, Kansas City, und Life Fellow des Royal Anthropological Institute London, nach mehr als zwanzigjähriger Arbeit hier vorgelegt hat, ist in vielerlei Hinsicht mit dem lexikographischen Werk des Mediziners Peter M. Roget vergleichbar, der rund fünfzig Jahre an und mit seinem" classed catalogue of words" gearbeitet hatte, ehe er ihn 1852 als seinen *Thesaurus* herausbrachte. Dieses nach rund 1000 Themen und Gegenstandsbereichen geordnete Wortverzeichnis des Englischen erwies sich bekanntlich in seiner Grundidee der inhaltlichen Gruppierung des Vokabulars mit Querverweisen zur wechselseitigen Erläuterung der Wortbedeutungen als so hilfreich in der Praxis des Formulierens und Schreibens, daß *Roget's Thesaurus* schon zu Lebzeiten des Autors Dutzende Neuauflagen erfuhr und bis heute ohne nennenswertes Risiko verlegt werden kann. Diesem dauerhaften Benutzer-Erfolg eines - wenn auch vielfach modernisierten-lexikographischen Werks des letzten Jahrhunderts steht nun seit nahezu zehn Jahren *Burger's Wordtree* gegenüber, dessen Erfolge sich bisher weniger in der großen Zahl seiner Benutzer als seiner Rezensenten zeigt! Deren durchweg positive bis überschwengliche Kritiken und Besprechungen haben dieses außerordentliche Wörterbuch seit seinem Erscheinen (1984) begleitet. Wieso - so wäre zu fragen - hat eine so umfassende und wohlwollende Aufnahme durch die lexikologische Fachkritik aber auch durch Vertreter der unterschiedlichsten wissenschaftlichen Disziplinen bisher nicht vermocht, *Burger's Wordtree* zu einem *Roget's Thesaurus* vergleichbaren Gebrauchs- und Verkaufserfolg zu machen?

1 Der Redaktion des *LDVForum* ging ein nummeriertes Rezensionsexemplar mit umfänglicher Dokumentation zu, was diese Vermutung nahelegt.

Seine umfängliche Einführung (S.16-50, C16-C257) - mit Erläuterungen der möglichen Fragestellungen, mit Hinweisen für die Benutzer des *Wordtree* und den Gebrauch der Wortlisten und Register - enthält auch ein den sprachtheoretischen Grundlagen gewidmetes Kapitel (S.25-40, C56-Cl72). In ihm entwickelt Burger seine Grundidee einer Klassifizierung nicht mehr nur der Materie, Substanzen und Objekte sondern vielmehr der Vorgänge und Veränderbarkeiten, welche die Welt als eine den Menschen umgebende Realität prägen.

> Roget conceptualized all vocabulary as unilinear. Such an arrangement fatally commingles cause and effect indiscriminately. [. . . Such] a dictionary seeks "clarity" rather than cause-and-effect. So it is necessarily circular in its definitions.(S.26, C62f)

Anders als die Gesamtheit der äußeren Erscheinungen nach analogen Formen und Strukturen zu gruppieren, um deren sprachlich-lexemische Repräsentationen als ein im wesentlichen statisches System begrifflicher Zuordnungen vorzustellen, versteht Burger Welt primär als Prozeß des Erscheinens, der Veränderung und des Einwirkens, dessen Ordnungen dynamisch, dessen Komponenten sich zunächst als sprachlich-lexikalische Unterscheidungen ergeben.

Gegen den (noch) vorherrschenden Deskriptivismus nicht-linguistischer und linguistischer Kodifikationen, die Strukturen und Prozesse vermischen (extant *nonlinguistic codifications commingle structure* & *process;* extant *linguistic codifications* err *likewise)* setzt Burger - mit Seitenhieben auf einen *structural functionalism* die Notwendigkeit ihrer klaren Unterscheidung mit der Möglichkeit der Klassifikation von Prozessen. Diese sei in der Fähigkeit des Menschen begründet, Resultate seiner Erfahrungen symbolisch zu repräsentieren und diese als natürlichsprachliche Kodifikationen zu vermitteln *(symboling).*

> Symbols are explanatory ways that can be appreciated without sensory enactment. Symboling enables the instant correlation of levels of activity without the need to retrace all the sub-branches. And vocabulary is a culture's stock of symbols that have withstood time. Why, then, should we not combine these insights to roster all the methods that the culture has painfully found for solving its problems?! (S.27, C72)

Trotz eines verkürzten, offenbar an den Sozialwissenschaften orientierten, rationalistischen Wissenschaftsverständnisses mit naiv ungebrochenem Kausalitätsbegriff wird hier der Welt-erschließende, Realitäts-determinierende Charakter sprachlich*semiotischer* Bedeutungskonstitutions- Prozesse hervorgehoben, als deren sprachlich*lexikalische* Resultate das Vokabular einer Sprache bzw. das Lexikon und seine durch Jahrhunderte des Gebrauchs etablierte, dynamische semantischen Strukturiertheit sich auszubeuten anbietet.

> Life is the process of finding ever more efficient ways to transform and control matter and energy. Subhuman organisms learn by experience and by genetic accident. But humans evolve experiences into predictions and institutions to accelerate the improvement of energy-harnessing. [. . .] Now, science is a statement of the conditions under which one event causes another. Therefore, the codification of science requires at least two elements: a roster of event causations, and a roster (within or beyond that first one) of circumstances. We argue, that the range of causative phenomena can be expressed

> by linguistic morphemes. Hence we can form a taxonomy of presently recognized scientific principles by rostering, parsing, and "actbranching" those words in the form that English calls transitives. (S.27 C68f)

Burger hat eine Methode entwickelt und umgesetzt, welche die produktive Flexibilität speziell des Englischen ausnutzt, nahezu beliebige Lexeme auch als *transitive Verben* verwenden zu können. Diese zunächst nur morpho-syntaktische Möglichkeit erlaubt es ihm, nahezu alle Substanzen und Eigenschaften, Materialien und Formen, Eigennamen und Abstrakta zu *prozessualisieren.* Gleichzeitig eröffnet sie eine lexiko-semantische Dimension, die nach den minimal unterscheidenden Spezifika solcher Prozesse zu fragen erlaubt. Die Zerlegung derart lexikalisierter Vorgänge und Handlungen (und ihrer Negationen) in lexikale Komponenten (sog. *acteme),* die als kleinste, einen so bezeichneten Vorgang differenzierende Unterscheidungen lexematisch bestimmt werden können, bildet das Ergebnis des Burger'schen Analyseverfahrens.

> The Wordtree [...] is a recursivefree simplifiel'. It analyzes the language into its components. A transitive cladistic does so by firmly defining each transitive verb as the combination of two and only two simpler transitives. Along the bottom branch rest some three dozen primitive, creation-based words, such as to elate which are clearly marked as having "no second step" [...]
>
> The Wordtree's hierarchy may, then, be followed from end to beginning to reveal processual simplification. But its principle is that of "add-on", so it may also be followed from beginning to end to produce a non-circular complexification of ideas. (S.21, C25f)

Burger hat das Lexeminventar des Englischen in derartige Acteme (24.600) zerlegt, die als ein baumartiger (d.h. zusammenhängender, kreisfreier) Graph mit 155.000 Knoten darstellbar sind. Dessen hierarchische Ordnungen-20.500 Wurzelbäume *(cladistica)* im Wordtree - wurden mit über 157.000 Nummerierungen indiziert. Zwei Verzeichnisse (Index aller analysierten transitiven Verben und Wörter, Hierarchie der Acteme) stellen über deren Nummern die Verweisstruktur der Knoten her, im wesentlichen in Form der 20.500 Acteme - Wurzelbäume mit ihren zugeordneten Sub- und Super-Actemen.

So führt beispielsweise der *alphabetische Index* den Benutzer unter to *fasten* auf: TO FASTEN = HOLD & STAY 12510 mit 37 Spezifizierungen der Art: FASTEN bound object = LASH; bundled object = NITCH; grounded object = STAKE; hooked object = CLASP; etc. und 9 Differenzierungen der Art: FASTEN & FEED = MAN GER; & FIRM = FIX; etc. sowie weitere 6 bedeutungsähnliche Transitiva wie etwa ADHERE, LOCK, SECURE, TIE, etc..Nummer 12510 bezeichnet in der *Wordtree- Hierarchie* den Actem-Knoten positiv FASTEN negativ UN-FASTEN mit seinen 38 Sub-Actemen von 12912 TO LATCH = ENTER (91) & FASTEN bis 12950 TO MAN GER = FASTEN & FEED (7641), wobei höhere Nummern auf komplexere Sub-Acteme, niedrigere Nummern auf einfachere Super-Actem im Baum verweisen. Wie etwa Super-Actem 7641 positiv FEED negativ HUNGER mit seinen 22 Sub-Actemen, darunter neg7731 TO UNDERFEED = FEED & FAlL (neg160) mit dem dadurch bewirkten Resultat UNDERFED. Sieht man Super-Actem 160 nach mit positiv EXCEED negativ FAIL, so finden sich 18 Sub-Acteme, worunter etwa auch Super-Actem 276 TO UNDERPOWER = POWER (67) & FAlL aufgeführt ist und zu Super-Actem 67 positiv POWER negativ WEAKEN

führt, etc. um schließlich bei 60 positiv **FREE** negativ **UNFREE** mit 4 weiteren Sub-Actemen etc.
Der Umstand, daß jedem unter einem Actem aufgeführten Sub- Actem als Komponenten zunächst einmal dieses Actem (oder sein Negat) plus einer minimalen Differenzierung in Form eines dem Actem übergeordneten Super-Actems zugehört, macht die» erklärende" *Verweisstruktur* der so komponentiell analysierten, lexikalisierten *Prozesse* aus. Das Durchmustern dieser Struktur sollte einladend sein, das Durchwandern der Listen aber ist aufwendig (den Blätterwald zu *durchblättern),* mühevoll (die aus Teillisten ermittelten Acteme aufzulisten) und ermüdend (durch Kleinstdruck Konzentrationschwächen).

Es scheinen daher vornehmlich praktische, ausschließlich den Gebrauch des *Wordtree* betreffende Gründe zu sein, die bisher verhindert haben, diese Wörterbuch-Struktur in einer Weise zu nutzen, die ihrer Vieldimensionalität einerseits, ihrem Informationsreichtum andererseits entspricht: die linearisierte Darbietung auf Papierseiten in Buchform eröffnet offenbar den Zugang zu dieser Informationsstruktur in ebenso begrenztem Maße, wie eine bloße Lektüre der Menuefolge den Geschmackssensationen nahekommt, die der Genuß der Speisen vermitteln wird.
Es wäre daher zu wünschen, daß Burger der den Zusammenhang von *Prozeß* und *Struktur* betont aber die Möglichkeiten der Zeitabstrahierten Repräsentation von Prozeß-Struktur-Zusammenhängen in Form von (formal-sprachlich repräsentierbaren) *Prozeduren* zu übersehen scheint bald die nötigen Mittel zur Verfügung hat, um seinen *Wordtree* in der in diesem Fall nicht nur zeitgemäßen, sondern auch sachlich begründeten Form einer Hypertext-Struktur (etwa als CD-ROM) anbieten zu können. Erst dann wird sich durch häufigen, weil bequemen Gebrauch und vielfacher, weil zugänglicher Nutzung erweisen können, welche von den beanspruchten Qualitäten diesem Werk tatsächlich zukommen, und aufgrund welcher Eigenschaften es möglicherweise als eben die epochale lexikographische Unternehmung gelten muß, als die zahlreiche Rezensenten *Burger's Wordtree* schon heute werten.

Burghard Rieger, Universität Trier

Stig Johansson/ Anna-Brita Stenström (Hrsg.):
English Computer Corpora:
Selected Papers and Research Guide, (1991) Vol. 3. In Jan Svartvik/ Hermann Wekker (Hrsg.): Topics in English Linguistics, Mouton de Gruyter

Die breite Verfügbarkeit von Rechenanlagen mit großen Massenspeichern hat der korpusbasierten Linguistik entscheidende Vorteile gebracht. Linguistische Korpora können nicht nur besser und schneller analysiert werden, sie sind auch ohne großen Aufwand anderen Forschungsgruppen zugänglich und erlauben Untersuchungen, die andernfalls unterbleiben. Hierzu zählt insbesondere die Arbeit mit riesigen Korpora (VLLC für *very large linguistic corpora),* deren Relevanz sich zunehmend deutlicher abzuzeichnen beginnt. Das im englischsprachigen Raum bekannte Brown Corpus etwa zählt mit wenig mehr als einer Million laufender Wörter längst nicht mehr zu den wirklich großen, die zwei- bis dreistellige Millionenbeträge an laufenden Wörtern *(running words)* umfassen.

Diese Textmengen sind jedoch auch zu untersuchen, denn durch das bloße Ansammeln linguistischer Daten entsteht kein Erkenntnisgewinn. Untersuchungen in dieser Größenordnung können jedoch manuell nicht mehr durchgeführt werden, noch nicht einmal das Markieren von Wortarten

- beim Brown Corpus noch mühsam per Hand durchgeführt - könnte in annehmbarer Zeit durchgeführt werden. An dieser Stelle sind somit grundlegende Aufgaben der Sprachanalyse zu algorithmisieren, um diese maschinell durchführen zu können.

Bisher bewegte sich die Computerlinguistik überwiegend in den höheren Ebenen der Sprachverarbeitung, vor allem Syntax und Parsing, man behalf sich dort wegen der geringen Anzahl von Testsätzen mit rudimentären morphologischen Komponenten oder gar Vollformenlexika. Bei großen Textmengen mit hunderttausenden von Types sind neue Verfahren und Paradigmen sind gefragt, so z.B. die Anwendung statistischer Methoden beim automatischen Markieren von Wortarten *(Tagging)* in Massentexten.

In dem von Stig Johansson und Anna-Brita Stenström herausgegebenen Band *English Computer Corpora: Selected Papers and Research Guide,* der 1991 bei Mouton de Gruyter als Nummer 3 der Reihe *Topics in English Linguistics* (herausgegeben von Jan Svartvik und Hermann Wekker) wird in zwanzig Beiträgen insgesamt ein guter Einblick in die Arbeiten gegeben, die momentan im Bereich der englischsprachigen Korpora durchgeführt werden. Zumeist handelt es sich um Kongreßbeiträge der 10. ICAME *(International Computer Archive 01 Modern English)* Konferenz in Bergen 1989. Abgedeckt werden die Bereiche probabilistische Analyse, Syntax, Lexis, gesprochene Sprache, regionale und soziale Varianten, spezialisierte Korpora und Anwendungssoftware. Dazu kommen eine Liste maschinenlesbarer Korpora des Englischen und eine umfassende Bibliographie mit über 650 Einträgen.

Im ersten Teil vergleicht Steven J. DeRose verschiedene probabilistische Tagging-Algorithmen, Geoffrey Leech und Roger Garside berichten von ihren Arbeiten bei der halbautomatischen Erstellung syntaktischer Datenbanken (sogenannter *tree banks),* aus denen inkrementell eine probabilistische Phrasenstrukturgrammatik abgeleitet werden soll. Dabei gehen sie zum Teil sehr detailliert auf alltägliche Probleme ein, mit denen sie sich beschäftigen mußten. Den Abschluß bilden Clive Souter und Tim F. O'Donoghue, die in ihrem Beitrag zum probabilistischen Parsen unter anderem feststellen, daß nicht nur Lexeme, sondern auch Phrasenstrukturregeln den durch Zipfs Gesetz bestimmten Verteilungen folgen.

Pieter de Haan geht im zweiten Abschnitt mehr ins Detail, wenn er die Verteilung postmodifizierender Teilsätze in englischen Nominalphrasen untersucht. Christian Mair stellt die Frage *Quantitative* or *qualitative corpus analysis'?* und beantwortet sie mit 'sowohl als auch'. Man dürfe über den großen Textmengen nicht vergessen, daß Korpora auch Sammlungen authentischer Daten darstellen, deren qualitativ-exakte Untersuchung realistischere Ergebnisse liefert als die bloße linguistische Introspektion.

Im Abschnitt Lexis geht Magnar Brekke auf lexikale Ambiguität ein, die im Englischen ein großes Problem darstellt. Sein Beitrag *Automatic Parsing meets the Wall* schildert die Probleme, die die verschiedenen Bedeutungen von *wall* bei der maschinellen Übersetzung aufwerfen. Der Kontext bietet hier zwar einige Anhaltspunkte, prinzipiell ist aber bei so marginalen Problemen schon ein immens großer Teil an Weltwissen erforderlich. Im folgenden Artikel von Piek Vossen geht es um einen Aspekt der Gewinnung semantischer Informationen aus einer maschinenlesbaren Version von *Longmans Dictionary of Contemporary English.* Es wird versucht, das Problem der Polysemie mithilfe der ausführlichen Kodierungen in diesem Wörterbuch zu bewältigen, was mit Ausnahme der abstrakten *top-level*- Wörter brauchbare Ergebnisse liefert. Dabei kommen als Seiteneffekt Inkonsistenzen zutage, wie bei-

spielsweise zirkuläre Definitionen von kleinen Wortgruppen.
Bengt Altenberg berichtet im Anschluß über Ergebnisse bei der Untersuchung verstärkender Adjektive im London-Lund-Korpus gesprochener Sprache. Es unterscheidet zwischen *Maximizern* und *Boostern,* wobei er bezüglich Restriktionen und Präferenzen innerhalb von Kollokationen große Differenzen bei diesen bei den Klassen feststellt. In einem weiteren Beitrag im Bereich gesprochener Sprache beschäftigt sich Gerry Knowles mit der Frage, wie Prosodie in Korpora gesprochener Sprache markiert werden kann. Der erste Schritt, die Kennzeichnung von Tongruppen, wird zusammen mit Problemen und möglichen Anwendungen näher erläutert. Ebenfalls mit Prosodie beschäftigt sich Anne Wichmann. Sie geht näher auf die Verteilung und Funktion von Grundtonerhöhungen ein.

Peter Collins vergleicht den Gebrauch von Modalverben in australischem Englisch mit dem in britischem und amerikanischem Englisch. Er stellt dabei Unterschiede fest, die z.B. registerabhängig sind. Zwei Register, gehobenes und dialektales Englisch werden im Anschluß von Ossi Ihalainen in bezug auf die Kategorie des Subjekts verglichen. Mit Kategorie meint er beispielsweise Personalpronomen, Relativpronomen und Subjektsellipsen. Er stellt große Unterschiede fest und zeigt außerdem, daß er mit einer relativ kleinen Stichprobe hinreichende Ergebnisse bekommt. Den Abschluß des Kapitels über regionale und soziale Varianten bildet Gerhard Leitners Beitrag über Diversifikationserscheinungen in Varianten des Englischen in verschiedenen Teilen der Welt, vor allem in ehemaligen Kolonien. Er verdeutlicht am Beispiel der Lexik des indischen Kolhapur Korpus, daß sowohl das Englische als auch das Amerikanische großen Einfluß auf Sprachvarianten in den Entwicklungsländern haben.

Um spezialisierte Korpora geht es im Artikel von Dorrit Faber und Karen M. Lauridsen. Sie beschreiben das Vorgehen ihrer .Projektgruppe bei der Erstellung eines dreisprachigen Korpus (dänisch/englisch/französisch) im Bereich der Vertragsgesetzgebung. Sie nennen zwar ihre Kriterien bei der Auswahl der verwendeten Texte, theoretische Fragen der Repräsentativität werden aber leider nicht erörtert. Weiterhin geben sie die Konditionen an, zu denen man ihren Korpus erhalten kann.
Magnus Ljung vergleicht die Anforderungen des TEFL *(Teaching English as a Foreign Language)* mit denen in der Realität (hier den 18 Millionen COBUILD-Wörtern). Dabei spricht er vielfältige Probleme an, die sich beim Vergleich zweier Korpora ergeben. Nicht nur macht die unterschiedliche Größe der Korpora Schwierigkeiten (das Korpus der untersuchten Lehrbücher umfaßt lediglich 1,5 Millionen Wörter, d.h. noch nicht einmal zehn Prozent des COBUILD-Korpus), auch die Suche nach methodisch akzeptablen Vergleichsverfahren ist nicht einfach. Er kommt durch seine Vergleiche zu dem Ergebnis, daß schwedische Schulabgänger im Englischunterricht nur unzureichend auf sprachliche Tätigkeiten vorbereitet werden, zu denen das Verständnis von abstrakten Texten notwendig ist.

Im Software-Teil werden drei Ansätze zur maschinellen Bearbeitung von Korpora vorgestellt. Dazu gehört PC Beta, ein Tool, das in vielen Fällen alleine zur vollständigen Bearbeitung korpusorientierter Aufgabenstellungen unter DOS ausreicht. Benny Brodda gibt neben einer kurzen Programmbeschreibung auch Beispiele hierzu, unter anderem Textnormalisierung, Konkordanz-Erstellung und Tagging. Knut Hofland stellt einige Konkordanzprogramme für Personalcomputer vor und vergleicht deren Laufzeiten auf verschiedenen Rechnermodellen. Zum Abschluß beschreibt Jacques Noel, wie das Unix-Tool *awk* zur musterbasierten Weiterverarbeitung unterschiedlicher Korpusty-

pen gebraucht werden kann, um die Fähigkeiten der spezialisierteren Textbearbeitungsprogramme sinnvoll zu ergänzen. Allerdings bezieht er sich auf die 'Magerversion', die unter MS-DOS läuft, was Konsequenzen hinsichtlich des verfügbaren Speichers nach sich zieht. Allgemein ist es verwunderlich, in welch hohem Maße das im Vergleich zu Unix doch sehr restringierte Betriebssystem DOS für Massentextverarbeitung genutzt wird, wo es doch schon verschiedene Unix-Implementationen für Personalcomputer gibt, die besseren Gebrauch von den Resourcen des Computers machen.

Das krönende Ende des Bandes bilden eine nützliche Liste von 36 maschinenlesbaren (überwiegend englischsprachigen) Korpora nebst Kurzbeschreibungen und die anfangs erwähnte Bibliographie. Von bei
den Beiträgen sind aktualisierte Versionen erhältlich, unter anderem auch in maschinenlesbarer Form per e-Mail-Fileserver. Ich meine, eine lohnende Lektüre für alle am Thema Interessierten.

Oliver Jakobs, Trier

Reinhard Köhler / Andreas Janßen: PASCAL Programmieren für die Sprach- und Textwissenschaften. UTB (Francke), 1991, IX und 250 Seiten.

Zielsetzung des Buches

Einführungen in PASCAL gibt es hunderte, weshalb bei jeder Neuerscheinung zu fragen ist: War eine weitere wirklich notwendig? Das vorliegende Buch hebt sich in sofern von den bisher vorliegenden Einführungen ab, als es sich besonders an Studenten aus dem Bereich der Geisteswissenschaften, speziell der Computerlinguistik richtet. Wer Computerlinguistik unterrichtet, vor allem in solchen Studiengängen, die in einem geisteswissenschaftlichen Umfeld angesiedelt sind, weiß, daß eine solche Einführung tatsächlich ein Desideratum darstellt. GeisteswissenschaftlerInnen haben oftmals immer noch eine hohe Schwellenangst vor dem Computer und haben natürlich auch andere Voraussetzungen und Bedürfnisse als z.B. MaschinenbauerInnen oder ElektrotechnikerInnen.

Inhalt

Um den speziellen Bedürfnissen von Geisteswissenschaftlern Rechnung zu tragen, haben die Autoren einen Einleitungsteil vorangestellt, der ausführlicher ist als in anderen Pascal-Einführungen gewohnt. Er besteht aus einer *Einführung* (9 *S.)* zu den Themen Computer, Programmieren und Pascal, dem *Kapitel* 2 *"Grundbegriffe"* (7 *S.)* zu Algorithmen und Variablen und dem *Kapitel* 3 *"Programmiertechnik"* (6 *S.)* zu den Themen Programmentwicklung und Prinzipien der Softwaretechnik.

Die Einführung in Pascal beginnt dann im *Kapitel* 4 *"Die einfachen Sprachelemente von Pascal"* (51 *S.).* Wie auch im Rest des Buches werden die Sprachelemente durch natürlichsprachliche Erläuterungen, EBNF - Regeln und Beispiele vorgestellt. Zu den "einfachen Sprachelementen" zählen die Autoren (in der Reihenfolge der Behandlung): Bezeichner, Zahlen, Zeichenketten, Operatoren, Begrenzer, Kommentare, Programmaufbau, Variablenvereinbarung, einfache Datentypen, Aufzählungstypen, Teilbereichstypen, Standardtypen, Konstantenvereinbarung, Ausdrücke, arithmetische Operatoren und Ausdrücke, Vergleichsoperatoren, logische Operatoren und Ausdrücke, Elementare Ein- und Ausgabe, Zuweisung, WHILE-Anweisung, REPEAT- Anweisung, FOR-Anweisung, Verbundanweisung, IF-Anweisung, CASE-Anweisung,

Typvereinbarung.

Das *Kapitel* 5 *"Funktionen und Prozeduren"* (48 *S.)* stellt die im Titel genannten Sprachelemente im Zusammenhang und recht ausführlich und vollständig dar. Behandelt werden u. a.: Rekursion, indirekte Rekursion, Gültigkeitsbereiche und Lebensdauer von Vereinbarungen.

Im *Kapitel* 6 *"Strukturierte Datentypen"* (48 *S.)* geht es um ARRAY, RECORD, SET und FILE, während dem Datentyp POINTER ein eigenes *Kapitel* 7 *"Dynamische (Rekursive) Datenstrukturen"* (55 *S.)* gewidmet ist, in dem es außer um POINTER auch um Listen, Bäume und Netze geht, die sich mittels POINTER-Strukuren darstellen lassen.

Der *Anhang* (18 *S.)* enthält noch einmal die Syntax von PASCAL in EBNF-Form und in Syntaxdiagrammen, eine Übersicht über vordefinierte Bezeichner und Standardprozeduren und -funktionen, so wie eine ASCII-Tabelle und ein extrem kurzes Literaturverzeichnis (2 Titel).

Kritik

Bei der Bewertung des Buches muß oberstes Kriterium sein, inwieweit das Buch seinem eigenen Anspruch, speziell für Geisteswissenschaftler zugeschnitten zu sein, genügt. Leider muß man sagen, daß dies nur mäßig der Fall ist. Gut ist zunächst die ausführliche Einleitung, die auch solche Dinge einführt, die - wie die Autoren zu Recht sagen - zumeist als selbstverständlich vorausgesetzt werden, es aber oftmals gar nicht sind, etwa das Variablenkonzept. Gut auch, daß die Autoren sich bemühen, ihre Beispiele dem Erfahrungsbereich der Leser, etwa der Linguistik, zu entnehmen, auch wenn dies zum Teil etwas bemüht wirkt, so bei folgendem Beispiel zur Illustration der IF - Anweisung:

"IF ((Endung = 'chen') OR (Endung = 'lein'))
AND (Stamm <> 'Mäd') THEN Funktion :=
Diminutiv
[So 69]

Wenig gelungen ist jedoch der sonstige didaktische Aufbau des Buches. In Kapitel 4 scheint es beispielsweise, daß man die Syntaxdiagramme eines nach dem anderen "abgearbeitet" hat, statt den Leser behutsam in einer sinnvollen Reihenfolge mit den wichtigsten Konstrukten bekannt zu machen. Ob der Leser Aussagen zum Thema Aufzählungstypen wie

> *"Die so aufgelisteten Konstanten sind ordinalskaliert;. sie besitzen in der Reihenfolge der Angabe aufsteigenden Wert, ein Intervall ist jedoch nicht definiert."* [S. 37]

wirklich zu würdigen weiß, bevor er auch nur ein lauffähiges Mini- Programm gesehen hat, erscheint fraglich. Bezeichnend,daß man erst 60 S. später etwas über ORD, SUCC und PRED erfährt. Auch die so sehr aufeinander abgestellten Konstrukte FOR-Anweisung und ARRAY werden in einem Abstand von 60 Seiten vorgestellt; die TYPE-Vereinbarung ist eingeklemmt zwischen die CASE-Anweisung und das FUNCTION-Konzept. Das Lernen in Sinnzusammenhängen wird dadurch nicht gerade gefördert und die Hemmschwelle vor dem ersten Programm wird nicht dadurch abgebaut, daß man zunächst eine Menge nebensächlicher Detailregelungen zur formatierten Ausgabe oder zur Bindungsstärke von logischen Operatoren zur Kenntnis nehmen muß.

Zur Verteidigung könnte vorgebracht werden, daß späteres Nachschlagen bei der gewählten Reihenfolge leichter fällt, doch ein Stichwortverzeichnis (es fehlt leider) hätte da sicherlich mehr gebracht.

Insgesamt erscheinen die Kapitel 5, 6, und 7 besser gelungen als Kapitel 4; vor allem die Darstellung der Funktionsweise von POINTER-Variablen, oft eine Klippe für PASCAL-Anfänger, ist recht gut gelungen und mit vielen Abbildungen versehen. An anderer Stelle würde man sich jedoch ein paar mehr Abbil-

dungen wünschen, etwa zum Aufbau von Rechenanlagen oder zur Sichtbarkeit von Deklarationen in geschachtelten Prozeduren. Durchgängig werden Aufgaben gestellt, Musterlösungen werden nicht gegeben.

Gut gemeint ist sicherlich die Fülle von Beispielen, die die Autoren bringen, doch auch hier gilt wieder: "Weniger wäre mehr gewesen". Denn die Autoren haben sich entschieden, immer wieder neue Beispiele zu bringen. Vielleicht wäre es sinnvoller gewesen, einige wenige, durchgängige Beispiele nach und nach auszubauen. So stehen die Beispiele oftmals etwas zusammenhanglos im Text; oftmals werden Konzepte nur kurz erläutert und dann nahezu kommentarlos ein (oder mehrere) Beispiele angefügt. Eine schrittweise Entwicklung der Beispiele und die Hervorhebung dessen, worauf es ankommt, würde dem Anfänger ihr Verständnis erleichtern. Durch die Fülle der Beispiele wird der Leser nicht ermutigt, sie alle einzutippen und auszuprobieren. Doppelt frustrierend für den, der es trotzdem tut, wenn die Beispiele fehlerhaft sind. Beispiel:

```
WHILE NOT EOF(Schritte) DO BEGIN
  CASE Schritte .Befehl OF
    Plus: Register := Register + Schritte .Zahl;
    Minus: Register:= Register - Schritte. Zahl;
    Mal: Register := Register * Schritte .Zahl;
    Durch: Register:= Register / Schritte .Zahl;
  END; (* CASE *)
  WRITELN (Register: 10:2)
END; (* WHILE *)
```

[So 163]

Die entstehende Endlosschleife (es fehlt ein GET) dürfte den Anfänger sehr verwirren.

Der letzte Kritikpunkt, der hier vorgebracht werden soll, ist sicherlich der problematischste, da man über ihn geteilter Meinung sein kann und er die schwerwiegendsten Konsequenzen hat. Bekanntlich gibt es nicht nur sehr viel PASCAL-Einführungen, sondern auch sehr viele PASCAL-Dialekte. Die Autoren lösen dieses Problem sehr rabiat, indem sie sich auf das ursprüngliche Standard-Pascal beschränken. Diese Entscheidung ist jedoch für mich heute weder sinnvoll, noch notwendig. Sie ist nicht sinnvoll, weil die Zeit seit 1970 nicht stehengeblieben ist. Wesentliche Neuerungen, die inzwischen auf dem Gebiet der Informatik erfolgt sind und ihren Weg in neuere Pascal-Varianten gefunden haben, werden so ausgespart: Modularisierung (Unit-Konzept) und Objektorientierung, um nur die wichtigsten zu nennen. In Standard-Pascal fehlt sogar noch String- Verarbeitung, die gerade für den erklärten Leserkreis von besonderer Relevanz ist. Diese Selbstbeschneidung ist auch nicht notwendig, da - bei allen Unterschieden - die verbreitetsten Pascal-Varianten (z.B. Turbo-Pascal und Macintosh-(Object-)Pascal) über diese Erweiterungen in vergleichbarer Weise verfügen, zumal sie den Ahnen UCSD-Pascal gemeinsam haben. Wer hingegen arbeitet heute noch mit Standard-Pascal?! Vielleicht ist es sowieso garnicht sinnvoll eine Einführung in PASCAL implementierungsunabhängig versuchen zu wollen, da es die von Köhler/Janßen geschilderten PASCAL-Implementierungen, bestehend aus unabhängigem Editor und Compiler [So 6] heute kaum noch gibt (es sei angemerkt, daß das Buch auf Vorlesungsskipten aus dem Jahr 1985 zurückgeht). StudentINNen werden heute eher auf Umgebungen wie Turbo-Pascal treffen, in denen integrierte Werkzeuge wie syntaxgesteuerte Editoren, Source- Level- Debugger o. ä. zum Standard gehören. Evtl. ist daher eine integrierte Einführung, die in PASCAL und die jeweilige Implementierung einführt, sinnvoller.

Zusammenfassung

Wer eine Einführung in Standard-PASCAL sucht, findet in diesem Buch eine, die

auf GeisteswissenschaftlerInnen besser zugeschnitten ist als manche andere auf dem Markt. Günstig ist auch der erschwingliche Preis (da UTB). Zum Selbstlernen oder als alleinige Grundlage eines Kurses ist es jedoch kaum geeignet, da der Leser oft überfordert sein dürfte und daher zusätzliche Erläuterungen benötigt und außerdem eine Einführung in die jeweilige Implementierung benötigt wird. DIE Einführung in die Programmierung in PASCAL für GeisteswissenschaftlerInnen bleibt nach wie vor ein Desideratum.

Nils Lenke, Universität Duisburg

Jürgen Krause/ Ludwig Hitzenberger (Hrsg.): Computer Talk Sprache und Computer, Band 12. Hildesheim, Zürich, New York: Olms, 1992. 184 S.

Diese Publikation besteht aus acht Beiträgen:

1. Natürlichsprachliche Mensch-Computer-Interaktion als technisierte Kommunikation: Die computer talk-Hypothese (J. Krause) .

2. Computer talk-Merkmale in den USLStudien (J. Krause).

3. Modellbildung, Versuchsaufbau und Durchführung in DICOS (L. Hitzenberger).

4. Programmdokumentation (L. Hitzenberger, F. Kireh).

5. Experimentelle Grundlagen und statistische Auswertung von Hypothesentests zur Mensch-Computer-Interaktion (Chr. Womser- Hacker).

6. Abweichungen und Überspezifikationen als mögliche Merkmale von computer talk in den DICOS-Versuchen (B. Mielke, Chr. Womser- Hacker).

7. Unterschiede zwischen Mensch-Computer-Interaktion und zwischenmenschlicher Kommunikation aus der interpretativen Analyse der DICOS-Protokolle (H. Kritzenberger)

8. Fazit und Ausblick: Registermodell vs. metaphorischer Gebrauch von Sprache in der Mensch-Computer-Interaktion (J. Krause).

In diesen Beiträgen werden, ausgehend von der USL- Evaluierung im Rahmen der KFG-Studie, ALP-Studie und weiteren kleineren Evaluierungsstudien (1980/81 ff.), die Ergebnisse des 1988 - 1990 am Fachgebiet "Linguistische Informationswissenschaft" der Universität Regensburg unter der Leitung von Jürgen Krause durchgeführten DICOS-Projektes im Zusammenhang vorgestellt, soweit sie über die für den SPICOS- Verbund zu liefernden spezifischen empirischen Daten (Wortschatz, Satzstrukturen, Besonderheiten für die beiden Scenarios Bahn- und Bibliotheksauskunft) hinausgehen.

Der Generalisierungsaspekt des DICOS-Projektes kreist um die Frage, ob bei der natürlichsprachlichen MCI (NL-MCI) Besonderheiten der Sprachverwendung zu erwarten sind, die sich nicht aus einer Analogie zur Mensch-Mensch-Kommunikation herleiten lassen.

Globales Thema des Bandes ist also "weder die Fachsprache der Computernutzer noch der Einfluß technologisch affizierter Interaktionsformen auf die zwischenmenschliche Kommunikation", sondern die Frage, "wie sich Menschen ausdrücken und verhalten, wenn sie Computersystemen, die natürlichsprachliche Eingaben zulassen, Anweisungen geben und sie befragen, und was daraus für die Realisierung von Com-

putersystemen folgt". (S. 1)

Wie diese Frage überhaupt entstehen konnte, darüber gibt Krause im ersten Beitrag Auskunft. Ausgangspunkt dabei ist, ob der wissenschaftliche Erkenntnis- und Forschungsanspruch der Computerlinguistik und der sprachorientierten KI-Forschung (NL-MCI als Simulation von zwischenmenschlicher Kommunikation) auf eine "Benutzer-Realität" trifft, oder anders ausgedrückt: ob ein Benutzer die Interaktion mit einem natürlichsprachlichen System als der zwischenmenschlichen Kommunikation zum Verwechseln ähnlich auffaßt, oder aber eben als eine Interaktion, die eine andere eigenartige Sprachverwendung erfordert.

Experimentelle und kommerzielle NL-Systerne standen von Anfang an - aufgrund der technologischen Entwicklung - in einer besonderen Konkurrenzsituation zu Systemen mit einen formalsprachlichen Zugang. Als "Zweitgeborene" mußten sie ständig nicht nur ihre Nützlichkeit, sondern auch ihre "Natürlichkeit" unter Beweis stellen. Die Natürlichkeit stieß und stößt aber dort an ihre Grenzen, wo ihre Machbarkeit als Voraussetzung nicht gegeben war (und ist): die Erkenntnisse von Linguistik, Kognitionspsychologie, etc. sind auch heute keineswegs so, daß sie Systeme zu implementieren gestatteten, die dem Anspruch der "Natürlichkeit" gerecht werden (daher ist der Ausdruck "natürlichsprachliche Systeme" für sich genommen ein Euphemismus, der mehr verstellt als erhellt; er ist nur historisch begreifbar als Gegensatz zu "formalsprachliche Systeme" - ein Ausdruck, der so in die Fachliteratur aber kaum Eingang gefunden hat). Dies hat nun aber Auswirkungen auf die sprachlichen Interaktionsmöglichkeiten zwischen Benutzer(n) und System und die Einschätzung dieser Möglichkeiten durch den (die) Benutzer.

Die These von Krause lautet: "Menschen verhalten sich bei der natürlichsprachlichen MCI anders als in der zwischenmenschlichen Kommunikation." (S. 6) Krause geht es nun um den Nachweis von "computer talk" - so die tentative Bezeichnung dieses besonderen Sprachregisters - im Sinne eines "Existenzbeweises" (so an mehreren Stellen), nicht aber schon darum, wie "die Wirkungsweise selbst adäquat modelliert werden (soll)" (S. 23). Der Nachweis eines "computer talk" hätte erhebliche Konsequenzen für die Modellbildung im Bereich der NL-MCI wie für die Konstruktion praxisrelevanter Systeme.

Ehe dieser Nachweis selbst geführt werden kann, werden einerseits Begriffe wie "sublanguage", "subset", "Mächtigkeitsgrammatik" und "Sprachregister" (die alle schon von der Linguistik her für die "Beschreibung" der NL-MCI importiert worden waren) geklärt und für die Analyse der NL-MCI präzisiert und muß andererseits die methodische Frage beantwortet werden, "wie empirische Tests aussehen können, die eine Verifizierung unserer These ermöglichen" (S. 7). In Abschnitt 1.2 *(Sublanguage und subset)* wird die gemeinsame Basis des sublanguage-Konzepts (seit Harris 1968) herausgearbeitet und als an thematische Restriktionen gebunden charakterisiert. Demgegenüber ist das subset-Konzept primär bezogen auf das Charakteristikum "habitability" im Kontext der ersten Entwicklungen von natürlichsprachlichen Frage-Antwort-Systemen: "einerseits soll der Sprachumfang möglichst klein sein (subset), damit er realisierbar bleibt und ökonomisch arbeitet, und andererseits muß der Benutzer in der restringierten Anfragesprache seinen Informationswunsch problemlos ausdrücken können." (S. 12) Demgegenüber ist das Sprachregister- Konzept (vgl. Abschnitt 1.3: *Sprachregister)* zwar auch auf die Merkmale 'Bezug zur Standardsprache' und 'eingeschränkte inhaltliche Domäne' bezogen; diese sind aber nur zwei unter vielen anderen situativen Faktoren wie Zweckbestimmung, Kommu-

nikationsziel, soziale Rollen und Eigenschaften der Dialogpartner, Kanalfaktoren, Zeit- und Raumbeschränkungen, soziale Distanz, beschränkte Sprachkompetenz bei einem Dialogpartner.

Abschnitt 1.4: *Methodologische Fragen und empirische Basis* beleuchtet die Notwendigkeit empirischer Forschung für die NL-MCI im Lichte der Registerforschung, macht zugleich aber auch deutlich, mit welchen Imponderabilien diese Art von Forschung (derzeit noch 7) zu rechnen hat und wie den daraus erwachsenen Problemen durch einen zweistufigen Testaufbau (Hypthesengenerierung und nachfolgende prüfstatistische Verifikation bzw. Falsifikation) begegnet werden kann - eine Methode, die eben im DICOS-Projekt verfolgt wurde.

Im zweiten Beitrag arbeitet Krause mittels interpretativer Protokollanalysen zu den bei den großen USL-Studien und unter Berücksichtigung einiger Beobachtungen in der einschlägigen Literatur heraus, daß sich die typischen Registermerkmale 'Abweichung', 'Simplification', 'Clarification', 'Upgrading' und 'Expressiveness' mit Ausnahme des letzteren auch bei NL-MCI feststellen lassen und damit deutliche Hinweise auf das Vorliegen eines Sprachregisters 'computer talk' liefern. Die dabei ermittelten Merkmale geben die Grundlage ab für die Hypothesenbildung für die prüfstatistischen Verfahren der zweiten Stufe.

Ehe im fünften Beitrag Womser-Hacker die statistischen Untersuchungsmethoden, ihre Auswirkungen für den Versuchsaufbau und die Auswertungsmethode, das experimentelle Design und die Ergebnisse der Hypothesentests beschreibt, werden im dritten und vierten Beitrag von Hitzenberger und Kirch die Aspekte ,der Modellbildung, der Versuchsaufbau mit der Methode der Hidden-Operator-Simulation, die technische Durchführung und die Dokumentation der für die technische Durchführung und die Auswertung entwickelten Programme gedrängt dargestellt. Ziel war die Schaffung einer konsistenten, stabilen und für den hidden operator auch handhabbaren Versuchsanordnung zur Ermittlung von NL-MCI mit vier qualitativ unterschiedlichen simulierten (Computer)systemen bei zwei verschiedenen Domänen mit je zwei verschiedenen Kommunikationskanälen (geschrieben/gesprochen).

Die von Womser- Hacker referierten Ergebnisse der Hypothesentests besagen, daß Computer talk bei der Interaktion mit Computersystemen häufiger zu beobachten ist als bei zwischenmenschlicher Kommunikation, daß Computer-talk-Eigenschaften in der Interaktion mit Computersystemen um so mehr zunehmen, je restringierter diese Systeme sind. Darüber hinaus geht die Verwendung partnerorientierter Dialogsignale - wie sie in zwischenmenschlicher Kommunikation ganz selbstverständlich sind - bereits bei der Interaktion mit einem "optimalen" Computersystem (d.i. einem System, das sich einem Menschen gleichwertig verhält) deutlich zurück. Gerade der letzte Befund läßt sich auf den Einfluß des Computerbildes aufseiten des Benutzers zurückführen: "d.h. wenn Menschen mit Computern kommunizieren, die sich ebenso kooperativ wie Menschen verhalten, variieren sie dennoch ihre sprachliche Ausdrucksweise" (S. 104).

Der sechste und siebte Beitrag sind der interpretativen Protokollanalyse des DICOS-Materials gewidmet. Als wichtigste Ergebnisse bleiben festzuhalten:

(a) Es gibt starke Indizien dafür, daß die für Computer talk potentiellen Registereigenschaften 'Sprachliche Abweichung' und 'Überspezifizierung' als tatsächliche Eigenschaften des Registers Computer talk anzusehen sind (vgl. S. 120).

(b) Die Eigengesetzlichkeit der M CI ge-

genüber der zwischenmenschlichen Kommunikation demonstriert sich darin, daß "Benutzer ihr Sprachverhalten gegenüber dem Dialogpartner Computer (verändern), um ihm das Verstehen zu erleichtern, z.B. durch Einschränkung der syntaktischen Vielfalt oder durch Formulierungen, die gegenüber dem normalen Sprachgebrauch als abweichend oder überspezifiziert erscheinen, durch verstärkte Kontrolle der Sprachproduktion, oder durch den Verzicht auf pragmatische Mittel" (S. 156).

(c) "Weiter brachte die interpretative Protokollanalyse auch deutliche Hinweise darauf, daß sprachliche Restriktionen der Informationssysteme sowie Modalitätsunterschiede [geschriebener vs. gesprochener Input; H.D. Lutz] ebenfalls zu Sprachvarianten führen oder Veränderungen im Sprachgebrauch verstärken." (ebd.)

Im abschließenden Beitrag unternimmt Krause den Versuch, eine über das linguistisch motivierte Registerkonzept hinausgehende Modellvorstellung für die MCI zu begründen. Die Notwendigkeit dafür liegt in Dialogsequenzen begründet, die mit dem Registermodell nicht mehr erklärbar sind. Ziel ist eine einheitliche kognitiv orientierte Modellvorstellung, die das Registerkonzept einschließt und die gleichzeitig als Grundlage für die beiden Formen "natürlicher" MCI, nämlich die grafisch-direktmanipulative wie die natürlichsprachliche, dienen soll. Krause knüpft an an die Vorstellung von mentalen Modellen im Umkreis der direktmanipulativen Interaktionform und an die Methode, den Aufbau von adäquaten mentalen Modellen aufseiten des Benutzers durch die Verwendung einer lebensweltlich fundierten Metapher zu unterstützen. So wie Benutzer in einer derartigen Metapher-Umgebung sich im klaren darüber seien, daß sie sich in einer 'als-ob'-Situation bewegen, so ließen die Beispielsequenzen, die sich nicht mehr mithilfe des Sprachregisterkonzepts erklären ließen, erkennen, daß die Benutzer auf eine "als-ob'-Nutzung von Sprache" (S. 186) ausgewichen seien.

Die "Natürlichkeit" der natürlichsprachlichen MCI liegt zwar darin, daß der Benutzer diesen Kommunikationsmodus bereits beherrscht. MCI ist jedoch Analogiebildung im *Sinne des Metapherngebrauchs, nicht Gleichsetzung. Der Benutzer weiß wie bei der Schreibtischmetapher -, daß es nicht um Sprachbeherrschung* im *Sinne der menschlichen Kommunikation geht. Die elektronische Welt läßt sich zwar leichter durch die Metapher der zwischenmenschlichen Kommunikation erschließen, erlaubt aber Unterschiede. Der Benutzer tut so, als ob der Computer die natürliche Sprache beherrscht. Er verhält sich so, als ob der Computer ein ganz besonderer Gesprächspartner mit spezifischen Eigenschaften ist (wie z.B. auch beim Ausländerdeutsch). Es überrascht ihn jedoch nicht weiter, wenn die Analogie nicht mehr trägt. Metaphern gelten nicht absolut; sie lassen Abweichungen zu. Das gehört zu ihrem Wesen. In Teilbereichen können sie durch Elemente ergänzt werden,. die zur Basisanalogie selbst nicht unbedingt kompatibel sein müssen.*

... Computer talk beinhaltet den potentiellen, partiellen Bruch mit der Analogie zwischenmenschlicher Kommunikation, genau dies erfaßt das Metaphernkonzept. (S. 167)

Computer talk - nun aufgefaßt als metaphorischer Gebrauch zwischenmenschlicher Kommunikation - kommt in Berührung mit der Benutzermodellierung aus der Sicht einer (kognitiv orientierten) Software-Ergonomie und aus der Sicht der KI. Dabei ist zu beachten, daß 'Benutzermodell' in den beiden Bereichen durchaus verschiedenes meint. Wird in der KI unter Benutzermodell ein Modell verstanden, das sich ein intelligentes Programm von der mit ihm interagierenden Person konstruiert, so wird

aufseiten der Software-Ergonomie darunter ein (mentales) Programm verstanden, das sich der Benutzer von dem mit ihm interagierenden System macht, oder auch auch ein generalisiertes Modell eines "typischen Benutzers", das der Konstrukteur eines Systems als Hilfe für den Entwurf eben dieses
Systems entwickelt. Krause plädiert nun wenn ich ihn richtig verstanden habe - für eine Annäherung der beiden Positionen, allerdings unter der Voraussetzung der o.g. einheitlichen kognitiv orientierten Modellvorstellung, und stellt am Ende seines Beitrages einige kurze Überlegungen dahingehend an, wie sich Register- bzw. computer talk-Überlegungen im Benutzermodell eines Computersystems bezogen auf den Bereich der Sprachgenerierung einerseits und bezogen auf den Bereich der M CI andererseits auswirken können.

Dieser Forschungsbericht - über die Arbeit von ca. 10 Jahren - gibt einen faszinierenden Einblick in der Problematik der NLMCI und ihrer (noch jungen) wissenschaftlichen Bewältigung. Allerdings möchte ich hierzu kritisch anmerken, daß diese Faszination (für mich) nicht nur positive Seiten hat. So hätte ich dieser Publikation gewünscht

=> ein Vorwort mit einer Leserführung und mit Hinweisen auf den forschungsstrategischen und forschungspraktischen Hintergrund der Regensburger Arbeiten
(jemand, der den DICOS-Hintergrund nicht kennt, wird Schwierigkeiten haben, bestimmte Vorgriffe im ersten Beitrag von Krause (v. a. S. 26ff.) zu verstehen und entsprechend einzuordnen),

=> einen Sachindex, der das Querlesen erleichtert (so gehört schon einige Konzentration und sehr gutes Erinnerungsvermögen dazu zu bemerken und zu verifizieren, daß sich eine Passage auf S.
168 (zur Sprachgenerierung) auf eine einschlägige Bemerkung auf S. 6 rückbezieht),

=> ein Literaturverzeichnis, das auch die SPICOS-Literatur aufführt, auf die auf den Seiten 70 und 95 Bezug genommen wird,

=> eine Liste mit der Auflösung aller benutzten Akronyme auf einen Blick,

=> eine Einordnung der Forschung in den internationalen Kontext (so fehlen etwa die Arbeiten, die am Department of Computer and Information Science der Linköping Universität zur NL-MCI durchgeführt wurden/werden, vollständig, was um so bedauerlicher ist, als (a) viele der dort erarbeiteten empirischen Befunde sich mit denen aus Regensburg decken, (b) der methodische Ansatz durchaus vergleichbar ist und (c) die Ableitungen bzgl. der Konstruktion von NL-MCI-Systemen in dieselbe Richtung gehen 2,

=> eine detaillierte Diskussion des Metaphernkonzepts, die u. a. auch zu berücksichtigen hätte, daß natürliche Sprachen bereits in ihrer primären Verwendungsumgebung, der zwischenmenschlichen Kommunikation, durchgängig tropisiert sind3, daß wir es im Kontext der NL-MCI also mit einer Metaphorisierung zweiter Stufe (oder besonderer Art) zu tun hätten oder aber, daß das Metaphorisierungskonzept - da bereits im primären Verwendungskontext feststellbar - gar nicht die differentia specifica der NL-MCI ist, womit dann aber der ganze Ansatz, MCI und damit auch NL-MCI als Analogiebildung im Sinne

2 vgl. etwa Jönsson, Arne; Dahlbäck, Nils, Talking to a computer is not like talking to your best friend, Research Report LITH-IDA-R-88-34, Linköping, Sept. 1988 oder Jönsson, Arne, A Natural Language Shell and Tools for Customizing the Dialogue in Natural Language Interfaces, Research Report LITH- IDA-R-91-10, Linköping, April 1991.

3 vgl. etwa Ungeheuer, Gerold, Vor-Urteile über Sprechen, Mitteilen, Verstehen; in: ders., Kommunikationstheoretische Schriften I: Sprechen, Mitteilen, Verstehen. Hrsg.v.G.J.Juchem (Aachener Studien zur Semiotik und Kommunikationsforschung. 14), Aachen 1987, S. 290-338.

des Metapherngebrauchs zu sehen, als gescheitert zu betrachten wäre.

Doch über diesen kritischen Anmerkungen möchte ich nicht die Positiva dieser Publikation vergessen:

- die Explizierung der Methodik, die Frage zu beantworten, was unter Computer talk verstanden werden kann bzw. soll und wie weit dieses Konzept trägt,
- die empirisch begründeten Hinweise für eine Methodologie zur Konstruktion von NL-MCI-Systemen,
- den Aufweis, daß es methodisch und instrumentell möglich ist, die Rahmenbedingungen für ein konkret zu entwickelndes NL- MCI-System bereits vor dem Entwurf "hart" zu machen, ohne sich dabei auf die Intuition der Systemkonstrukteure zu verlassen,
- den Brückenschlag zwischen einer Benutzungsmodellierung aufseiten der (kognitiv orientierten) Software-Ergonomie und einer Benutzermodellierung aufseiten der Kr.

Hans-Dieter Lutz, Universität Koblenz-Landau

Burghard Rieger: Unscharfe Semantik. Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten. Peter Lang Verlag Frankfurt, Bern, New York und Paris, 1989, 346 Seiten, DM 85.

Das Phänomen der semantischen Vagheit ist in letzter Zeit immer stärker in das Zentrum 1inguistischen Forschungsinteresses gerückt. Dies ist jedoch keine isolierte Erscheinung, wenn man an das große Interesse an sog. fuzzy- Themen innerhalb anderer, z. T. benachbarter Disziplinen wie z.B. der KI denkt. Das Buch von Burghard Rieger bietet eine ausführliche, interdisziplinär eingebettete Darstellung des Phänomens der Vagheit natürlichsprachlicher Wortbedeutungen aus linguistischer Sicht.

Untermauert durch eine Vielzahl von Quellen zeigt Rieger in den ersten beiden Kapiteln seines Buches den Stellenwert des semantischen Vagheitsphänomens im Laufe der sprachwissenschaftlichen und sprachphilosophischen Entwicklung. Interessant ist, daß dieses Phänomen seit dem klassischen Altertum bemerkt, aber, wenn nicht ganz als Thema ausgeklammert, so doch innerhalb der gesamten sprachwissenschaftlichen Tradition recht negativ gesehen wurde. Die Divergenz zwischen der Zeichenstruktur und den von ihr bezeichneten "Dingen" wurde als Defizit der natürlichen Sprache angesehen. Erst in diesem Jahrhundert wurde die Vagheit zum eigentlichen Untersuchungsgegenstand sprachwissenschaftlicher Forschung. In diesem Zusammenhang werden Lösungsvorschläge zur Erklärung des Vagheitsphänomens aus kognitions- und formaltheoretischen Blickwinkeln diskutiert. Allen diesen Explikationsversuchen steht jedoch meist noch als Pendant das Ideal der Präzision gegenüber.

In Kapitel 3 erfolgt die "Rehabilitierung" des Vagheitsphänomens, indem sein Wert innerhalb realer Äußerungszusammenhänge dargestellt wird. Interessant ist dieser erste empirische Versuch, dem Sprachsystem empirisch beschreibbare Spracherscheinungen zuzuordnen und diese zu quantifizieren. Die Vagheit wird dabei als Abweichungsgrad zwischen beobachteten Regularitäten und den entsprechenden Regelvarietäten und den dazugehörigen Regeln interpretiert. Eine Neuorientierung in bezug auf das Vagheitsphänomen wird durch die Auffassung Wittgensteins vollzogen. Die Vorstellung von einer prinzipiellen Vagheit aller natürlichsprachlichen Ausdrücke tritt bei Quine in den

Vordergrund. Eine Präzisierung vager Aussagen erfolgt durch die pragmatischen Rahmenbedingungen des situativen Kontexts. Die sich daraus ergebende Konsequenz war die sog. Situationssemantik, die Bedeutungen natürlichsprachlicher Ausdrücke in Abhängigkeit von ihren Kontexten analysierte, und zwar den Kontext nicht nur ergänzend einbezog, sondern zur Grundlage des gesamten Ansatzes machte. Trotz der möglichen Modellierung des Vagheitsphänomens ist dieser Ansatz auch kritisch zu sehen, da die Situationssemantik auf die Einheit des Satzes beschränkt bleibt und noch keine Hinweise auf eine Rekonstruktion der Grundeinheit 'Situation' gegeben werden.

Mi t Beginn der 70er Jahre ist innerhalb der Sprachwissenschaft eine verstärkt performanzorientierte und empirisch ausgerichtete Arbeitsweise vorherrschend, die der Sprachverwendung innerhalb tatsächlicher Kommunikationssituationen den Vorrang vor dem idealisierten Sprachsystem gibt. Der verstärkte Empiriebezug äußert sich auch durch den Rückgriff auf verbundene Rede enthaltende Sprach-Corpora, die im statistischen Sinne als Stichproben aus der Grundgesamtheit aller möglichen Äußerungen angesehen werden.

Kapitel 4 des Buches verfolgt nicht zuletzt eine methodische Absicht. Hier wird die Rolle der Statistik innerhalb der Linguistik näher beleuchtet, die anfangs eng mit der Corpusproblematik verbunden ist. Die Entwicklung von der reinen Belegsammlung bis hin zur repräsentativen Stichprobe aus der unendlichen Grundgesamtheit sprachlicher Äußerungen wird im Detail erörtert. Anders formuliert: der Übergang von der deskriptiven Statistik zur Inferenzstatistik, die das Prüfen und Generieren von Hypothesen zuläßt. Hier werden die zu erfüllenden statistischen Axiome diskutiert und mit den entsprechenden Gegebenheiten auf linguistischem Gebiet konfrontiert (z.B. Was ist hier die so oft zitierte Grundgesamtheit? Wann ist eine Stichprobe im statistischen Sinne repräsentativ?). Riegers Resüme: Die Statistik hat auf dem Gebiet der Linguistik als Methode höchste Relevanz, da es mit ihr möglich ist, aus unvollständiger oder unsicherer Information Aussagen abzuleiten und deren Gültigkeit mit bestimmten Wahrscheinlichkeiten zu versehen.

Kapitel 5 beschäftigt sich mit der Behandlung unsicheren Wissens (insbesondere semantischer Vagheit) vor dem Hintergrund kognitionspsychologischer Theorien und Modelle. Die Spannbreite reicht von dem Kantschen Schemabegriff, über die Carnapsche Vorstellung von linguistischer Disposition bis hin zum Frame-Konzept Minskys. Die Bedeutungskonstitution innerhalb des kognitiven Prozesses ist auch innerhalb der KI im Hinblick auf das Problem der Repräsentation von Wissensstrukturen eine zentrale Komponente. Neben allen Kontroversen innerhalb dieser Disziplin (z.B. um die deklarative und/oder prozedurale Form) ist das Fehlen der Empirie beim Aufbau von Wissensbasen bezeichnend. Die Ermittlung des Wissens, was heute oft mit knowledge engineering bezeichnet wird, verläuft meist auf intellektuellintrospektive Weise. Die Folge dieser Methode ist, daß nur bestimmte Wissensausschnitte repräsentiert werden, nämlich meist die, die sich durch den gewählten Formalismus ausdrücken lassen. D.h., vorhandene Wissensmodelle verfügen noch über keine Repräsentationsmöglichkeiten unsicheren Wissens, da die Kanten im Sinne von Bedeutungsbeziehungen zwischen den Bedeutungselementen als weitgehend statische Relationen aufgefaßt werden. Diesen bei den zentralen Fragen, wie kann vages oder unscharfes Wissen repräsentiert werden und wie kann man derartiges Wissen erheben, geht Rieger im weiteren Verlauf des Kapitels 5 nach. Ausgangspunkt sind die Modellvorschläge aus dem Bereich der Gedächtnisforschung über Art und Aufbau

semantischer Entitäten und deren Verarbeitung durch den Menschen. Hier setzt der Autor mit seiner Kritik an in bezug auf die weitgehende Beschränkung der Art der Experimente auf das konventionelle Paradigma des behavioristischen Ansatzes. Auf der Suche nach einer adäquaten Algebra für ein semantisches System führt der Autor die Unterscheidung zwischen prädikativem Wissen, dessen Strukturierung auf lexikalisierten, gesetzmäßigen Beziehungen basiert, und assoziativen Bedeutungsbeziehungen ein, deren Regelhaftigkeit weniger streng determiniert ist. Ziel ist es, Repräsentationsformen zu finden, die unterschiedliche Arten von Wissen erfassen und darstellen können.

In Kapitel 7 wird der semantische Neuansatz beschrieben, der, basierend auf der Theorie der unscharfen Menge, die seit 1965 von L. Zadeh auf linguistische Phänomene angewendet wird, die Unschärfe von Bedeutungen mit einbezieht. Dargestellt wird zunächst der Formalismus der sog. fuzzy set theory, was mit Hilfe von Beispielen auf verständliche Weise geschieht. Anschließend wird die Anwendung dieses Ansatzes innerhalb eines Bedeutungsmodells und dessen Einfluß auf neuere se*mantische Ansätze* untersucht. Ziel ist die Entwicklung einer theoretisch untermauerten Systemstruktur, welche durch die Abbildung von Wortverwendungsunterschieden eine empirisch-fundierte und prozedural definierbare Bedeutungsrekonstruktion zuläßt. Riegers Ansicht nach kann die Forderung nach adäquaterer Beschreibung vager Wortbedeutungen innerhalb der natürlichen Sprache durch das Konzept der unscharfen Menge wie es Zadeh vorschlägt nur zum Teil erfüllt werden, weil die informationelle Bedeutung lexikalischer Einheiten entscheidend für die Bedeutungskonstitution ist. Darunter versteht der Autor z.B. Anwendungsregularitäten und Ko- bzw. Kontextinformation. Rieger schlägt hier den direkten Rückgriff auf tatsächlich produzierte Äußerungen vor, um auf objektiver Grundlage die Regularitäten der Bedeutungsverwendung realer Sprecher/Hörer /Schreiber/Leser ermitteln zu können. Die Bedeutung eines Wortes kann folglich empirisch-quantitativ bestimmt werden "als Funktion aller Unterschiede aller seiner Verwendungsregularitäten zu sämtlichen anderen Einheiten des verwendeten Vokabulars in den analysierten Texten eines Gegenstandsbereichs" (S. 179). Die Theorie der unscharfen Menge kann als Formalismus dienen für eine derartige lexikalisch-semantische Bedeutungsnotation.

Die Kapitel 8 und 9 umfassen die Darstellung des Modells und zugleich den empirischen Teil des Buches. In Kapitel 8 wird zunächst das statistische Modell vorgestellt, dessen Kern auf der Verteilung der Lexeme auf Texte innerhalb eines Corpus und deren Korrelationen basiert. Augenfällig ist hier die entscheidende Rolle der Textauswahl für das Corpus, die zur bestimmenden Größe wird. Der Autor betont hier die Notwendigkeit, diese Modellvorstellung empirisch zu überprüfen, indem tatsächliches Textmaterial zugrundegelegt wird, damit die Resultate beurteilbar werden. Für sog. Bedeutungspunkte (semantische Entitäten) werden semantische Räume aufgespannt, die als unscharfe Mengen interpretiert werden können. Dabei wird die Algebra der unscharfen Mengen eingebracht und eine Interpretation im Sinne semantischer Relationen vorgenommen, wobei das Modell auch neue Bedeutungen (durch Verknüpfung mit schon gegebenen) erzeugen kann. Dabei geht der Autor von der These aus, daß die systematische Lage der Bedeutungspunkte zueinander die semantischen Ähnlichkeiten der entsprechenden Bedeutungen abbildet. Zur gezielten Analyse und Beschreibung wer den Cluster-analytische Verfahren eingesetzt, deren Grundlage zunächst detailliert erläutert wird. Ergebnis dieser Methode

ist Aufschluß über die interne Struktur von Bedeutungspunktgruppen und deren Ähnlichkeitsniveau.

Im folgenden 9. Kapitel werden, aufsetzend auf das beschriebene Modell des Repräsentationssystems für Wort-bedeutungen, die dynamischen Prozeduren zur Ermittlung spezifischer Relationen zwischen den Bedeutungspunkten im semantischen Raum beschrieben. Die Dynamik wird durch einen Algorithmus modelliert, der den Selektionsprozeß der Bedeutungselemente steuert und in Form von Dependenzstrukturen ausgibt. Trotz aller statistisch-wahrscheinlichkeitstheoretischer Raffinesse bleibt als Zweifel bestehen, ob derartige Analysen semantischer Strukturen überhaupt in der Lage sind, die Vielfalt natürlichsprachlicher Strukturen adäquat zu modellieren. Eine Fortführung des empirischen Ansatzes im Sinne einer repräsentativen Evaluierung des Modells innerhalb eines bestimmten Kontexts könnte für viele Disziplinen von Interesse sein.

Abschließend findet der Leser ein umfangreiches Literaturverzeichnis und ein Namens- und Stichwortverzeichnis.

Mag der Titel des Buches auch sehr genuin "linguistisch" erscheinen, so hat man es dennoch mit einem höchst interdisziplinären Buch zu tun, das seine Leser in höchstem Maße interdisziplinär fordert. Das Buch umfaßt ca. 350 Seiten und enthält eine vollständige Darstellung des Phänomens der unscharfen Semantik inkl. theoretisch-interdisziplinärer Wurzeln. Der Stil des Autors stellt höchste Anforderungen an den Leser, da sehr viele entlinearisierenden Stilmittel gebraucht werden, aber im Hypertextzeitalter ist das aller Wahrscheinlichkeit nach kein Problem mehr. Was mancher Leser (z.B. aus dem Bereich der Informationswissenschaft oder der Computer Science) vielleicht als Mangel empfinden mag, ist der Übergang von der Theorie zur praktischen Integration des Ansatzes innerhalb eines Computersystems. Hier darf man gespannt sein auf die weiteren Arbeiten des Autors, welche den Einsatz in der Praxis zeigen werden.

Christa Womser-Hacker, Universität Regensburg

Eileen Cornell Way: Knowledge Representation and Metaphor. Kluwer Academic Publishers, Dordrecht 1991 (271 S., geb.)

Mit dieser Monographie liegt der bisher anspruchsvollste Versuch vor, Metaphern computerlinguistisch zu erfassen. Was könnte daran so aufregend sein?

Computerlinguisten untersuchen Aufbau, Bedeutung und Verwendung menschlicher Sprache mit dem Ziel, Computer zu intelligenten Werkzeugen menschlicher Kommunikation zu machen. Grundsätzlich geht das nur soweit, wie Sprache algorithmisierbar ist, also nach letzten Endes eindeutig und im vorhinein formulierbaren Regeln funktioniert. In vielen Fällen sprechen Menschen aber gar nicht regelmäßig, sondern unmittelbar aus der Situation heraus und frisch auf den einzelnen Kontext bezogen, und das stürzt den regelorientierten Linguisten und Computerlinguisten in Schwierigkeiten. Was macht man zum Beispiel mit Fällen, in denen Sprecher gar nicht 'wörtlich' verstanden werden wollen? Meist gehen Computerlinguisten stillschweigend davon aus, daß maschinell erzeugte bzw. analysierte Texte 'wörtlich' zu verstehen seien, ohne Hintersinn, nicht ironisch, nicht metaphorisch. Denn Metaphern sind produktive Anarchisten in der Sprache; sie artikulieren ad hoc Erkenntnis, indem sie eingefahrene Regeln verwirren.

Angesichts derartiger Probleme wendet sich Way gegen das übliche Ziel der

Künstlichen- Intelligenz- Forschung, natürliche (menschliche) Sprache schnurstracks in ein formalisiertes Kalkül zu übersetzen. Sie möchte sich vielmehr zunächst theoretisch mit den Eigenschaften der menschlichen Sprache, hier also der Metapher, befassen, um die theoretischen Ergebnisse erst dann in eine formale Darstellung zu bringen. Folgerichtig diskutiert sie zunächst (Kapitell) das Verhältnis von wörtlicher und metaphorischer Bedeutung und sichtet dann (Kapitel 2) einige wichtige Grundzüge verbreiteter Metapherntheorien (leider ohne wichtige kontinentaleuropäische Ansätze wie Derrida, Eco, Ricour, Weinrich u. a.). Dabei entwickelt sie gute Argumente dafür, daß die kognitiven Mechanismen, die den metaphorischen Prozeßtragen, die menschliche Art, mit Wissen umzugehen, besser zeigen als die in der KI-Forschung meist benutzte Prädikatenlogik erster Stufe. Die Auseinandersetzung mit der Metapher bringt sie also zum Kernproblem der maschinellen Simulation menschlicher Intelligenz, nämlich der formalen Wissensrepräsentation. Ihr ist das dritte Kapitel gewidmet, zugleich eine gute Einführung in dieses Thema. Das vierte (Representation Schemes and Conceptual Graphs) führt einige später gebrauchte Spezialitäten aus.

Auf diesem Wege hat sich die Verfasserin mit allerlei Argumenten für eine Kombination zweier Positionen entschieden, im Bereich der Metapherntheorie nämlich für Max Blacks Interaktionstheorie und im Bereich der Wissensrepräsentation für John Sowas Conceptual Graph Theory. Beide gehen in ihren eigenen Ansatz ein, den sie Dynamic Type Hierarchy Theory of Metapher nennt und in Kapitel 5 vorstellt. Kapitel 6 zeigt überzeugend, wie konkurrierende computerlinguistische Ansätze den Kern des metaphorischen Prozesses verfehlen. In Kapitel 7 und 8 werden einige benachbarte sprachphilosophische Implikationen diskutiert, von der Struktur semantischer Hierarchien über das Verhältnis von Ideal and Ordinary Language Philosophy zur sprachorientierten KI-Forschung. Kapitel 9 schließlich gibt einige technische Details zur Programmierung des vorgeschlagenen Konzeptes.

Wie geht Way selbst nun vor? Wie üblich in der KI-Forschung modelliert sie Wissen in Typhierarchien, also in Netzwerken von Begriffen, die wie in porphyrischen Bäumen nach Allgemeinheitsgraden geordnet sind; die Bedeutung eines Begriffs hängt dann von seiner Position in der Typhierarchie ab. Way erkennt nun aber an, daß Wissen (also die Weltvorstellung von Sprechern/Hörern) sich ändert. Deshalb müssen Typhierarchien dynamisch gestaltet werden: abhängig von Zeit und Kontext. Genauer gesagt: wenn sprachliche Äußerungen auf Typhierarchien abgebildet werden, ist ein Zeitparameter einzufügen, der in die Typhierarchie eingreift (S. 126).

Im metaphorischen Prozeß nun sieht sie eine Technik der Dynamisierung von Wissen. Der Satz "Das Auto ist durstig" beispielsweise weckt Way (S. 135f) zufolge im Hörer bestimmte Teile seines begrifflichen Netzes, die 'vor der Metapher' nicht zueinander passen. Denn 'Auto' läßt an unbelebte Dinge denken, 'durstig' aber an Lebewesen. Um diese Unstimmigkeit zu bereinigen, muß ein Konzept gefunden werden, das Autos und Lebewesen in Bezug auf 'Durst' gemeinsam haben könnten, und siehe da: 'nach der Metapher' sind Autos und Lebewesen gleichermaßen 'bewegliche Einheiten, die Flüssigkeiten brauchen' (S. 137). Der Prozeß der Metaphorisierung besteht also darin, neue Konzepte und neue Beziehungen in begrifflichen Netzwerken zu erzeugen (vgl. ebd. xvii). Und das geschieht so:

Der Kontext des Satzes, in dem die Metapher vorkommt, legt sozusagen eine Maske über die an sich kontextneutrale Typhierarchie (vgl. S. 126). Diese Maske läßt die übliche Unterscheidung zwischen 'be

~

lebt' und 'unbelebt' in den Hintergrund treten, so daß Eigenschaften, die eigentlich nur Lebewesen zugeschrieben werden, nun auch einem toten Gegenstand zugesprochen werden können. So hat alles wieder seinen Platz. Und der Unterschied zwischen Menschen und Autos verschwindet, obwohl doch jeder Hörer weiß, daß Autos eben nicht im gleichen Sinne durstig sind wie Menschen.

Eigentlich also, das heißt außerhalb des Kontextes, sieht die Typhierarchie so und so aus; der konkrete Fall fügt ihr aber etwas hinzu oder blendet etwas aus. Entweder so oder so, aber nicht bei des zugleich! Auf diese Weise wird die Metapher wieder einmal entschärft. Denn tatsächlich erzeugen Metaphern ja doch eine unstimmige Spannung: Autos sind zugleich durstig und doch nicht durstig. Diese tatsächlich empfundene Spannung aber wird auch in Ways Ansatz aufgelöst zugunsten einer immer unterstellten sicheren Ordnung eindeutiger und letzten Endes unveränderlicher Identitäten. So kann der einzelne Fall entweder nur im nachhinein ad hoc beschrieben werden (als Ausnahme: wir tun so, als wären Autos durstig), oder es wird ihm das Widerspenstig- Einzigartige genommen (als Regelfall: Autos und Menschen brauchen halt Flüssigkeiten). Das ist genau das klassische Problem rationalistischen Denkens: wie fügt sich der widerspenstige einzelne Fall in ein vorab gültiges allgemeines Regelsystem? Oder in Ways Worten: wie werden in lebendigen Äußerungen (a la, "die Uhr holt Atem, du Esel, Verkehrsinfarkt, durstiges Auto") jeweils welche (von tatsächlich unendlich vielen) Masken erzeugt oder ausgewählt und warum gerade diese? Diese entscheidende Frage bleibt offen (S. 126). Way wiederholt einfach die althergebrachte Auffassung, daß eine Metapher die für einen Wissens bereich übliche Redeweise zum Teil auf einen anderen überträgt (vgl. S. 127). Sie schlägt vor, diese sprachliche Strategie bei der Konstruktion dynamischer begrifflicher Netze zu berücksichtigen. Sie bedenkt, daß Wissen dynamisch ist, weil es VOR unserer lebendigen Kultur abhängt. Aber sie modelliert die Wechselwirkung von Wissen und Leben nicht für die Maschine. Wie könnte sie auch? Sie ist ja unüberschaubar. Die Einbettung von Wissen in Leben läßt sich maschinell nicht wiederholen, weil wir dafür keine allgemeine Regel angeben können. Das hat Way, ohne es zu wollen, gezeigt. Ihr Buch liefert eine ungewöhnlich solide und gedankenreiche Annäherung an die Frage, wie situationsempfindliche menschliche Intelligenz auf situationsunabhängigen Maschinen simuliert werden kann. Sie kann es nur um den Preis ganz elementarer Restriktionen.

Ulrich Schmitz, Universität Duisburg

Veranstaltungen

VERANSTALTUNGSKALENDER

13.02.-14.02.1993, Vaasa, Finnland Tagung: Fachsprache und Computer-Nutzung Information: Prof. Dr. Walther von Hahn, Universität Hamburg, FB Informatik, Bodenstedtstr.16, D-2000 Hamburg 50

17.02.-19.02.1993, Hamburg, BRD 2. Deutsche Tagung über Expertensysteme XPS-93 Veranstalter: GI-FA 1.5 (Expertensysteme) Information: Prof. Dr. Wolff von Gudenberg, Universität Würzburg, Institut für Informatik, Am Hubland, D-8700 Würzburg

02.03.-03.03.1993, Zürich, Schweiz Hypermedia 93
Termine: Beiträge erbeten bis 31.07.1992 Veranstalter: GI, OCG, SI
Information: Prof: H. P. Frei, ETH Zürich, Informationssysteme, CH-8092 Zürich

03.03.-05.03.1993, Kiel, BRD
Sprachtechnologie: Methoden, Werkzeuge, Perspektiven Jahrestagung 1993 der Gesellschaft für Linguistische Datenverarbeitung GLDV
Information: GLDV-93, c/o Dr. Horst P. Pütz, Germanistisches Seminar, Universität Kiel, D2300 Kiel 1
Fax: 0431/880 1524, 0431/880 1512

03.03.-05.03.1993, Jena, BRD
Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft
Thema: Sprachvariation und Sprachgeschichte
Information: Rudolf Emons, Insstr. 40, D-8390 Passau

03.03.-05.03.1993, Braunschweig, BRD
BTW 93 Tagung Datenbanksysteme in Büro, Technik und Wissenschaft.
Information: Prof. Dr. H.-D. Ehrich, TU Braunschweig, Institut f. Programmiersprachen und Informationssysteme, Postfach 3329, D-3300 Braunschweig

23.03.-24.03.1993, Aachen, BRD 2. Anwendersymposium zu Fuzzy Technologien Veranstalter: MIT - Management Intelligenter Technologien GmbH
Information: MIT GmbH, Frau Karin Besting, Korneliuscenter, Promenade 9, D-5100 Aachen Tel.: 02408/94580
Fax: 02408/94582

22.04.-23.04.1993, Karlsruhe, BRD
6. Workshop Simulation und Künstliche Intelligenz
Information: AK "Simulation und KI", Dr. J. Krauth, BIBA, Postfach 330560, D-2800 Bremen; AK "Simulations-Soft-und-Hardware", Dr. J. Halin, Inst. f. Energietechnik, CH-8092 Zürich

29.04.-30.04.1993, Bamberg, BRD Modellierung von Informationssystemen Veranstalter: GI-FB 5 Wirtschaftsinformatik Information: Prof. Dr. Elmar Sinz, Universität Bamberg, Lehrstuhl für Wirtschaftsinformatik, Feldkirchstr. 21, D-8600 Bamberg

11.05.-14.05.1993, Berlin, BRD
ICCV'93 Fourth International Conference on Computer Vision. Humbold Universität Berlin
Information: Deutsche Informatik-Akademie, Frau Trapp, Ahrstr.45, D-5300 Bonn 2
Tel.: +49-721-6091-210
Fax: +49-721-6091-413
e-mail: hhniitb.fhg.de

24.05.-28.05.1993, Avignon, France Thirteenth International Conference Artificial Intelligence, Expert Systems, Natural Language
Information: Jean-Claude Rault, EC2, 269-287 Rue de la Garenne, F-92024 Nanterre Cedex, France
Tel.: +33 1 47 80 70 00
Fax: +33 1 47 80 66 29

12.06.-18.06.1993, Berkeley, USA 5. Kongress der Internationalen Vereinigung für Semiotik(IASS/AIS). Information: Irmengard RAuch, Department of German, University of California, Berkeley, CA 94720, USA.
.

27.06.-30.06.1993, Pittsburg, USA Conference on Research and Development in Information Retrieval Information: Prof. Dr. N. Fuhr, Universität Dortmund, Lehrstuhl Informatik VI, D4600 Dortmund

12.07.-16.07.1993, Soest, BRD Working Conference on Computer Mediated Education of Information Technologie Professionals and Advanced End-Users Information: Rudolf Hambusch, Landesinstitut für Schule und Weiterbildung, Paradieser Weg 64, D-4770 Soest

18.07.-23.07.1993, Leuven, Belgien 3rd International Cognitive Linguistics Conference Information: Dirk Geeraerts, ICLA93, Department of Linguistics, Katolieke Universiteit LeuYen, Blijde Inkomstraat 21, B-3000 Leuven.

20.08.-25.08.1993, Trondheim, Norwegen International Conference on TeleTeaching 93: Learning and working independant of time and distance Information: Kersti Larsen, Norwegian Computer Society, P.O. Box 6714, Roddelokka, N-0503 Oslo

25.08.-27.08.1993, Köln, BRD TKE'93 Terminologie and Knowledge Engineering Information: TKE'93, Prof. Dr. Klaus Dirk Schmitz, Fachhochschule Köln, Fachbereich Sprachen, Mainzer Str. 5, D-500 Köln 1

06..09.-08.09.1993, Prague, Czechoslovakia
DEXA 93 4th International Conference on Database and Expert Systems Applications Information: Research Institute for Applied Knowledge Processing, University of Lintz, Altenbergstr. 69, A-4040 Linz, Austria
Tel.: 0732 2468/791
Fax: 0222 732 24 68/93

08.09.-1-0.09.1993, Aachen, BRD
1st European Congress on Fuzzy and Intelligent Technologies
Veranstalter: ELITE-Foundation
Information: Prof. Dr. H.J. Zimmermann, ELITE-Foundation, Korneliuscenter, Promenade 9, D-5100 Aachen

13.09.-15.09.1993, Regensburg, BRD Information Retrieval
Veranstalter: Gesellschaft für Informatik(GI), Fachgruppe: Information Retrieval
Termine: Beiträge erbeten bis 28.02.93 Information: Prof. Dr. G. Knorz, Technische Hochschule Darmstadt, FB Informatik, FG Datenverwaltungssysteme, Alexanderstr.l0, D6100 Darmstadt
Tel.: +(49) (6151) 16-2953
Fax: +(49) (6151) 16-5489
e-mail: knorzdvs2.informatik.th-darmstadt.de

21.09.-23.09.1993, Berlin, BRD Eurospeech 93 Veranstalter: Europeen Speech Communication Association Information: COMTESS Messe-Service, B. Wulson, Biedermannweg 5, D-I000 Berlin

27.09.-29.09.1993, Maseru, Lesotho 2th International LiCCA Conference (Languages in Contact and Conflict Africa) The development and empowerment of indigenous languages in Southern Africa
A one-page abstract of your intended paper should be submitted by November 1, 1992 to: Prof. Zach Matsela Address: Faculty of Education, The National University of Lesotho, P. O. Roma 180, Lesotho, Africa Voice: 09266/340601 (W), 09266/3840258 (H)
Telex: 4303 LO
Fax: 0926 /340000

4.10.-7.10.93 Frankfurt/M, BRD
7. Internationaler Kongreß der Deutschen Gesellschaft für Semiotik.
Auskunft: Brigitte Schlieben-Lange, Mittelweg IB, D-6368 Bad Vilbel.

21.10.-22.10.1993, Braunschweig, BRD Workshop Fuzzy-Systeme - Management unsicherer Informationen
Veranstalter: GI-FA 1.2 (Inferenzsysteme)
Information: Prof. Dr. R. Krause, Technische Universität Braunschweig, Institut für Betriebssysteme und Rechnerverbund, Blütenweg 74/75, D-3300 Braunschweig

Mitteilungen aus der GLDV

1. Im Rahmen der diesjährigen GLDV-Jahrestagung in Kiel (03.-05.03.) wird auch die Mitgliederversammlung 1993 abgehalten und zwar am

Donnerstag, 04.03.1993, 16.00 Uhr;
Senatzsitzungsraum, Auditorium Maximum,

wozu hiermit alle Mitglieder der GLDV satzungsgemäß eingeladen sind.

2. Die Tagesordnung umfaßt bisher folgende Punkte:

1. Regularien
2. Bericht des Vorstands mit Kassenberichten 1991-92 und 1992-93
3. Entlastung des Vorstands
4. Wahl von Kassenprüfern
5. Bericht des Beirats
6. Berichte der Arbeitsgruppen und Arbeitskreise
7. Vorbereitung der Neuwahlen:
   - 7.1 Kandidatenliste: Vorstand
   - 7.2 Kandidatenliste: Beirat
8. Nächste Jahrestagungen
9. Arbeitsprogramm 1993-94
10. Verschiedenes

Die Beantragung weiterer TOPs durch Mitglieder (gemäß § 17) muß bis zum 24.02.1993 schriftlich beim Vorstand erfolgt sein.

3. Für die laut § 16(5) zu beschließenden Kandidatenlisten für die Neuwahlen (TOP 7) werden hiermit gemäß c) der Wahlordnung vom 09.09.1988 vom Vorstand folgende Kandidaten vorgeschlagen. Gleichzeitig wird daran erinnert, daß laut Wahlordnung bis zum Beginn der Mitgliederversammlung von GLDV-Mitgliedern dem Vorstand weitere Kandidaten mit deren Einverständniserklärungen vorgeschlagen werden können.

Kandidatenliste Vorstand:
1. *Vorsitzender:* Prof. Dr. Winfried Lenders (Bonn)
2. *Vorsitzender:* Prof. Dr. Hans Haller (Saarbrücken)
Schatzmeister: Prof. Dr. Roland Hausser (Erlangen)
Schriftführer: Prof. Dr. Jürgen Rolshoven (Köln)
Informationsreferent: Dr. Ludwig Hitzenberger (Regensburg)

Kandidatenliste Beirat:
Dr. Hans Billing, GMD St.Augustin
Dr. Karin Haenelt, GMD Darmstadt Prof. Dr. Christa Hauenschildt, Uni Hildesheim
Hans Haugeneder, Siemens München Prof. Dr. Georg Knorz, FH Darmstadt Prof. Dr. Jürgen Krause, Uni Regensburg Prof. Dr. Burghard Rieger, Uni Trier Dr. Dietmar Rösner, FAW Darmstadt Prof. Dr. Burkhard Schaeder, GHS Siegen Dr. Reinhard Wonneberger, ESC Rüsselsheim

4. Es wurde angeregt, zur bevorstehenden Briefwahl von Vorstand und Beirat nicht nur den Namen der Kandidaten zu nennen, sondern auch deren Vorstellungen für ihre Arbeit in diesen Gremien zu formulieren, mit Kurzvita und Bild zu ergänzen und den Wahlunterlagen als zusätzliche Informationen (max. je ! Seite) beizufügen. Den derzeit rund 350 wahlberechtigten Mitgliedern der GLDV wird hierdurch eine Wahlentscheidung vielfach erst möglich, oftmals auch erleichtert. Der Vorstand der GLDV bittet daher die Kandida-

ten hiermit sehr nachdrücklich, diese Informationen rechtzeitig (bis zum 02.03.1993) an die Vorsitzende des Wahlausschusses, Frau Prof. Dr. U. Klenk (Seminar für Romanische Philologie der Universität, Humboldtallee 19, 3400 Göttingen) zu leiten, die für die Erstellung und den Versand der Wahlunterlagen der diesjährigen Neuwahlen zuständig ist.

5. Mit dem Wunsch nach schnellerem Informationsaustausch, nach GLDVinterner Diskussionen wichtiger Themen, Standpunkte und Vorschläge, aber auch nach aktueller Unterrichtung über kurzfristig aufkommende Termine, auslaufende Fristen, fällige Daten, Erinnerungen, etc. vermöge des elektronischen Mediums GLDV-Newsletter *(gldv-nl)*, erneuert der Vorstand die Bitte an alle Mitglieder der GLDV, ihre Email Anschriften elektronisch mitzuteilen an:

ldvforum@ldv0l.Uni-Trier.de

Denn Dietmar Rösner ist weiterhin bemüht, alle GLDV-Mitglieder in ein *gldvnI-Abonnement* aufzunehmen und hofft, daß die inzwischen ja beträchtlich vergrößerte Zahl der Mitglieder mit Emmail-Anschluß auch in größeren Aktivitäten innerhalb dieses von ihm betreuten GLDV-Angebots sich umsetzen wird. Und darüber hinaus würde es auch der Vorstand begrüßen, wenn er (schon aus Gründen der Arbeitsvereinfachung) diejenigen der (meist lang"jährigen) GLDV-Mitglieder elektronisch erreichen könnte, die inzwischen über Email-Adressen verfügen, diese (oder etwaige Änderungen) dem Vorstand aber bisher nicht bekannt gaben.

6. Zur COLING-94 in Kyoto (Japan) wäre die GLDV-genügende Teilnehmerzahl vorausgesetzt-bereit, eine Sammelreise zu verbilligten Tarifen zu organisieren. Interessenten werden daher gebeten, sich sobald als möglich mit dem Informationsreferenten im Vorstand der GLDV, Prof. Dr. Hans Haller, (IAI, Martin-Luther-Str. 14, 6600 Saarbrücken, Email: hans@iai.unisb.de) in Verbindung zu setzen.

7. Leider haben wir wieder zahlreiche Postrückläufer von Mitgliedern, die ihre gültigen Anschriften dem GLDV- Vorstand nicht rechtzeitig mitgeteilt haben. Wer die neue Anschrift eines der nachfolgend aufgeführten GLDV- Mitglieder kennt, möge sie bitte dem Vorstand bekanntgegeben:

Barth, Thomas, Rohrbacherstr. 75, 6900 Heidelberg
Bhedassek, Thomas, EWH Koblenz, Seminar für Informatik, Rheinau 3-4, 5400 Koblenz
Blumenthai, Andreas, M.A., Universität Heidelberg, Germanistisches Seminar, Hauptstr. 207 209, 6900 Heidelberg
Brun, Georg, Kanzlerrain 19, CH-5430 Wettingen, Schweiz
Caspary, Christoph, Vogelsbergstr.22, 6000 Frankfurt 1
Dierks, Dr. Karin, Franckstr. 35, A-8010 Graz/Österreich
Diestelmann, Martin, Arno-Assmann-Str. 15, 8000 München 83
Eckert, Karin, Saarbrücker Str. 208, 6602 Dudweiler
Ehlich, Ute, Universität Erlangen-Nürnberg, IMMD5, Markusstr. 3, 8520 Erlangen
Eicke, Christine, Koppel 23, 2000 Hamburg 1
Engelberg, Klaus-Jürgen, Fraunhofer-Institut IAO, Holzgartenstr. 17, 7000 Stuttgart
Esa, Mohamed, Friedrich-Ebert-Anlage 47, 6900 Heidelberg
Firzlaff, Beate, Roonstr. 34, 5400 Koblenz
Guckler, Dr. Gudrun, Software AG, Haardtring 100, 6100 Darmstadt
Hanke, Manfred, Nordstr. 10, 5300 Bonn 1
Hirschmann, Astrid, Augsburger Str. 93a, 8400 Regensburg
Jacob, Daniel, Schroederstr. 37, 6900 Heidelberg
Kellner, Marianne, Sterbergstr. 26,8400 Regensburg
Kraiß, Martin, Möhlstr. 14, 6800 Mannheim 1
Krauss, Margit, Rohrbacher Str. 75,6900 Heidelberg
Krenn, Monika, Therese-Giese-Allee 50, 8000 München 50
Kroupa, Edith, Hesselauweg 103, 7000 Stuttgart 80
Kuhnert, Klaus-Dieter, Dipl.-Ing., Robert-Koch Str. 14, 8012 Ottobrunn
Malchow, Carsten, Grosse Barlinge 43,3000 Hannover

Marhenke, Ralf, Dyroffstr. 3,5300 Bonn 1
Müller-Zantop, Susanne, Mauerkircherstr. 29, 8000 München 80
von Oettingen, Edgar, c/o Oettingen GmbH, Friedrich-Ebert-Str. 27, 4000 Düsseldorf
Ortmann, W.D., Dr., Goethe-Institut, Mailand, Italien
Philippi, Julia, Kalckreuthweg 89,2000 Hamburg
Pröpper, Ellen, Georg-Treber-Str. 59, 6090 Rüsselsheim
Ripp, Volker, Birkenstr. 12, 1000 Berlin 21
Ritzke, Johannes, Nixdorf Computer AG, Abt. E 054, Pontanusstr. 55,4790 Paderborn
Rudolf, Klaus, IABG/SzFF, Einsteinstr., 8012 Ottobrunn
Ruge, Nina, Volksgartenstr. 22,4000 Düsseldorf
Rychly, Vladislav, Mühlheimer Str. 63,4300 Essen 1
Schlögell, Volker, Sternstr. 39,4400 Münster
Schmidt, Arno, Manteiffelstr. 6,4600 Dortmund
Schreiber-Schwenkgle, Almut, M.A., Kleiststr. 1,6200 Wiesbaden
Steffens, Petra, Taunusstr. 72, 7030 Böblingen
Strehlitz, Birgit, Am Gonsenheimer Spiess 87, 6500 Mainz
Wagner, Franc, Postfach 10 61 51, 6900 Heidelberg
Wirtz, Guido, Endenither Allee 110, 5300 Bonn 1
Zamurovic-Heller, Nada, Fraunhofer Institut, Holzgartenstr. 17, 7000 Stuttgart 1
Zock, Michael, 1, rue A. Guilpin, F-9425 Gentilly, Frankreich

Protokoll der Mitgliederversammlung der GLDV vom 08.10.1992 in Nürnberg

Beginn: 16.30 Uhr; Ende: 17.50 Uhr; Sitzungsleitung: B. Rieger

Tagesordnung

I> Regularien
I> Bericht des Vorstands mit Kassenbericht
I> Entlastung des Vorstands
I> Wahl von Kassenprüfern
I> Wahl eines Wahlvorstandes für 1993
I> Bericht des Beirats
I> Berichte der Arbeitsgruppen und Arbeitskreise
I> Satzungsänderung: Neuformulierung des § 6
I> Jahrestagung 1993 in Kiel
I> Nächste Jahrestagungen
I> Arbeitsprogramm 1993
I> Verschiedenes

TOP 1 Regularien

B. Rieger stellt fest, daß die Einladung zur Mitgliederversammlung fristgerecht ergangen ist. Es sind zunächst 24 Mitglieder anwesend; die Öffentlichkeit wird von der MV (=Mitgliederversammlung) bei später anwesenden 3 Gästen in TOP 8 einstimmig zugelassen. Die Tagesordnung wird in der von B. Rieger vorgelegten Form angenommen. Anträge auf Stimmübertragung wurden nicht vorgelegt. Das Protokoll der letzten Mitgliederversammlung in Trier, das durch das LDV-Forum allen Mitgliedern zugegangen ist, wird einstimmig von der MV genehmigt.

TOP 2 Bericht des Vorstands mit Kassenbericht

Da B. Schaeder nicht anwesend ist, kann kein Kassenbericht und damit auch keine Entlastung des Vorstandes stattfinden; vgl. TOP 3. B. Rieger berichtet, daß seit der letzten

Mitgliederversammlung je zwei Treffen von Vorstand und Beirat stattgefunden haben. Ein kombiniertes zweitägiges Vorstands- und Beiratstreffen am 28./29.Februar 1992 in Darmstadt und je eine Sitzung des Vorstandes und ein Treffen des Beirats im Rahmen der KONVENS vor der Mitgliederversammlung. Dabei wurden die folgenden Punkte behandelt:

Jahrestagung 1993 in Kiel

Es wird angestrebt, daß der Tagungsband bereits zur Tagung vorliegt. Inzwischen zeichnen sich die folgenden Sektionen ab: Bearbeitung großer Corpora, Fuzzy-Ansätze und Quantitative Linguistics.

LDV-Forum

Die Redaktion von Heft 9/2 findet in Trier statt. Die Redaktion geht ab Heft 1/1993 wieder an G. Knorz über. Wie verabredet, übernimmt Saarbrücken/IAI von diesem Heft ab die Herstellung.

Newsletter

Der Aufruf im LDV-Forum zur Mitteilung der e-mail-Nummern ergab eine einzige Rückmeldung.

Europäische Aktivitäten

Auf europäischer Ebene war ein unabhängiges Komitee für den Bereich der Forschung geplant: European Language Technology Agency (ELTA). Die Aktivitäten waren bis zum Einholen von Stellungnahme gediehen (Koordination in Deutschland: Rohrer) als das Unternehmen durch die EG gestoppt wurde.

Sonstiges

Die Akademie in Moskau ersuchte um die Lieferung von zwei GLDV-Tagungsbänden (der GLDV-Tagung in Siegen). Zur EUROSPEECH 93 wurde wieder eine Cosponsorship der GLDV vereinbart, was den Mitgliedern der GLDV eine verbilligte Teilnahme an der Eurospeech 93 ermöglicht. Im September 1992 wurde die Neuauflage der Blätter zur Berufskunde Computerlinguistik" unter der Federführung von J. Krause fertiggestellt.

Mitgliederentwicklung

Die Befürchtungen, die anläßlich der Erhöhung der Mitgliedsbeiträge geäußert wurden, haben sich nicht bewahrheitet. 1991 zur MV in Trier zählte die GLDV 361 Mitglieder. Seitdem sind 29 Eintritte zu verzeichnen, 2 Austritte und 47 Postrückläufer. Das ergibt einen Bestand von 342 Mitgliedern (ohne Postrückläufer). Im Laufe des Jahres sind 10 Kündigungen eingegangen, die 1993 wirksam werden.

TOP 3

entfällt für diese MV und wird vertagt auf die nächste MV.

TOP 4

Herr Lenders und Frau Spielmanns-Rome werden zu Kassenprüfern vorgeschlagen und mit zwei Enthaltungen gewählt.

TOP 5

Der Wahlvorstand für die Wahl des Vorstands und des Beirats 1993 sollte jetzt schon bestimmt werden. Frau U. Klenk, Herr H.J. Weber und Frau U. Weiß werden vorgeschlagen. Die Kandidaten werden mit drei Enthaltungen gewählt und nehmen die Wahl an.

TOP 6 Bericht des Beirats

Der Beiratsvorsitzende J. Rösner ist bei einer Systemvorführung unabkömmlich. Der Bericht des Beirats ist jedoch mit dem Bericht des Vorstandes inhaltlich insofern identisch, da eine gemeinsame Arbeitstagung mit dem Vorstand stattgefunden hat und der Beirat sonst nicht weiter aktiv war. W. Lenders erkundigt sich nach dem Konzept für die nächste Jahrestagung, wonach die Mitglieder des Beirats jeweils eine Sektion der Tagung betreuen sollten. B. Rieger berichtet, daß dieses Konzept nicht zu bemerkenswerten Reaktionen geführt hätte und die Konstitution der Sektionen (vgl. TOP 9) insofern ein Ausfallskonzept darstellen.

TOP 7 Bericht der Arbeitsgruppen und Arbeitskreise

AK Quantitative Linguistics

Es fanden zwei Arbeitssitzungen in 'Trier und in Münster statt. Die Anzahl der Mitglieder des Arbeitskreises pendelt sich von anfänglich 24 Interessenten auf 14 aktive Mitglieder ein. Die zweite Nummer des QL-Rundbriefs ist bereits in Vorbereitung. Sie wird allen AK-Mitgliedern zugehen.

AK Ausbildung und Berufsperspektiven

Die künftige Leitung des AK ist noch ungeklärt, die Aktivitäten ruhen. M. Lutz-Hensel ist allerdings beteiligt an der Neuauflage der Blätter zur Berufskunde. In diesem Zusammenhang ist mitzuteilen, daß an der LMU (München) und in Heidelberg neue Studienmöglichkeiten (verabschiedete Studiengänge) bestehen.

Ansonsten liegen keine weiten Mitteilungen von Arbeitskreisen vor.

TOP 8 Satzungsänderung

Die Satzungsänderung war notwendig geworden, da ein Erlöschen der Mitgliedschaft ohne Mitteilung an das Mitglied unwirksam blieb. Der Wortlaut der Änderung wurde satzungsgemäß allen Mitgliedern mit der Einladung zur Mitgliederversammlung zugeleitet. Die

MV läßt einstimmig die Öffentlichkeit zu (da Gäste anwesend sind). Die Satzungsänderung wird einstimmig (24 Ja-Stimmen) angenommen. 6 Abs. 3 der Satzung lautet nun wie folgt:

> Ein Mitglied kann durch Beschluß des Vorstands von der Mitgliederliste gestrichen werden, wenn sein Jahresbeitrag sechs Monate nach Beginn des Geschäftsjahres noch nicht gezahlt ist und trotz Mahnung durch eingeschriebenen Brief, der den Hinweis auf das Erlöschen der Mitgliedschaft enthalten muß, auch innerhalb eines Monats nicht eingegangen ist.

TOP 9 Jahrestagung 1993 in Kiel

Das in Darmstadt aufgestellte Konzept, wonach die Mitglieder des Beirats sich um die einzelnen Sektionen annehmen, wurde bisher nicht realisiert. Aus dieser Lage entstand das Konzept für neue Sektionen: demnach betreut R. Köhler die Sektion Quantitative Linguistics, W. Lenders die Sektion Große Corpora und B. Rieger eine Fuzzy-Sektion.
Der Anmeldeschluß für Tagungsbeiträge war auf den 15. Oktober festgelegt worden. Auf Vorschlag von W. Lenders wird die Anmeldefrist um einen Monat verlängert. H. Haller wird dies in geeigneter Form publizieren.

TOP 10 Nächste Jahrestagungen

Durch die Teilnahme der GLDV am Konvens-Tagungsschema finden alleinige GLDV-Tagungen in zweijährigem Turnus statt. Die MV einigt sich darauf, daß die GLDV eine KONVENS- Tagung 1994 in Saarbrücken anbietet. Saarbrücken wäre als Tagungsort auch später noch möglich. Regensburg kommt als Tagungsort erst ab 1996 infrage. Ein romanistischer Schwerpunkt wäre denkbar für eine nur-GLDV-Tagung entweder 1995 in Saarbrücken oder bei Prof. Figge. Eine Tagung in den neuen Bundesländern wäre zwar wünschenswert - eine weitere Konsolidierung scheint dennoch bis dahin erforderlich.

TOP 11 Arbeitsprogramm 1993

Eine Neuauflage des Studienführers als bekannte und erfolgreiche Publikation der GLDV sollte in Angriff genommen werden. Die Fortschreibung dieser Dienstleistung für die wissenschaftliche Kommunität muß sichergestellt werden. Dazu müßte vor allem geklärt werden, ob das Arbeitstreffen Computerlinguistik zu einer Fortführung der Arbeiten evtl. bereit wäre.

Zu den Jahrestagungen der GLDV vgl. TOP10. Allgemein ist festzustellen, daß breite Konferenzangebote existieren. Daraus sollte die Konsequenz gezogen werden, eher thematisch engere Angebote zu machen (Workshops).

Das Schulungsangebot wurde seit der Herbstschule 1990 nicht weitergeführt. Organisatoren für die nächste Herbstschule gesucht!.

Mit EAGLE ist von der EG ein Gremium geplant, das bei EG-Forschungsprogrammen beratende Funktion haben soll. Mitglieder dieses Gremiums werden den Planungen nach Forschungsinstituitionen sein. H. Haller wird einen Vorstoß unternehmen, dieses Gremium auch auf wissenschaftliche Gesellschaften zu erweitern. Die MV unterstützt diesen Vorstoß einstimmig.

TOP 12 Verschiedenes

VI/. Lenders gibt bekannt, daß die COLING 94 in Kyoto stattfinden wird. Es wurde angeregt, im LDV-Forum eine Umfrage zur Organisation einer Sammelreise zu unternehmen.

Ch. Schneider
(Schriftführung)

B. Rieger
(Sitzungsleitung)