TAGUNGSBERICHT

JAHRESTAGUNG 1993 DER GLDV IN KIEL

"SPRACHTECHNOLOGIE: METHODEN, WERKZEUGE, PERSPEKTIVEN"

3.3. - 5.3.1993

Die Tagung wurde begonnen mit einem Einleitungspanel unter der Leitung von W. Lenders (Bonn), das in diesem Band dokumentiert wird. Das Thema der Tagung wurde in 4 Sektionen (Quantitative Linguistik: R. Köhler, Fuzzy Linguistik: B. Rieger, Maschinelle Übersetzung: J. Haller, Maschinelle Korpora: W. Lenders) sowie in einigen sektionsfreien Vorträgen abgehandelt.

Sektion Quantitative Linguistik (Leitung: Reinhard Köhler): R. Köhler und M. Galle (Trier) verdeutlichten in ihrem Beitrag "Dynamische Eigenschaften von Textmaßen" die Problematik des Ausdrucks von Texteigenschaften mit traditionellen statistischen Methoden. Weder die Fixierung solcher Eigenschaften in einer einzigen Zahl noch Darstellungen von Häufigkeits- oder Rangverteilungen ermöglichten es bislang, die textuelle Dynamik angemessen auszudrücken; die exemplarisch vorgeführten Verlaufskurven zur Handlungsintensität mit ihrem flachen bzw. auffällig gleichmäßigen Verlauf verdeutlichten den Mangel an Ausdrucksfähigkeit dieser Methodik eindrucksvoll. Köhler / Galle schlagen daher auf der Grundlage eines Modells von G. Altmann eine differenzierte Vorgehensweise vor, wobei sie in Abhängigkeit von der Position des jeweils ausgewählten Text-ausschnitts Korrekturglieder in die statistische Berechnungsformel des Verhältnisses von Adjektiven zu Verben einführen. Ergebnis dieser Modifikation sind wesentlich ausdrucksfähigere Aktionsquotientenkurven, anders gesagt: erst nach dieser Veränderung sind die Kurven überhaupt aussagekräftig.

- P. Schmidt (Trier) stellte in seiner Untersuchung zu "Metrisierungsmöglichkeiten in der Morphosyntax" Möglichkeiten einer" vergleichenden Analyse des relativen Gewichts und der Interaktion" verschiedener "Faktoren bei der (menschlichen oder maschinellen) syntaktischen Analyse und ihrer typologischen Beziehungen" vor. Der Vortrag war in seiner Dokumentation von Breite und Tiefe des Wissens beeindruckend, dem Nicht-Spezialisten (so z.B. dem Berichterstatter) wurde damit der Nachvollzug des Vorgetragenen allerdings nur schwer möglich. Der Referent legte Wert auf die Feststellung, daß "die. . . vorgestellten Maße offensichtlich ohne Schwierigkeiten für die Messung analoger Eigenschaften in anderen sprachlichen Subsystemen, wie der Flexionsmorphologie oder dem Lexikon, adaptiert werden" können.
- Die weiteren in dieser Sektion angekündigten Vorträge fielen wegen Erkrankungen der Referenten aus.

Am Nachmittag des ersten Tages fanden dann in parallelen Sitzungen "freie Vorträge" statt. So berichtete U. Seewald (Hannover) über "Objektorientierte Programmierung als Werkzeug der Linguistischen

Datenverarbeitung - Überlegungen anhand von Lexikon- und Datenbankdesign" . Die Referentin benutzt die Programmiersprache Objectworks\ Smalltalk von Parc Place Systems. Die darin zur Verfügung stehende Einheit eines "Objekts" kann sich aus dem jeweils aktuellen Zustand, den Eigenschaften und Operationen des zu beschreibenden Objekts zusammensetzen. Auf die Werte der Variablen dieses Objekts kann durch Nachrichten an das Objekt zugegriffen werden. Am Beispiel eines Morphemlexikons wurde gezeigt, wie eine hierarchische Ordnung der zu beschreibenden Einheit geschaffen werden kann, wobei die übergeordneten Einheiten als Klassen jeweils ihre Eigenschaften an ihre Unterklassen vererben, soweit sie dort nicht durch aktuelle Werte überschrieben werden; so lassen sich ohne Schwierigkeiten auch Ausnahmen von Default-Werten realisieren. An sich unterstützt die gewählte Programmiersprache keine Mehrfachklassifikation (polyhierarchische Vererbung), wie sie jedoch gerade in der Lexikographie oft benötigt wird (z.B. *Leber:* ORGAN und SPEISE). Über die Definition von Abhängigkeitsrelationen hat die Referentin jedoch eine Struktur aufgebaut, die einem polyhierarchisch geordneten System nahekommt. Sie demonstrierte die Möglichkeiten dieser Programmierung an einer Datenbank zu Morphemen des Französischen. Die Abhängigkeits(baum)struktur konnte sie durch hintereinander angeordnete Teilfenster der Lexikonoberfläche sichtbar machen.

- Hans-Dieter Lutz zeigte in seinem Vortrag über "Software-Ergonomie für Sprachsoftware" zunächst die Gründe für eine grundlegende Veränderung des Verhältnisses von Software-Entwicklung und Software-Ergonomie auf. Resultat dieser Entwicklung ist, daß heute Software generell "von außen nach innen" zu entwickeln ist. Im Zentrum des Interesses stehen die Funktionalitäts-, die Korrektheits- und die Schnittstellen-Ergonomie. Lutz differenzierte anschließend den Begriff der Sprachsoftware und versuchte eine Beschreibung der Benutzerklassen dieser Produkte. Auf der Grundlage von EG-Richtlinien zur ergonomischen Gestaltung von Bildschirmarbeitsplätzen betrachtete Lutz abschließend die Konsequenzen für 3 Typen von Sprachsoftware: Textbe- und -verarbeitung, Software auf natürlichsprachlicher Kommunikationsbasis und Sprachlernprogramme. Dabei zeigte sich, daß die Computerlinguistik (CL) auch zukünftig noch wichtige Aufgaben zu erfüllen hat, daß sie aber auch das Spektrum ihres Grundlagenforschungsansatzes um Fragebereiche der Sprachverwendung zu erweitern haben wird.

- Jutta Marx (Regensburg) referierte über den "Einsatz natürlichsprachlicher Komponenten in einer multimodalen Benutzerschnittstelle für Werkstoffdatenbanken" und berichtete von Erfahrungen mit unterschiedlichen Datenbankabfragekonzepten allgemein und Kombinationen dieser Grundtypen bei Tests anläßlich der Entwicklung einer Werkstoff-Datenbank (Projekt WING). Ein Schwerpunkt lag dabei auf der Evaluierung der Einsatzmöglichkeit natürlicher Sprache, die in der aktuellen Realisierung des Projekts bei der Zustandsanzeige, im Korrekturmodus und im Hilfesystem verwendet wird. Weitere Tests, die auch ein aktualisiertes Retrieval-Modul einschließen sollen, sind für Anfang 1994 vorgesehen.

- I. Batori (Koblenz) plädierte in einem gemeinsam mit M. Volk (ebd.) erarbeiteten Beitrag dafür, "Workbenches auch für die Forscher in der theoretischen Linguistik anzubieten". Batori führte die verschiedenen Phasen des Linguistic Engineering vor (Problemanalyse, Problemdefinition, Systemspezifikation, Implementierung) und beschrieb die Beziehungen zwischen CL und Sprachforschung. Kernstück des Referats war jedoch eine Kritik an der Verfahrensweise linguistischer Forschung, die - jeweils veranlaßt durch die Erkenntnis der Unzulänglichkeit einer Grammatik

- ständig neue Grammatikmodelle entwirft, welche zwar die zuvor entdeckten Probleme zu lösen versprechen, ohne daß jedoch die Deckungsgleichheit mit der Reichweite des vorhergehenden Modells überprüft und ohne daß das im früheren System enthaltene korrekte Wissen weiterverwendet wird. Es wurde gezeigt, daß die CL hingegen "die Vererbung des linguistischen Wissens im Falle eines Modellwechsels zu gewährleisten" vermag.
- Marc Domenig als Vertreter einer Arbeitsgruppe (Basel) beschrieb "Werkzeuge zur Akquisition und Verwaltung von morphologischem und phrasealem Wissen" einschließlich Navigationsmöglichkeiten im Gesamtwissensbestand eines in Basel entwickelten Datenbanksystems. Dieses ist auf der Basis eines Client/Server-Modells konstruiert und erlaubt dem Benutzer Eingriffe in Detailbereiche, ohne daß er sich um die Position oder Editionsformate der benutzten Subdateien zu kümmern braucht. Die verwendeten (Fenster-)Editoren "kennen" die jeweiligen Spezifikationen der Dateien, und ggf. vorgenommene Eingabeänderungen werden vom Compiler automatisch entdeckt und verarbeitet. Dieses hochspezifizierte Werkzeug eignet sich entsprechend der Vorführung während der Tagung vorzüglich für moderne Lexikonarbeiten.
- H. Elsen und J. Hartmann (Bonn) berichteten in einem kurzfristig noch eingefügten Vortrag über "Satzbandanalyse und Aufbereitung des Defintionswortschatzes eines deutschen Wörterbuchs" über die Vorarbeiten zur Extraktion von Informationen aus einem maschinell lesbaren Teil des ‚DUDEN - Deutsches Universal Wörterbuch' sowie die erforderlichen Arbeiten zur Überführung dieser Daten in ein SGML-Format. Sie bedienten sich bei dieser Arbeit teilweise der auf dem Satzband vorgefundenen internen Markierungen und entwickelten eigene Programme, die sie zusammen mit kommerzieller Software (FLEX, BISON und den in Bonn bereits vorhandenen lexikalischen Ressourcen (Bonner Wortdatenbank, Wortanalytisches Wörterbuch) zur Segmentierung lexikalischer Einheiten auf dem Magnet band einsetzen konnten.
- Im Zusammenhang des gleichen Bonner Projekts führte N. Weber die Beschreibung der (computer-)linguistischen Probleme einer "Computergestützte(n) Analyse von Definitionstexten in einem deutschen Wörterbuch" weiter aus. Ziel dieser Arbeiten ist eine formalisierte Repräsentation von Lexembedeutungen, die aus den maschinenlesbaren Lexikoneinträgen des DUDEN nach umfangreichen Vorarbeiten gewonnen werden konnten, damit diese dann anderen sprachverarbeitenden Systemen zur Weiterverarbeitung bereitgestellt werden können.

Sektion Fuzzy Linguistik (Leitung: B. Rieger): Arbeitsteilig berichteten B. Rieger ("LLAMA - ein Pilotsystem zum referentiellen Sprachlernen mit unscharfem Bedeutungserwerb"), B. Badry ("Sprachliche Unschärfe und ihre experimentelle Modellierung in LLAMA") und M. Reichert ("Cluster-Strukturen der Zwischenrepräsentationen in LLAMA") - alle aus Trier - über Teilaspekte des gleichen großen Projekts "Language Learning And Meaning Acquisition". Am Anfang standen grundsätzliche Überlegungen zu Grundlagen und Problemen des Erwerbs und der Verarbeitung von Wissen in kognitiven Systemen, zu denen auch der Mensch gehört. Der Forschungsansatz des Projekts beruht variierend auf Erkenntnissen des International Computer Science Institute, Berkeley, in dessen MLA-Projekt ("Miniature Language Acquisition"). Das in Trier entwickelte System besitzt weder ein Regelwissen noch eine Kenntnis der Bedeutungen der zu verarbeitenden Zeicheneinheiten, vermag aber nach Verarbeitung der eingegebenen Texte und kurzer Lernphase, sich in einer ‚Situation' zu orientieren und z.B. die Lage bestimmter Objekte mit Hilfe vorgegebener Richtungsparameter zu bestimmen.
- M. Galle (ebenfalls Trier) beschrieb die

Erfahrungen, die man im Rahmen eines studentischen Studienprojekts mit der Aufbereitung von sehr großen Zeitungskorpora (dpa, taz) gemacht hat. Dabei ging es nicht nur um die Eliminierung von störenden Satzsteuerzeichen oder die einheitliche Umsetzung von Umlauten und anderen Sonderzeichen; unbedingtes Ziel war es vor allem, die im Originalkorpus enthaltenen Informationen auch in der aufbereiteten Korpusversion möglichst vollständig beizubehalten. Zu den zu bewältigenden Problemen zählte u. a. der Aufbau von Routinen, welche die enormen Datenbestände auch Forschern zugänglich machen, die von der textinternen und speichertechnischen Struktur der Daten keine Kenntnis haben.

Sektion Maschinelle Übersetzung

(Leitung: J. Haller): Aufgrund von Absagen der METAL- und LOGOS- Referenten sowie einer weiteren kurzfristigen Absage reduzierte sich das Programm der Sektion beträchtlich. Anstelle der LOGOS-Referentin lieferte R. Nübel (Saarbrücken) einen Erfahrungsbericht über "Möglichkeiten zur Evaluierung eines kommerziellen MÜ-Systems". In der Vortragsfassung berichtete die Referentin zunächst über die frustrierenden Einsichten in die Schwierigkeiten einer Verständigung zwischen Systementwickler (konkret: LOGOS) - Bewahrung von Entwicklungs- /Firmengeheimnissen - und wissenschaftlichen Evaluieren. Im zum Druck vorliegenden Beitrag beschränkt sich die Autorin auf grundsätzliche Fragestellungen im Rahmen von Evaluierungsaufgaben bei kommerziellen MÜ -Systemen, die ohne Einblick in die Entwicklungsinterna des Systems erfolgen müssen (black box-Evaluierung). Dabei werden ausführlicher drei Typen von Evaluierungsstrategien behandelt: Testen der linguistic coverage, der text type coverage und der textsortenspezifischen (Übersetzungs-)Ergebnis-Adäquatheit. Abschließend machte die Autorin deutlich, daß eine sinnvoll-produktive Zusammenarbeit zwischen Systementwickler und externen (Computer-)Linguisten den Zugang zu Systeminterna zur unbedingten Voraussetzung hat.

- J. Haller (Saarbrücken) referierte anschließend über Verarbeitungsstrategien und linguistisches Konzept von CAT2 (" Vom Forschungssystem zum präindustriellen Prototyp"). Im Unterschied zur EUROTRA-Entwicklung wird bei CAT2 ein SICSTUS-Prolog-Compiler anstelle eines YAP-Compilers verwendet. Die auf drei Ebenen repräsentierte Grammatik ist stark lexikonorientiert. Ein besonderes Charakteristikum ist auf der Interface-Ebene ein großer sprachunabhängiger Teil, der eine relativ problemlose Integration weiterer Zielsprachen zu ermöglichen verspricht. Auf der Software-Seite werden eine auf Sun-View aufbauende Dialogversion, eine (für Testzwecke geeignete) Batchversion sowie einen für die linguistische Evaluierung wichtige Debug-Version angeboten. Die geplante Auslagerung des Lexikons in eine externe Datenbank soll den weiteren Ausbau des Lexikons erleichtern. Der Referent kündigte an, daß das System von einem industriellen Pilotanwender im Medizinbereich sowie im Rahmen von EUROLANG bei der anvisierten Entwicklung eines modernen Übersetzerarbeitsplatzes eingesetzt werden wird. Die Tagungsteilnehmer hatten mehrfach Gelegenheit, sich das System ausführlich vorführen zu lassen.

- K. Schubert (Flensburg) sprach über die Beziehung "Zwischen Benutzerschulung und Wissenschaft. Sprachtechnologie in der Übersetzerausbildung. "Er beschrieb dabei zunächst den Wandel in der berufsorientierten und daher notwendigerweise praxisnahen Übersetzerausbildung. Hervorgehoben wurde, daß sich eine verantwortungsbewußte zukunftsorientierte Ausbildung nicht mit der Schulung an jeweils aktuell verfügbaren Systemen der Textbearbeitung! mit Datenbankaufbau und maschinellen Übersetzungshilfsmitteln begnügen darf. Die Tendenz zur Automatisierung und die potentielle Vielfalt im Übersetzer-

arbeitsbereich (z.B. als Berater und Evaluierer bei Modernisierungsentscheidungen durch das Firmenmanagement) zwingen zu einer auch systemunabhängigen Ausbildung, welche die Grundlagen und Arbeitsweisen moderner automatisierter Sprachverarbeitungssysteme verstehen läßt.

In der Sektion Maschinelle Korpora (Leitung: W. Lenders) wurden vier Vorträge angeboten (die oben bereits erwähnten Referate von H. Elsen/ J. Hartmann und N. Weber ließen sich hier allerdings ebenfalls einordnen). Den Anfang machte K. Wothke (Heidelberg) mit einem Vortrag über "Statistisch basiertes Wortklassentagging von deutschen Textkorpora. Einige Experimente." Wothke berichtete von Experimenten des Wissenschaftlichen Zentrums der Fa. IBM Deutschland GmbH im Rahmen der projektierten Entwicklung einer deutschsprachigen Version des Spracherkennungssystems TANGO RA. Ziel war dabei zunächst die Entwicklung einer statistisch orientierten Prozedur, die mit hoher Treffsicherheit deutsche Textkorpora mit Wortklassenangaben zu versehen vermag (Tagging). Längerfristig soll mit dem so entwickelten Algorithmus ein großes Textkorpus teilautomatisch annotiert und aus dem Ergebnis Auftretenswahrscheinlichkeiten von Wort klassen abgeleitet werden. Diese Erkenntnisse wiederum sollen der zukünftig besseren Differenzierung von akustisch ähnlichen Wörtern durch TANGORA dienen. Die vorgelegten Prozentzahlen über die bei kleinen Textmengen erreichte Trefferquote zeigten ein erstaunlich positives Bild, bei größeren Textmengen allerdings war das Ergebnis deutlich schlechter. Als Ursache vermutete Wothke einen unzureichenden Umfang des zugrundeliegenden Trainingskorpus, wodurch die Auftretenswahrscheinlichkeiten nicht hinreichend realitätskonform berechnet werden konnten.

- B. Schröder (Bonn) trug anschließend seine Überlegungen zu "Fragen der Repräsentativität linguistischer Korpora" vor. Zunächst beschrieb er die grundsätzliche Problematik der gegenseitigen Abhängigkeit von Empirie und Theorie. Bei der näheren Betrachtung von Korpora als linguistischen Datensammlungen ging Schröder dann auf Fragen der Bestimmung von Datentyp, Datenbereich und Auswahlprinzipien im Verhältnis zur angestrebten Repräsentativität ein, hierbei insbesondere auf unterschiedliche Verfahren der Zufallsauswahl und der Schichtenbildung. Abschließende Betrachtungen galten der Anlage sehr großer Korpora, die für unterschiedliche Zwecke verwendbar sein sollen und daher auch eines besonderen Augenmerks bei der Gewichtung der diversen Auswahlparameter bedürfen.

- Auf der Grundlage eines bereits vorhandenen Korpus wandte sich G. Willee (Bonn) dann den "Erfahrungen mit morphologischem Tagging am Beispiel des LIMAS-Korpus" zu. Er stellte fest, daß zwar die Annotation lexikalischer Elemente in großen Korpora als Desiderat gilt und Voraussetzung für die Nutzung der Korpus-Daten in linguistischer Forschung ist, daß jedoch über die Art und Weise der Annotation noch keineswegs Einigkeit herrscht. Für seinen Test (zum Zwecke der Feststellung des benötigten Zeitaufwands) hat Willee das am IKP /Bonn zur Verfügung stehende Lemmatisierungsprogramm LEMMA2 benutzt. Für die Verarbeitung der ca. 1.1 Millionen laufenden Wortformen des LIMAS-Korpus wurde ein Zeitaufwand von insgesamt ca. 2050 Stunden errechnet, d.h. in etwa ein Jahr Arbeitszeit für einen einzelnen Bearbeiter. Bei einem Korpusumfang wie dem des Trierer dpa/taz-Korpus wären somit ca. 70 Jahre (bei einem einzigen Bearbeiter) erforderlich.

- In einem Vortrag zu "Tagging - Formen und Tools" beschäftigte sich W. Lenders (Bonn) mit unterschiedlichen Formen sowie mit einigen vorhandenen Taggingtools. Nach einer knappen Analyse der Begrifflichkeit und der Tagging-Typologie cha-

rakterisierte er eingehend entwicklungsgeschichtlich wichtige Taggingsysteme je nach ihrer Zielanwendung (Wortklassentagging, Syntaxtagging) und ihrer Methodik (Lexikonbasis, Umgebungsanalyse, morphologische Analyse, Vorkommenswahrscheinlichkeit). Zur Veröffentlichung des Referats wird ein dreiteiliger Anhang gehören, worin ein Überblick geboten wird über die typologische Einordnung existierender Taggingsysteme, über Detailcharakteristika dieser Tagger sowie über zwei Tagsets für das Englische (Penn Treebank und SUSANNA formtags).

Daß neben den genannten Vorträgen eine Begrüßung der Teilnehmer durch den GLDV - Vorsitzenden sowie die Rektorin der Universität Kiel anläßlich eines kleinen Empfangs, eine Mitgliederversammlung, ein Konzert und ein gemeinsames (und trotz der Eiseskälte draußen gemütliches) Abendessen stattfand, sei nur am Rande und der Vollständigkeit halber erwähnt. Schade nur, daß die Teilnehmerzahl nicht sehr groß war, die Referentinnen und Referenten hätten ein zahlreicheres Publikum verdient gehabt.

Horst P. Pütz (Kiel)