9. JAHRESTAGUNG DER GESELLSCHAFT FÜR LINGUISTISCHE DATENVERARBEITUNG 30. und 31. März 1995 in Regensburg

Gerhard Knorz

Gleißende Schneeflächen, strahlendblauer Himmel und ein fernes Glockengeläut aus der Altstadt begleiteten den Weg vieler Computerlinguisten} die sich am 30. A1ärz 1995 *zur* 9. *Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung trafen. Bei insgesamt knapp einhundert TeilnehmerInnen - keineswegs eine deprimierende Anzahl in einer schwierigen Zeit für große Teilnehmeranzahlen - und einer reibungslosen Organisation durch das Fachgebiet Linguistische Informationswissenschaft* [1] *war das passende Ambiente für eine erfolg reiche Tagung durchaus gegeben, leicht getrübt nur durch die gefühllose Kälte von Betonmauern, die viele Wochen unbeheizter Semesterferien hinter sich hatten.*

Die Begrüßungsrunde, eingeleitet durch Ludwig Hitzenberger, der die Hauptlast der örtlichen Vorbereitung der Tagung getragen hatte, verbreitete zeitweise die Atmosphäre eines Nachrufes, wird doch Jürgen Krause nach ca. 20 Jahren linguistischer Informationswissenschaft in Regensburg ab Mai auf einen Informatik-Lehrstuhl an der Universität Koblenz wechseln und das IZ Sozialwissenschaften (Bonn) leiten. In diesem Sinne war es schon richtig (und doppeldeutig), wenn Winfrid Lenders, Vorsitzender der GLDV, nach den Grußworten des Prorektors bemerkte, es sei höchste Zeit für die Regensburger gewesen, eine GLDV-Jahrestagung auszurichten. Ansonsten war es die Aufgabe des GLDV-Vorsitzenden gewesen, anläßlich des 20-jährigen Bestehens der Gesellschaft kurz den Werdegang der GLDV von den Anfängen noch unter dem unkonventionellen Namen LDV-Fittings - bis heute darzustellen und die positiven Aspekte herauszustellen.

1 Maschinelle Lexikographie

Der fachliche Teil der Tagung wurde mit dem eingeladenen Hauptvortrag von Willy Martin eröffnet: *1 Maschinelle Lexikographie: Ein Blick in die Zukunft.* Martin vertritt das Fach Lexikographie an der Universität Amsterdam 1, und er gab sich bescheiden *"angesichts der Schwierigkeit in einer fremden Sprache als Prophet aufzutreten, wo doch schon die Darstellung der Gegenwart schon schwierig genug ist* 11. So erarbeitete er in seinem sehr klaren und auch für den nicht speziell Interessierten problemlos verständlichen Vortrag zunächst den status quo, den er dann aus seiner Sicht kommentierte und recht vorsichtig in die Zukunft fortschrieb.

Er definierte die Besonderheit der *maschinellen* Lexikographie/Lexikologie anhand von 4 Einflußparametern.

1 Vormals war es das Fach *Computerlinguistik*

Das Mittel ist eben der Computer.

Bei dem Objekt kann es sich um ganz unterschiedliche Formen von Wörterbüchern handeln, vom maschinenlesbaren Dictionary bis zur lexikalischen Datenbank.

Die Benutzung ist nicht auf verschiedene Zielgruppen beschränkt, sondern kann auch *auf den Computer* hinzielen.

Die Orientierung ist nicht auf ein Produkt ausgerichtet, sondern Produkte entstehen durch benutzerorientierte Front-Ends für die mehrfach nutzbaren Wörterbücher.

Für Martin lautet die Frage nicht, *ob* es zukünftig noch gedruckte Wörterbücher geben wird, sondern *wie* man zu diesen Wörterbüchern kommen wird.

In Bezug auf Korpora vertrat Martin die Auffassung, daß

> das gegenwärtige Streben nach immer größeren Korpora unnötig sei. Gebraucht würden vergrößerbare Korpora.

> die Unterscheidung von geschriebenen und gesprochenen Korpora immer wichtiger werde.

> ein semantisches Tagging auf probabilistischer Basis letztlich nicht angemessen sei: Sprache sei regelbasiert, auch wenn nicht alle Regeln leicht zu finden und zu formulieren seien.

> der Bottom-up-Ansatz (Ansatz der "Strandgut-Räuberei") bei der Bearbeitung von Korpora durch einen Top-down-Ansatz abzulösen oder zu ergänżen sei.

Die Zukunft maschineller Lexikographie sieht Martin in der Verbindung von mehreren Sprachen, wobei Editoren, die nicht nur auf Form- sondern auch auf der Bedeutungsebene arbeiten, sich als Wörterbuch-Generatoren eignen werden: Hat man zwei Wörterbücher für die Sprachpaare A - B und B - C, so sollte sich das Wörterbuch A - C automatisch generieren lassen.

Die kommerzielle Lexikographie gibt sich konservativ und bietet gegenwärtig elektronische Wörterbücher an, die konzeptionell nicht innovativ sind. Zwar geht die Entwicklung in Richtung eines mehr dynamischen Zugriffs und hin zu Multimedia, aber erst wenn die Bedeutungsebene im Lexikon repräsentiert ist, wird es wirklich dynamisch sein und z.B. im Satzkontext von 7 Bedeutungsvarianten eines Wortes die richtige anbieten können.

Bis zum Jahr 2000 wird sich die maschinelle Lexikographie hin zur *maschinellen Lexikologie* entwickelt haben.

Eines der wichtigsten Anliegen der maschinellen Lexikographie wird die Erforschung ihrer empirischen Grundlage sein: die induktive und dann auch die deduktive Bearbeitung von Korpora.

Lexikalische Datenbanken werden als zentrale Ressource entsprechender Verlage eingerichtet sein. Die Anwendungen werden durch die Stichwörter *intelligenter Gebrauch, Hypermedia und dynamisch* zu charakterisieren sein. Im übrigen sei es an der Zeit, auch nichtkommerzielle Impulse anzumahnen, um die Beschäftigung mit Lexika als Investition in eine *sozio-kulturelle Infrastruktur* zu fördern.

2 Vorträge in den. einzelnen Sektionen

Das Tagungsprogramm war durchgängig mit parallelen Sektionen angelegt. Die Eindrücke beziehen sich dementsprechend nur auf maximal 50 Prozent des Tagungsprogramms - und ich nutze die Gelegenheit, aus meiner ganz persönlichen Sicht einzelne Beiträge zu kommentieren.

2.1 Informationsysteme und Information Retrieval

2.1.1 Das Regensburger Projekt WING

Stoff-Datenbanken mit ihren unterschiedlichen Datentypen, ihrer Vielzahl von Attributen und Attributabhängigkeiten, ihren systembedingten "Datenlücken" und Unschärfen ("gut schweißbar bei niedrigen Temperaturen") bieten wissenschaftliche Betätigung in verschiedenen Disziplinen: Nonstandard-Datenbanken, Information Retrieval - und Computerlinguistik.

Zwei Vorträge beschäftigten sich mit unterschiedlichen Aspekten des Regensburger Projektes WING, das ein multimodales Werkstoff-Informationssystem entwickelt.

Sprachgenerierung und Analyse

Der Beitrag *„Sprachgenerierung und Analyse* im *Rahmen eines multimodalen Werkstoffinformationssystems"* (J. Marx) behandelte die Designprinzipien, die Metaphern und Werkzeuge der objektorientierten Oberfläche. Marx erläutert speziell die Komponente, die eine natürlichsprachige Paraphrasierung der über Formulare und besondere Werkzeuge (Kennwert-Präzisierer) eingegebenen Anfrage übernimmt.

Die Paraphrasierung geschieht in kleinsten Schritten begleitend (bereits beim Selektieren einer Werkstoffbezeichnung wird formuliert: *Daten zu Werkstoff X).*

Besonders interessant ist die Tatsache, daß der Benutzer diese Paraphrasierung frei editieren kann mit dem Effekt, daß damit die Frageformulierung durchgängig im System geändert ist. Das Parsing profitiert davon nicht - eine komplette neue Analyse hat sich als einfacher herausgestellt als eine Identifikation und Interpretation der Änderung -, wohl aber der Benutzer, der eine textuelle Vorgabe hat und damit nicht vor dem Problem steht, "irgend etwas" in ein leeres Fenster schreiben zu müssen 2.

Analyse und Generierung sind bisher durch ad-hoc-Algorithmen realisiert (regelbasiertes Keyword-Parsing), in Kürze wird eine FUG-Lösung eingesetzt werden.

Vagheit und Unschärfe

Der zweite Vortrag *Fuzzy Logic und neuronale Netze als Werkzeuge zur Handhabung von Vagheit und Ähnlichkeit in Informationssystemen* (Th. Mandl) motivierte zunächst das Problem vager Anfragen und stellte die Ähnlichkeitssuche als Möglichkeit vor, das Problem von Datenlücken zu überwinden. Inwieweit die elementare Einführung in die Thematik der Fuzzy Sets und der neuronalen Netze notwendig war, mag ich nicht beurteilen. Die Realisierung stellt eine menügesteuerte Möglichkeit zur Verfügung, linguistische Variablen mit Werten zu belegen (z.B. Temperatur mit *hoch, niedrig,* etc.), wobei eine Kontextabhängkeit der Zugehörigkeitsfunktion (das Problem der *großen* Maus) noch nicht verwirklicht ist.

Die Ähnlichkeit zwischen Werkstoffen (hinsichtlich gemeinsamen Anwendungsmöglichkeiten) wurde nach dem Scheitern eines regelbasierten Ansatzes mittels eines neuronalen Netzes trainiert und bestimmt. Nach zwei Anläufen war schließlich bei ca. 20 bis 30 Eingangsparametern und *einer hidden layer* ein befriedigendes Resultat erreicht.

Die Mitarbeiter boten eine Demonstration des Systems WING IIR an.

2.1.2 Natürlichsprachiges Interface zu einem Bibliothekssystem

Mertens, Schulz und Helbig von der Fernuniversität Hagen stellten mit ihrem Beitrag *„Analyse mit Wortagenten* im *NL-System LINAS"* einen Ansatz für ein natürlichsprachiges Interface zu einem Bibliotheks-Retrievalsystem vor. Die Systemarchitektur unterscheidet die drei sequentiellen Phasen morphologische Analyse, semantische Repräsentation und Datenbankanfrage. Es handelt sich um eine wortzentrierte Analyse mit Wort- und Wortklassenagenten, deren Prinzipien und Konzeption detailliert vorgestellt wurden.

Auf die Anwendung wurde nicht näher eingegangen, aber mein Eindruck ist der, daß hier wieder einmal ein völlig naiver Zugang zu einem Information Retrieval Problem versucht wird. Die Problemstellung wird so behandelt, als ob es um die Generierung von SQL-Abfragen aus natürlichsprachigen Datenbank-Queries ginge. Man mag in dem vorgestellten Rahmen einen sinn vollen Beitrag zum Problem der Literatursuche leisten können, aber diesbezügliche Lösungsansätze stehen noch vollständig aus. Und ob man das bisher Geleistete für die Literatursuche tatsächlich benötigt, wage ich zu bezweifeln.

2 "Müssen" ist eine mißverständliche Formulierung, denn der Effekt ist der, daß Benutzer ohne Hilfestellung den Modus der natürlichsprachigen Eingabe schlichtweg ignorieren.

2.2 Fuzzy Linguistik

Insgesamt 4 Vorträge wurden zu einer Sektion *Fuzzy Linguistik* zusammengefaßt. Einer davon ist bereits unter Abschnitt 2.1.1 beschrieben. Bei dem ersten Beitrag " *Warum Fuzzy Linguistik? Überlegungen und Ansätze zu einer Neuorientierung"* handelte es sich um ein von B. Rieger (Universität Trier) ungemein engagiert vorgetragenes Plädoyer für die Weiterentwicklung der quantitativen Linguistik hin zur Fuzzy Linguistik, für das Primat einer neu zu definierenden *Performanz* gegenüber der bisher vorherrschenden *Kompetenz*, für eine Modellierung durch *Funktionsräume* anstelle von kategorialen Ansätzen. Die angeregte Diskussion läßt hoffen, daß der neuzugründende gleichnamige GLDV-Arbeitskreis mit Leben zu füllen sein wird.

Fundierung von Fuzzy Werten

Mit seinem Beitrag *„zur Präzisierung des Begriffs der Unschärfe"* gab S. Mehl (Universität Duisburg) zunächst Beispiele für den Einsatz von Fuzzy Werten in Syntax, Semantik und Stilistik. Seine Kritik an der mangelnden Fundierung von Fuzzy Werten motivierte er an einem Beispiel, bei dem Finken und Rotkehlchen in unterschiedlicher Weise" typische Vögel" sein sollten, und auch Fledermäuse noch irgendwie als Vögel gelten - operationalisiert durch Probanten- Urteile auf mehrstufigen Skalen. Mehls Ansatz dagegen erprobte anhand von Testsätzen, inwieweit sich ein Begriff durch verschiedene Oberbegriffe ersetzen läßt, um den graduellen Charakter von Kategorien zu operationalisieren. Die *Dampfwalze* (Fahrzeug, Baumaschine) ist insofern kein typisches Fahrzeug, als man schlecht sagen kann: „Der *Teerbelag vor der Dampfwalze war noch nicht geglättet. Doch das Fahrzeug rührte sich nicht."* ,
Ein anderer Testsatz, der die notwendige Tiefe der notwendigen Überlegungen offensichtlich macht, hat sich mir nachhaltig eingeprägt:

„Immer, wenn ich ein Huhn sehe,
das einen Wurm frißt, habe ich
Mitleid mit dem Tier."

In der Diskussion zeigte es sich, daß Mehl vielleicht gegen eine Einstellung argumentiert hatte, die bei den ZuhörerInnen vielleicht gar nicht präsent war - daß er also nur "offene Türen eingerannt hatte", wie er selbst feststellte.

Harris revised

Eine Revision bzw. Modifikation des Verfahrens von Harris zur korpusbasierten Konstruktion unscharfer Einheiten schlug R. Wagner (Universität Trier) unter dem Titel *Harris revised* vor. Es mag sein, daß ich den Vortrag nicht richtig zu würdigen weiß - oder aber daß es sich um ein Versehen der Programmkommission gehandelt hat.

2.3 Studentische Sektion

Ein Experiment wagte die GLDV mit einer studentischen Sektion, die studentischen Beiträgen eine Chance geben sollte und mit einem kleinen Preis positiv auf die Ausbildungssituation wirken sollte. Die *Be*dingungen waren leider für ein erstes Mal nicht allzu gut: Die Ausschreibung wurde dem Call for Paper erst nachgeschoben, die Sektion war im vorläufigen Programm noch nicht enthalten, die Wettbewerbsbedingungen gaben noch Auslegungsspielraum, eine "politische Komponente" in Form eines besonderen Hinweises auf der Tagung oder eines leibhaftig anwesenden Vorstandsmitglieds fehlte und die Sektion wurde doch tatsächlich parallel zu zwei "normalen" Sektionen gelegt!

Sah es zu Beginn ganz nach einer *leeren Menge* von Zuhöre rinnen aus, füllte sich doch nach und nach der kleine Raum und es entwickelte sich nach den zwei Vorträgen eine rege Diskussion. U. Koch (Universität Koblenz) hielt einen Vortrag über *„Deutsche Relativsätze in HPSG"*, eine Arbeit, bei der er einen in Konstanz entwickelten Prolog-Ansatz für eine unifikationsbasierte Analyse erprobte.

A. Daunensteiner und A. Hechtbauer berichteten über ihre Überlegungen und Vorarbeiten für ein *Studieninformationssystem* im *World Wide Web* an der Universität Regensburg. In der Diskussion wurde insbesondere diskutiert, inwieweit sich der

GLDV-Studienführer für computerlinguistische Studiengänge als dezentrales Informationssystem im WWW eignen würde.

Ein dritter Vortrag fiel leider aus zwei verschiedenen Gründen aus, ohne eine analoge Wirkung wie die doppelte Verneinung entwickeln zu können.

Fazit: Beim nächsten Mal wird man zur Belebung von Interesse und Engagement der Studierenden einiges gelernt haben. Schließlich hat man anderswo, etwa beim *Hochschulverband für Informationswissenschaft* und seinen ISI-Tagungen in Bezug auf Niveau und Resonanz wirklich gute Erfahrungen mit einem studentischen Wettbewerb gemacht.

2.4 Gesprochene Sprache

Drei Vorträge bildeten die Abschlußsektion über *gesprochene Sprache.*

Bahnauskunft

Zwei Beiträge bezogen sich auf jeweils eine Anwendung zur *telefonischen Bahnauskunft.* In beiden Fällen können in fließender Sprache ohne kontrollierende Rahmenbedingungen Auskünfte von einem Informationssystem eingeholt werden, die in synthetischer Sprache beantwortet werden. Beide Systeme lösen das schwierige Problem der sprecherunabhängigen Erkennung von Spontansprache u. a. auch dadurch, daß kleinere Erkennungsfehler ohne Belang sind, sofern nur die Satzbedeutung noch rekonstruierbar ist.

Firmen-Entwicklung

Der erste Vortrag behandelte das Thema *Stochastische Sprachmodellierung* aus Sicht einer industriellen Entwicklergruppe. R. Kneser (Philips, Aachen) gab eine Einführung in die Grundstruktur heutiger Spracherkennungssysteme, die zur Worterkennung eine probabilistische akustische Modellierung mit einer probabilistischen Modellierung des Sprachsystems (auf der Basis von n-Grammen3 kombinieren.

Er erläuterte Vorteile und Nachteile der statistischen Sprachmodellierung, und zeigte die Richtung von Weiterentwicklungen an. Wer den Versuch nicht scheut, kann unter der Telefonnummer 0241-604020 sein nächstes Bahnproblem zu lösen versuchen4.

Universitätsentwicklung

Unter dem provokativen Titel *„Der Benutzer -- ein Störfaktor? Erfahrungen beim Einsatz eines Dialogsystems mit spontansprachlicher Eingabe"* trug W. Eckert (Universität Erlangen) über Architektur, Anspruch und Leistungsfähigkeit des Erlangener IC-Auskunftssystems vor. Der seit Januar 1994 laufende Feldversuch (Anschluß ans Telefonnetz unter 09131-16287) hat nicht nur Evaluierungsdaten geliefert, sondern durch Neuadaption der Worterkennung an diesen Trainingsdaten auch die Worterkennungsrate von 59,5% auf 72% gesteigert. Eindrucksvoll war der Tonbandmitschnitt eines mißglückten Auskunfts-dialogs, der im übrigen auch verständlich macht, daß 5 % der Dialoge Flüche enthalten 5. Als entscheidende Frage formulierte Eckert: *"Verhält sich der Nutzer so, wie das Modell es erwartet?"*. Die Antwort wird empirisch Arbeitende nicht überraschen:

Nein, denn Modelle sind immer unvollständig.

Benutzer ändern ihr Verhalten, wenn sich das Modell ändert.

Experimente mit verschiedenen Modellen sind deshalb nicht vergleichbar.

Als Fazit formulierte Eckert: Worterkennung (unter. den hier betrachteten Bedingungen) liefert gegenwärtig noch zu schlechte Ergebnisse, tatsächliche Anwendungen brauchen große Feldversuche, lernende Verfahren sind essentiell.

3 Normalerweise für n = 2 oder 3

4 Angabe ohne Gewähr. Ich persönlich bin noch nicht durchgekommen.

5 Vielleicht ist dies ein Mißverständnis und es ist tatsächlich so, daß es sich bei 5 % der Wörter um Flüche handelt?

Sprachdiktieren

Die Erfahrungen der IBM mit dem wissenschaftlichen und dem auf Produkte zielenden Thema Spracherkennung bearbeitete K. Mohr (IBM Heidelberg) in seinem Beitrag *„8pracheingabe beim Computer. Stand, Weg, Ziel.* ", Vom Projekt *Tangora,* Sept. 88, über eine Workstation/Unixbasierte Anwendung bis zum heutigen Produkt *Voice Type Dictation* (Windows/OS2) führte ein Weg, auf dem manche Lektion gelernt werden mußte: insbesondere, daß Spracherkennung in die Hardware-Umgebung der Kunden hineinpassen muß und daß nur die Integration in die Kundenanwendung Akzeptanz schafft.

Bei heutiger Technologie ist eine Einschränkung auf bestimmte Bereiche erforderlich, bekannt sind die Anwendungen in Medizin und Juristerei. Voice Type kommt mit *32000* Wörtern zurecht6, die mit kurzen Sprechpausen isoliert gesprochen werden müssen. Auf diese Weise schafft man (und das System) ca. 70 Wörter/Minute. Außerdem kann man einen persönlichen Wortschatz von ca. *2000* Wörtern trainieren. Um das statistische Sprechermodell des Systems an seinen Nutzer anzupassen, muß dieser ca. 1 Stunde Sprechertraining absolvieren. Beim Betrieb läßt sich damit eine Erkennungsgenauigkeit von ca. 95 % erreichen, das bedeutet jeweils einen Fehler in jedem 20-ten Wort. Das System löst weitgehend das Problem der Homophone (*„Viel Regen fiel vom Himmel"),* sowie das der Groß- und Kleinschreibung.

Eine Demonstration des Systems schloß den Vortrag ab.

2.5 20 Jahre GLDV. Kein Grund zum Feiern (?)?

Anläßlich des 20-jährigen Bestehens des GLDV wurde von Hans-Dieter Lutz eine Podiumsdiskussion moderiert, bei der das Fragezeichen im provokativ gemeinten Titel fraglich war - zumindest ließ es sich nicht feststellen, wer es zu verantworten hatte. Von den Gründungsmitgliedern der LDV-Fittings (so der ursprüngliche Name der GLDV) saßen Schweisthal, Lutz, Lutz-Hensel, und Lenders auf dem Podium, wobei wie beabsichtigt die Diskussion von früheren Vorsitzenden (Krause, Rieger) und anderen (Batori, Ott, Knorz. . .) auch aus dem Plenum engagiert geführt wurde.

Es wurde eingebracht, daß die mit der Namensänderung einhergehende Neuformulierung der Satzung die Ziele so definiert hat, daß auch eine verwaltete Gesellschaft sich noch als erfolgreiche Gesellschaft verstehen kann. Der These des gegenwärtigen GLDV-Vorsitzenden, daß nämlich *"die GLDV ihre ursprünglich gesetzten Ziele verfehlt habe* **",** wurde z. T. heftig widersprochen. Insbesondere die Formulierung, daß *„sich die GLDV unnötigerweise bei der Computerlinguistik angebiedert hätte* ", führte zu energischen Reaktionen. Der ursprüngliche Anspruch der LDV-Fittings, die Schwerpunktsetzungen der einzelnen Vorsitzenden und das gegenwärtige Bild der Gesellschaft wurden z.T. kontrovers und engagiert diskutiert, und es war doch recht unvermutet, daß der Vorsitzende alle Anwesenden schließlich mit der Gegenthese zu seiner These überraschte: *"daß die GLDV ihre Ziele doch recht gut erreicht hat* **".**

3 Fazit

Alles in allem eine gelungene Tagung, so lautet das Fazit. Die lokale Organisation (Hitzenberger, Womser-Hacker und Krause; dazu weitere Mitglieder der Regensburger Informationswissenschaft) lief reibungslos, das *Social Event* im alten Rathaus der Stadt Regensburg und anschließend im historischen Haus Heuport war interessant, lehrreich, anregend und sättigend. Das Schlußwort von J. Krause soll denkwürdig gewesen sein - trug man mir zu, als ich Tagungsteilnehmer eine Woche später auf der HIM'95 in Konstanz wiedertraf, aber dazu kann ich aus eigener Erfahrung nichts mehr beitragen.

Bleibt nachzutragen, daß in den Tagungsunterlagen alle Beiträge mit kurzem bis extended Abstracts zu finden waren und daß die Proceedings von L. Hitzenberger herausgegeben und bei Olms erscheinen werden.

6 Da jede flektierte Wortform ein eigenes Wort darstellt, ist dies im Deutschen längst nicht so weitreichend wie im Englischen.