

LDV-FORUM

Forum der Gesellschaft für linguistische Datenverarbeitung GLDV

LDV-Forum 12.1 (1995) Forum der Gesellschaft für Linguistische Datenverarbeitung e.V.

Herausgeber

Prof. Dr. Gerhard Knorz; Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)

Anschrift: Fachhochschule Darmstadt, Fachbereich Information und Dokumentation (IuD), Schöfferstr. 1-3, D-64295 Darmstadt Tel.: (06151)168490; Fax: (06151)16-8980; Email: knorz@fhda.com2.fhrz.fh-darmstadt.de

Redaktion

Gerhard Knorz, Ute Hauck

Wissenschaftlicher Beirat

Dr. Karin Haenelt, Prof. Dr. Christa Hauenschild, Prof. Dr. Gerhard Knorz, Prof. Dr. Jürgen Krause, Prof. Dr. Burghard Rieger, Dr. Dietmar Rösner, Prof. Dr. Burkhard Schaefer

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



Editorial

Freitag nachmittag ist es, Vatertag war gestern - nein, nicht wie Sie meinen, Vorlesungen habe ich vorbereitet, bis 01:00 Uhr, und erholt habe ich mich bei der sommernächtlichen Radfahrt nach hause. Das Editorial ist für Montag versprochen und meine to-do-Liste für das Wochenende ist noch lang: so sitze ich also durchaus etwas erschöpft, aber wild entschlossen vor dem PC, um ein weitgehend pünktliches Erscheinen der ersten 95er Ausgabe des LDV-Forum in die Wege zu leiten.

Anderen Personen muß es wohl ähnlich ergehen. Ich zitiere aus dem Protokoll der Vorstands- und Beiratssitzung vom 28.3.95 in Regensburg:

W. Lenders berichtet über die Aktivitäten der Arbeitskreise: Der Arbeitskreis "Lexikographie" unter der Leitung von Nico Weber hat im Januar in Bonn ein Kolloquium zum Thema "Theorie der Semantik und Theorie der Lexikographie - Angewandte Semantik und Praxis der Lexikographie" organisiert, das auf großes Interesse gestoßen ist.

Wie ich meine, eine erfreuliche Aktivität der GLDV und aus diesem Grunde - und überhaupt - ein Arbeitskreis-Engagement, über die es sich zu berichten lohnt. Der Sprecher des Arbeitskreises hatte bei unserem Gespräch über den entsprechenden LDV-Forum-Beitrag auch keine Zweifel, daß er ihn schreiben würde - aber: es hat nicht geklappt. Entschuldigen wir Nico Weber mit dem Wechsel seines Tätigkeitsbereichs und den vielen zusätzlichen Arbeiten, die ein Stellenwechsel mit sich bringt, und wünschen ihm in seiner neuen Position als Professor an einer Fachhochschule alles Gute.

Auf eine weitere bemerkenswerte Veränderung der Personallandschaft durch die Berufung von Prof. Dr. Jürgen Krause auf eine Professur der Universität Koblenz/Bonn, verbunden mit der Leitung der IZ Sozialwissenschaften. Somit hat sich Jürgen Krause als früherer Vorsitzender der GLDV (damals noch unter dem Namen LDV-Fittings) von der *Linguistischen* Informationswissenschaft über die nicht weiter attributierte Informationswissenschaft zur Informatik aufgemacht. Der Informationswissenschaft wird er wohl erhalten bleiben (schließlich wurde er erst kürzlich zum Vorsitzenden des HI gewählt). Über die verbleibenden Verbindungen zur Computerlinguistik muß man

Teilnehmer der Jahrestagung in Regensburg fragen, die bis zur letzten Minute ausgeharrt haben². Wünsche für einen guten Start von dieser Stelle! Wobei wir damit beim Thema der Jahrestagung in Regensburg wären, einer insgesamt geglückten Veranstaltung. Über Verlauf und Bewertung der Tagung berichte ich ausführlich unter der Rubrik " Tagungsberichte". Meine Anregung zielt darauf, einem Wettbewerb studentischer Arbeiten in Zukunft einen größeren Stellenwert beizumessen. Ich verspreche mir davon u. a. auch etwas (mehr) jugendlichen Schwung bei der Tagung. Apropos *Schwung*. Den glaube ich in dem Aufruf zur Interessensbekundung für einen neuen Arbeitskreis *Hypermedia* (S. 91) verspürt zu haben³. Ich bin gespannt, was daraus wird. . .

So, und nun hoffe ich; daß auch Sie gespannt auf den weiteren Inhalt des Heftes sind. Sollten Sie sich positiv oder negativ durch das Angebot an Beiträgen angesprochen fühlen - jede angemessene Form von Feedback ist herzlich willkommen. Und wenn Sie sich fragen, ob Sie nicht selbst. . . Die Antwort lautet Ja! Schreiben Sie einen Fachbeitrag, oder einen Leserbrief dazu. Verfassen Sie einen Tagungsbericht oder eine Rezension oder schicken Sie ein paar Zeilen mit Nachrichtenwert! Versprochen?

Titelgestaltung
Markus Allgayer, Saarbrücken

Fachbeiträge
Unaufgefordert eingesandte
Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte (dreifach) sollten daher möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall zusätzlich auch noch auf Diskette (5 1/4" bzw. 3 1/2") als ASCII oder U-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 12.2/95: 31. Okt. 1995; für
G.K. Heft 13.1/96: 30. Apr. 1996

Herstellung IAI,
Saarbrücken

Druck
reha GmbH, Saarbrücken

Auflage
400 Exemplare

Anzeigen
Preisliste und Informationen: Prof. Dr. Johann Haller, Institut für Angewandte Informationsforschung (IAI), Martin-Luther-Straße 14, D-66111 Saarbrücken; Tel.: (0681) 39313; Fax: (0681) 397482; Email: hans@iai.unisb.de

Bankverbindung LDV-Forum
(Prof. Haller): SaarLB Saarbrücken
(BLZ 590 500 00) KtoNr. 20 00 21
43

GLDV-Anschrift
Prof. Dr. Winfried Lenders, Institut für Kommunikationsforschung und Phonetik (IKP), Poppelsdorfer Allee 47, D-53115 Bonn; Tel.: (0228) 735638, Fax: (0228) 735639; Email: lenders@uni-bonn.de

PS. Eine Richtigstellung für das LDV-Forum 94/2 ist nachzutragen: Aufgrund einer technischen Panne ist der Autor des Berichtes über die „3. Internationale Fachtagung für Computereinsatz in der Historischen Sprachwissenschaft" (S. 66) falsch angegeben. Richtig muß es heißen: Johann Tischler (Dresden). Ich bitte um Entschuldigung.

²Meine Bahnverbindung zog mich eine Stunde zu früh aus dem Geschehen ab.

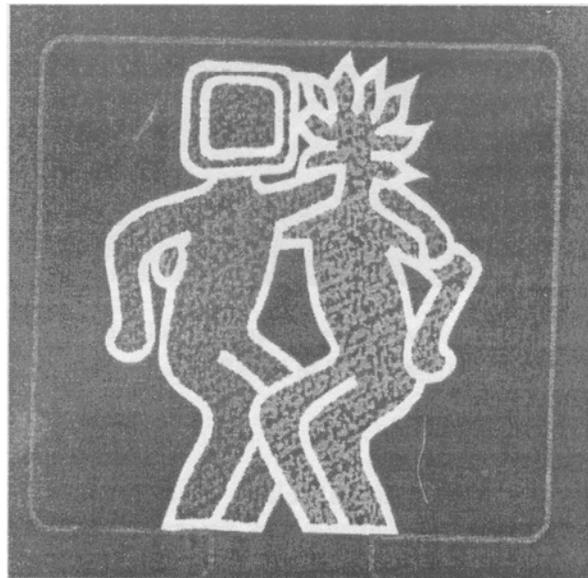
³Diese Hervorhebung darf keinesfalls in der Weise mißverstanden werden, daß anderen AK-Aufrufen der Elan fehlen würde.

KOMMUNIKATIVES HANDELN:

MENSCH - COMPUTER

Helmut Sagawe

Der Computer als quasi „selbständig denkende und handelnde“ Maschine hat ohne Zweifel unsere alltägliche Lebenswelt erobert. Er kann nicht mehr nur als technischer Organersatz des Menschen angesehen werden, er ist zum Agenten im kommunikativen Handlungsprozeß geworden. Handlungs- und Verhaltensänderungen sind die Folge. Der Sprache, die heute immer noch als das elementare Kommunikationselement zwischen Mensch und Maschine angesehen wird, kommt durch die Anthropomorphisierung des Computers und seiner Metaphorik eine bedeutende Rolle zu.



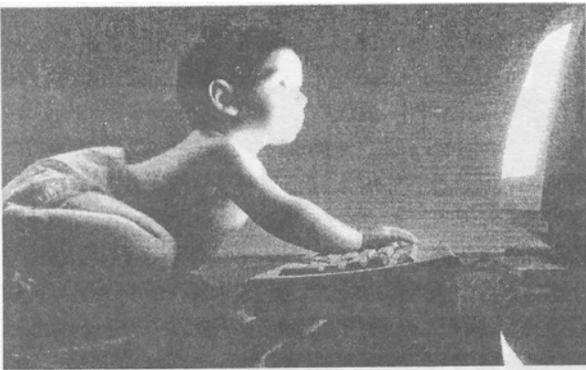
Die Beziehung zwischen dem Menschen und der Maschine rückt heute immer mehr ins wissenschaftliche Forschungsinteresse. Kritiker und Befürworter informationstechnischer Innovation befassen sich in der Regel mit ergonomischen und arbeitspsychologischen Auswirkungen und Problemen beim Einsatz von Computern, verkennen dabei aber die subjektiv wahrgenommene soziotechnische Evolution des neuen "Partners COMPUTER".

Broschüre CeBIT, Hannover 16.-23.3.1994, Chancen 2000, Computer und Umwelt, Deckblatt, Deutsche Messe AG, Hannover.

Diese unerforschte Situation gilt es nun aus wissenschaftlicher Sicht empirisch zu erfassen bzw. für den wissenschaftlichen Erkenntnisprozeß zugänglich zu machen.

Bereits seit einigen Jahren werden im Bereich der informationstechnischen Innovation vermehrt theoretische Überlegungen dazu angestellt, wie die Beziehung Mensch/Gesellschaft versus Maschine aus wissenschaftlicher Sicht analysiert werden könne. In der bis heute geführten Diskussion zeichnen sich im wesentlichen nur wenige grundsätzliche Positionen ab: die Befürworter einer Informatisierung der Gesellschaft möchten mit ihrer internen Kri-

tik, die auf die Funktionsweise der Computer bezogen ist, eine Verbesserung der sozio-technischen Systeme erreichen, die Gegner der Computerisierung hingegen möchten durch eine externe Kritik das Eindringen der *Künstlichen Intelligenz* in alle Lebensbereiche des Menschen verhindern, und die allgemeine Technologiekritik macht auf die gesellschaftlichen Risiken und Folgen des Einsatzes der Computer-Technologie aufmerksam.



Baby vor dem Computer: Anzeige: "BTX special", in BILDSCHIRMTEXT MAGAZIN 4/94, Seite 61, Ulm.

In sozialwissenschaftlichen Analysen wird im allgemeinen gerne von krisenhaften Erscheinungen gesellschaftlicher Entwicklungen gesprochen, wie bei der Bildungskrise oder insbesondere heute bei der Krise des Umweltbewußtseins. All diese Misereen werden als Beziehungskrisen zwischen Wissenschaft und Technik einerseits, Wissenschaft und Gesellschaft andererseits analysiert. Sie beruhen auf Legitimationsverlust bzw. Funktionsdefizit der organisierten Umsetzung von Wissenschaft und Technik in sozialen Systemen. Die technologische Evolution sozio-technischer Systeme wird hier als der Hauptverursacher verantwortlich gemacht. Bisher hatte man Technikwissen überwiegend als Form sozialer Machtausübung, als Form struktureller Herrschaftssicherung oder als Raster für soziale Prozesse wahrgenommen. *Herbert Marcuse* etwa wies 1967 darauf hin, daß die Technik nicht erst durch ihre konkrete Anwendung, sondern schon durch ihre Logik und Methodik zur Herrschaftsausübung benutzt werden könne. Die Einführung der

Mikroelektronik in alle Lebensbereiche darf als weiterer Eingriff in die Autonomie des Menschen gesehen werden, der zur Folge hatte, daß "*Individuen erstmals in die Lage versetzt werden können, auch Handlungen interaktiver Art ohne Einbezug anderer menschlicher Personen zu vollziehen*", so der Soziologe *Hans Geyer* (Zürich), da der Computer Fähigkeiten aufweise, dank derer der Mensch ihn eher als Interaktionspartner denn als bloßes Werkzeug ansehen könne. Wie das menschliche Gehirn, so kann auch er Informationen speichern, rezipieren, transformieren und kombinieren, er ist lernfähig und sorgt gelegentlich sogar für Überraschungen. Aufgrund dieser Eigenschaften des Computers stellt *Geyer* der sozialen Evolution der Menschheit die *Koevolution* intelligenter Maschinen zur Seite, die sich zwischen humanen und elektronischen Interaktionspartnern mit all ihren Auswirkungen vollziehe. Lege man einen objektivistischen, behavioristischen Verhaltensbegriff bei dieser Interaktion zwischen Mensch und Computer zugrunde, müsse man feststellen, daß zwischen beiden rückgekoppelte Interaktionssequenzen stattfänden, die sich von zwischenmenschlichen Interaktionssequenzen nicht mehr wesentlich unterschieden. Aus phänomenologischer Sicht dürften daher dem Computer anthropomorphe und soziomorphe Attribute zugesprochen werden, die die Grundlage emotionaler Bindungen zu ihm bilden könnten.

Zum ersten Mal in der technischen Evolution ergeben sich aufgrund solcher Überlegungen spezifische Sozialbeziehungen zu einem Arbeits- oder gar *Denkzeug* (*Klaus Haefner*, Informatiker / Bremen), das die Einseitigkeit des früheren formalen Handelns mit technischem Gerät ablöst und kommunikatives Handeln evoziert. Mit den früheren Tätigkeiten in beruflichen Sozialgemeinschaften vergleichbar, hat der agierende Mensch nun wieder die Möglichkeit, ein auf Interaktionssequenzen basierendes Sozialverhalten zu praktizieren, er hat im Computer ein menschenähnliches Substitut gefunden, das die am modernen Arbeitsplatz nicht mehr zustandekommenden Sozialbeziehungen ausgleichen kann.

Ist damit ein neues Zeitalter der Mensch

heitsgeschichte angebrochen? Es wäre nicht das erste Mal, daß die Menschen wieder einmal umdenken und sich einen neuen Standort suchen müßten: nachdem sie durch *Kopernikus* erfahren hatten, daß die Erde und damit der Mensch nicht im Mittelpunkt der Welt stehen, durch *Darwin*, daß der Mensch nicht durch einen besonderen Schöpfungsakt von Gott geschaffen wurde, sondern nur Endglied eines natürlichen Evolutionsprozesses ist, durch *Freud*, daß der Mensch nicht völlig Herr seiner selbst ist, sondern maßgeblich von unbewußten Trieben und Kräften gesteuert wird, so erhebt sich gegenwärtig die Frage, ob der Mensch womöglich nicht mehr alleine die Fähigkeit zur Intelligenz besitzt, ja unter Umständen sogar von seinen selbst erschaffenen Maschinen übertroffen werden könnte. Ob dies als subjektive Meinung der Menschen oder objektive Tatsache in der Realität wahrgenommen wird, sei dahingestellt, problematisch aber mag es sein, geeignete Indikatoren zu finden, die es ermöglichen, technische Evolutionsprozesse im Bewußtsein der Menschen aufzuzeigen und zu beurteilen. Die in der Sprachwissenschaft oft thematisierte verbale und visuelle Metapher mag solch ein Indikator sein.

Es gibt in der Soziologie bis heute kaum Ansätze, welche die sog. *Künstliche Intelligenz* in ihre Erklärungsmodelle einer *Mensch-Maschine-Beziehung* zu integrieren vermögen, auch wird kaum über "unterschiedliches" Denken bei Menschen und "intelligenten" Maschinen reflektiert. Sollen aber Risiken oder Nutzen dieser Technologien untersucht werden, könnte dies auf verschiedenen Ebenen durchgeführt werden. Ökonomischer Nutzen kann durch Mechanismen der Betriebs- und Volkswirtschaftslehre gemessen werden, sozialer Nutzen etwa durch die Aufsummierung von verlorengegangenen und hinzugekommenen Arbeitsplätzen und psychologische Auswirkungen etwa in Veränderungen beim Sozialverhalten und der Kommunikationsfähigkeit einzelner sowie bei Subgruppen. *Kommunikative* und soziale Verhaltensänderungen können aber am ehesten an Sprache und Sprachhandlungen beobachtet werden. Soziale Veränderungen machen zuerst in

den Medien wie Presse, Rundfunk und den visuellen Kommunikationsmitteln von sich reden. Durch Sprache werden neue Sachverhalte verbalisiert und im zweiten Schritt realisiert, d.h. in eine formale oder metaphorische Sprache transformiert.

Was ist jedoch der Indikator Metapher und welche Funktion hat er? - Die Metapher gehört zu den Sinn- oder Namensänderungen (Tropen), wird aber von einem Vergleich unterschieden. Während beim Vergleich zwei Sachverhalte in ihrer Ähnlichkeit explizit nebeneinander gestellt werden, läuft dieser Prozeß bei der Metapher implizit ab, wodurch eine sehr viel stärkere, zum Teil unbewußt ablaufende Identifizierung beider Sachverhalte suggeriert wird. Im Bereich der elektronischen Datenverarbeitung, und zwar sowohl in der Umgangssprache als auch in der Fachsprache, bei Produktwerbung und Imagepflege, wird durch die Metaphorisierung häufig ein tiefgreifender anthropomorphisierender Prozeß eingeleitet, der sich verbal und visuell darstellen läßt. Die Metapher ist soziologisch gesehen also sehr viel mehr als nur im linguistischen Sinne ein bloßer Redeschmuck, sie erfüllt zwar als Ausdruck sprachlicher Kreativität eine wesentliche sprachästhetische Funktion, muß darüber hinaus aber als Indikator für einen Verinnerlichungsprozeß fachspezifischer Sachverhalte angesehen werden. So scheint der Computer mit allen menschlichen Eigenschaften versehen, zu einem menschlichen Wesen geworden zu sein. Der Mensch wird in überwältigender, ja intimer Weise angesprochen, wie er es bisher selten erlebt hat. Beeinflußt werden seine Persönlichkeit, seine Identität, ja sogar seine Sexualität. Es werden Gefühle geweckt und sexueller Lustgewinn erzeugt. Die sozialen Grenzen zwischen Mensch und Maschine werden aufgehoben.

Die Ebene der Metaphorik ist jedoch nur eine erste Stufe zur Erklärung des Verinnerlichungsprozesses, auf der sich der Umgang mit der informationsverarbeitenden Technologie in der heutigen Gesellschaft widerspiegelt. Anthropomorphisierende Metaphern wie beispielsweise "*read (l* oder "*write (l* werden aber inzwischen auf einer zweiten Stufe entweder überhaupt nicht mehr als

Metaphern empfunden oder sind bereits fachsprachlich lexikalisiert (*tote Metapher*). Der allgemeine Eindruck läßt derweilen darauf schließen, daß eine "konzeptuelle" Metapher immer mehr Raum gewinnt, auf deren Basis der Computer *animistisch* als *intelligente Maschine* erlebt wird und als Projektionsfläche für *eine ganze Skala von Gefühlen* dient, somit personell und sozial verinnerlicht wird.

Betrachtet man die Mensch-Maschine-Beziehung nun vor diesem anthropomorphen Hintergrund als überaus vielschichtige Kommunikationssituation, so ist der bisherigen Auffassung einer zielgerichteten und erfolgsorientierten *Kommunikation Mensch-Maschine* der erweiterte Begriff vom *kommunikativen Handeln Mensch*

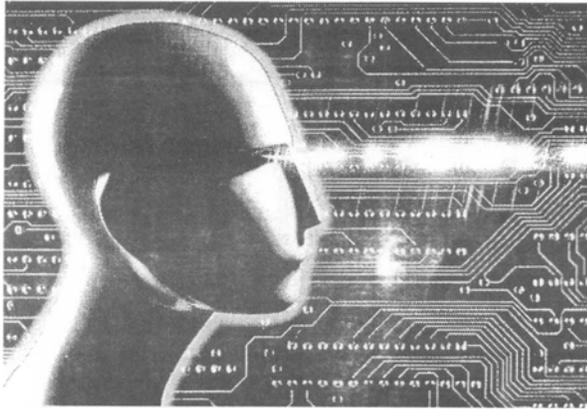
Maschine gegenüberzustellen. Die traditionelle Anschauung von der Mensch-Maschine-Kommunikation basierte auf der Handlungsebene eines teleologischen (zielgerichteten) Handelns, das sowohl als ein rein auf die Maschine bezogenes instrumentelles Handeln angesehen wie auch als soziales strategisches Handeln betrachtet wurde. Beide Fälle beziehen erfolgsorientierte Kalküle als Handlungsziele mit ein. Solch instrumentell-strategisches Handeln wird durch Zweck- und Erfolgsorientierung wie egozentrische Erfolgskalküle und die Koordination individueller Handlungspläne begründet. Die Zweck-Mittel-Relation kennzeichnet hier das *Rationalisierungsprinzip*.

Der Mensch hat zudem aber auch *Geltungsansprüche* an eine "intelligente" Maschine. Diese beruhen auf den Prinzipien der allgemeinen Verständlichkeit und der Wahrheit. Beide können aus der objektiven Richtigkeit der Daten und deren Verarbeitung nach naturwissenschaftlichen Gesetzen und Logik abgeleitet werden. Als sprachliche Kommunikationselemente werden zwischen Mensch und Maschine Befehle, Lemmata, Assoziationen und Daten, aber inzwischen auch Icons (Sinnbilder, z.B. in WINDOWS) ausgetauscht, die jedoch einen abgesteckten Handlungsrahmen bestimmen.

Kommunikatives Handeln zwischen Mensch und Maschine kann nun im "Rah-

men eines erweiterten Kommunikationsmodells dargestellt werden. Es basiert zwar auch auf dem Konzept eines normenregulierten Handelns und ist durch kognitive Erfahrungen bestimmt, antizipiert aber Erwartungen aus der Handlungssituation, da sich das Handeln auf die Maschine und deren Umwelt ebenso bezieht wie auf den Menschen als Handelnden selbst. In beiden Fällen darf die Handlungssituation als sozial charakterisiert werden, da Werte, Normen und das antizipierte Einverständnis der zugehörigen sozialen, aber nicht direkt an der Kommunikationssituation beteiligten Gruppe (User) einer gesellschaftlichen Bestätigung dienen, und zudem Regeln und Syntax von Programmiersprachen und Anwender-Menüs die Erwartungshaltung und somit das kommunikative Handeln mitbestimmen.

Erstmalig kann hier nun auch ein *reflexiver Modus von Interpretation* als Rationalitätsprinzip gedeutet und die Handlungssituation - in Anlehnung an die Theorie des Kommunikativen Handelns, die *Habermas* ausführlich anhand der zwischenmenschlichen Kommunikation dargestellt hat miteinbezogen werden. Beide Agierenden, Mensch und Computer, sind in einen sozialen Handlungszusammenhang integriert, nehmen Bezug auf ihr erfahrenes und gespeichertes Wissen, ihre subjektiven Interpretationen (wenn dies beim Computer auch nur durch logisches Kombinieren geschehen kann) und ihre normativen Orientierungen, beanspruchen aber innerhalb ihrer realen Welten und visuellen Scheinwelten die Geltungsansprüche der Verständlichkeit der sprachlichen Ausdrücke, der objektiven Wahrheit der geäußerten Behauptungen (Daten, Befehle), der subjektiven Wahrhaftigkeit von zum Ausdruck gebrachten Intentionen (Übereinstimmung mit den gespeicherten Daten) sowie der Richtigkeit der vollzogenen Sprachhandlung mit den ihr zugrundeliegenden Normen (Syntax der Programmiersprache, Verwendung der vorgegebenen Befehle und Icons).



Künstlicher Kopf mit Lichtreflexen zur elektronischen Leiterplatte: in Bildschirmtext magazin 1/94, Seite 6.

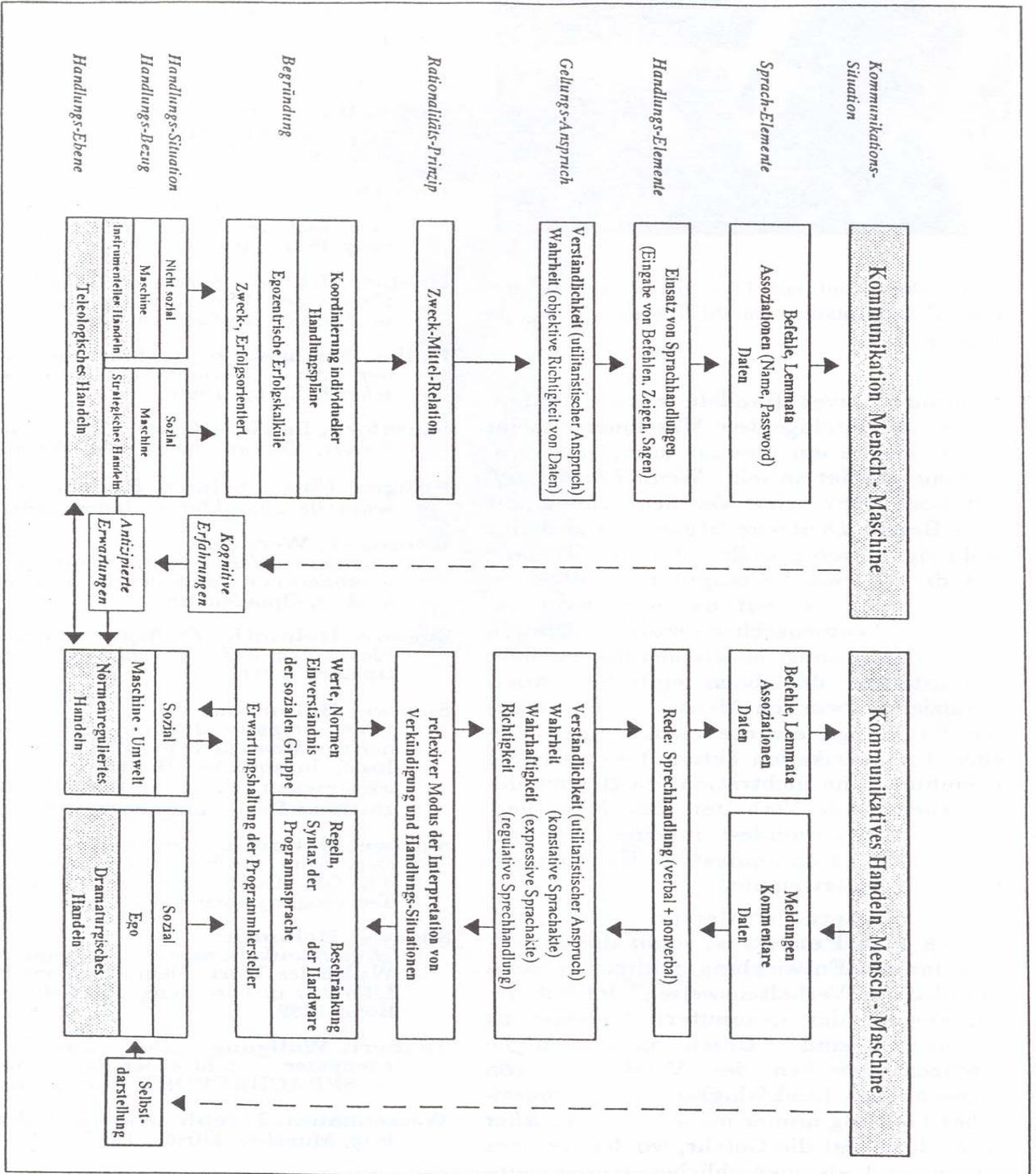
Kommunikatives Handeln zwischen Menschen und intelligenten Maschinen scheint inzwischen in ein überaus komplexes Umfeld eingebettet zu sein. Wenn Klaus Haefner noch 1987 seine Veröffentlichung mit dem Begriff *Denkzeug* titulierte und damit wohl signalisieren wollte, daß der Umgang mit dem Personal Computer im allgemeinen Verständnis weit davon entfernt ist, eine zwischenmenschliche soziale Situation zu evozieren, so ist inzwischen immer mehr festzustellen, daß beim einfachen Kommunizieren zwischen Mensch und Computer Handlungselemente zwischenmenschlicher Kommunikation aktiviert werden. Sie beeinflussen im nichtrationalen Bereich dahingehend das Verhalten des Menschen, daß er sich zumindest in eine zwischenmenschliche kommunikative Handlungssituation versetzt glaubt.

Die "konzeptuelle" Metapher THE COMPUTER IS A PERSON ist somit die Grundlage für die Entwicklung zahlreicher kommunikativer Verhaltensweisen, die seit der Einführung der Computertechnologie zu beobachten sind. Durch sie werden die Grenzen zwischen der Vorstellung von menschlicher Denkfähigkeit und rechnerischer Leistung immer mehr beseitigt. Aber nicht dort liegt die Gefahr, wo der Rechner zunehmend als menschliches Wesen empfunden wird, sondern in der Umkehrung dieses Sachverhalts, in der Verinnerlichung

einer neuen konzeptuellen Metapher HUMANS ARE MACHINES, die den Menschen zunehmend als funktionales Wesen darstellt, wie viele Darstellungen und berufliche Anforderungen es heute schon beweisen.

Literatur

- Geser, Hans. *Der PC als Interaktionspart*
ne~ ZEITSCHRIFT FÜR SOZIOLOGI~, Jg. 18, Heft 3, Juni 1989.
- Habermas, Jürgen. *Theorie des kommunikativen Handelns, Handlungsrationalität und gesellschaftliche Rationalisierung*, Frankfurt 1981.
- Haefner, Klaus. *Denkzeuge, Was leistet der Computer? Was muß der Mensch selber tun?*, Basel-Boston-Stuttgart 1987.
- Haller, Michael. *Weizenbaum contra Haefner. Sind Computer die besseren Menschen?*, Zürich 1990.
- Mumford, Lewis. *Mythos der Maschine, Kultur, Technik und Macht*, Wien 1980.
- Pflüger, Jörg / Schurz, Robert. *Der maschinelle Charakter*, Opladen 1987.
- Rammert, Werner.
Computerwelten - Alltagswelten. Wie verändert der Computer die soziale Wirklichkeit, Opladen 1990.
- Sagawe, Helmuth. *Einfluß intelligenter Maschinen auf menschliches Verhalten*, Opladen 1994.
- Sagawe, Helmuth.
„Der Computer als Kommunikationspartner des Menschen“ in: Wallmannsberger, Josef, Innsbrucker Beiträge zur Kulturwissenschaft (Universität Innsbruck), Anglistische Reihe, Innsbruck 1990.
- Sagawe, Helmuth. „Ordenadores y usuarios, Una Probemtical relation“, in: TELOS, Cuadernos de Comunicacion, Tecnologia y sociedad 20, Madrid 1990.
- Sagawe, Helmuth.
„Kommunikations-Partner Computer“, in: Wirkendes Wort, Deutsche Sprache und Literatur in Forschung und Lehre, Nr. 2 Bonn 1989.
- Teubert, Wolfgang. *„Die Sprache der Computer - bald unser aller Sprache?“*, in: SPRACHREPORT, Mannheim 1986.
- Weizenbaum, Joseph. *Kurs auf den Eisberg*, München-Zürich, 1987.



SPRACHVERARBEITUNG AUF DER BASIS DER SEMANTIKORIENTIERTEN GRAMMATIK

Maria Theresia Rolland

Zusammenfassung

Zunächst steht die semantikorientierte Grammatik im Blickpunkt, wobei deren Komponenten mit zentralen vorhandenen Ansätzen verglichen und die Unterschiede verdeutlicht werden. Im Anschluß daran wird als Anwendung ein auf der semantikorientierten Grammatik basierendes Dialogsystem beschrieben. Schließlich folgten Hinweise auf die wohl bedeutendste Anwendung, die maschinelle Übersetzung.

Abstract

*Language processing based on the semantic-oriented grammar
First we focus on the semantic-oriented grammar, comparing this new approach with some existing approaches, whereby the differences become visible. As an application we describe a query system that is based on this new grammar. Finally, we sketch machine translation as the most important application.*

1 Ausgangslage

Im Bereich der Sprachverarbeitung kann man grundlegend verschiedene Anwendungstypen unterscheiden. Einerseits ist es so, daß der Computer sprachlich zu Aktionen, zu Handlungen unterschiedlichster Art aufgefordert wird. Da er nun aber die Sprache nicht "verstehen" kann, wie der Mensch sie versteht, hat man zu Repräsentationen der sprachlichen Äußerung gegriffen, also zu Bedeutungsrepräsentationen, zu Wissensrepräsentationen, und hat auf diese Weise die Befehle für den Computer "verstehbar" gemacht und kann somit erfolgreich eine Steuerung des Computers bewirken.

Ein anderer Typ der Anwendung ist der, bei dem sprachliche Pendanten im Mittelpunkt stehen, so Frage- und Antwort-Entsprechungen, wie z.B. in Dialogsystemen, oder Aussageentsprechungen in der Ausgangssprache und in der Zielsprache, wie bei der maschinellen Übersetzung. Für diesen Anwendungstyp ist es nunmehr möglich, sprachimmanent zu verfahren und die Sprache direkt, und zwar vollautomatisch, zu verarbeiten, d.h. ohne die üblichen Repräsentationsformalismen [z.B.: Hellwig, 1989]. Welches sind die Voraussetzungen dafür?

Die erzielten Ergebnisse gehen zurück auf Grundgedanken des großen Sprachphilosophen Wilhelm von Humboldt [1830-35], die in unserem Jahrhundert aufgegriffen und spezifiziert worden sind von Leo Weisgerber [1962], lange Jahre Professor an der Universität Bonn, dem Begründer der Sprachinhaltforschung. Auf dieser Theorie aufbauend und

sie weiterentwickelnd, sind die Grundzüge der semantikorientierten Grammatik aufgestellt [Rolland, 1994], die ihrerseits die Voraussetzung für die vollautomatische Verarbeitung der Sprache in den genannten Anwendungsbereichen darstellen.

Die Grundüberlegungen sind folgende: Der Computer kann nur dann jeden eingelesenen Satz identifizieren und im angesprochenen Sinne adäquat verarbeiten, wenn die Regeln und Gesetzmäßigkeiten, nach denen Sprache "funktioniert", d.h. die der Mensch unreflektiert und ganz selbstverständlich richtig handhabt, die also *implizit* in ihm sind, *explizit* gemacht werden. Es geht daher darum, im Vorfeld der Sprachverarbeitung aufzuweisen, welches Ausgangswort mit welchem abhängigen Wort in welcher Beziehung (Relation) semantisch sinnvoll verknüpft werden kann.

Wenn der einzelne Mensch spricht, kann er die Sprache nicht manipulieren, sondern er muß, wenn er verstanden werden will, die Wörter so wählen und miteinander kombinieren, wie dies in der Sprache als möglich vorgegeben ist. Der Mensch beherrscht zwar die Sprache - er kann sie handhaben, in Wirklichkeit aber beherrscht (regiert) die Sprache den Menschen, da er sich ihren Gesetzmäßigkeiten unterwerfen muß. Wenn der Mensch spricht, wählt er also lediglich aus den in der Sprache vorgegebenen Möglichkeiten aus. Um diese Möglichkeiten explizit greifbar zu machen, ist es notwendig, an der Sprache selbst, der *langue* [Saussure, 1967] anzusetzen, und nicht, wie bisher in der Regel geschehen, bei der Sprachverarbeitung von Sätzen und Texten auszugehen; denn Sätze und Texte sind individuelle Sprachrealisierungen, und es ist bis heute nicht gelungen, von diesen individuellen Realisierungen aus die korrekten Rückschlüsse zu ziehen auf die in der Sprache generell geltenden Regeln.

Das Kernelement der Sprache ist das Wort. Dieses steht daher im Mittelpunkt der Betrachtung. Demgemäß heißt die neue Methode, auf der alles basiert, d.h. die der semantikorientierten Untersuchung der Sprache dient und eine vollautomatische Sprachverarbeitung ermöglicht: Logotechnik ("Worthandhabung"). Das heißt natürlich nicht, daß im Zuge der Erforschung der Sprache nicht gerade auch das Wort als solches, auch in seiner semantischen Ausprägung [Stechow. et al., 1991], bzw. seine Rolle in der Sprachverarbeitung [z.B.: Batori et al., 1989; Guenthner et al., 1986] untersucht worden wäre. Im folgenden werden daher zentrale Ansätze dieser Art skizziert und die Unterschiede zu dem neue Ansatz der Logotechnik verdeutlicht. So stehen - sprachlich gesehen im Blickpunkt: Homonyme, Wortarten, Flexionsbezug, Relationen, Wortinhalt, auch der Text. Schließlich wird gezeigt, wie auf der Basis der erzielten Ergebnisse eine vollautomatische Sprachverarbeitung möglich ist, und zwar detailliert beschrieben am Beispiel eines Dialogsystems, angedeutet für die maschinelle Übersetzung, die wohl die bedeutendste Anwendung darstellen dürfte.

2 Wort und Homonym

Zunächst ist es wichtig, sich darüber klar zu werden, was ein Wort ist. Ein Wort ist eine Ganzheit aus Lautform und Inhalt, wobei der Inhalt maßgebend ist für die Bestimmtheit des Sprachmittels [Weisgerber, 1962].

Wort	
Lautform	Inhalt

Abb. 1 Strukturformel des Wortes

Ein Wort kann aus einer einzigen Zeichenfolge bestehen (z.B.: *Computer*), oder sich als Wendung ausdragen (z.B.: *ein X für ein U machen*) oder ein Homonym darstellen. Homonyme sind Wörter gleicher Lautform, aber verschiedenen Inhalts, z.B. *Schloß* (*Gebäude*), *Schloß* (*Türschloß*). Daß 'in der Sprache zahlreiche Homonyme vorliegen, ist

bekannt. Die Bedeutungs differenzierung, die vorgenommen wird, ist jedoch derart (vgl. den Aufbau der Wörterbuchartikel), daß (bis auf vereinzelte Ausnahmen) üblicherweise *ein* Wort mit mehreren Bedeutungen angenommen wird, d.h. die Lautform wird als ausschlaggebend angesetzt, statt daß der Inhalt den Maßstab darstellt, wie es die Sprache erfordert [Weisgerber, 1962]; denn nicht ein Wort hat mehrere Bedeutungen, sondern mehrere Wörter haben die gleiche Lautform.

Wenn man sich streng an diese sprachliche Gesetzmäßigkeit hält, ergeben sich umfassende und entscheidende Unterschiede u.a. auch im Bereich der Konjugation und der sogenannten adverbial verwendeten Adjektive.

Wir sind gewohnt, von den 3 *Personen* Singular und Plural zu sprechen. Die *Sache* wird [Duden, 1984: §311] unter die 3. Person subsumiert. Da *Person* (vereinfacht gesagt statt: Lebewesen) und *Sache* aber durchgehend als differenzierende Komponenten in der Sprache vorliegen, müssen sie entsprechend auch semantikorientiert unterschieden werden. Es gibt also - wie man jetzt sagen muß - 8 *Subjektkategorien*: *ich, du, er-sie-es* (Person), *er-sie-es* (Sache), *wir, ihr, sie* (Person), *sie* (Sache). Ein Verb mit Personen- und Sachsubjekt wie: *ziehen* weist also pro Tempus und Modus 8 Konjugationsformen auf. Diese Differenzierung in 8 Subjektkategorien hat - wie ohne weiters einleuchten dürfte - eine entscheidende Bedeutung für die Klassenbildung der Verben [Details: Rolland, 1994: 164ff.].

Darüber hinaus gibt es nicht 3 Hilfsverben, sondern 4: *sein, haben, werden-Futur* (er wird rufen), *werden-Passiv* (er wird gerufen). In der Form: *er wird gerufen werden* ist *wird* = *werden-Futur*, *werden* = *werden-Passiv*.

Neben den Vollverben: *sein* 1(existieren), *sein* 2(gehören), *sein* 3(sich befinden), *haben* (besitzen), *werden* (sich verwirklichen) und den gleichlautenden Hilfsverben gibt es bekanntermaßen noch die Kopulaverben: *sein, haben, werden*, z.B.: *groß sein, fertig werden, Glück haben*. Die aus einem Kopulaverb und einer Ergänzung bestehenden Verben bilden kein Zustandspassiv [Duden, 1984: §187], sondern es handelt sich um Aktivformen, z.B.: *verzaubert sein: er ist verzaubert, er war verzaubert, er wird verzaubert sein, er ist verzaubert gewesen, er war verzaubert gewesen, er wird verzaubert gewesen sein* im Gegensatz zu: *verzaubern: er wird verzaubert, er wurde verzaubert, er wird verzaubert werden, er ist verzaubert worden, er war verzaubert worden, er wird verzaubert worden sein* [vgl. Rolland, 1994: 63].

Außerdem sind 4 Partizipien zu unterscheiden und nicht 2, nämlich: im Aktiv und Passiv je die Partizipien Präsens und Perfekt, z.B.: *ziehend, gezogen (habend), gezogen (werdend), gezogen (worden seiend)*.

Es gibt jeweils 2 eigenständige Infinitive: *loben, zu loben*, also den einfachen Infinitiv und den Modalinfinitiv (= Infinitiv mit *zu*), der eine Ganzheit darstellt [vgl. Rolland, 1994: 70ff.].

Was Adjektiv bzw. Adverb angeht (das *schöne* Buch - das Buch ist *schön*), so ist zur Unterscheidung der Wörter folgendes zu sagen [vgl. Bergenholtz/ Schaefer, 1977: 108; Weinrich, 1976: 230]: Der Sprachgebrauch, die spezifischen Steigerungsformen und die parallele Erscheinung im Bereich: Verb - Substantiv (z.B.: *genehmigen* - *Genehmigung*) machen deutlich, daß es sich um zwei Wörter handelt. Das Adjektiv wird gesteigert: das *schöne, schönere, schönste* Buch, das Adverb dagegen: *schön, schöner, am schönsten* [Details, s. Rolland, 1994: 95ff.].

Die hier skizzierte Andersartigkeit in zentralen Bereichen der Sprache im Vergleich zum herrschenden bisherigen Verständnis hat, wie deutlich sein dürfte, entscheidende Folgen. Durch die Erkenntnis, was ein Wort ist und welche Differenzierungen notwendig sind, ist die Möglichkeit gegeben, die Sprache so abzubilden, wie sie wirklich ist, d.h. wie sie auch der Mensch gebraucht, wenn er spricht. Daher ist alles, was auf diesen Ergebnissen aufbaut, anders, als es bisher gehandhabt wurde. Von besonderer Bedeutung ist überdies,

daß nach einer detaillierten Homonymenklärung im Gegensatz zur bisher vielfach geübten Praxis [Beispiel bei: Wahlster, 1982: 245] genau gesagt werden kann, welche Wörter mit welchen Ausgangswörtern in welchen Beziehungen semantisch sinnvoll verknüpft werden können [vgl.u.5].

3 Wortarten und Satzglieder

Ein Wort liegt immer als wortartlich geprägtes Wort vor. Die Differenzierung der Wortarten im Deutschen (der Terminus *Wortart* wird beibehalten und statt *Wortklassen* verwendet [vgl. Rolland, 1994: 54]) ist von der griechischen bzw. lateinischen Grammatik übernommen [z.B.: Weisgerber, 1962: 298]. Inzwischen ist die Literatur zu diesem Problem uferlos [vgl. Kaltz, 1983] und die Meinungen über Art und Anzahl der Wortarten sind sehr geteilt. Maximal werden 51 Wortarten angenommen [Bergenholtz-Schaeder, 1977: 73-75], mindestens aber werden drei unterschieden: Nomen, Verb, Partikel [Hakett = Kaltz, 1983: 43]. Die Duden-Grammatik [1984: §114] nennt 9 Wortarten: *Verben, Substantive, Adjektive, Pronomen, Artikel* und die *Partikel: Adverbien, Präpositionen, Konjunktionen* sowie die *Interjektionen*. Bünting [1987: 105] setzt 10 an, indem er zusätzlich noch das *Numerale* angibt. Im Gegensatz zu all diesen Annahmen hat die detaillierte und nachvollziehbare Sprachuntersuchung gezeigt [zur Begründung s. Rolland, 1994: 53ff.], daß es sechs, sich als Oppositionspaare gliedernde, in sich spezifizierte Wortarten gibt:

Verben -	Substantive (einschließlich Pronomina und Artikel)
Adjektive -	Adverbien (einschließlich der Interjektionen)
Präpositionen -	Konjunktionen (spezielle: weil, damit, obwohl, ...) (generelle: und, oder, als(Rolle), sowie, ...)

Abb. 2 Wortarten

Zu den Wortarten in eindeutiger Wechselbeziehung stehen die Satzglieder. Neu ist u.a. [anders: Duden, 1984: §1063], daß der Terminus: *Attribut* ausschließlich für das Adjektiv als Satzglied verwendet wird. Von einem Substantiv abhängige Satzglieder werden entsprechend als Objekt (z.B.: die Größe *des Computers*) oder als Umstandsbestimmung (z.B. das Buch *gestern*) bezeichnet.

Konjunktionen verknüpfen gleichartige Satzglieder. So ergibt eine Konjunktion zusammen mit einem Ursprungssatzglied, also zusammen mit Prädikat, Subjekt, Objekt, Umstandsbestimmung, Attribut ebenfalls ein Satzglied, z.B.: ..., *weil* die Sonne *schien* = Umstandssatz; ..., *daß* er *gekommen ist* = Subjektsatz; er kaufte Brot *und* Butter = Objekt-"Konjunkionalbestimmung" - wie dies terminologisch gefaßt ist [vgl. Rolland, 1994: 113].

Erst das Durchschauen der wirklichen Wortarten im Deutschen und der entsprechenden Satzglieder macht einen adäquaten Aufweis der Sprachstruktur möglich.

4 Flexionsbezug

Was die Flexion betrifft, so hat man die Wörter in flektierbare (Verb, Substantiv, Artikel, Pronomen, Adjektiv) und nicht flektierbare (Adverb, Präposition, Konjunktion, Interjektion) eingeteilt [z.B.: Duden, 1984: §114]. *Flexion* heißt: *Abwandlung* und bezieht sich auf die Formunterschiede des einzelnen Wortes. Diese Unterschiede müssen aber nun nicht am Wort direkt greifbar sein, sondern können auch durch den Bezug des Wortes zu anderen Wörtern, also durch seinen *Flexionsbezug*, deutlich werden. Jedes Wort hat solch eine Flexion bzw. einen Flexionsbezug, durch den die *Verknüpfung zum Satzglied* erfolgt. Das kann geschehen:

- ▷ durch mehrere Wörter: (z.B.: *hat beschafft*: Prädikat; *der Computer*: Subjekt, *den Computer*: Akkusativobjekt)
- ▷ durch Endung bzw. Ablaut, also wortimmanent: (z.B.: *stell-te*; *sang*: dem schöne-n Buch: Attribut Prädikat; (Adjektiv = schöne))
- ▷ durch den Flexionsbezug:
 - bei Präpositionen: (z.B.: *auf dem Tisch*: Umstandsbestimmung)
 - bei Konjunktionen (durch den Bezug zu den Formen von Tempora, Kasus, Komparation, da in der Kombination von Konjunktion und Ergänzung ein neues Satzglied entsteht, z.B.: Kasus: *Er als Experte* = Subjektkonjunkionalbestimmung)

Danach sind als in Wechselbeziehung zueinanderstehende Satzglieder und Wortarten zu unterscheiden [Zur Kennzeichnung der spezifizierten Wortarten sind z.T. neue Termini eingeführt. Es werden daher in Klammern Beispiele angegeben. Wiederum ist eine Homonymenklärung erforderlich, unter anderem bei: *wobei?*, *woneben?* *wozu?* usw. als Modaladverbien im Vergleich zu den Zustandsadverbien: *wobei*, *woneben*, *wozu* usw. Zur detaillierten Erläuterung s. Rolland, 1994: 101ff.]:

Satzglieder	Wortarten
1. Prädikat	Verb
2. Subjekt - Subjektsatz	Substantiv, Adverb Modalssubstantiv (interrogativ: wer?), Modaladverb (wann?), Zustandskonjunktion (daß-absolut)
3. Objekt - Objektsatz	Substantiv, Adverb, Präposition + Substantiv Verlaufsadverb (wobei), Modalsubstantiv (interrogativ: was?), Modaladverb (wann?), Zustandskonjunktion (daß-absolut)
4. Umstandsbestimmung - Umstandssatz	Substantiv, Adverb, Präposition + Substantiv/ Adverb Basiskonjunktion (weil) + Prädikat, Zustandsadverb (woneben)
5. Attribut - Attributsatz	Adjektiv Modalsubstantiv (relativ, attributiv: der)
6. Konjunkionalbestimmung - komplex	generelle Konjunktion (und) + Prädikat/ Subjekt/ Objekt/ Umstandsbestimmung/ Attribut generelle Konjunktion (und) + Subjektsatz/ Objektsatz/ Umstandssatz/ Attributsatz

Abb. 3 Wechselbeziehung zwischen Satzgliedern und Wortarten

Während natürlich bestimmte grundlegende Angaben innerhalb dieses Schemas Allgemeingut sind (z.B.: Prädikat – Verb, Substantiv: Subjekt, Objekt, Umstandsbestimmung und andere), sind zahlreiche Spezifikationen neu. Aber erst durch die Gesamtdarstellung erhält man ein Abbild der wirklichen in der Sprache existierenden Strukturkategorien.

5 Relationen

Bei den Beziehungen oder Relationen zwischen den Wörtern werden die vielfältigsten Arten unterschieden. Unter dem Aspekt der Erkenntnisse der Logotechnik wird eingegangen auf: die paradigmatischen Relationen, die Sinnrelationen, die syntagmatischen Relationen mit einer spezifischen Auseinandersetzung mit: Valenztheorie, semantischen Merkmalen,

Kasustheorie, semantischen Rollen sowie auch Hinweisen zu Parsingverfahren, auch dem Word Expert Parsing.

Paradigmatische Relationen

Im Rahmen der Wortfeldtheorie wird angenommen, daß sich die Wörter eines Wortfeldes wechselseitig bedingen und damit *paradigmatisch* abgrenzen (z.B.: *erlauben, genehmigen, stattgeben, gestatten, zustimmen* usw.). Daß das nicht haltbar ist und daß man die Wortfeldtheorie nicht als Ergebnis, sondern als Methode verwenden sollte, ist in der Literatur hinreichend begründet [u. a. : Kandler, 1971; vgl. Rolland, 1994: 155].

Sinnrelationen

Darüber hinaus nimmt man an, daß zwischen den Wörtern die verschiedensten *Sinnrelationen* bestehen [z.B.: Lyons, 1980: 281ff.; Grewendorf et al., 1987: 299ff.], z.B.: *Hyponymie* (Laubbaum), *Hyperonymie* (Baum), *Synonymie* als "bedeutungsgleiche" Wörter, die es aber außer bei Fremdwörtern (Fernsprecher, Telefon) nicht gibt [vgl. Rolland, 1994: 51f.], *Inkompatibilität* ("nüchtern - sternhagelvoll"), *Komplementarität* ("ledig - verheiratet") sowie die unterschiedlichsten "Gegenwörter" [Agricola et al., 1992], z.B.: *Antonyme* ("laut - leise"), *skalare Antonyme* ("arm - bettelarm"), z.T. *Gegenwörter mit Alternativen* ("befehlen - gehorchen/ sich weigern"; "fragen - antworten/ schweigen"), *Kontrast im allgemeinen* ("Sonne - Mond; Sonne - Erde; Sonne - Schatten; Sonne - Regen" ...) usw.

Es steht jedem frei, je nach Bedarf und Zweck die verschiedensten Beziehungen zwischen Wörtern anzusetzen, nur sollte man sich dabei bewußt sein, daß es sich hierbei meist um von außen an die Sprache herangetragene Beziehungen handelt; denn das *einheitliche* Kriterium, nach dem sich alle Wörter aller Wortarten in der Sprache gliedern, ist das der *inhaltlichen Ähnlichkeit in Abstufungen*, von gerade noch ähnlich in die sogenannten *Sinnklassen*, in 4 weiteren Stufen in die *Großklassen*, die *Kernklassen*, die *Kleinklassen* und die *Einzelwörter*, die sich nur noch in Nuancen unterscheiden. Sinnklassen, Großklassen und Kernklassen lassen sich als Konträrpositionspaare zusammenstellen, so daß man z.B. erhält:

Sinnklassen: "Sprachverben" - "Verstandesverben";

Großklassen: "Verben der Wissensäußerung" - "Verben der Meinungsäußerung";

Kernklassen: "loben" - "tadeln".

Daß dies die sprachimmanente Ordnung des Wortschatzes darstellt, ist am Beispiel der über 1600 Verben des Sprechens und Sagens an Hand nachvollziehbarer linguistischer Kriterien detailliert beschrieben und bewiesen [Rolland, 1969]. So ist auch im o. a. Beispiel zu *befehlen* der Gegensatz nicht: *gehorschen* bzw. sich *weigern*, sondern es gilt:

befehlen - *bitten*; sich *weigern* - sich *bereit erklären*; *gehorschen* gehört nicht hierhin, da es ein "Handlungsverb", kein "Sprachverb" ist. Zu *fragen* ist der Gegensatz nicht: antworten/ schweigen, sondern: *fragen* - *abhören*; *antworten* - *schweigen* ist korrekt.

Syntagmatische Relationen

Schließlich unterscheidet man noch die *syntagmatischen* Relationen, die zwischen den Wörtern bestehen, "die mit einem gegebenen Wort sinnvoll verknüpft werden können" [Porzig, 1957: 125]. Es handelt sich also um genau das, was Tesniere [1959; vgl. Helbig et al., 1978; Helbig, 1992; Somers, 1987] unter dem Begriff: *Valenz* (Wertigkeit) gefaßt hat, wobei er allerdings die *Aktanten*, die Mitspieler im engeren Sinne, also Subjekt und Objekte, von den *Adjunkten*, den freien Angaben, den Umstandsbestimmungen, trennt und außerdem noch *obligatorische* und *fakultative Aktanten* unterscheidet, z.B.: "Er wohnt

bei seinen Eltern". Hier ist *bei seinen Eltern* obligatorisch, da: er *wohnt* kein vollständiger Satz ist. Im Satz: "Er versteckte sich bei seinen Eltern" ist *bei seinen Eltern* fakultativ, da die Aussage: *Er versteckte sich* einen vollständigen Satz darstellt. Daß das gleiche Satzglied einmal obligatorisch, einmal fakultativ sein kann, bezeichnet auch Pohlenz [1988: 131] als Problem innerhalb der Valenztheorie. Die Lösung ist die: Wenn man nicht das Satzglied (*bei seinen Eltern*) in den Mittelpunkt stellt, sondern vom Ausgangswort, den unterschiedlichen Verben, ausgeht, dann ist es so, daß *wohnen* neben dem Subjekt aus einer Menge *variierbarer Satzglieder* eines erfordert, welches, das hängt von dem ab, was man sagen möchte, z.B.: "Er wohnt *in Köln*. Er wohnt *komfortabel*. Er wohnt *bei seinen Eltern*." usw. Bei *obligatorisch* kommt es darauf an, ob das Ausgangsverb neben dem Subjekt *notwendig dasselbe* bzw. *notwendig dieselben Satzglieder* erfordert, z.B.: *den Freund* belügen, *dem Verräter das Bürgerrecht* aberkennen. Der in der Sprache existierende Gegensatz ist nicht auf die Aktanten beschränkt (z.B.: Er fährt *gut/ gern/ nach Köln* usw.) und lautet nicht: obligatorisch und fakultativ, sondern gilt generell und heißt: *notwendig dasselbe/ notwendig dieselben Satzglieder - variierbare Satzglieder*, wobei im Grenzfall das Verb absolut steht (z.B.: er *hat sich verabschiedet*) [vgl. Rolland, 1969]. Innerhalb der Valenztheorie nimmt man Gruppen von Verben an [Tesniere, 1959: 238], die bezeichnet werden als: avalent (ohne Subjekt: es *regnet*), monovalent (mit Subjekt: er *kommt*), divalent (mit Subjekt und einem Objekt: er *sagte nichts*), trivalent (mit Subjekt und zwei Objekten: er *gab ihm das Buch*). Da nunmehr erkannt ist [vgl. o. 2], daß es ein eigenes Sachsubjekt es gibt, wird deutlich, daß es auch avalente Verben nicht geben kann, da dieses es ein eigenständiges Sachsubjekt darstellt.

Im Valenzrahmen werden also *Satzglieder* angesetzt. Bei dieser Vorgehensweise wird der entscheidende semantische Unterschied nivelliert; denn zunächst kommt es auf die Relation zwischen Ausgangswort und dem abhängigen Wort an, erst dann auf die Ausprägung als Satzglied, z.B.: Wenn es heißt: *die Rede des Direktors, das Buch des Direktors*, dann ist *des Direktors* in beiden Fällen Genitivobjekt, wobei das gleiche Wort in der gleichen Flexionsform den gleichen Inhalt hat, nämlich: *des Direktors*, wie auch immer der Zusammenhang lautet. Den entscheidenden Unterschied erkennt man erst an der *Relation*, die sich in *spezifizierten Fragewörtern* ausprägen muß, also: *die Rede wessen, die wer hält?*; *das Buch wessen, das wem gehört?* Es handelt sich bei der *Relation*, da diese vom jeweiligen Ausgangswort (*Rede, Buch*) ausgeht und in seinem Inhalt beschlossen ist, um eine *Inhaltskomponente des Ausgangswortes* und nicht - wie bisher üblicherweise angenommen [z.B.: Duden, 1984: §1075] - um die Funktion des Genitivs, gefaßt als Genitivus subjektivus und Genitivus possessivus. Damit wird auch deutlich, daß es ein inadäquater Ansatz ist, wenn man die *abhängigen* Wörter, um sie semantisch zu kennzeichnen, auf einer *Metaebene* entweder mit *semantischen Merkmalen* (z.B.: Mann = Lebewesen, menschlich, erwachsen, männlich) oder mit den *semantischen Rollen* der Kasustheorie [Fillmore, 1968; vgl. Rolland, 1994: 23ff.] charakterisiert (z.B.: *Agent, Objekt, Source, Goal, Time, Location*), da es sich nicht um eine Rolle o. dgl. des abhängigen Wortes handelt, sondern um eine *Relation* im Inhalt des Ausgangswortes. Zur Ermittlung der semantischen Merkmale gibt es bekanntermaßen keine Kriterien, die Anzahl der semantischen Rollen variiert innerhalb der verschiedenen Theorien erheblich, sie sind z. T. für das gleiche Wort austauschbar, und man fragt sich zu Recht, welchen Aussagewert die Kennzeichnung mit semantischen Rollen hat [vgl. Seyfert, 1981: 155]. Überdies dürfte ja um es nochmals zu betonen - aus dem o. a. Beispiel (*Rede/ Buch des Direktors*) deutlich geworden sein, daß es nicht auf die Markierung des *abhängigen* Wortes ankommt, sondern daß es sich bei der Beziehung um eine *Relation* handelt, die im Inhalt des *Ausgangswortes* verankert ist.

Wenn es im Hinblick auf die Valenz u. a. heißt [Hörmann in Pohlenz, 1988: 130]: "Wer 'schneiden' hört, stellt Patient, Agent, Instrument bereit, welche die in der Umgebung dieses Verbs agierenden Nomina spielen müssen ...", so trifft das "müssen" nicht die

Sprachwirklichkeit; denn auch ohne Instrument gibt es korrekte Sätze, z.B.: "Die Mutter schneidet den Kuchen." "Der Gärtner hat gestern den Rasen geschnitten." "Der Metzger schneidet die Wurst dünn." usw.

Pohlentz [a.a.O.] hält es im Hinblick auf die Angabe der Abhängigkeitsbeziehungen für problematisch, "entsprechende grammatikalische Regeln für die Ergänzungen von Prädikatausdrücken zu formulieren". Die Abhängigkeitsbeziehungen sind, wie soeben betont [s. a. Rolland, 1994: 21 Off.], in spezifizierten Fragewörtern aufweisbar. Die Fragemöglichkeiten sind begrenzt. Im Einzelfall ist zu testen, welche dieser Möglichkeiten bei welchem Ausgangswort zutreffen - Ausgangswort kann jedes Wort jeder Wortart sein. Man geht also nicht von Texten aus, die trotz eines großen Umfangs keine Vollständigkeit garantieren, sondern setzt an der Ebene der langue an. In der Behauptung [Pohlentz, 1988: 131], die "freien... Angaben" seien von der Wahl des Ausgangswortes "völlig unabhängig" und es handele sich um eine "offene Liste von Bezugsstellen-Rollen" [159], wird verkannt, daß der Inhalt des Ausgangswortes alle Abhängigkeiten bestimmt, z.B.: möglich: "Er beschaffte ihm das Buch *innerhalb von 2 Wochen*" (*wann, innerhalb welchen Zeitraums?*); nicht möglich: *"Die Sitzung dauerte *innerhalb 2 Wochen*". Der Inhalt von *beschaffen* bzw. von *dauern* bestimmt die abhängigen Relationen, die ihrerseits die Konkretisierungen implizieren. Es handelt sich hierbei nicht um eine "offene Liste" von Rollen als Kennzeichnungen der abhängigen Wörter, sondern jeweils um eine bestimmte Anzahl möglicher Relationen, also um Inhaltskomponenten des Ausgangswortes, die auf Grund des Inhalts des Ausgangswortes *ausgangswortspezifisch* sind.

Der für eine Verbgruppe geltende Valenzrahmen wird üblicherweise auch als *Bauplan* dieser Verben bezeichnet und meint die jeweilige kleinste Einheit von Aktanten, die für viele Verben gelten [vgl. Duden, 1984: §1081ff.; Schumacher: 1986]. Auf Grund der Tatsache, daß jedes Wort kraft seines Inhalts seine potentielle Wortumgebung selbst bestimmt, hat jedes Wort seinen eigenen Bauplan. *Bauplan* meint also, der Intention der "inhaltbezogenen Grammatik" entsprechend [Weisgerber, 1962], die größtmögliche Einheit pro Ausgangswort, also die Relationen einschließlich ihrer Konkretisierungen. Jede Äußerung, jeder Satz ist demnach ein Extrakt aus den in einem solchen Bauplan angelegten Möglichkeiten. Faßt man alle Wörter pro Wortart, die in ihren Relationsbezeichnungen gleichartig sind, zusammen, so erhält man *Bauplantypen*. Sie bestehen also, um es nochmals zu betonen, nicht aus Satzglied-Leerstellen mit nachträglich draufgepfropften semantischen Rollen bzw. einer offenen Liste für die Adjunkte, sondern aus einer vom Inhalt des Ausgangswortes bestimmten festen Anzahl von Relationen, die, den sprachimmanenten Gesetzmäßigkeiten gemäß, ausschließlich in spezifizierten Fragewörtern als Pendants zur Feststellung greifbar sind und ihre Konkretisierungen implizieren. Diese Konkretisierungen können sich, da die Sprache ein lebendiges Gebilde ist, im Laufe der Zeit ändern - es können vorhandene Wörter vergehen bzw. neue Wörter hinzukommen -, aber beim Ausgangswort muß die Struktur der abhängigen Wörter erhalten bleiben; denn diese Struktur bedingt die Geltung des Ausgangswortes [vgl. Rolland, 1969].

Parsing, Word Expert Parsing

In den vielen Grammatiktheorien, auf denen *Parser* operieren [Hellwig, 1989], hat man "künstliche" Grammatiken geschaffen, künstlich deshalb, weil sie nur für bestimmte kleine Anwendungsbereiche gelten, aber bei jeder Erweiterung des Anwendungsbereichs ebenfalls erweitert werden müssen. Daher wird von Hellwig die Empfehlung ausgesprochen, Grammatik und Parser zu trennen, damit nicht jede Änderung der Grammatik, die ja unweigerlich notwendig erscheint, auch eine Änderung des Parsers bedingt.

Durch die detaillierte Analyse der Sprachstruktur im Vorfeld jeglicher Verarbeitung, wovon im vorausgehenden zentrale, sich von bisherigen Ansätzen unterscheidende Komponenten vorgestellt worden sind, ist "es möglich, die Grammatik des Deutschen generell

aufzuweisen [zur Gesamtkonzeption s. Rolland, 1994]. Damit sind nachträgliche Änderungen grundsätzlicher Art nicht notwendig - lediglich neues Wortgut ist einzuarbeiten bzw. veraltetes zu eliminieren.

Im Gegensatz zu den genannten und üblicheren Parsingverfahren steht beim *Word Expert Parsing* (WEP) [Small, 1981; vgl.: Reddig, 1984; Eimermacher, 1984] das Wort im Mittelpunkt der Betrachtung. Das Wort hat 2 Funktionen: Es ist Träger der lexikalischen Information und zugleich Initiator im Analyseprozeß. Was die Semantikorientiertheit des WEP betrifft, so erübrigt es sich, näher darauf einzugehen, da diese doch wieder nur auf der Kennzeichnung der Wörter mit semantischen Merkmalen und semantischen Rollen beruht, deren Tragfähigkeit im vorausgehenden schon als unzureichend charakterisiert ist. Außerdem wird bei dem logotechnischen Ansatz zwar vom Wort ausgegangen, aber dann vom Inhalt des Wortes, konkret: von seinen Relationen, wobei bei jeder Relation die zugehörigen Wörter in ihren zugehörigen Formen angegeben sind. Die Satzidentifikation des eingelesenen Satzes besteht also im Kern aus dem Vergleich, welche der beim Ausgangswort im jeweiligen Bauplan *vorgegebenen möglichen* Wörter im konkreten Satz *tatsächlich* ausgewählt sind. Trotz dieser entscheidenden Unterschiede wird noch auf einige Einzelheiten hingewiesen [vgl. Mertens et al., 1995].

Beim WEP geht von den einzelnen Wörtern eine Kontexterwartung aus, die keineswegs den sprachlich geltenden Gesetzmäßigkeiten entsprechen muß. Beim Links-Rechts-Parsing kann z.B. der *Artikel* ein *Nomen* (Subjekt), dieses ein *Prädikat* erwarten, die Beziehung ist also: *Artikel* --+ *Nomen* (Subjekt) --+ *Prädikat*. In der Sprache ist die Beziehung, und so ist sie im Bauplan auch fest gehalten, genau umgekehrt: *Prädikat* --+ *Nomen* (Subjekt) --+ *Artikel*.

Kongruenzüberprüfungen für Genus, Kasus, Numerus bei *Artikel*, *Adjektiv*, *Nomen* können entfallen, da beim *Nomen* als Ausgangswort des entsprechenden Bauplans in der zugehörigen Relation die semantisch und flexionsmäßig korrekten Wörter und Formen schon vorgegeben sind.

Auch das Rechtsassoziationsprinzip als Default-Regel kann entfallen, da die Relationen im Satz von ihrem jeweiligen Ausgangswort bedingt werden und durch Festhalten der Wortstellung die Zusammengehörigkeit deutlich wird, z.B.:

a) Er kaufte mir gestern ein Buch.

Gestern kaufte er mir ein Buch. Ein

Buch kaufte er mir gestern.

In allen Fällen gilt vom Bauplan: *kaufen*:

wer; wem; wann, während welchen Zeitraums; was?

b) Gestern das Buch hat mir gefallen.

Es gilt vom Bauplan: *gefallen: was; wem?*

Es gilt vom Bauplan: *Buch: das Buch wann, während welchen Zeitraums'?*

Die Wortstellung von *gestern* im Verhältnis zum *Verb* wäre anders, wenn *gestern* von *gefallen* abhängig wäre, nämlich:

Gestern hat mir das Buch gefallen (heute nicht mehr).

Das Buch hat mir *gestern* gefallen (heute nicht mehr).

c) Die Disambiguierung erfolgt - wie betont - im Prinzip nicht erst im konkreten Satz, da im Bauplan pro Relation die korrekten Konkretisierungen vorgegeben sind. Es gibt jedoch Einzelfälle, bei denen in der gleichen Relation Homonyme stehen, z.B.: *Er kaufte das Schloß (Gebäude)*. - *Er kaufte das Schloß (Türschloß)*.

Wenn der Satz so lautet und kein Zusammenhang existiert, weiß auch der Mensch nicht, was gemeint ist. Wenn aber nun im gleichen Satz oder im vorausgehenden Satz Angaben vorhanden sind wie:

- *Der Manager, der Baron* o. ä. kaufte das Schloß oder:
- Er kaufte das Schloß *zum Preis von X Millionen* oder:
- *Der Baron* hatte sich *das Anwesen* angesehen. Er kaufte das Schloß.

dann, so wird per Regel festgelegt, gilt: *Schloß (Gebäude)*. Falls keins der angegebenen Wörter vorliegt, gilt die Regel: gib eine Fehlmeldung. Dann wird nachträglich entschieden, welches Homonym zutrifft. Auf diese Weise ist aber sichergestellt, daß der Satz auf jeden Fall korrekt identifiziert ist und damit einer entsprechenden Abfrage zugänglich gemacht werden kann.

Im übrigen ist die grundlegende Vorgehensweise beim WEP und bei dem logotechnischen Verfahren eine andere. Beim WEP hat jedes Wort zugleich eine Parsingfunktion. Bei der Logotechnik werden im Vorfeld jeder Verarbeitung die möglichen Relationen einschließlich ihrer Konkretisierungen aufgewiesen. Damit ist das sprachliche Relationsgefüge dargestellt. Damit dieses für den Computer formal greifbar wird, werden die Wörter mit Kennungen versehen, die als ganz wesentliche Komponente die sprachlichen Relationen wiedergeben.

Wenn nun ein Satz eingelesen wird, wird auf schnellstem Wege das Prädikat als Ausgangspunkt des Satzes gesucht. Dann wird der auf diesem Verb basierende *Satzbauplan* abgearbeitet. Dabei macht man sich - wie oben erwähnt - das in der Sprache geltende Prinzip zunutze, daß jedem Feststellungssatztyp (Aussage, Befehl, Ausruf, Wunsch) als Pendant ein bestimmter Fragesatztyp entspricht [Rolland, 1994: 290ff.]. Das Verb bleibt konkret, z.B. (noch zu spezifizieren):

<i>Der Vater</i>	hat	seinem Sohn	<i>gestern</i>	<i>in Köln</i>	<i>ein Buch</i>	gekauft.
<i>Wer</i> hat		<i>wem</i>	<i>wann</i>	<i>wo</i>	<i>was</i>	gekauft?

Falls von einem vom Verb bedingten Wort weitere Angaben abhängig sind - zu beachten ist die Wortstellung, z.B. "das *teure* Buch" -, so geht man in den Bauplan von Buch, den *Syntagmenbauplan*. Da bestimmte Konjunktionen, wie: *und, oder, sowie* usw. den vorhandenen Bauplan erweitern können, wird beim Vorliegen einer solchen Konjunktion (was beim ersten Durchgang durch den Satz erkannt wird, [vgl. Rolland, 1994]) der *erweiterte Bauplan* abgearbeitet. Das geschieht in der Weise, daß zunächst wieder das Verb des erweiterten Bauplans aufgesucht wird, danach der Bauplan dieses Verbs abgearbeitet wird usw.

Die Gesamtheit aller Baupläne ergibt das *potentielle Relationsgefüge* der Sprache. Jeder Satz ist ein Extrakt aus diesen Vorgaben. Das potentielle Relationsgefüge stellt die *semantikorientierte Grammatik* des Deutschen dar. Die an der deutschen Sprache erarbeiteten Prinzipien gelten generell, d.h. auch für andere Sprachen. Einzelheiten dazu s. [Rolland, 1994].

6 Wortinhalt

Die entscheidende Einsicht, die die vorliegenden Ergebnisse ermöglicht hat, ist die von der Ganzheit des Wortes und die damit mögliche Erkenntnis vom Aufbau des Wortinhalts - fern jeder sogenannten Referenzsemantik bzw. Inhaltssemantik [z.B.: Bünting, 1987: 199ff. vgl. Rolland, 1994: 121f.]. Es wird rein sprachimmanent und damit semantikorientiert verfahren, indem man die sprachlichen Pendants feststellt: Der Wortinhalt ist bedingt von dem von ihm abhängigen Wörterkomplex *in seiner Struktur*, wobei die Schwierigkeit darin besteht, diese Struktur zu ermitteln; denn die syntagmatischen Relationen stellen nur einen Inhaltzug im Gesamtinhalt eines Wortes dar. Im einzelnen verhält es sich mit dem Aufbau des Wortinhalts so [vgl. Rolland, 1994]:

Der Wortinhalt besteht aus 2 Teilinhalten, dem *speziellen Inhalt*, der das Eigenständige im Wortinhalt meint und dem *generellen Inhalt*, den das Wort mit anderen Wörtern der gleichen Wortart teilt. Der generelle Inhalt wiederum besteht aus zwei generellen (Inhalt-) Zügen:

- ▷ der 1. *generelle Zug* meint den Flexionsbezug und bedingt die Verknüpfung zum Satzglied;
- ▷ der 2. *generelle Zug* meint den Konstruktionsbezug, also die Relationen, und bedingt die *Verknüpfung zwischen Satzgliedern*.

Gesamtinhalt		
Teilinhalte		
spezieller Inhalt	genereller Inhalt	
	1. genereller Zug (Flexionsbezug)	2. genereller Zug Konstruktionsbezug)

Abb. 4 Die Struktur des Wortinhalts

Das Ausgangswort bedingt das abhängige Wort. Nun ist es so, daß bei der Verknüpfung der Satzglieder die Relation vom 2. generellen Zug des Ausgangswortes zum 1. generellen Zug des abhängigen Wortes führt, wobei als Voraussetzung die speziellen Inhalte der betroffenen Wörter ohnehin zueinander passen müssen. Bei der Verknüpfung: Prädikat — Subjekt fallen Flexionsbeziehung und Konstruktionsbeziehung zusammen, so daß die Relation vom 1. generellen Zug des Prädikats zum 1. generellen Zug des Subjekts führt. Flexionsbeziehungen erfolgen zwischen den 1. generellen Zügen (z.B. Lehrer → der) [Details s. Rolland, 1994]. Die in der folgenden Abbildung dargestellten Kästchen symbolisieren Wörter. Das linke Kästchen meint den speziellen Inhalt, das mittlere den 1. generellen Zug, das rechte den 2. generellen Zug. Das Beispiel verdeutlicht die Beziehung zwischen Prädikat, Subjekt und einem Akkusativobjekt:

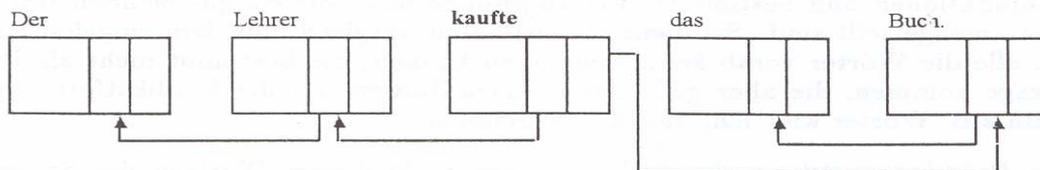


Abb. 5 Verknüpfungen im Satz

7 Text

So wie das Wort den Satzaufbau bestimmt, so ergibt eine Abfolge der Sätze, die einen Sinnzusammenhang darstellen, einen Text. Wichtig ist also die Reihenfolge der Sätze. Je nach der Wortwahl ist es dann oftmals so, daß man statt des spezifischen Substantivs ein allgemeineres Wort, ein Pronomen, verwenden kann, z.B.: „Der Lehrer kaufte das Buch. Es gefiel ihm.“ Der Computer kann nicht „verstehen“, daß mit es „das Buch“ und mit ihm „der Lehrer“ gemeint ist. Aber bei *gefallen was(Nominativ)?* sind u.a. Wörter wie: *Buch, Bild, Schuh, es* usw. möglich, bei *gefallen wem?* Wörter wie: *dem Lehrer, Schüler, Präsidenten, ihm* usw. Auf diese Weise kann der Satz zunächst einmal identifiziert werden.

Außerdem aber kann man per Regel, wenn man will, im voraufgehenden bzw. im nachfolgenden Satz suchen (z.B.: *Es hatte dem Jungen gefallen. Das Buch war spannend.*),

welches Wort, das bei *gefallen* in dieser bestimmten Relation möglich ist, auch beim Verb des anderen Satzes im Bauplan vorgegeben und im Satz tatsächlich aufgeführt ist. Ist dies identifiziert, könnte man es im Satz mit dem Pronomen hinter diesem zur Verdeutlichung in Klammern anfügen. [Weitere Details zum *Text* s. Rolland, 1994].

8 Dialogsystem

Bekanntermaßen steht und fällt ein sprachverarbeitendes System im genannten Anwendungstyp mit der zugrunde liegenden Grammatik. Da nunmehr die semantikorientierte Grammatik, deren Besonderheiten oben skizziert sind [die Gesamtkonzeption ist auf 300 Seiten dargelegt, Rolland, 1994] -da diese Grammatik als Abbild der Sprache dargestellt werden kann, die der Mensch spricht, kann jeder eingelesene Satz entsprechend identifiziert werden. Im folgenden wird der Aufbau eines Dialogsystems beschrieben. Die zentralen Komponenten hierfür sind: der Aufbau der *Relationsbasis*, der auf dieser basierende Aufbau der *Wissensbasis*, die wiederum ihrerseits als Grundlage für die *Wissensabfrage* dient. Im folgenden werden die wichtigsten Gesichtspunkte hierzu umrissen.

9 Aufbau der Relationsbasis

Der entscheidende Schritt im Hinblick auf die Sprachverarbeitung ist der, aus dem potentiellen Relationsgefüge durch Hinzufügen von Kennungen, den sogenannten Kategorieneinträgen, die *Relationsbasis* zu machen, d.h. zunächst jedes Ausgangswort mit den abhängigen Wörtern in der jeweiligen Relation in der zutreffenden Flexionsform zu kennzeichnen, so daß man aus diesen *potentiellen* Relationen, je nach konkretem Satz, die *relevanten* Relationen ermitteln kann. Da die Relationen im Satz zunächst vom Verb als Prädikat bedingt sind, kommt alles darauf an, möglichst schnell das Prädikat festzustellen, um von diesem als dem Träger des Satzbauplans ausgehen zu können. Im einzelnen sind dazu folgende Komponenten erforderlich:

1. Die *Häufigwortliste*, in der als häufige Wörter: Artikel, "Pronomina", Präpositionen, Konjunktionen und bestimmte Adverbgruppen bzw. die entsprechenden Homonyme zusammengestellt sind. Sie dient als Liste zum Vergleich mit dem eingelesenen Satz, um alle die Wörter vorab kennzeichnen zu können, die bestimmt nicht als Prädikat infrage kommen, die aber ggf. schon einen Hinweis auf die Prädikatform zulassen, wenn z.B. Wörter wie: *ich, du* o. ä. vorkommen.
2. Die *Prädikatermittlung*, durch die aus den verbliebenen Wörtern des Satzes (Verben, Substantiven, Adjektiven, bestimmten Adverbien) das Prädikat festzustellen ist. Hierzu sind zunächst die rein formalen Listen der Hilfs-, Kopula- und Modalverben erforderlich, ferner dann Regeln und Listen mit den weiteren Formen.
3. Die *Formgruppen*, in denen die Flexionsausprägung (Flexion bzw. Flexionsbezug) der einzelnen Wörter der übrigen Wortarten festgehalten ist. Es genügt, sie *einmal* aufzuweisen und dann jeweils durch Verweisungen auf sie zurückzugreifen.
4. Die *Wortklassen* (nicht zu verwechseln mit *Wortarten*), die aus den speziellen Inhalten der Wörter pro Wortart zu ermitteln sind. Je nach inhaltlicher Ähnlichkeit unterscheidet man jeweils, wie betont: *Sinnklassen* (z.B. Geräte, Möbel, Lebensmittel, Sprache), *Großklassen* (Wissensäußerung, Meinungsäußerung), *Kernklassen* (Wortfeldebene: Lob - Tadel; Bitte - Befehl; Erlaubnis - Ablehnung) und die *Kleinklassen* mit den Einzelwörtern (genehmigen, die Genehmigung geben, die Genehmigung erteilen).

Inhaltlich ähnliche Wörter stehen auf Grund dieser Tatsache, je nach Struktur des Ausgangswortes, im gleichen Abhängigkeitsverhältnis zu diesem Ausgangswort. Kann man z.B. sagen: *Möbel bestellen*, dann ist auch sprachlich korrekt: *Tische bestellen*, da das Wort: *Tische* auf Grund seiner inhaltlichen Ähnlichkeit zur semantischen Klasse von *Möbel* gehört. Die Wortklassen, ggf. um neues Wortgut zu erweitern, sind *einmal* zu ermitteln, können dann aber generell verwendet werden; d.h. man kann von den verschiedensten Ausgangswörtern auf die relevanten Klassen zugreifen. Will man daher die bei einem Ausgangswort in einer bestimmten Relation möglichen Wörter feststellen, so genügt es, da ganze Wortklassen, und zwar auf verschiedenen Ebenen, hierher gehören können, lediglich ein *Leitwort* als Stellvertreter der jeweiligen Klasse anzuführen.

Die Wortklassen sind mit Wortklassennummern (WKnr) gekennzeichnet. Diese sind entsprechend der Klassengliederung spezifiziert, z.B.: WK1: 5.4.71.3 = Wortklasse 1 (= Substantive), Sinnklasse 5, Großklasse 4, Kernklasse 71, Kleinklasse 3. Die Wörter innerhalb einer Kleinklasse erhalten keine Nummer, da es sich um sehr wenige Wörter, oft nur um ein *Wort* handelt.

Bei jedem Wort der Wortklasse ist die mögliche Relationsangabe in Form eines Fragewortes vermerkt und zugleich ein Hinweis auf die zugehörige Formgruppe gegeben, z.B.:

WKnr: Computer:	(wer oder) was (Nom)?	→ Formgruppe ...
	(wessen oder) welcher Sache (Gen)?	→ Formgruppe...
	(womit oder) mit welcher Sache?(Dat)?	→ Formgruppe Präp: mit &
		→ Formgruppe...
	was (Akk)?	→ Formgruppe... usw.

5. Die *Baupläne*, unterschieden nach *Satzbauplänen*, *Syntagmenbauplänen* und *erweiterten Bauplänen*. Hier wird jedes Ausgangswort mit seinen Relationen und den zugehörigen Konkretisierungen angegeben. Man braucht aber nicht jeweils alle Wörter im einzelnen anzuführen, sondern es genügt — wie im vorausgehenden beschrieben —, das *Leitwort* der jeweiligen Wortklasse anzugeben, z.B.:

erlauben:	wer?	Leitwort1: Wortklassennummer
		Leitwort2: Wortklassennummer usw.
	wem?	Leitwort1: Wortklassennummer
		Leitwort2: Wortklassennummer usw.
	was?	Leitwort1: Wortklassennummer
		Leitwort2: Wortklassennummer usw.

6. Die *Zugriffsliste*, die zur schnelleren Identifikation des im eingelesenen Satz und in dem zugehörigen Bauplan vorkommenden Wortes dient, die die Formen von Substantiven, Adjektiven und bestimmten Adverbgruppen einschließlich der Wortklassennummer enthält, so daß auf diese Weise in kürzester Zeit innerhalb des Bauplans die Relation aufgefunden ist, zu der das betreffende Wort gehört. So sind die Formen von z.B. *Computer* rein formal: *Computer*, *Computers*, *Computern*. Diese 3 Formen erhalten die gleiche Wortklassennummer. Angenommen, eine dieser Formen steht im Ausgangssatz, so kann auf schnellstem Weg mit Hilfe der Zugriffsliste die Wortklassennummer gefunden werden, mit Hilfe der Wortklassennummer die Relation im Bauplan und mit Hilfe der hierarchisch zusammengesetzten Wortklassennummer das zugehörige Wort innerhalb der Leitwörter.

Als Komponenten der Relationsbasis erweisen sich somit:

Häufigwortliste
Prädikatermittlung
Satzbaupläne Syntagmenbaupläne
Erweiterte Baupläne
Zugriffsliste
Wortklassen (pro Wortart)
Formgruppen (pro Wortart)

Abb. 6 Struktur der Relationsbasis

Auf der Basis dieser Ergebnisse erfolgt der Aufbau der Wissensbasis, der nunmehr im einzelnen umrissen wird.

10 Aufbau der Wissensbasis

Im folgenden werden die verschiedenen Arbeitsstufen vom Einlesen eines Satzes über die Verarbeitung bis hin zur Speicherung der Ergebnisse in ihren wesentlichen Punkten dargestellt .

Der Mensch äußert sich einerseits - wie oben (5) angeführt - in *Feststellungen*, wie man Aussage, Befehl, Ausruf, Wunsch nennen könnte, und in *Fragen*. Wenn nun zum Aufbau eines Frage-Antwort-Systems ein Satz eingelesen wird, ist der Computer auf Grund der Vorgaben, d.h. der Fakten (= dem mit Kennungen versehenen potentiellen Relationsgefüge) und zusätzlicher Regeln, die auf den Fakten operieren, in der Lage, den Satz automatisch zu identifizieren. Es handelt sich also um eine echte automatische Wissensakquisition. Diese vollzieht sich in folgenden Schritten:

1. Kennzeichnung der Position der Wörter des Satzes

Damit wird die Reihenfolge der Wörter festgehalten. Das ist sehr wichtig zur Feststellung der Abhängigkeiten. So ist z.B. in dem Satz: "Der Vortrag *heute* war ausgezeichnet", *heute* von *Vortrag* abhängig; im Satz: "Er wird heute kommen", ist *heute* von *kommen* bedingt.

2. *Linearer Vergleich der Wörter mit der Häufigwortliste und Kennzeichnung der Treffer* Die Häufigwortliste enthält: Artikel, Pronomina, Präpositionen, Konjunktionen, bestimmte Gruppen von Adverbien sowie die zugehörigen Homonyme o. ä. Dieser Schritt dient dazu, die Wörter, die mit Sicherheit nicht als Prädikat in Frage kommen, als solche zu kennzeichnen, um aus den verbliebenen, nämlich: Verben, Substantiven, Adjektiven, bestimmten Adverbien, das Prädikat ermitteln zu können; denn da ein Satz in seiner Grundstruktur vom Prädikat bedingt wird, kommt alles darauf an, das Prädikat zu identifizieren.

3. Prädikatermittlung

Die Prädikatermittlung erfolgt in der Weise, daß mit Hilfe von Listen und Regeln zunächst rein formal Hilfs-, Kopula- bzw. Modalverben als mögliche Elemente einer zusammengesetzten Verbform aufgewiesen werden und danach die komplexe Verbform vervollständigt wird bzw. ggf. die einfache Verbform festgestellt wird. Liegt das Prädikat vor, dann ist damit das Ausgangsverb für die weitere Identifikation des Satzes ermittelt.

4. Abarbeitung der Baupläne

Ausgehend vom Verb, wird zunächst der zugehörige Satzbauplan abgearbeitet. Aus der Fülle der bei dem festgestellten Prädikat möglichen Relationen und ihren Konkretisierungen werden diejenigen Wörter in den entsprechenden Relationen festgehalten, die der eingeleseene Satz aufweist. Ggf. sind darüber hinaus die relevanten *Syntagmenbaupläne* (z.B.: Das Buch *des Lehrers* lag auf dem Tisch) bzw. die erweiterten Baupläne zu berücksichtigen (z.B.: Das Buch *und die Tasche* des Lehrers lagen auf dem Tisch). Bei der Abarbeitung der Baupläne greift man mittels der Leitwörter auf die *Wortklassen* sowie auf die *Formgruppen* zurück. Außerdem ist an dieser Stelle die *Zugriffsliste* einzubeziehen, mit deren Hilfe auf Grund der Wortklassennummer ein schneller Zugriff auf das relevante Wort in der betreffenden Relation im Bauplan hergestellt werden kann.

Insofern bei der Identifikation der Wörter und ihrer Relationen die Beziehungsangaben (also z.B.: kaufen (noch zu spezifizieren): *wer, was, wem, wann, wo, wie, warum?* usw.) in Form von Fragewörtern vorliegen, erhält man zusätzlich zum Feststellungssatz einen Fragesatz, der gleichsam seinen Analogsatz darstellt, insofern die Relationen gleich sind, z.B.:

Feststellung: Der Onkel kaufte seinem Neffen gestern das Buch.
Frage: Wer kaufte wem? wann? was?

Man erhält also 2 Sätze mit den gleichen Relationen, einmal konkretisiert durch in diesen Relationen mögliche Wörter, einmal explizit in Form der Relationen.

5. Speicherung der Ergebnisse

Als Verarbeitungsergebnis werden Satz und Analogsatz eingespeichert und stellen damit die erste Komponente der Wissensbasis dar. In einer Übersicht stellt sich der Ablauf der Wissensaufnahme folgendermaßen dar:

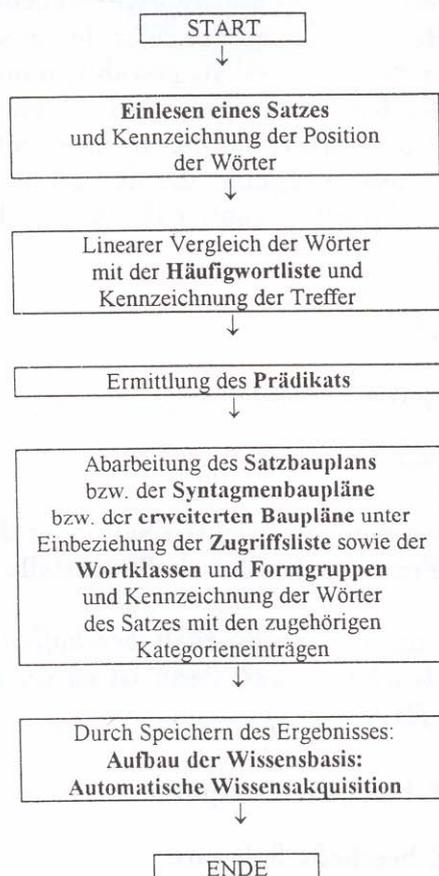


Abb. 7 Wissensaufnahme

Der Aufbau der Wissensbasis erfolgt automatisch, und zwar durch die Speicherung der jeweiligen analysierten Sätze und ihrer zugehörigen *Analogsätze*. In der Wissensbasis werden die Sätze alphabetisch nach Verben und hier nach Verbformen angeordnet, so daß z.B. innerhalb der Sätze mit *beschaffen* alle Sätze zusammengestellt sind, die die Verbform: *hat beschafft* oder *wird beschaffen* aufweisen.

Entsprechend verfährt man mit den Syntagmen, wie sie z.B. häufig in Buchüberschriften vorkommen; d.h. insofern u. a. auch Substantive allein als Ausgangswörter dienen können, werden die mit ihnen gebildeten Syntagmen einschließlich ihrer *Analogsyntagmen* pro Wortart alphabetisch und hier auch wiederum, nach Flexionsformen gegliedert, angeführt, also z.B. alle Syntagmen mit dem Nominativ des Wortes: *Computer*, wie: *der Computer* im 20. Jahrhundert; *der Computer* als Unterstützungsinstrument; *Computer*, Sprache und Denken.

Eine eventuell später notwendige Erweiterung der Wissensbasis erfolgt ebenfalls auf die gleiche beschriebene Art und Weise. Bei einer Eliminierung von Wissen aus der Wissensbasis werden die entsprechenden Sätze bzw. Syntagmen einschließlich der Analogangaben gelöscht.

11 Wissensabfrage

Die Wissensabfrage erfolgt auf der Grundlage der Wissensbasis. Man braucht also nicht mehr in das mit Kennungen versehene potentielle Relationsgefüge, in die Relationsbasis, einzusteigen.

Der Benutzer hat einen Sachverhalt im Kopf, zu dem er Wissen aus dem System erfragen möchte. Damit der Benutzer nun nur *richtig* formulierte Fragen stellen kann, d.h. diejenigen Formulierungen verwendet, in denen das Wissen in der Wissensbasis enthalten ist, werden ihm Wörter (Verben bzw. Substantive) angeboten sowie abhängige Wörter in den jeweils zugehörigen Relationen, aus denen er den gewünschten Sachverhalt zusammenstellt. Hat der Benutzer den Sachverhalt gewählt, dann wird in der Wissensbasis in den Sätzen mit dem betreffenden Sachverhalt gesucht, und es erscheinen auf dem Bildschirm alle Sätze, in denen die Wörter, die den Sachverhalt betreffen, konkret sind, und zwar in der jeweiligen Flexionsausprägung, die übrigen Relationen liegen aber noch in Form von Fragewörtern vor. So erhält man z.B. zum Sachverhalt *Computer beschaffen* Fragesätze wie:

1. Wer beschafft Computer?
2. Wer wird ab wann Computer beschaffen?
3. Wer hatte wem wann einen Computer beschafft?

Hieraus wählt der Benutzer den gewünschten Satz oder die gewünschten Sätze aus und erhält z.B. als Antwort auf Frage 2: "Die Beschaffungsstelle Y wird ab 1.1.1995 Computer beschaffen. "

Geht es dem Benutzer nur um den Sachverhalt *beschaffen*, und er wählt aus den angebotenen Fragen aus: *Wer beschafft was?*, dann ist es so, daß er auf die eine Frage alle vorliegenden Antworten erhält, z.B.:

- > Die Beschaffungsstelle X beschafft Computer. t> Die Beschaffungsstelle Z beschafft Software.
- > Die Zentralbibliothek beschafft Fachbücher.

Außerdem kann der Benutzer von jeder Bildschirmseite auf die vorausgehenden zurückspringen und die Frage erweitern, spezifizieren, abändern, neu anfangen. Ferner ist es möglich, jedes Wort der Antwort anzuklicken und zum Ausgangspunkt einer neuen Frage zu machen. Schließlich wird dem Benutzer die Möglichkeit geboten, in einem eigenen Bildschirmfenster, in dem er scrollen kann, den Antwortsatz im Kontext einzusehen. Damit ist der Benutzer optimal informiert.

12 Maschinelle Übersetzung

Der genannte Ansatz ist auch speziell für den Bereich der *maschinellen Übersetzung* von besonderer Tragweite (Copeland et al, 1991; SCS, 1990); denn er bietet gute Chancen für eine korrekte vollautomatische Übersetzung.

Dazu ist es erforderlich, daß man für jede der betroffenen Sprachen gesondert das potentielle Relationsgefüge erstellt und mit Kennungen versieht, wodurch man die jeweilige *Relationsbasis* erhält. Diese kann einerseits in der nationalen Sprache die Grundlage für den Aufbau eines Dialogsystems darstellen. Zugleich dient die gleiche Grundlage, nämlich die Relationsbasis, als nationale Ausgangsbasis für den Aufbau eines Übersetzungsystems.

Der entscheidende Schritt ist der, daß man die Relationsbasen parallel setzt - eine Aufgabe für Spitzenübersetzer und ein gewaltiges Unternehmen, das aber zunächst am Kernwortschatz mehrerer Sprachen durchgeführt werden könnte. Dabei stellt man dann fest, daß der Relation A in der Ausgangssprache gerade nicht die Relation A' in der Zielsprache entspricht, sondern die Relation X. Es ist auch möglich, daß einem Wort in der Ausgangssprache ein Nebensatz in der anderen Sprache entspricht. Ggf. kann es erforderlich sein, so spezifisch zu verfahren, daß man einem Ausgangswort in einer speziellen Relation mit speziellen Konkretisierungen ein eigenes Wort in der anderen Sprache zuordnen muß [vgl. Rolland, 1995]. Diese Parallelisierung ist nur *einmal* durchzuführen und gilt dann als Grundlage für die Übersetzung beliebiger Texte im Bereich des gewählten Wortschatzes. Es gibt eine endliche Anzahl dieser grundsätzlichen Möglichkeiten, aber in jedem Einzelfall muß zunächst die entsprechende Relation einschließlich der zugehörigen Konkretisierungen und Flexionsformen aufgewiesen werden. Dann kann der Computer "übersetzen", indem er, wenn ein Ausgangswort in einer bestimmten Relation und Form vorliegt, gezielt das parallel gesetzte Wort in der zugehörigen Relation und Form auffindet [vgl. Rolland, 1994].

13 Ergebnis

Im vorausgehenden sind zunächst einige zentrale vorhandene Ansätze über die Struktur der Sprache, speziell auch im Hinblick auf die Sprachverarbeitung, unter dem Aspekt des neuen methodischen Ansatzes auf der Grundlage der Logotechnik diskutiert und verglichen worden. Dabei hat sich gezeigt, daß und in welcher Weise das Wort der Baustein aller Zusammenhänge ist. Ferner ist deutlich geworden, daß nur durch eine Besinnung auf die Ganzheit des Wortes und die notwendige Homonymenklärung, auf die Struktur der Wortarten und Satzglieder, auf die flexionsmäßige Ausrichtung der Wörter, auf die Relationen und damit die Bedingungen der Satzkonstruktion, auf den Aufbau des Wortinhalts und den Zugang zu ihm die semantikorientierte Grammatik aufgewiesen werden kann.

Ferner ist auf der Basis dieser Erkenntnisse der Aufbau eines Dialogsystems skizziert. Außerdem sind Hinweise für die Nutzung dieser Ergebnisse im Bereich der maschinellen Übersetzung gegeben.

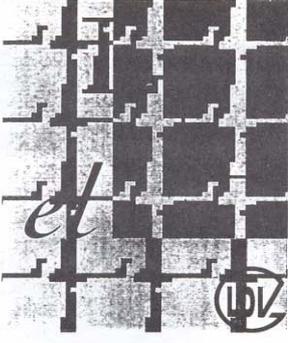
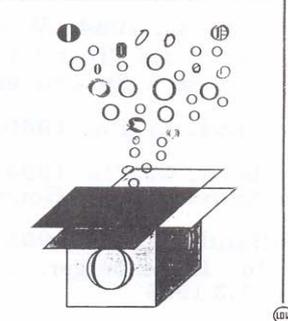
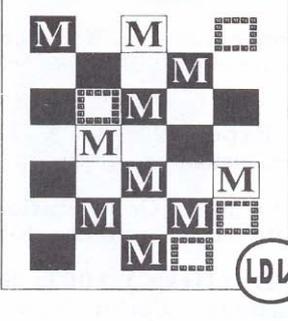
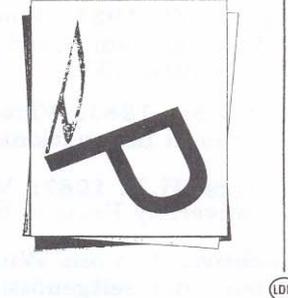
So erweist sich das Durchschauen der Sprachstruktur als die entscheidende Voraussetzung zur Lösung zentraler Fragen im Bereich der Handhabung der natürlichen Sprache

mit dem Computer und ermöglicht in den genannten Bereichen die angestrebte vollautomatische Sprachverarbeitung.

Literatur

- Agricola, Ch.; Agricola, E. ²1992:** Wörter und Gegenwörter. Wörterbuch der sprachlichen Gegensätze: Mannheim, Leipzig, Wien, Zürich: Dudenverlag = Duden Taschenbücher 23
- Bátori, I.S.; Lenders, W.; Putschke, W. 1989 (Hrsg.):** Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin, New York: de Gruyter = Handbücher zur Sprach- und Kommunikationswissenschaft 4
- Bergenholtz, H.; Schaefer, B. 1977:** Die Wortarten des Deutschen. Versuch einer syntaktisch orientierten Klassifikation. Stuttgart: Klett
- Bünting, K.-D. ¹21987:** Einführung in die Linguistik. Frankfurt/ Main: Athenäum = Linguistik, AT 2011
- Copeland, C.; Durand, J.; Krauwer, S.; Maegaard, B. (eds) 1991:** The EUROTRA Linguistic Specifications. Luxembourg: Office for Official Publications of the European Communities, 2 vols = Studies in Machine Translation and Natural Language Processing 1/2
- DUDEN ⁴1984:** Grammatik der deutschen Gegenwartssprache. DUDEN, Band 4. Drosdowski, G. (Hrsg.). Mannheim, Wien, Zürich: Dudenverlag
- Eimermacher, M. 1984:** Word Expert Parsing in PROLOG. In: Rollinger, C.-R. (Hrsg.): Probleme des (Text-) Verstehens. Ansätze der Künstlichen Intelligenz. Tübingen: Niemeyer, S.89-103 = Sprache und Information 10
- Fillmore, Ch. J. 1968:** The case for case. In: Bach, E.; Harms, R.T. (eds): Universals in Linguistic Theory. London, New York, Sydney, Toronto: Holt, Rinehart and Winston, Inc.
- Grewendorf, G.; Hamm, F.; Sternefeld, W. 1987:** Sprachliches Wissen. Eine Einführung in moderne Theorien der grammatischen Beschreibung. Frankfurt/ Main: Suhrkamp = stw 695
- Guenther, F.; Lehmann, H. 1986:** Verarbeitung natürlicher Sprache - ein Überblick. Informatik-Spektrum 9 (1986) 162-173
- Helbig, G. 1992:** Probleme der Valenz- und Kasus-theorie. Tübingen: Niemeyer = Konzepte der Sprach- und Literaturwissenschaft 51
- Helbig, G.; Schenkel, W. ⁴1978:** Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig: Bibliographisches Institut
- Hellwig, P. 1989:** Parsing natürlicher Sprachen. Grundlagen/ Realisierungen. In: Bátori, I.S.; Lenders, W.; Putschke, W. (Hrsg.): Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin, New York: de Gruyter, 348- 377/ 378-432 = Handbücher zur Sprach- und Kommunikationswissenschaft 4
- Humboldt, W. von 1830-35:** Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts = Werke VII, Leitzmann, A. (Hrsg.) Berlin: Behr's 1907 (reprint: Bonn: Dümmler)
- Kaltz, B. 1983:** Zur Wortartenproblematik aus wissenschaftsgeschichtlicher Sicht. Hamburg: Buske. Diss. Mainz = Hamburger Philologische Studien 57
- Kandler, G. 1971:** Dogmatismus und Empirie in der Sprachforschung. In: Schweisthal, K.G. (Hrsg.): Grammatik, Kybernetik, Kommunikation. FS Hoppe. Bonn: Dümmler = Dümmlerbuch 8376
- Lyons, J. 1980:** Semantik. München: Beck. 1. Bd. = Beck'sche Elementarbücher

- Mertens, A.; Schulz, M.; Helbig, H. 1995:** Analyse mit Wortagenten im NLP-System LI-NAS. In: Hitzenberger, L. (Hrsg.): Proceedings der 9. GLDV Jahrestagung, Regensburg, 30.-31.3.1995
- Pohlentz, P. von ²1988:** Deutsche Satzsemantik. Grundbegriffe des Zwischen-den-Zeilen-Lesens. Berlin, New York: de Gruyter = Sammlung Göschen 2226
- Porzig, W. ²1957:** Das Wunder der Sprache. Probleme, Methoden und Ergebnisse der modernen Sprachwissenschaft. Bern: Francke = Sammlung Dalp
- Reddig, C. 1984:** Word Expert Parsing -Ein Überblick. In: Rollinger, C.-R. (Hrsg.): Probleme des (Text-) Verstehens. Ansätze der Künstlichen Intelligenz. Tübingen: Niemeyer, S.77-88 = Sprache und Information 10
- Rolland, M.Th. 1969:** Zur Inhaltbestimmung der Sprachverben. Diss. Bonn
- Rolland, M.Th. 1994:** Sprachverarbeitung durch Logotechnik. Sprachtheorie, Methodik, Anwendungen. Bonn: Dümmler
- Rolland, M.Th. 1995:** Ein semantikorientierter Ansatz im Bereich der Sprachverarbeitung. In: Hitzenberger, L. (Hrsg.): Proceedings der 9. GLDV Jahrestagung, Regensburg, 30.-31.3.1995
- Saussure, F. de ²1967:** Grundfragen der allgemeinen Sprachwissenschaft. Bally, Ch.; Secheyne, A. (Hrsg.). Berlin: de Gruyter
- SCS, 1990:** Scientific Control Systems GmbH: Maschinelle Übersetzung. Grundlagen, Stand, Perspektiven. Hamburg: SCS-Informationstechnik GmbH = SCS-Studie
- Schumacher, H. 1988:** Valenzbibliographie (Stand: Juni 1988). Mannheim: Institut für deutsche Sprache
- Seyfert, G. 1981:** Eine Wiederbelebung der Kasusgrammatik? In: Pleines, J. (Hrsg.): Beiträge zum Stand der Kasustheorie. Tübingen: Narr, 149-159 = Tübinger Beiträge zur Linguistik 133
- Small, St. 1981:** Viewing Word Expert Parsing as Linguistic Theory. Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI 1981)
- Somers, H.L. 1987:** Valency and Case in Computational Linguistics. Edinburgh: Edinburgh University Press = Edinburgh Information Technology Series 3
- Stechow, A. von; Wunderlich, D. (Hrsg.) 1991:** Semantik. Ein internationales Handbuch der zeitgenössischen Forschung. Berlin, New York: de Gruyter = Handbücher zur Sprach- und Kommunikationswissenschaft 6
- Tesnière, L. 1959:** *Élément de syntaxe structurale*. Paris: Klincksieck (²1965)
- Wahlster, W. 1982:** Natürlichsprachliche Systeme. Eine Einführung in die sprachorientierte KI-Forschung. In: Bibel, J.; Siekmann, J.H. (Hrsg.): Künstliche Intelligenz. Frühjahrsschule Teisendorf, März 1982. Berlin, Heidelberg, New York: Springer, 202-283
- Weinrich, H. 1976:** Sprache in Texten. Stuttgart: Klett
- Weisgerber, L. ³1962:** Grundzüge der inhaltbezogenen Grammatik. Düsseldorf: Schwann = Von den Kräften der deutschen Sprache I (⁴1971)

<p>BAND 7 NUMMER 1/2 Dez. 1990 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 	<p>BAND 9 NUMMER 2 DEZEMBER 1992 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 	<p>BAND 11 NUMMER 1 JUNI 1994 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 
<p>BAND 8 NUMMER 1/2 November 1991 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 	<p>BAND 10 NUMMER 1 AUGUST 1993 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 	<p>BAND 11 NUMMER 2 DEZEMBER 1994 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 
<p>BAND 9 NUMMER 1 März/April 1992 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> 	<p>BAND 10 NUMMER 2 DEZEMBER 1993 ISSN 0172-9926</p> <p>LDV-FORUM Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV</p> <p>Li<u>ngu</u>istische</p> <p>Date<u>u</u> </p> <p><i>www.werphologie</i> Verarbeitu<u>u</u>g</p>	<p>Ältere Ausgaben des LDV-Forums können zum Preis von DM 20,- bezogen werden über:</p> <p>IAI Martin-Luther-Str. 14 66111 Saarbrücken Tel.: +49 681 39313 Fax: +49 681 397482</p>

Aus der Lehre für die Lehre

COMPUTERLINGUISTIK UND MASCHINELLE ÜBERSETZUNG IN EINEM STUDIENGANG FÜR ÜBERSETZER UND DOLMETSCHER*

Johann Haller

1 Einleitung

An den Studiengängen für Übersetzer und Dolmetscher hat die meist im Studienplan vorgeschriebene Linguistik keinen leichten Stand: alles, was über eine grundlegende Einführung hinausgeht, wird von den sehr zielgerichtet studierenden Kommilitonen danach hinterfragt, welchen konkreten Nutzen es für die spätere Berufstätigkeit bringt. Auch die entsprechenden Mitglieder des Lehrkörpers neigen oft zu der Ansicht, daß Ausdauer und Übung wichtiger sind als ein theoretisches Bewußtsein und Reflexion über das eigene Tun. Nicht einmal die ausführlichen Arbeiten kontrastiver Linguisten waren nah genug an der speziellen Vermittlungstätigkeit der Übersetzer, bei deren Aufgabe oft Faktoren wie Auftraggeber, Adressat und Zweck alle theoretischen Ergebnisse einer kontrastiven Analyse überspielen können. Dies war (und ist zum Teil immer noch) jedenfalls die Hauptthese der eigentlichen Übersetzungswissenschaft, die kontrastive Linguistik als eine zwar wichtige, aber eben nur Teile des Problems abdeckende Hilfsdisziplin betrachtete. Erst in den letzten Jahren hat die Übersetzungswissenschaft sich auf empirische Methoden wie die Analyse des lauten Denkens verlegt, die die beim Übersetzen ablaufenden Prozesse einer empirischen Analyse zugänglich

machen soll.

Diese Fakten legen den Schluß nahe, daß die weitaus mehr mathematischen Charakter tragenden Modelle der Computerlinguistik im Kontext der Übersetzerausbildung noch schwieriger zu vermitteln sind als die griffigen Modelle der Zeichentheoretiker, Semantiker und Pragmatiker.

Computer allgemein sind jedoch in der Ausbildung und Berufstätigkeit des Übersetzers ein Thema, das unter vielen verschiedenen Gesichtspunkten schon ausführlich diskutiert wird: in der Form von sogenannten CAT -Systemen (Integration von Terminologieverwaltung und Textverarbeitung), die bereits heute schon mehr oder weniger zum Handwerkszeug gehören, wird ihre Einsatzfähigkeit eigentlich nicht mehr in Frage gestellt. Eine Frage ist hier höchstens der Preis, den ein einzelner Übersetzer für Hard- und Software sowie für Fachwörterbücher zu zahlen hat bzw. in welchen Zeiträumen und bei welchen Umsätzen sich eine solche Investition bezahlt machen kann. Die zunehmende Vernetzung unserer Gesellschaft führt auch dazu, daß man sich in weltweiten Datenbanken über alle Sachgebiete schnell informieren kann, zu denen man eine Übersetzungs- oder Dolmetschaufgabe gestellt bekommt.

Die Sinnfälligkeit von " Translation Memories", die mit statistischen Methoden nach gleichen oder ähnlichen Sätzen suchen, die schon einmal übersetzt wurden, ist vor allem bei technischer Dokumentation sofort

Dieser Beitrag ist eine stark überarbeitete und aktualisierte Fassung eines Vortrags auf dem Symposium "Sprachdatenverarbeitung" an der Universität des Saarlandes 1992 (erschieden in den entsprechenden Proceedings).

zu erkennen. Aus diesem Grund werden solche Werkzeuge auch an den meisten Ausbildungsstätten für Übersetzer (und Dolmetscher) während des Studiums präsentiert und diskutiert. Schließlich bieten viele Studiengänge auch eine kurze Einführung in die Grundkonzepte und die Geschichte der eigentlichen maschinellen Übersetzung (fortan MÜ), d.h. in den alten Menschheitstraum von der vollautomatischen Übertragung eines Textes von einer Sprache in die andere; hier sind jedoch die Reaktionen der professionellen und zukünftigen Übersetzer sehr viel zurückhaltender, besonders wenn zuwenig darauf hingewiesen wird, daß es auch bei diesen Systemen heute nur mehr darum geht, eine Rationalisierung des Zeitaufwandes für das Übersetzen in Höhe von einigen Prozentpunkten zu erreichen. Daß dies nur in engbegrenzten Bereichen überhaupt möglich ist, zeigt die aktuelle Orientierung vieler heutiger Forschungsprojekte, die von der EU-Kommission oder US-Unternehmen gefördert werden. "Visionäre" Projekte wie das VERBMOBIL (automatisches Dolmetschen für Terminvereinbarungen) werden von der Gemeinde der Übersetzer mit skeptischen Augen betrachtet; dies liegt zum einen an den stark plakativen Aussagen, die während der Phase der Projekterstellung verwendet und mittlerweile längst durch eine Verlagerung des Akzents auf die erwarteten Spin-Off-Effekte abgelöst wurden. Zum anderen sieht man natürlich auch den hohen finanziellen Aufwand, der für die Erforschung aller nötigen Feinheiten der künstlichen Intelligenz zu treiben ist, wenn ein Dialog in gesprochener Sprache auch nur annähernd verstanden und in einer anderen Sprache akustisch wiedergegeben werden soll.

Hat es denn für den angehenden Übersetzer (und auch für den, der sich neben seiner Arbeit weiterbilden will) überhaupt Sinn, sich mit den computerlinguistischen Grundlagen von Systemen wie SYSTRAN, LOGOS und METAL und von Forschungssystemen wie EUROTRA und CAT2 sowie den Versuchen der Carnegie-Mellon-Universität und der Japaner zu befassen? Welche Erkenntnisse über die Einsetzbar

keit und die Funktionsweise können ihm von Nutzen sein, wenn er an seinem Arbeitsplatz mit einem dieser Systeme konfrontiert wird? Gibt es darüber hinaus andere positive Erkenntnisse für seine Arbeit?

In einem Übersetzerstudiengang wie z.B. an der Universität des Saarlandes geht es zunächst um (kommerzielle und forschungsorientierte) Systeme zur maschinellen Übersetzung, die zwar (möglicherweise) mit Prä- und Postedition, jedoch ohne Intervention des Benutzers während des Übersetzungsvorganges Zielsprachentexte erstellen; diese Systeme werden in einer Einführungsvorlesung vorgestellt und in speziellen Seminaren vor allem als mögliche Arbeitsinstrumente für den Übersetzer behandelt. Beispiele hierfür sind Übungen zur Postedition maschineller Texte, zur Überprüfung der Verständlichkeit von Rohübersetzungen und zu Evaluierungsmethoden allgemein. Auch hier werden wichtige Fragen für die spätere Tätigkeit mit MU-Systemen erörtert: die Eignung bestimmter Textsorten für die maschinelle Verarbeitung, die Integration in Büro-Umgebungen und die Frage der generellen Rentabilität unter bestimmten Bedingungen.

In unserem Beitrag nun soll darüber hinaus die (oft kontrovers diskutierte) Frage untersucht werden, inwiefern der Einblick in die computerlinguistischen Grundlagen, die Funktionsweise und das eigene praktische Arbeiten in Entwicklung und Wartung solcher Systeme einen Nutzen für die universitäre Ausbildung von Übersetzern und anderen mit Sprache befaßten Berufen bringen können.

2 Argumente für computerlinguistische Elemente im Übersetzerstudium

Die Gründe, die für die Beschäftigung mit den computerlinguistischen Grundlagen und der Funktionsweise von MÜ-Systemen während des Übersetzerstudiums sprechen, können unter folgenden Stichworten zusammengefaßt werden:

- > Anforderungen des Arbeitsmarkts
- > Einführung neuer Technologien

- > Ergänzungsfach EDV
- > mathematisch-logische Denkweise
- > Formalisierung von MÜ - Problemen
- > Einfluß auf die Entwicklung neuer MÜ Systeme
- > Verbreiterung der Qualifikation

2.1 Anforderungen des Arbeitsmarkts

Seit zwei Jahren führt die Fachrichtung 8.6. unserer Universität eine Statistik der eingegangenen Stellenangebote, die nach verschiedenen Kriterien ausgewertet wird. Die Auswertung der Anforderungen an die Kandidaten zeigt, daß einerseits der Anteil der Stellen, bei denen EDV- und SDV-Kenntnisse explizit gefordert werden, von ca. 30% auf ca. 80% und mehr gestiegen ist, daß aber auch zunehmend explizite Hinweise auf die maschinelle Übersetzung ihren Platz finden. Dies wird zwar meist nur als Wunsch und noch nicht als Voraussetzung genannt, solche Kenntnisse stellen aber in jedem Fall für den Bewerber ein Plus dar.

2.2 Einführung neuer Technologien am Arbeitsplatz

Bei dem Einsatz neuer Softwaretechnologien in der Industrie geht die Initiative meist vom Management aus, das Rationalisierungsmöglichkeiten sieht; beraten werden die Manager dabei vom Rechenzentrum und seinen Mitarbeitern, die schon von ihrer Funktion her an der Anschaffung neuer Soft- (und damit auch Hard-)ware interessiert sind. Gerade bei MÜ - Systemen sind hier einige Beispiele von mißlungenen Einführungen bekannt, die nicht nur zu finanziellen Belastungen für das Unternehmen, sondern auch zur Frustration der Übersetzer beigetragen haben.

Der Übersetzer dagegen, der Kenntnisse über die Funktionsweise solcher Systeme erworben hat, kann die Unternehmensleitung bei Akquisition und organisatorischer Einbettung eines solchen Systems als Fachmann und künftiger Anwender selbst besser

beraten; er kann erkennen, wo geeignete Bedingungen vorliegen oder welche Maßnahmen ergriffen werden müssen, um solche zu schaffen.

Werbeaussagen von MÜ-Herstellern können als solche erkannt und in ihrer Bedeutung relativiert werden; der Dialog mit dem Hersteller kann fachlich besser geführt werden, wenn der Partner auch Kenntnisse über das "Innenleben" des Systems hat.

Der entsprechend ausgebildete Übersetzer kann auch einmal eine fundierte Ablehnung eines solchen Systems formulieren; besitzt er diese Kenntnisse nicht, geschieht die Ablehnung meist nur auf emotioneller Basis.

2.3 Ergänzungsfach EDV

Der zukünftige Übersetzer belegt neben Übungen zu Sprache, Sprachwissenschaft, Landeskunde und Übersetzung ein sogenanntes Ergänzungsfach; er spezialisiert sich (z.B. in Saarbrücken) auf Recht, Wirtschaft oder Technik. Innerhalb des letzteren (meistgewählten) Ergänzungsfaches kann er sich z.B. auf die elektronische Datenverarbeitung spezialisieren.

Bei der Beschäftigung mit MÜ-Systemen erweitert der Übersetzer zugleich seine Kenntnisse in der EDV allgemein; er fügt der DOS-Welt, die er bei der Beschäftigung mit den PC- Werkzeugen ausführlich kennengelernt hat, andere Betriebssysteme hinzu (Großrechner, UNIX); es liegt auf der Hand, daß dies mit Programmen leichter ist, die dem Übersetzer vertraute Aufgaben ausführen als mit abstrakten Tabellen und Zahlenproblemen, wie sie in Einführungen für Informatiker verwendet werden. Diese umfangreichen Kenntnisse können beispielsweise bei Übersetzungsarbeit in der oder für die EDV-Branche nutzbringend angewendet werden. Ein großer Übersetzungsbedarf besteht etwa bei Hard- und Software-Beschreibungen und Manualen. Auch auf anderen Gebieten (z.B. Maschinenbau oder Medizin) dringt die Datenverarbeitung immer weiter vor, und ihr Vokabular muß das entsprechende fachspezifische Wissen ergänzen.

2.4 Mathematisch-logisch Denkweise

Eine schulische und/oder universitäre Ausbildung, die das Schwergewicht nicht auf naturwissenschaftliche Fächer wie Mathematik und Physik legt, führt oft zu einem weniger intensiven Training mathematisch-logischer Denkweisen, wie sie in vielen technischen und naturwissenschaftlichen Berufen zum unbedingt notwendigen Handwerkszeug gehören. Obwohl in manchen Teilen etwa der Übersetzungswissenschaft für eine stark individualisierte und persönlichkeitsbezogene Arbeitsweise plädiert wird (wie sie sicherlich für die Übertragung von poetischen oder auf Werbewirkung ausgerichteten Texten ideal ist), dürfte für einen großen Teil der in der Praxis anfallenden Übersetzungen eine Einsicht auch in mathematisch-naturwissenschaftliche Argumentation sowie formale Logik von Nutzen sein.

Unter anderem kann der Übersetzer sich mit solchen Vorgängen vertraut machen, indem er die Datenverarbeitung intensiver studiert als nur in dem Maße, wie ihm Programme als Werkzeug dienen. Dies geschieht durch die Auseinandersetzung mit Menütechniken, Kommandosprachen, Softwarewerkzeugen beispielsweise zur Stringverarbeitung sowie Einführungen in prozedurale (z.B. BASIC, C) und deklarative (z.B. Prolog) Programmiersprachen. Bei diesen Übungen können einfache Teilprobleme der Textverarbeitung, der Terminologieverwaltung und der maschinellen Übersetzung vom Studenten selbst in algorithmische Form gebracht, programmiert und getestet werden.

2.5 Formalisierung von Übersetzungsproblemen

Die Formulierung von Analyse-, Synthese- und Transferregeln sowie von mono- und bilingualen Lexikoneinträgen in MÜ-Systemen setzt eine starke Formalisierung linguistischer Sachverhalte und die Auseinandersetzung mit linguistischen Theorien voraus, wie sie in den Veranstaltungen zur Sprach- und Übersetzungswissenschaft "auf dem Papier" eingeführt worden sind. .

Dies geschieht in einführenden Seminaren zur Computerlinguistik, die auf die erwähnten theoretischen Vorkenntnisse zugeschnitten sind. Einzelne Themen können durch den Besuch weiter spezialisierter Seminare vertieft werden (Morphologie, Syntax, Semantik, Pragmatik, Dialogsysteme).

Insofern stellt dies eine Ergänzung zur Studienkomponente Sprach- und Übersetzungswissenschaft dar, die den an der Universität ausgebildeten Übersetzer dazu befähigen soll, nicht nur aufgrund langen Trainings und guten Sprachgefühls richtig zu übersetzen, sondern die Schritte seiner Tätigkeit auch mit adäquatem und präzisiertem Vokabular zu begründen und darüber zu reflektieren.

Er kann hierdurch einerseits schneller Rationalisierungsmöglichkeiten in seiner eigenen Tätigkeit erkennen; andererseits kann er auch den gesamten Ablauf des Dokumentationsprozesses im Unternehmen kritisch untersuchen. So kann er z.B. fundierte Kritik an Ausgangstexten üben bzw. Schreibregeln in verständlicher Form für die Autoren entwerfen. In der Regel wird ein (technischer) Text auch für den menschlichen Leser leichter verständlich, wenn gewisse Regeln für Einfachheit, Kürze, Satzbau und logische Abfolge eingehalten werden.

2.6 Mitwirkung des Übersetzers an der Entwicklung von MÜ-Systemen

Seit dem Beginn der Geschichte der MÜ ist festzustellen, daß Systeme in der Hauptsache von Informatikern und später von Linguisten konzipiert und entwickelt wurden; Übersetzer selbst waren nur in sehr geringem Maße beteiligt. Heute scheint sich das Gewicht auf spezialisierte Computerlinguisten, aber auch zu Modellen der Künstlichen Intelligenz zu verschieben.

Bereits während ihres Studiums einschlägig ausgebildete Übersetzer könnten die Lücke füllen, die zur Erstellung von besseren und vor allem auf ihre Arbeitssituation zugeschnittenen Systemen unbedingt geschlossen werden muß.

Hier schließt sich auch der Kreis der Argumentation: wenn er selbst einmal an der Entwicklung neuer Systeme mitwirken

kann, wird der Übersetzer an seinem Arbeitsplatz nicht mehr mit Systemen konfrontiert werden, die seine Belange in der Entwicklung nur am Rande berücksichtigt haben.

2.7 Verbreiterung der Qualifikation

Das früher einheitliche Berufsbild des Übersetzers und Dolmetschers ist heute in der Wandlung begriffen. So werden an mehreren Übersetzerbildungsstätten in Universitäten und Fachhochschulen neue Studiengänge eingerichtet wie Technische Kommunikation, mehrsprachige Wirtschaftsfachleute etc. Diese Auffächerung bzw. Spezialisierung des Berufs zeigt sich auch in der an der in Saarbrücken alle zwei Jahre stattfindenden Veranstaltung "Praxis und Studium"; sie beinhaltet immer einen Arbeitskreis "Fachfremde Tätigkeiten", in denen Absolventen über ihre Tätigkeit berichten, die nicht mehr (oder nicht mehr ausschließlich) die des Übersetzers oder Dolmetschers ist. Terminologie, Technische Normung, Export, Präsentation, Organisation: dies alles sind Tätigkeitsfelder, in denen zu der Orientierung auf sprachliche Kommunikation (die in der heutigen politischen Lage mehrsprachig sein muß) Fähigkeiten wie Flexibilität, technisches Verständnis und das Umgehen mit neuer Technologie allgemein hinzutreten müssen. Diese können u. a. durch die intensive Beschäftigung mit MÜ-Systemen und deren linguistischen und technischen Grundlagen während des Studiums erworben werden.

Zu einer solchen Neuorientierung ist jedoch nicht immer die bürokratisch aufwendige Einrichtung von neuen Studiengängen mit klangvollen Namen, wirtschaftsnah formulierten Ausbildungszielen und schulisch straff organisierten Auslandssemestern mit automatisch anerkannten Scheinen nötig; inhaltlichen Wandel läßt das heutige Universitätsstudium mit Pro- und Hauptseminaren auch jetzt schon zu, Auslandsaufenthalte können durch Kontakte vermittelt und vom Studenten individuell gestaltet werden.

3 Computerlinguistik und Maschinelle Übersetzung in der Übersetzerbildung in Saarbrücken

In diesem letzten Abschnitt soll kurz vorgestellt werden, wie das Gebiet der Maschinellen Übersetzung an unserer Fachrichtung angeboten wird. Die Basiskenntnisse werden in einer Studienkomponente "Sprachdatenverarbeitung" vermittelt, die den Gebrauch von PC- Werkzeugen vermittelt: spezielle Textverarbeitungsprogramme, die den Gebrauch mehrerer Sprachen in einem Dokument erlauben, Terminologiedatenbanken, sogenannte Translation Memories und PC-basierte Übersetzungssysteme.

Wie bereits oben erwähnt, werden in einer Einführungsvorlesung neben einer Übersicht auch kurze Einblicke in die technische Funktionsweise der größeren kommerziellen Systeme gegeben.

Seit mehreren Semestern werden daneben Seminare zu computerlinguistischen Grundlagen, zu forschungsorientierten Systemen wie EUROTRA (mit seiner "sideline" CAT2) und anderen unifiktionsbasierten Systemen angeboten. Hier haben die Studenten eine umfangreiche Grundlektüre zu bewältigen. Nach einer anfangs mühevollen praktischen Ausbildung, bei der auch oft UNIX-Kenntnisse nachgeholt werden müssen, werden jedoch von einzelnen Teilnehmern beachtliche Erfolge erzielt: sie bearbeiten einzelne kontrastive Probleme und erweitern das System selbst um die entsprechenden Regeln und Lexikoneinträge. Systematische Tests und eine ausführliche Dokumentation sind das konkret dokumentierte Ergebnis.

Da EUROTRA und seine Folgeprojekte im Förderprogramm LE (Linguistic Engineering) der Kommission der Europäischen Union mit allen offiziellen Sprachen der Gemeinschaft arbeiten und auch einige Experimentierversionen mit Sprachen wie Russisch, Koreanisch und Japanisch existieren, können Studenten aller Sprachen, die in Saarbrücken angeboten werden, hier Erfahrung erwerben. Für besonders Interessierte stehen Praktikumsplätze sowohl in

Saarbrücken (am Institut für Angewandte Informationsforschung, IAI) als auch an den Partneruniversitäten in den anderen westeuropäischen Ländern, aber auch z.B. in Moskau und Mexico, zur Verfügung. Die bisher durchgeführten Praktika zeigten eine große Akzeptanz sowohl bei den Studenten als auch bei den aufnehmenden Institutionen. Ins besondere ist das Ziel einer Integration der Forschungsarbeiten am IAI mit der Ausbildung an der Universität erreicht.

Kleine Gruppen nehmen an Spezialseminaren zu einzelnen Aspekten wie syntaktische Theorien (GB, HPSG u.a.), Semantik, Wissensrepräsentation und Formalismen teil; Ergänzungsmöglichkeiten bestehen hier auch in Zusammenarbeit mit der Fachrichtung 8.7. Computerlinguistik. Die Studenten dieser Fachrichtung können ihrerseits Maschinelle Übersetzung als Ergänzungsfach wählen, was bereits häufig wahrgenommen wird und woraus sich interessante Zusammensetzungen von Seminaren ergeben.

Durch Lehraufträge an Vertreter von kommerziellen MÜ-Systemen wird periodisch die Möglichkeit gegeben, solche Systeme näher kennenzulernen; auch hieraus entwickeln sich oft Praktikumsmöglichkeiten.

Schließlich steht seit SS 92 das System Logos in einer UNIX-Version zur Verfügung. In einer regelmäßig stattfindenden Übung werden in Gruppen Testsatzbatterien in verschiedenen Sprachen entwickelt und in die Hauptfremdsprache der Teilnehmer übersetzt. Durch den Studiengang mit Französisch als Muttersprache, durch die jährlich ca. 60 ERASMUS-Studenten aus allen europäischen Ländern sowie Austauschstudenten aus der ganzen Welt können meist alle in den Zielsprachen erzeugten Texte auch durch Muttersprachler überprüft werden. Nach einer ersten Übersetzung entsteht eine linguistische Fehlerbeschreibung; das Kodieren unbekannter oder falsch analysierter Wörter wird geübt sowie die Regelmäßigkeit auftretender Fehler durch Variieren der Sätze überprüft. Hierzu muß die Handhabung des Gesamtsystems sowie der Kodiersysteme ALEX (für einfache

Wörter und morphosyntaktische Informationen) und SEMANTHA (für semantische Informationen und Wortverbindungen) erlernt werden. Schlußtests dokumentieren die Entwicklung des Übersetzungsergebnisses sowie die hierzu notwendigen Schritte.

Obwohl durch die politisch bedingten Sparmaßnahmen an der Universität des Saarlandes schon seit dem Wintersemester 1994 weniger Studienanfänger im Übersetzen und Dolmetschen aufgenommen werden, wächst die kleine Gruppe der an dieser neuen Technologie interessierten Studenten kontinuierlich weiter an; auch wenn sich mittelfristig nur ein kleiner Teil der Übersetzer mit dem "Innenleben" solcher Systeme beschäftigen wird, kann gesagt werden, daß sich das 1990 eingerichtete Fach Maschinelle Übersetzung an der Universität des Saarlandes erfolgreich etabliert hat. Neben die Übersetzern des Diplom-Studienganges treten derzeit mehrere Magisterkandidaten für dieses Fach, das als Haupt- und Nebenfach studiert werden kann. Inzwischen liegen auch schon einige Diplomarbeiten sowie zwei Dissertationen im Fach Maschinelle Übersetzung vor, die sich mit den angegebenen Themen ausführlich auseinandersetzen; weitere stehen kurz vor der Fertigstellung.

TITUS

DAS PROJEKT EINES INDOGERMANISTISCHEN THESAURUS

Jost Gippert Vergleichende Sprachwissenschaft Universität Frankfurt
gippert@em.uni-frankfurt.d400.de <http://www.rz.uni-frankfurt.de/home/ftp/pub/titus/public.html>

Auf der indogermanistischen Fachtagung in Leiden (September 1987) berieten einige der Teilnehmer darüber, inwieweit es möglich sei, im Hinblick auf die an verschiedenen Forschungsstätten des In- und Auslands angelaufenen oder laufenden Projekte, die die Einspeicherung von für die Vergleichende Sprachwissenschaft relevanten Texten auf Datenträger zum Ziel hatten, zu einer Zusammenarbeit zu kommen. Man war sich einig, daß es wenig sinnvoll sei, wenn jeder einzelne Forscher in diesem Bereich allein und für sich arbeiten würde, weil dies immer wieder zu einer überflüssigen Duplizierung des Aufwands führen müsse; statt dessen sei es zweckmäßig, die Eingabe von Texten von vornherein zu koordinieren.

Damit war die Idee zu einem - innerhalb der Indogermanistik beispiellosen - Gemeinschaftsprojekt geboren, das schon bald darauf, mit einem Aufruf in "Die Sprache" 32/2, 1987, unter dem Namen eines "Thesaurus indogermanischer Textmaterialien auf Datenträgern" ins Leben gerufen wurde. Nach einer nunmehr achtjährigen Laufzeit ist die primäre Zielsetzung des Projekts, nämlich über den gesamten die Grundlage der Vergleichenden Indogermanischen Sprachwissenschaft darstellenden Textbestand aus altüberlieferten Sprachen wie dem Altindischen (Sanskrit), dem Altiranischen (Avesta, Altpersisch), ~dem Altgriechischen, dem Lateinischen, altanatolischen Sprachen wie dem Hethitischen, alt germanischen Sprachen wie dem Althochdeutschen oder Altenglischen in einer für die elektronische Analyse zugänglichen Form zu verfügen, in greifbare Nähe gerückt. Im Zuge der ständig wachsenden Kapazitäten von Hard- und Software haben sich Zielsetzungen und Perspektiven des Projekts im gleichen Zeitraum jedoch ebenfalls erheblich ausgeweitet, was nicht zuletzt die neue Namensgebung "Thesaurus indogermanischer Sprach- und Textmaterialien (TITUS)" reflektieren soll, unter der es seit der 3. Tagung für Computereinsatz in der Historisch-Vergleichenden Sprachwissenschaft (Dresden 1994) geführt wird. Aus Anlaß dessen, daß sich auf der letztgenannten Tagung zugleich auch ein eigener Arbeitskreis der GLDV mit dem Titel "Historisch-vergleichende Sprachwissenschaft" konstituierte, dessen Mitglieder zum größten Teil am TITUS-Projekt beteiligt sind, erscheint es angebracht, die Entwicklung, den gegenwärtigen Stand und die nächsten Vorhaben des Projekts hier kurz darzustellen.

Das primäre Ziel des TITUS-Projekts bestand, wie bereits gesagt, in der koordinierten Erfassung der relevanten Originaltexte altüberlieferter indogermanischer Sprachen. Natürlich war die Idee, derartige Texte dem sich seinerzeit eher mühsam durchsetzenden neuen elektronischen Medium "anzuvertrauen", nicht wirklich neu; tatsächlich hatte es ja schon seit den 60er Jahren verschiedene Projekte (v.a. in USA) gegeben, die auf eine (vollständige oder teilweise) Erfassung von Textdaten auch altüberlieferter Sprachen abzielten. Unter ihnen ist zunächst das Projekt des "Thesaurus Linguae Graecae" zu nennen, das eine Erfassung sämtlicher klassisch- und spätklassisch-griechischer Textmaterialien seit dem Beginn der Überlieferung (Homer) bis in die frühmittelalterliche byzantinische Zeit anstrebt und inzwischen das gesteckte Ziel zu ca. 80% erreicht hat. Da das Griechische innerhalb der indogermanischen Sprachfamilie als einer der bedeutendsten Vertreter gelten kann, war mit dem TLG-Projekt von vornherein ein möglicher Kooperationspartner gegeben, der in mancherlei Hinsicht auch als Vorbild dienen konnte; im Bereich des Griechischen reduzierten sich die für den indogermanistischen Thesaurus verbleibenden Aufgaben auf die vorklassische Überlieferung (der mykenischen Epoche) sowie auf die Überlieferung von Textmaterialien aus den altgriechischen Dialekten, die freilich bis heute noch nicht in befriedigendem Umfang bearbeitet sind.

Den eigentlichen Grundstock der Textsammlung innerhalb des Thesaurus bildeten demgegenüber zwei Texte aus dem indoiranischen Sprachzweig, die von höchstem indogermanistischem Interesse sind und deren elektronische Bearbeitung der Konstituierung des Projektes vorausging. Es handelt sich zum einen um die altindische Ṛgveda-Saṃhitā, die bereits in den 70er Jahren unter der Leitung von W.P. LEHMANN an der University of Texas eingegeben worden war, zum anderen um das altiranische Avesta-Corpus, das im Hinblick auf die Erstellung einer Textkonkordanz im Rahmen eines von der DFG geförderten Projekts durch S. GIPPERT-FRITZ an der Freien Universität Berlin elektronisch aufbereitet wurde. An diesen beiden Einzelprojekten lassen sich bereits die wesentlichsten Probleme aufzeigen, die während der Anfangsphase des Thesaurus-Projekts zu lösen waren.

Der Ṛgveda-Text war, wie damals nicht anders möglich, ursprünglich auf einer Großrechneranlage eingegeben worden. An eine Eins-zu-eins-Wiedergabe in der schriftlichen Form, in der der Text selbst überliefert worden ist, d.h. in einem indischen Alphabet wie der Devanāgarī, war damals nicht zu denken; es wäre allerdings seitens der (sprachwissenschaftlich orientierten) Fachwelt auch gar nicht unbedingt erwünscht gewesen, da sich diese seit dem Vorliegen der Edition von Th. AUFRECHT ("Die Hymnen des Rigveda", 1877) an den transkribierten Text gewöhnt hat. Aber auch eine solche Transkription war nicht eins zu eins auf den Großrechner abbildbar, da sie von zahlreichen diakritischen Buchstabenkombinationen geprägt ist, die (bis heute) in keinem Codierungsstandard vorgesehen sind (z.B. Kombinationen von Vokalbuchstaben mit Makron und Akzent wie \acute{a} oder \grave{u} oder Kombinationen von Konsonantenbuchstaben mit subskribiertem Punkt oder Kringel wie $\underset{\cdot}{m}$, $\underset{\cdot}{n}$ oder $\underset{\cdot}{r}$). Ein ganz gleich gelagertes Problem betraf auch das Avestacorpus, das in einer völlig einzigartigen, in der maßgeblichen Ausgabe K. GELDNERs ("Avesta. Die heiligen Bücher der Parsen", 1895) beibehaltenen Originalschrift ("Avestaschrift") überliefert ist: dessen Eingabe konnte zwar bereits mit einem PC begonnen werden, doch gab es auch hier a priori keine Möglichkeit, die Originalschrift oder auch nur die übliche Transkription, die mit diakritischen Kombinationen wie $\underset{\cdot}{t}$ oder griechischen Buchstaben wie γ , δ arbeitet, auf Bildschirm und Drucker darzustellen. So mußte in beiden Fällen zunächst auf eine ersatzweise Darstellung auf der Grundlage des ASCII-Codes zurückgegriffen werden, bei der z.B. die Diakritika durch adskribierte Zeichen wie \wedge , \backslash , \sim repräsentiert wurden.

Auch wenn eine derartige Darstellung für die maschinelle Analyse kein schwerwiegendes Hindernis bedeutet, so wurde doch die Möglichkeit einer der wissenschaftlichen Gepflogenheit entsprechenden Wiedergabe zumindest transkribierter "Sonderzeichen" mit steigender Leistungsfähigkeit von Rechnern (PCs) und Druckern mehr und mehr als ein Desiderat empfunden. Tatsächlich bedeutete die Schaffung von Programmeinheiten, die eine Ausgabe von Transkriptionssystemen oder Originalschriften auf Bildschirm und Drucker ermöglichen sollten, seit dem Beginn des Projekts eine ständige Herausforderung. Nachdem sich erste "Gehversuche" in dieser Richtung noch auf das Drucken mit einem 24-Nadel-Drucker beschränkten¹, scheint in diesem Bereich jetzt, wo "maßgeschneiderte" Pakete vektorisierter Fonts (Postscript, True Type u.a. Formate) für alle in Frage kommenden Schriftsysteme erarbeitet sind, alles nötige getan. Dennoch kann im Zusammenhang mit der Darstellung und der mit ihr stets verknüpften rechnerischen Codierung der Zeichen aus mehreren Gründen noch keine "Entwarnung" gegeben werden: Zum einen sind die verschiedenen verbreiteten Rechnersysteme nach wie vor weit davon entfernt, eine einheitliche Zeichendarstellung zu benutzen. Gemeinsame Grundlage ist nach wie vor lediglich der sog. ASCII-Standard, der wohl die 26 Zeichen des latein. Alphabets, aber keinerlei diakritische Kombinationen wie *ä*, *á* oder *ā* enthält. Natürlich gibt es in jüngerer Zeit Versuche, die Codierungslücke durch die Schaffung einheitlicher, systemunabhängiger Standards zu schließen. Die bisher vorliegenden Lösungsversuche sind jedoch für die besonderen Aspekte der im Rahmen des TITUS-Projekts zu berücksichtigenden Schriftphänomene in keiner Weise ausreichend, ganz abgesehen davon, wieweit sie überhaupt schon von marktgängigen Betriebssystemen und Anwenderprogrammen unterstützt werden.

Das gilt zunächst für den Standard der sog. WordPerfect-Zeichensätze, der, seit er vor ca. sechs Jahren eingeführt (und über die Jahre nicht unerheblich erweitert) wurde, immerhin einen wichtigen ersten Schritt in die richtige Richtung darstellte. Das hinter dem WP-Standard stehende System baut zwar im wesentlichen auf der Grundlage einer 16-Bit-Codierung auf und ermöglicht damit theoretisch einen Zeichenvorrat von 65536 Zeichen, ausgenutzt ist es jedoch nur zu einem äußerst geringen Teil (ca. 2000 Zeichen), und daß es sich durchsetzen wird, erscheint aus marktpolitischen Gründen eher unwahrscheinlich.

Ähnliches gilt auch für den sog. "Unicode"², dessen (ebenfalls in 16 Bit codierte) 65536 Zeichen — wie schon zuvor die verschiedenen systeminternen 8-Bit-Zeichensätze und Codepages — in viel zu starkem Maße an heute gebräuchlichen Nationalalphabeten ausgerichtet sind, als daß sie zur Bearbeitung der — im TITUS-Projekt vorrangig interessierenden — schriftlichen Zeugnisse altüberlieferter Sprachstufen, sei es in Originalschriften, sei es in Transkription, geeignet wären. Eine bessere Ausgangsposition würde demgegenüber die geplante ISO-Norm 10646 darstellen, die — bei einer 32-Bit-Codierung — mit insgesamt rund 3 Milliarden Zeichen tatsächlich einen für alle je von Menschen benutzten Schriftsymbole ausreichenden Vorrat darstellen könnten. Leider deutet im Moment aber nichts darauf hin, daß sich diese Norm irgendwann einmal zur Basis handelsüblicher Rechnersysteme entwickeln könnte.

Auch ein dritter Lösungsweg, der momentan an Aktualität gewinnt, ist derzeit noch nicht genügend ausgereift, um für die Verarbeitung von Sprachmaterialien der genannten Art eine geeignete Grundlage zu bilden. Es handelt sich um den sog. HTML-Standard ("HyperText Markup Language"), dessen Bedeutung im Zusammenhang mit dem "Weltweiten Netz" ständig zunimmt. Tatsächlich wird die Benutzbarkeit dieses Standards durch den in ihm vorgesehenen Zeichenvorrat,

¹ Den Generierungsmöglichkeiten war u.a. mein auf dem Internationalen Orientalistenkongreß (ICANAS) in Hamburg (28.8.1986) gehaltener Vortrag "Probleme der Textverarbeitung in der Orientalistik" gewidmet.

² Cf. dazu die von "Unicode Consortium" herausgegebene Darstellung "The Unicode Standard. Worldwide Character Encoding. Version 1.0.", Volume 1-2, 1991.

der neben den reinen "ASCII"-Zeichen, ähnlich wie die durch MS-Windows verbreitete ANSI-Norm, lediglich die in westeuropäischen Nationalalphabeten gebräuchlichen Zeichenkombinationen wie *Á*, *ç*, *Ñ* kennt, erheblich beeinträchtigt. Es ist eines der dringendsten Desiderate, daß die Entwickler beim ständigen weiteren Ausbau des HTML-Standards die Codierbarkeit beliebiger Akzentkombinationen sowie nichtlateinischer Schriften vorsehen. Da der HTML-Standard (entsprechend den Vorgaben des sog. SGML-Verfahrens ["Standard Generalized Markup Language"], als dessen Derivat er gelten kann) eine systemunabhängige 7-Bit-Grundlage verwendet, bei der nicht zur ASCII-Norm gehörende Zeichen explizit benannt werden (z.B. "´" für *á*, "ß" für *ß*), stellt er ein a priori offenes System dar, dessen Ausnutzung die Aufgabe des jeweiligen Anwendungsprogramms ist; somit wäre er für die avisierte Erweiterung bestens geeignet.

Es sei noch einmal ausdrücklich festgehalten, daß für eine sprachwissenschaftlich-philologische Auswertung gegebener Textmaterialien in beliebigen Sprachen die Eins-zu-Eins-Wiedergabe der Daten in Originalschrift bzw. Transkription gar nicht unbedingt die Voraussetzung ist. Entscheidend ist vielmehr, daß eine **umkehrbar-eindeutige** Codierung gegeben ist, die das gemeinte Zeichen jederzeit exakt abrufbar zu machen gestattet. Zu warnen ist in diesem Zusammenhang v.a. vor einem allzu leichtfertigen Umgang mit den Möglichkeiten einer freien Umdefinition von Zeichen innerhalb von 8-Bit-Systemen wie denjenigen von Windows-TrueType oder Macintosh-Postscript, denn hier ist gerade die Eindeutigkeit nicht immer gewährleistet (v.a. bei einem Transfer über die Systemgrenzen hinweg): Wenn in einem gegebenen TrueType-Font dasjenige Zeichen, das im zugrundeliegenden Windowssystem als *a* erscheint (Zeichen Nr. 97), durch ein griechisches *alpha* (α) ersetzt ist, dann hängt die Übermittlung der entscheidenden Information, daß eben ein griechisches *alpha* und kein lateinisches *a* gemeint ist, davon ab, ob der korrekte Font vorhanden und anwählbar ist oder nicht; eine Information, die beim Datentransfer zwischen verschiedenen Systemen, aber auch zwischen verschiedenen Anwenderprogrammen auf ein und demselben System viel zu leicht verloren geht. Besonders katastrophal wirkt sich dies aus, wenn innerhalb eines Textes mehrere derartige Fonts mit "überlappenden" Codierungen nebeneinander benutzt sind (wenn also z.B. sowohl griechisches *alpha* als auch lateinisches *a* vorkommen und beide den Bytewert 97 haben). Im Zweifelsfall erscheint demgegenüber jede "unelegant" oder sogar "unbeholfen" wirkende Mehrbytecodierung, wenn sie Eindeutigkeit mit sich bringt (z.B. \$a für griechisches α), nach wie vor zweckmäßiger.

Solange wir also einer eindeutigen und systemunabhängig-einheitlichen Codierungsmöglichkeit für sämtliche in Frage kommenden Schriften und Transkriptionszeichen nicht näher gekommen sind, ist es für die im TITUS-Projekt erfaßten Textmaterialien noch nicht sinnvoll, ein bestimmtes "endgültiges" Format anzustreben. Statt dessen werden die Texte derzeit noch in verschiedenen Formaten "nebeneinander" gespeichert, wobei lediglich das Prinzip der eindeutigen Codierung obwaltet, durch das die Konvertierbarkeit bedingt ist. Unabhängig von dem jeweiligen "Ausgangsformat", d.h. demjenigen Format, in dem die Texte von den Beitragenden selbst erstellt werden, sollen die Texte in absehbarer Zeit v.a. in eine für ein sprachwissenschaftliches Retrieval geeignete Codierung gebracht werden. Von den hierfür zur Verfügung stehenden Systemen wird (auf DOS-Ebene) derzeit das von der Brigham-Young-University (Utah) entwickelte "Wordcruncher"-System bevorzugt, das — nach einmal erfolgter Durchindizierung auch großer Textmengen — einen enorm schnellen Zugriff auf einzeln oder im Verbund zu suchende Wortformen sowie eine bequeme Erstellung von Konkordanzen, Indizes etc. ermöglicht³. Ob das Wordcruncher-System das System

³ Die umfangreichste Textsammlung, die im Rahmen von TITUS bisher mit dem Wordcruncher aufbereitet wurde, ist das altenglische Corpus der University of Toronto, das einen Gesamttext von 25 MB (25 Mio. Zeichen) repräsentiert.

der Zukunft sein wird, dürfte allerdings wieder von marktpolitischen Faktoren abhängen, die nicht zuletzt das zugrundeliegende Betriebssystem (DOS) betreffen.

Obwohl die Probleme der Codierung und, davon abhängig, der Konvertierbarkeit den Fortgang des Projekts bis heute immer wieder behindert haben, konnte der Bestand an zur Verfügung stehenden Texten und Textsammlungen seit den Anfängen doch kontinuierlich ausgeweitet werden, so daß heute die Zielsetzung nicht mehr unrealistisch erscheint, bis zum Jahre 2000 über sämtliche für die Vergleichende Indogermanische Sprachwissenschaft relevanten Textmaterialien in elektronisch codierter Form zu verfügen. Während die Textfiles, die für eine Integration gewonnen werden konnten, in der Anlaufphase des Projekts eher "zufällig" dadurch bestimmt waren, daß sich jemand — ein Institut, ein Projekt, eine Privatperson — die Mühe gemacht hatte, sie — meist von Hand — einzugeben, hat sich in der Zwischenzeit in viel größerem Maße der Gedanke der Koordination und Kooperation durchgesetzt; im Normalfall bedeutet dies, daß sich potentielle Beitragende vor dem Beginn ihrer Mitarbeit nach den offenstehenden "Desiderata" erkundigen und ihren Beitrag dementsprechend auswählen. Dies hat auch den Vorteil, daß bestimmte Prinzipien, die die Eingabe betreffen — insbesondere dasjenige der eindeutigen Codierung, aber auch Fragen der Formatierung — vorab geklärt werden können, um so eine nachträgliche Anpassung, die oft recht mühsam "von Hand" erfolgen muß, vermeidbar zu machen. Wie sinnvoll ein solches Verfahren ist, mag ein Beispiel illustrieren:

Bereits seit einiger Zeit wird der Fachwelt das umfangreiche Corpus des altindischen Epos, des Mahābhārata, das unter der Leitung von M. TOKUNAGA an der Universität von Kyōtō / Japan eingegeben wurde, über das Internet zur Verfügung gestellt. Die zugehörigen Textfiles (insgesamt ca. 10 MB) enthalten (in einem 16-Bit-Code) den transkribierten Text entsprechend der meistbenutzten kritischen Ausgabe (ed. V.S. SUKTHANKAR / S.K. BELVALKAR, Poona 1933-1959) ohne Variantenapparat. Abgesehen von zahlreichen Transkriptionsfehlern, die nicht ausbleiben können, wenn eine derartige Eingabe von Hand gemacht wird⁴, hat die so verbreitete Fassung, die nunmehr auch in die TITUS-Sammlung integriert werden konnte, den schwerwiegenden Nachteil, daß in ihr nicht zwischen Wortgrenzen und den zwischen Kompositalgliedern bestehenden Morphemgrenzen unterschieden worden ist; man vgl. etwa die Notierung der Komposita *dhṛtarāṣṭro* (Eigenname, Nom.Sg.) und *mahārājah* ("Großkönig", Nom.Sg.) in folgender Verszeile (mit gegenübergestellter "normaler" Transkription):

0110010013/ 11,1,1c	dhRta.raaSTro.mahaa.raajah.zrutvaa.kim.akaron.mune.// <i>dhṛtarāṣṭro mahārājah śrutvā kim akaron mune //</i>	vs.
------------------------	---	-----

Diese Unterscheidung, die in verschiedenen anderen elektronisch bearbeiteten Sanskrittexten unterschiedlich durchgeführt worden ist (z.B. durch die Anwendung von Zeichen wie +, ^ oder _ : *dhṛta+rāṣṭro*, *mahā+rājah*), kann nun nicht mit automatischen Verfahren nachträglich appliziert werden; sie setzt statt dessen eine — bei der Größe des Gesamttexts höchst aufwendige — manuelle Weiterbearbeitung voraus, durch die zumindest alle vorkommenden Komposita als solche vorweg bestimmt werden müssen. Ein zweites Problem betrifft den sog. Sandhi, d.h. die in altindischen Texten grundsätzlich auftretende Erscheinung, daß Wortformen, wo sie im Satz aufeinander stoßen, lautlich aneinander angepaßt werden, wobei sich meist der Auslaut des ersten, seltener der Anlaut

⁴ Für die Anwendung von optischen Scannern gibt es bei der in indischen Ausgaben verwendeten Devanāgarī-Schrift bisher m.W. noch keine zielführenden Verfahren, und selbst wenn dies so wäre, würde auch das automatische Lesen erfahrungsgemäß zu zahlreichen Eingabefehlern führen.

des zweiten Wortes ändert. Bei den bisher verfügbaren elektronisch verarbeiteten Sanskrittexten werden in dieser Hinsicht wiederum völlig unterschiedliche Verfahren angewendet: Teilweise wird, wie auch in manchen gedruckten Texteditionen, der Sandhi "aufgelöst", d.h. es wird eine Normalform eingesetzt, die die kontextuellen Veränderungen nicht aufweist und somit von der tatsächlich überlieferten Wortgestalt abweicht (im o.g. Fall: *dhṛtarāṣṭraḥ mahārājaḥ śrutvā kim akarot mune*); oder aber es wird der Sandhi "beibehalten", was dazu führt, daß ein und dieselbe morphologische Einheit (z.B. der Nom.Sg. *mahārājaḥ* "der Großkönig") in mehreren verschiedenen lautlichen Formen im Text erscheint (z.B. *-rājas, -rājo, -rāja*), die nicht ohne weiteres gemeinsam abrufbar sind. Da die Äquivalenz zwischen einer gegebenen Sandhivariante und der zugrundeliegenden Normalform ("Pausaform") nicht in allen Fällen umkehrbar-eindeutig ist, ist wiederum keine völlig automatische Transposition der einen in die andere Form möglich. Wünschenswert wäre es in dieser Hinsicht, eine doppelte Repräsentation der Texte (mit beibehaltenem **und** aufgelöstem Sandhi) zu haben. Eine solche Paralleladaptation liegt innerhalb der TITUS-Sammlung inzwischen für die älteste altindische Textsammlung überhaupt, die Ṛgveda-Saṃhitā, vor, wobei dem (inzwischen mehrfach überarbeiteten und korrigierten) eins-zu-eins transkribierten Ausgangstext (s.o.) der von A. LUBOTSKY im Hinblick auf eine zu publizierende Textstellenkonkordanz erarbeitete, dem sog. "Pada-Pāṭha" nahekommende interpretative Text Vers für Vers gegenübergestellt ist; man vgl. folgendes Beispiel:

RV 1,1,6	{a}	(yád aṅgá dásúṣe tvám)
		[yát aṅgá dásúṣe tvám]
	{b}	(ágne bhadrám kariṣyási /)
		[ágne bhadrám kariṣyási /]
	{c}	(távét tát satyám aṅgiraḥ)
		[táva ít tát satyám aṅgiraḥ //]

An diesem Beispiel dürfte bereits klar geworden sein, daß für die elektronische Textdatenbank, auf die das TITUS-Projekt abzielt, weit mehr als eine bloße elektronische Erfassung der Texte in ihrer kanonischen, z.B. in maßgeblichen Editionen schriftlich niedergelegten Form von Interesse ist. Eine solche Encodierung kann in den meisten Fällen allenfalls eine erste Arbeitsgrundlage sein, auf der weitere Bearbeitungsschritte, die eine sprachwissenschaftliche Auswertung vorbereiten, aufzubauen haben. Derartige Weiterbearbeitungen werden den zukünftigen Ausbau des TITUS-Thesaurus im wesentlichen bestimmen. Es wird z.B. vielfach, wie im Falle des Altindischen, darum gehen, auf verschiedenen Ebenen gelagerte, durch unterschiedliche Transkriptionsmodalitäten abbildbare Interpretationen schriftlich überlieferter Formen aufeinander beziehbar zu machen. Neben dem Altindischen betrifft das in hohem Maße z.B. die in Keilschrift überlieferten Textmaterialien der Sprachen Altanatoliens, des Hethitischen und seiner Schwestersprachen, die für eine sprachwissenschaftliche Auswertung in ganz erheblichem Maße einer — gegenüber der in den meisten gedruckten Texteditionen gepflogenen rein transliterierenden Wiedergabe — interpretativ-analytischen Transkription hinsichtlich der zugrundeliegenden Lautgestalt bedarf.

Des weiteren ist daran zu denken, daß elektronische Verfahren im gegebenen Zusammenhang auch dort sinnvoll zu einer Erkenntniserweiterung führen können, wo es um die Erschließung einer authentischen Textgestalt aus divergierenden handschriftlichen Traditionen selbst geht. So ist z.B. wahrscheinlich, daß eine vollständige Auswertung des Handschriftenmaterials mit elektronischen Verfahren bei Corpora wie dem (altiranischen) Avesta in zahlreichen Punkten zu einer Verbesserung des Kenntnisstands gegenüber der manuell erarbeiteten gedruckten Textedition führen wird, da es mit elektronischen Mitteln wesentlich leichter ist, sämtliche divergierenden Schreibweisen, die

in den Handschriften auftauchen, auf ihre Konsistenz hin zu überprüfen und damit zwischen singulären, für die Textherstellung relevanten Lesarten und etwaigen, weniger relevanten, Marotten oder Eigenheiten bestimmter Schreiber zu differenzieren.

Um gezielte Auswertungen im Hinblick auf morphosyntaktische Fragestellungen zu ermöglichen, wird es darüber hinaus erforderlich sein, parallel zu den eigentlichen Textmaterialien abrufbare grammatische Angaben zu integrieren. Im Hinblick auf derartige Angaben ("morphologisches Tagging"), wie sie außerhalb des TITUS-Projekts etwa zu verschiedenen Bibeltexten verfügbar sind, ist freilich zu berücksichtigen, daß verschiedene der hier interessierenden Sprachen als ausgesprochene Spezialgebiete anzusehen sind, bei denen eine *communis opinio* im Hinblick auf die Bewertung morphologischer oder syntaktischer Verhältnisse in vielen Fällen nicht zu erzielen sein wird. Hier wird wiederum vieles davon abhängen, ein möglichst flexibles Retrievalsystem zu haben, das auch alternative Auffassungen und Interpretationen zu verarbeiten zuläßt. So sollte z.B. ein morphologisches Tagging im Falle des Avesta die durch die Homographie gegebene Mehrdeutigkeit von Formen wie (altavest.) *vohū* (Adjektiv "gut" im Nom.Akk.Sg.ntr., Instr.Sg.mask./ntr. oder Nom./Akk.Pl.ntr. oder Substantiv ntr. "das Gut" im Nom.Sg., Instr.Sg. oder Nom./Akk.Pl.) nicht unterdrücken, sondern als solche zu erkennen geben (vgl. das in Tafel 1 wiedergegebene Beispiel). Als "automatisches" Zusatzresultat einer derart aufbereiteten Textsammlung sind umfassende elektronische Wörterbücher zu erwarten, die nicht nur die tatsächliche Beleglage aller morphologischen Varietäten dokumentieren, sondern zugleich auch — im Sinne von Indizes — Referenzen auf die Textstellen selbst enthalten. Damit werden sich zugleich völlig neue Perspektiven für die vergleichend-sprachwissenschaftliche Beschäftigung mit dem Material eröffnen, insofern z.B. die lautlichen Verhältnisse, die zwischen einzelnen der beteiligten (Corpus-)Sprachen bestehen, durch eine in phonologischer, morphologischer und syntaktischer Hinsicht vollständige Erfassung des Materials in ganz anderem Maße überprüfbar werden als bisher (etwa im Hinblick auf Lautgesetze und Ausnahmen davon sowie im Hinblick auf die relative Chronologie sprachlicher Veränderungen).

Von einer umfassenden Bearbeitung der Textmaterialien im genannten Sinne ist das TITUS-Projekt momentan natürlich noch weit entfernt. Was in absehbarer Zukunft zum Abschluß gelangen soll, ist, wie gesagt, als erster Schritt zunächst eine elektronische Erfassung der relevanten Texte selbst, wobei auch die oben als notwendig dargestellte Einheitlichkeit und Eindeutigkeit zunächst noch nicht in allen Fällen gewährleistet ist. Eine Aufstellung, die den derzeitigen Stand der Texterfassung dokumentieren soll, ist im Anhang beigegeben.

Daß eine erschöpfende elektronische Bearbeitung von philologisch problematischen Textmaterialien der genannten Art in ganz erheblichem Maße über das hinausgeht, was etwa im Falle von Textcorpora moderner, gesprochener Sprachen gilt, und im Normalfall nach wie vor als eine eigene wissenschaftliche Leistung anzusehen ist, dürfte außer Frage stehen. Damit ist auch der wichtigste Grund gegeben, warum die im Rahmen des TITUS-Projektes erarbeiteten Textmaterialien derzeit nicht frei und für jedermann zugänglich, etwa über das internationale Netz, abrufbar sind, sondern lediglich dem Kreis der Beiträger (im Sinne einer Arbeitsgruppe) für ihre wissenschaftliche Arbeit zur Verfügung stehen. Einer Publikation von Einzelauswertungen und -Ergebnissen steht damit natürlich nichts im Wege.

Gerade in jüngster Zeit hat das TITUS-Projekt noch eine erhebliche Weiterung erfahren. Bei einem informellen Treffen an der Universität Wien wurde Anfang dieses Jahres die Möglichkeit eruiert,

Y. 29,1c:

<Textherstellung:>

{Y.29,1c:}

{Y.29,1c: Gdn.}

<Lesarten (Auswahl):>

{Y.29,1c: K5}

{Y.29,1c: J2}

{Y.29,1c: Pt4}

{Y.29,1c: S1}

<Textbearbeitungen:>

{Y.29,1c: Bthl.}

{Y.29,1c: Insl.}

{Y.29,1c: Kell.}

{Y.29,1c: Humb.2}

{Y.29,1c: Monna}

nōiṭ. mōi. vāstā. xšmaṭ. aniiō. aḍā. mōi. sṣtā. vohū. vāstriiā.ṽ

nōiṭ. mōi. vāstā.¹² xšmaṭ. aniiō. aḍā. mōi. sṣtā.¹³ vohū. vāstriiā.¹⁴ṽ

nōiṭ. mōi. vāstrā. xšmaṭ. /aniiō. aḍā. mōi. sṣstrā. vohū. vāstr/iiā.ṽ

nōiṭ. mōi. vāstā. xšmaṭ. aniiō. aḍā. mōi. sṣstrā. vohū. vāstriiā.ṽ

nōiṭ. mōi. vāst'ā. xšmaṭ. aniiō. aḍā. mōi. sṣtā. vohū. vāstriiā.ṽ

nōiṭ. mōi. .xšmaṭ. aniiō. aḍā. mōi. sṣtā. vohū. vāstriiā.ṽ

nōiṭ. mōi. vāstā. xšmaṭ. anyō. aḍā. mōi. sṣtā. vohū. vāstryā.

nōiṭ mōi vāstā xšmaṭ anyō aḍā mōi sṣtā vohū vāstr[y]ā (7+9)

nōiṭ mōi vāstā xšmaṭ aniiō, aḍā mōi sṣtā vohū vāstriiā/vāstrā? 7+10/9

nōiṭ mōi vāstā xšmaṭ aniiō aḍā mōi sṣtā vohū vāstriiā 7+10

nait mai vāstā (x)šmat anyah āt mai sṣta vahū vāstriyā 7+9/10

	Formenbestimmung	phonolog.	Lemma
nōiṭ	C	naṭ	nait
mōi	Eds	maṭ	az-am
vāstā *vāstrā	Nnsm (Nism) (Nisn) (Nnpn) (Napn)	uāstā uāstrā uāstrā uāstrā uāstrā	uāstār- uāstra-
xšmaṭ	Pbp	xšmat	ṭūž-am
aniiō	Ansm	anjah	anja-
aḍā	D	aḍā	aḍā
mōi	Eds	maṭ	az-am
sṣtā	V2pApA	sansta	√sand-
vohū	Aisn (Ansn) (Aasn) (Anpn) (Aapn) (Aism) (Nnsn) (Nasn) (Nisn) (Nnpn) (Napn)	uahu uahu uahu uahu uahu uahu uahu uahu uahu uahu	uahu-
vāstriiā *vāstrā	Nisn (Nnpn) (Napn) (Nisn) (Nnpn) (Napn) (Nism)	uāstriiā uāstriiā uāstriiā uāstrā uāstrā uāstrā uāstrā	uāstrija- uāstra-

Tafel 1

unter Ausnutzung der durch das Internationale Datennetz vorgegebenen Kapazitäten ein umfassendes Informationssystem einzurichten, das sich auf alle Bereiche aktueller Informationen aus der Indogermanistik bzw. Vergleichenden Sprachwissenschaft erstrecken soll. Über einige auf dem WWW-Server der Universität Frankfurt abgelegten HTML-Seiten (s. die o.g. URL) ist dieser Informationsdienst seit einigen Wochen erreichbar; in baldiger Zukunft werden sich sprachwissenschaftliche Institute der Universitäten Prag, Kopenhagen, Leiden, Wien u.a. beteiligen. Angeboten werden sollen dann — neben Mitteilungen über das eigentliche TITUS-Projekt — bibliographische, curriculare, personenbezogene u.ä. Informationen, die das gesamte Gebiet der Vergleichenden Indogermanischen Sprachwissenschaft wie auch angrenzende Gebiete betreffen. Das Gelingen dieses Vorhabens wird ganz wesentlich von der Bereitschaft der Benutzer zur Mitarbeit abhängen.

Anhang:

Das TITUS-Projekt:

Dokumentation der Texte (Stand: 1.6.95)

Alt-, mittel- und neuindisch:*vedisch:*

Ṛgveda-Saṃhitā:

Eingabe des Textes unter der Leitung von W.P. LEHMANN (Austin, Texas) durch H.S. ANANTHANARAYANA (Hyderabad); Überarbeitungen durch S.D. ATKINS (Pomona, California), G.E. DUNKEL (Zürich), J. GIPPERT und F.J. MARTÍNEZ GARCÍA (Frankfurt)

Ṛgveda-Pāṭha (sandhifreie Version):

Eingabe durch A. LUBOTSKY (Leiden)

Ṛgveda-Khīlāni:

Eingabe durch Carlos JORDÁN CÓLERA (Zaragoza; in Bearbeitung)

Aitareya-Brāhmaṇa:

Eingabe durch J. GIPPERT und F.J. MARTÍNEZ GARCÍA (Frankfurt)

Sāmaveda-Saṃhitā:

Eingabe durch F.J. MARTÍNEZ GARCÍA (Frankfurt; in Vorbereitung)

Gobhīla-Gr̥hya-Sūtra:

Eingabe durch Carlos JORDÁN CÓLERA (Zaragoza); Überarbeitungen durch F.J. MARTÍNEZ (Frankfurt)

Kāṭha-Saṃhitā:

Eingabe durch Chl.H. WERBA (Wien; in Bearbeitung)

Taittirīya-Saṃhitā:

Eingabe durch M. FUSHIMI (Ōsaka)

Taittirīya-Brāhmaṇa:

Eingabe durch M. FUSHIMI (Ōsaka); Kollationierung durch Y. IKARI (Kyōtō)

Taittirīya-Prātiśākhya:

Eingabe durch M. FUSHIMI (Ōsaka)

Āpastamba-Śrauta-Sūtra:

Eingabe durch M. FUSHIMI (Ōsaka)

Baudhāyana-Śrauta-Sūtra:

Eingabe durch M. FUSHIMI (Ōsaka)

Vājasaneyi-Saṃhitā:

Eingabe durch C.-M. BUNZ (Saarbrücken; in Bearbeitung)

Śatapatha-Brāhmaṇa (Mādhyam̐dina-Rezension):

Eingabe durch G. KNOLL (Frankfurt; in Vorbereitung)

Śatapatha-Brāhmaṇa (Kāṇva-Rezension):

Eingabe durch J. GIPPERT (Frankfurt; in Bearbeitung)

Atharvaveda-Saṃhitā:

Eingabe durch V. PETR; Kollationierung durch P. VAVROUŠEK (Prag; in Bearbeitung)

episches und klassisches Sanskrit:

Mahābhārata:

Eingabe durch Muneo TOKUNAGA (Kyōtō)

Rāmāyaṇa:

Eingabe durch Muneo TOKUNAGA (Kyōtō)

- Tantrākhyāyika (Buch 1 und 2):
Eingabe durch L.M. FOSSE (Oslo)
- Pañcatantra:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Hitopadeśa:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Kālidāsa, Kumārasambhava:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Kālidāsa, Meghadūta:
Eingabe durch J. GIPPERT (Frankfurt)
- Kālidāsa, Ṛtusamhāra:
Eingabe durch J. GIPPERT (Frankfurt; in Bearbeitung)
- Nalopakhyaṇa:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Dandin, Daśakumāracarita:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Harṣacarita:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
- Vikramacarita:
Eingabe durch P. OLIVIER (Frankfurt; in Bearbeitung)
- mittelindisch:*
Mahāvamśa:
Eingabe durch P. OLIVIER (Frankfurt; in Vorbereitung)
- Alt-, mittel- und neuiranisch:**
- avestisch:*
Avesta-Gesamtkorpus:
Eingabe durch S. GIPPERT-FRITZ und J. GIPPERT (Frankfurt)
- Nirangistān:
Eingabe nach der Edition WAAG durch H. KUMAMOTO (Tōkyō)
- altpersisch:*
Altpersisches Gesamtkorpus:
Eingabe unter der Leitung von G.E. DUNKEL durch S. GINDRO, S. SCARLATA, P. WIDMER (alle Zürich);
Ergänzungen und Korrekturen durch G. KEYDANA (Münster); Überarbeitung durch J. GIPPERT (Frankfurt);
- parthisch:*
Manichäische Texte:
Eingabe nach M. BOYCE, Reader durch J. GIPPERT; Überarbeitung durch D.N. MACKENZIE (Göttingen)
- mittelpersisch:*
Manichäische Texte:
Eingabe nach M. BOYCE, Reader durch J. GIPPERT; Überarbeitung durch D.N. MACKENZIE (Göttingen)
- Kārnāmag-i Ardašīr-i Pābagān:
Eingabe durch D.N. MACKENZIE (Göttingen)
- Mēnōg-i xrad:
Eingabe durch D.N. MACKENZIE (Göttingen)
- Arda-virāf-nāmag:
Eingabe durch P. VAVROUŠEK (Prag)
- Ayādgar-i Zarērān:
Eingabe durch A. CANTERA (Berlin; in Bearbeitung)
- Videvdād-Pahlavī-Übersetzung:
Eingabe durch A. CANTERA (Berlin; in Bearbeitung)
- Vizīdagihā-i Zādspram:
Eingabe durch L. PAUL (Göttingen; in Bearbeitung)
- khotan-sakisch:*
Khotan-sakisches Corpus:
(Khotanese Buddhist Texts; Khotanese Texts 1-5; Book of Zambasta): Eingabe durch R.E. EMMERICK (Hamburg); Überarbeitung durch H. KUMAMOTO (Tōkyō)
- Book of Zambasta:
Eingabe durch P. VAVROUŠEK (Prag)
- sogdisch:*
Sogdisches Corpus:
(Alles publizierte Material außer MÜLLER-LENTZ, Sogd.Texte II, Text 1-3): Eingabe durch N. SIMS-WILLIAMS (Cambridge)

NACHRICHTEN

Buddhistische und Manichäische Texte:
Eingabe durch Y. YOSHIDA (Tōkyō)

neupersisch:

Qabūli, Ġazals
Eingabe durch M. GLÜNZ (Bern / Seattle)

Anatolisch:

hethitisch:

Hethitische Ritualtexte:
Eingabe durch Chr. ZINKO (Graz)

Hethitisches Corpus:
Eingabe durch P. VAVROUŠEK (Prag) und H.C. MELCHERT (Chapel Hill, N.C.; in Bearbeitung)

luvisch:

Luvisches Corpus:
Eingabe durch J. TISCHLER (Dresden)

palaisch:

Palaisches Corpus:
Eingabe durch J. TISCHLER (Dresden)

lydisch:

Lydisches Corpus:
Eingabe durch J. TISCHLER (Dresden)

lykisch:

Lykisches Corpus:
Eingabe durch J. TISCHLER (Dresden)

Tocharisch:

A-tocharisch:

A-Tocharisches Corpus:
Eingabe durch O. HACKSTEIN (Halle; in Bearbeitung)

B-tocharisch:

B-Tocharisches Corpus:
Eingabe durch C. SCHAEFER (Berlin; in Bearbeitung)

Armenisch:

altarmenisch:

4 Evangelien und Psalter:
Eingabe nach der Zohrab-Bibel durch H. PALANDJIAN (Montreal)

4 Evangelien:
Eingabe nach der Edition KÜNZLE durch J. WEITENBERG (Leiden)

Agat'angelos:
Eingabe durch J. GIPPERT und R.-P. RITTER (Frankfurt; in Bearbeitung)

Movsēs Xorenac'i:
Eingabe durch H. PALANDJIAN (Montreal)

Patmowt'iw n vrac':
Eingabe durch H. PALANDJIAN (Montreal)

Saraknoc':
Eingabe durch H. PALANDJIAN (Montreal)

Baltisch:

litauisch:

Duonelaitis, Metai:
Eingabe durch G. KEYDANA (Münster)

Germanisch:

gotisch:

Gotische Bibel:
Eingabe durch W. GRIEPENTROG (Nürnberg)

altenglisch:

Altenglisches Corpus:
Eingabe an der University of Toronto

althochdeutsch:

Isidor:

Eingabe durch M^a Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. BRYSCH (Dresden) und R. SCHUHMANN (Gießen)

Benediktiner-Regel:

Eingabe durch M^a Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. BRYSCH (Dresden); Korrekturlesungen durch A. POTTHOFF-KNOTH und Roland SCHUHMANN (Gießen)

Tatian:

Eingabe durch M^a Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. Brysch und R. Schuhmann (Gießen); Korrekturlesungen durch S. ZEILFELDER (Jena) und A. POTTHOFF-KNOTH (Gießen)

Tatian: Præfatio Victoris Capuani / Übersicht der Capitelüberschriften / Pariser Fragmente:

Eingabe durch R. SCHUHMANN (Gießen); Weiterbearbeitung durch J. BRYSCH (Dresden)

Otfrid:

Eingabe unter der Leitung von R. LÜHR (Jena) durch R. SCHUHMANN und M. BAYER (Gießen); Weiterbearbeitung durch J. BRYSCH (Dresden); Korrekturlesungen durch A. POTTHOFF-KNOTH, R. SCHUHMANN, A. HOLZHÄUER (Gießen) und S. ZEILFELDER (Jena)

Notker:

Eingabe unter der Leitung von R. LÜHR (Jena) durch K. LEPPER (Gießen) und S. ZEILFELDER (Jena)

Griechisch:*klassisch:*

Septuaginta:

Eingabe an der University of Pennsylvania (CCAT)

Neues Testament:

Eingabe an der University of Pennsylvania (CCAT)

Italisich:*oskisch:*

Tabula Bantina:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Cippus Abellanus:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Oskische Inschriften:

Eingabe durch V. SLUNEČKO (Prag; in Bearbeitung)

umbrisch:

Tabulae Iguvinae:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Umbrische Inschriften:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Keltisch:*alt- und mittelirisch:*

Würzburger Glossen:

Eingabe durch J. GIPPERT (Frankfurt) und D. DURKIN (Münster; in Bearbeitung)

Aided Oenfir Aife

Eingabe durch D. DURKIN (Münster)

Compert Con Culainn

Eingabe durch D. DURKIN (Münster)

De chophur in da muccida

Eingabe durch D. DURKIN (Münster)

Esnada Tige Buched

Eingabe durch D. DURKIN (Münster)

Fled Dúin na nGéd

Eingabe durch D. DURKIN (Münster)

Fingal Rónáin

Eingabe durch D. DURKIN (Münster)

Orgain denna rí

Eingabe durch D. DURKIN (Münster)

Orgguin trí Mac Diarmata Mic Cerbaill
 Eingabe durch D. DURKIN (Münster)
 Scela Cano Meic Gartnain
 Eingabe durch D. DURKIN (Münster)
 Scela Mucce Meic Dathó
 Eingabe durch D. DURKIN (Münster)
 Serglige Con Culainn
 Eingabe durch D. DURKIN (Münster)
 Tainbó Froích
 Eingabe durch D. DURKIN (Münster)
 Togail Bruidne dá derga
 Eingabe durch D. DURKIN (Münster)

Rest- und Trümmersprachen:

phrygisch:

Phrygisches Corpus:
 Eingabe durch A. LUBOTSKY (Leiden)

Ein Dinosaurier im Internet: Maschinelle Übersetzung mit SUSY online

Seit dem Sommer 1994 existiert⁵ eine frei zugängliche Benutzungskennung auf einem Solbourne-Rechner (Betriebssystem UNIX) des Rechenzentrums der Universität des Saarlandes, über die man dem maschinellen Übersetzungssystem SUSY deutsche Sätze zur Übersetzung ins Englische (oder russische zur Übersetzung ins Deutsche) eingeben kann.

SUSY ist ein Dinosaurier unter den Maschinellen Übersetzungssystemen. Seine Ursprünge gehen zurück in die 60er Jahre, entwickelt wurde das System von 1972 - 1986 in einem Teilprojekt des Sonderforschungsbereichs 100 "Elektronische Sprachforschung" an der Universität des Saarlandes in Saarbrücken, der die Initialzündung für die Etablierung der Computerlinguistik in Saarbrücken gab. Praktische Ergebnisse des SFB 100 waren u. a. die Entwicklung der Programmiersprache COMSKEE und die Schaffung großer Computerlexika, die auch heute noch an verschiedenen Universitäten Grundlage computerlinguistischer Forschung sind und in große lexikalische Datensammlungen Eingang gefunden haben.

Die bekannteste Entwicklung aber war SUSY, ein "universelles, multilinguales maschinelles Übersetzungssystem", zu dem verschiedene Projektgruppen (für Deutsch, Englisch, Französisch und Russisch) beigetragen haben. Linguistische Grundlage war die Dependenzgrammatik, programmiert wurde das Ganze in FORTRAN, die linguistischen Daten wurden in einem hybriden Formalismus repräsentiert. Aufgrund dieser Konstellation versteht es sich von selbst, daß die SUSY-Algorithmen heute nicht mehr weiterentwickelt werden - aber in seiner heutigen Gestalt als "black box" taugt das System durchaus noch als Demonstrationsobjekt (und als Vergleichsobjekt für die modernen MÜ-Systeme), und die Wörterbücher und die linguistischen Daten sind nach wie vor erweiter- und veränderbar.

Wer sich auf dem open account SUSYTEST einloggt (vgl. unten), wird mit einigen Warnungen empfangen, Tenor: Man sollte nicht zuviel erwarten! Zunächst einfache Strukturen und im Erfolgsfall kompliziertere Strukturen eingeben! Und Geduld sollte man auch aufbringen!

⁵ dank freundlicher Mithilfe von Joachim Blum vom Rechenzentrum der Universität des Saarlandes

Für wen lohnt sich denn eine Beschäftigung mit SUSY, wenn die Ergebnisse - auch mit einigem Wohlwollen - alles andere als publikationsreif sind? Jeder, der SUSYs Fehlern nachspürt, kann mehr über das Funktionieren der natürlichen Sprache lernen, als der Fall wäre, wenn SUSY nur richtige Übersetzungen liefern würde. Wer der Ursache einer mißglückten SUSY-Übersetzung auf die Spur kommt (und von den systembedingten Fehlern abstrahiert), kann dabei z.B. etwas lernen über die Bedeutung lexikalischer Informationen oder die Schwierigkeit, grammatisches Wissen darzustellen, Wissen über sprachliche Elemente und Strukturen. Wenn dies auch heute in moderneren Formalismen dargestellt und mit benutzungsfreundlicheren Werkzeugen verarbeitet wird als vor 20 Jahren, so läßt die Arbeit mit SUSY doch erahnen, wie detailliert Regeln und wie umfangreich Wörterbücher sein müssen, damit auch nur ein winziger Ausschnitt der natürlichen Sprache vom Computer "verarbeitet" werden kann.

Es kann beim Arbeiten mit SUSY nicht darum gehen, die Übersetzungsergebnisse auf ihre Verwertbarkeit hin zu überprüfen. Die Möglichkeit, eine Kaskade von Fehlern bis zum auslösenden Fehler zurückzuverfolgen und evtl. durch Ergänzen eines Lexikoneintrags zu beseitigen, das könnte den Reiz von SUSY ausmachen.

SUSY ist modular aufgebaut, man kann einzelne Module für besondere Aufgaben einsetzen. Abgesehen davon, daß das Ergebnis jedes einzelnen Analyseschritts (morphologische Analyse, Satzsegmentanalyse, Nominalanalyse etc.) gesondert darstellbar ist, kann man SUSY außer für maschinelle Übersetzung auch für andere Zwecke, z.B. die automatische Indexierung einsetzen, d.h. für die Reduzierung deutscher Texte auf Grundformen. Hierfür ist das deutsche Lexikon mit seinen 142.000 Einträgen (vorwiegend Wortstämme) eine gute Grundlage. Und schließlich läßt sich das deutsch-englische Übersetzungswörterbuch mit seinen nach Fachgebieten differenzierten 350.000 Einträgen ganz unabhängig von der maschinellen Übersetzungskomponente als Terminologiepool nutzen.

Beispiele für SUSY-Anwendungen:

- > morphologische Analyse: Flexion, Derivation, Komposition
- > Disambiguierung auf Wort-, Syntagmen- und Satzebene
- > Analyse komplexer Sätze
- > Analyse von Nominalphrasen: elementare NPs, Attribute, Koordination
- > Prädikat/Argument-Strukturen
- > Übersetzungsprobleme:
 - Präpositionen
 - Wortstellung
 - PTZ- Konstruktionen

Seit dem Herbst 1994 haben sich ca. 180 Internet-Reisende aus allen Teilen der Welt auf dem Solbourne-Rechner eingeloggt, um SUSY zu testen, u.a. auch im Rahmen von Lehrveranstaltungen zur Einführung in Probleme der Maschinellen Übersetzung und der Syntaxanalyse (Universitäten Bonn und Göteborg).

Man erreicht SUSY mit der folgenden Prozedur:

```
telnet sbusol.rz.uni-sb.de
login: susytest
Password: susyguest
```

Wer sich mit SUSY beschäftigt, erfährt nichts über den **heutigen** Stand der Maschinellen Übersetzung. Einer der Unterschiede zu den jetzt eingesetzten bzw. in Entwicklung befindlichen Systemen ist der, daß man mittlerweile erkannt hat, daß "universelle", also

möglichst viele Sprachpaare, Anwendungen und Fachgebiete einkalkulierende Systeme wie SUSY zu weit greifen und vielmehr gezielte Anwendungen eher praktischen Erfolg versprechen.

Heinz-Dirk Luckhardt, (dlu@rz.uni-sb.de)

10 Jahre IAI

Im Mai 1995 feierten die Gesellschaft zur Förderung der Angewandten Informationsforschung (GFAI) und das gleichnamige Institut IAI ihren 10. Geburtstag.

Eingerichtet wurde das kleine Forschungsinstitut im historischen Wartburg-Gebäude im Herzen Saarbrückens zur Durchführung des deutschen Anteils am EU-Projekt EUROTRA zur maschinellen Übersetzung. Sein Gründungsdirektor Prof. Harald Zimmermann installierte damit eines der ersten sogenannten An-Institute der Universität des Saarlandes, deren Aufgabe die Umsetzung von Grundlagenforschung in anwendungsorientierte Konzepte und Prototypen ist. Zunächst wurden hier eine Reihe hochwertiger Arbeitsplätze für Wissenschaftler mit Forschungsgeldern von Bundesministerien und der EU geschaffen, die sich auf dem komplexen Gebiet der automatischen Verarbeitung geschriebener Sprache schnell einen internationalen Ruf erworben haben.

Mit spezialisierten Versionen einer maschinellen Übersetzungshilfe (z.B. für medizinische Diagnosen oder Auto-Reparaturmeldungen) gelingt es nun in den letzten Jahren, zunehmend auch den Schritt zum Transfer solcher Ergebnisse in die Industrie zu tun, in der Form gemeinsamer Projekte und einzelner Direkt aufträge. Ähnliches gilt für ein Programm zur Rechtschreib- und Grammatikprüfung deutscher Texte, das gegenüber den derzeit auf dem Markt verfügbaren Systemen eine weitaus größere linguistische Intelligenz aufweist. So ist das IAI heute ein kompetenter Ansprechpartner der Industrie für Problemlösungen auf dem Gebiet der Sprachtechnologie.

Die GFAI (1. Vorsitzender: Karl-Albrecht Graeber) und der derzeitige geschäftsführende Direktor des IAI, Prof. Johann Haller, wollen diesen Weg mit weiteren Pilotpartnern und neuen Anwendungen der Sprachsoftware z.B. in Lernprogrammen und Hilfen für Behinderte konsequent weitergehen. Gleichzeitig soll auch die Verbindung mit der Grundlagenforschung erhalten bleiben: so ist das IAI an dem visionären Projekt Verbomobil beteiligt, das bis zum Jahr 2000 den Prototyp eines automatischen Dolmetschers erstellen will.

Das IAI hat seinen Standort in der Mitte zwischen Grundlagenforschung und Produkterstellung. Seine Stärke liegt im Transfer der Ergebnisse eigener und universitärer Forschung in Richtung kommerziell nutzbarer Pakete.

Das IAI erstellt, bewertet und testet Forschungsprototypen. In Zusammenarbeit mit künftigen Anwendern werden sie zu präindustriellen Labor-Prototypen weiterentwickelt. Beispiele hierfür sind:

- > ein **Grammatikprüfer**, der Rechtschreib- und Grammatikfehler entdeckt und Korrekturvorschläge macht

- > ein Programm zum Textverstehen, das die wichtigsten Schlagworte aus Briefen und Dokumenten automatisch auswählt

- t> ein experimentelles System zur multilingualen Maschinellen Übersetzung (CAT2), das neben Deutsch, Englisch und Französisch auch für Russisch und Koreanisch verfügbar ist.

Außerdem führt das IAI für öffentliche Auftraggeber Studien und Bewertungen z.B. zu der Verarbeitung paralleler Texte in mehreren Sprachen oder zur kommerziellen Maschinellen Übersetzung durch.

Johann *Haller*, IAI Saarbrücken

10 Jahre IuD an der Fachhochschule Darmstadt

Zum Wintersemester 1985/86 begann der damals neugegründete Fachbereich *Information und Dokumentation* an der Fachhochschule Darmstadt mit dem Vorlesungsbetrieb unter sehr provisorischen Verhältnissen.

Untergebracht in einer umgerüsteten ehemaligen Verlagskantine, von den Betroffenen liebevoll *Hundehütte* genannt und mit einem Anfangsbestand an Hochschullehrern wurde damit der Startschuß für eine mittlerweile im Hochschulbereich und in der Berufspraxis anerkannte und etablierte Ausbildung gegeben. Daß an BewerberInnen für diesen NC-Studiengang kein Mangel ist und die "AbsolventInnen beim Arbeitsamt so gut wie nicht in Erscheinung treten" (Zitat des Arbeitsamtes) ist als ein Erfolg zu werten, denn der Fachbereich hat sich von Anfang an nicht in eingefahrenen Gleisen bewegt: Der akademische Abschluß des *Diplom-Informationswirtes* mußte erst in der Genehmigungsbürokratie durchgesetzt und in der Berufspraxis etabliert werden. Die strategische Entscheidung, die Anwendung moderner Informationstechnologien gegenüber traditionellen bibliothekarischen Lehrinhalten in den Vordergrund zu stellen, hat sich als erfolgreich herausgestellt. Informationsspezialisten Darmstädter Prägung finden sich nach ihrem 8-semesterigen Studium in Unternehmensberatungen (größter Abnehmer!), bei Banken, Versicherungen, Verlagen, Rundfunkanstalten, im Schulungsbereich, letztlich überall, wo professionell mit Information umgegangen wird. Nur ein Bereich muß auf Diplom-Informationswirte verzichten: Der öffentliche Dienst kann FachhochschulabsolventInnen kein konkurrenzfähiges Gehalt anbieten.

Unabhängig von der Branche gibt es drei prototypische berufliche Orientierungen für Informations-Spezialisten:

- > Informationsvermittlung bis hin zu Consulting
- > Informationsmodellierung und -produktion. Das "Produkt Information" löst sich von der Software: Nicht die Rechner, sondern die Daten und deren Verwendbarkeit sind das Entscheidende.

> Tätigkeit als Integrator, projektorientiert, konstruktiv.

Spezielle neuere Tätigkeitsfelder sind:

- > Information Consulting.
- > Organisationsberatung, einschließlich Informations- und Kommunikationsanalysen.
- > Konzeption und Entwicklungen im Bereich neuerer multimedialer Informationssysteme, was die informationelle Seite betrifft.
- > Dokumentenverwaltungs- und Vorgangsverwaltungssysteme. Hier speziell als Integratoren und Kommunikatoren zwischen Anwendern und Informatikern.

Es dürfte nur wenige Fachbereiche geben, die über sich selbst und über ihre Absolventen so gut informiert sind wie der Fachbereich IuD: Zwei empirische Untersuchungen über den Verbleib der AbsolventInnen im Berufsfeld und eine Evaluierung des Fachbereichs im Rahmen des hessischen Modellprogramms *Lehrberichte* liegen vor.

Der Fachbereich IuD wird anlässlich seines 10-jährigen Bestehens eine Festschrift in der Reihe des Hochschulverbandes Informationswissenschaft (HI) herausbringen und eine Veranstaltung vom 4. - 6. Oktober 1995 anbieten: Informationsmanagement - Informationsmärkte - Informationsgesellschaft. 10 Jahre Hochschulausbildung für einen Schlüsselbereich der wirtschaftlichen Entwicklung (siehe Seite 83).

Kontakt: Fachhochschule Darmstadt, Fachbereich IuD, Prof. Dr. G. Knorz (Dekan), Haardtring 100, D-64295 Darmstadt, Tel.: (06151) 16-8401, Fax: -8980, e-mail: knorz@fh-darmstadt.de.

Kultur - Zeichen - Raum

Deutsche Gesellschaft für Semiotik schreibt Förderpreis Semiotik aus

Mittlerweile ist es eine gute Tradition, daß auf dem alle drei Jahre stattfindenden Internationalen Kongreß der Deutschen Gesellschaft für Semiotik ein Förderpreis für Semiotik verliehen wird. Sinn dieses Preises ist es, Nachwuchswissenschaftler zu ermutigen, sich mit semiotischen Themen auseinanderzusetzen.

Einsendeschluß ist der 31. März 1996. Der Preis von DM 5.000,- wird im August 1996 vergeben.

Aufgabe

Das Thema des Förderpreises weist auf das Thema des Kongresses in Amsterdam 1996 "Kultur - Zeichen - Raum" voraus. Dieses Thema versteht sich einerseits als Gegenstück zum Tübinger Kongreßthema "Zeichen und Zeit", andererseits greift es die immer drängender werdende Frage nach der Verfaßtheit kultureller Räume auf, nach deren Identifizierbarkeit und Grenzen. Die drei Begriffe des Themas können sechsfach kombiniert werden:

- > Kulturraum
- > Raumkultur
- > Raumzeichen
- > Zeichenraum
- > Kulturzeichen
- > Zeichenkultur

Die Preisfrage bezieht sich auf diese mehrfach ausdeutbare Konstellation: Gibt es Raum/Räume ohne Zeichen? Sind Kulturen/Räume zeichenhaft konstituiert? Ein besonderer Grund für die Wahl des Kongreßthemas war die Tatsache, da der Kongreß in Amsterdam stattfindet und so eine Thematisierung der Stadtsemiotik, der Zeichen im urbanen Raum nahelag.

Art der Ausführung

Möglich sind sowohl beschreibende als auch theoretisch formulierte Arbeiten. Zugelassen sind auch Gemeinschaftsarbeiten von mehreren Verfassern. Grundsätzlich soll die Untersuchung semiotisch orientiert und inter- bzw. transdisziplinär angelegt sein; sie kann durchaus von der Perspektive eines bestimmten Gegenstandsbereichs bzw. Fachgebiets ausgehen.

Vorgaben

Die Arbeit sollte in deutscher Sprache abgefaßt sein. Sie sollte bislang nicht veröffentlicht worden sein. Verbale Ausführungen sollten sich auf höchstens zweihundert Seiten beschränken. Dreidimensionale Ausführungen sollten in Form von Fotografien und Zeichnungen eingerichtet werden. Die Arbeiten sollten in dreifacher Ausfertigung vorliegen. Auch audiovisuelles Material sollte nach Möglichkeit in drei Kopien vorhanden sein. Bei Filmen bitten wir um drei Kopien auf Video-Kassetten.

Die Teilnehmer

Die Bewerber sollten bei Einsendeschluß (31.03.1996) nicht älter als 35 Jahre sein.

Die Jury

Die vom Vorstand eingesetzte Jury besteht aus einem Gremium von 4 namhaften Hochschulprofessoren. Die Jury bewertet die anonym vorgelegten Arbeiten und schlägt einen Preisträger vor.

Preisverleihung

Die Preisverleihung wird im Rahmen eines Festaktes während des Kongresses vorgenommen, wobei die preisgekrönte Arbeit vorgestellt wird. Die Höhe des Preises beläuft sich auf DM 5.000,-. Gestiftet von der Bayrischen Hypotheken- und Wechselbank HYPOBANK. *Verwendung der Einsendungen*
Titel und Verfasser der ausgezeichneten Arbeit werden in der "Zeitschrift für Semiotik", dem offiziellen Organ der DGS, veröffentlicht. Die ausgezeichnete Arbeit wird anlässlich des Kongresses in einer Pressekonferenz bekannt gemacht. Nonverbale Arbeiten werden auf dem Kongreß ausgestellt, bzw. an bekanntgegebenen Terminen gezeigt, wenn es sich um Film, Videostreifen oder Dias handelt. Schriftliche Texte werden (zumindest auszugsweise) vorgetragen. Die nicht ausgezeichneten Arbeiten werden zu Lasten des Eigentümers nach der Entscheidung der Jury zurückgesandt. Rückgabe der ausgezeichneten Arbeit wird mit dem Eigentümer entschieden.

Zusätzliche Informationen können bei der Geschäftsführerin des DGS angefordert werden:

Prof. Dr. Irmela Neu
Fachhochschule München
FB 10, Betriebswirtschaft und Tourismus
Am Stadtpark 20
81243 München

Wilhelm Weisweber:

Termersetzung als Basis für eine einheitliche Architektur in der maschinellen Sprachübersetzung. Das experimentelle MÜ-System des Berliner Projekts der EUROTRA-D-Begleitforschung (KITFAST). Sprache und Information, Bd. 28, Niemeyer, Tübingen 1994.

Mit einigem zeitlichen Abstand zum Ende des EUROTRA-Projekts und seiner "Begleitforschungsprojekte" in Bonn, Bielefeld, Stuttgart und Berlin erscheinen nunmehr Monographien, die Überblicke über die Systeme bzw. über Einzelthemen oder zusätzliche Versuche bieten. So wurde in einem der letzten LDV-Foren eine Arbeit über die Anwendung der EUROTRA-Philosophie (und -Software) auf die exotische Sprache Persisch rezensiert; heute liegt eine Studie von Wilhelm Weisweber zum KIT-FAST-Projekt an der TU Berlin vor, die im wesentlichen den Aspekt des Termersetzungsverfahrens behandelt. Dieses Verfahren sollte im experimentellen MÜ-System der Berliner Gruppe den meisten der Analyse- und Syntheseschritte zugrundegelegt werden. Obwohl der Rezensent für die formale Seite des Architekturkonzepts nur teilweise kompetent ist, kann gesagt werden, daß die im Buch geschilderte einheitliche Architektur lediglich zum Teil im System ausgeführt worden ist; aus diesem Grund hat das innovative Kernstück der Arbeit, das Kap. 6.3. (Rolle der Termersetzungssysteme in der MÜ) auch nur theoretischen Charakter.

Doch beginnen wir von vorne und systematisch: Die Kapitel 1 (Einführung) und 2 (Maschinelle Sprachübersetzung) sind eine kondensierte Einführung in das Berliner Projekt und seine Überlegungen zur MÜ-Architektur, über die sich der interessierte Leser hier rasch und vollständig informieren

kann; das Kap. 3 (Produktionssysteme) behandelt einige alternative Ansätze, deren vergleichende Wertung jedoch äußerst knapp ausfällt. Jedenfalls ist nicht direkt ersichtlich, inwieweit die dann in den Kapiteln 4-6 sehr ausführlich dargestellten Termersetzungssysteme die Schwierigkeiten der anderen Systeme überwinden können. Insbesondere für das Problem der Abhängigkeiten unter den TE-Regeln werden nur theoretische Lösungen angeboten. Warum muß z.B. dem Knuth-Bendix-Algorithmus ein ganzes Teilkapitel gewidmet werden, wenn er in der MÜ gerade nicht verwendet werden kann?

Überhaupt verliert man manchmal die Übersicht, ob der gerade behandelte (und formal immer sehr detailliert dargestellte) Teil-Algorithmus nun in der MÜ bzw. im KIT-FAST-System tatsächlich eingesetzt wird, theoretisch einsetzbar wäre oder aufgrund irgendwelcher prinzipieller Schwierigkeiten gar nicht einsetzbar ist. Viele Abschnitte sind so als Einzelkapitel interessant zu lesen, z.B. die Abschnitte 6.3. Termersetzungssysteme in der MÜ, 6.4. TE-Systeme in KIT-FAST: hier wird ein interessanter Ansatz vorgestellt, GPSG-Analyse als TE-System zu formalisieren. Aber gerade da, wo es spannend wird, bricht der Text ab und wird durch die Dokumentation der Implementierung (von Teilen!?) abgelöst.

Diese besteht aus mehreren Seiten Prolog Klauseln mit vielen sehr technischen Kommentarzeilen, die nur für den Spezialisten verständlich sind. Von größerem allgemeinem Interesse wären die Editoren und Compiler, deren Beschreibung jedoch auf 2 Seiten zusammengedrängt ist. Ebenso isoliert steht die Schlußseite über das MÜ-System.

Das Kap. 8 (Zusammenfassung und Ausblick) präsentiert relativ zusammenhanglos einige Ideen zur Weiterentwicklung, z.B. die

Einführung von Alternationen, was in allen derartigen Projekten und Systemen ein wichtiger Faktor ist. Die anderen Ideen werden nur allgemein angerissen (variable Analysetiefe, Desambiguierung, Reversibilität).

Die beiden Anhänge (Installations- und Benutzungshandbuch) - 100 von insgesamt 240 Seiten - sind nach Meinung des Rezensenten unnötigerweise auf teures Papier gebracht worden - wieso haben die Herausgeber das zugelassen? Das experimentelle MÜ-System ist (sehr lobenswert!) über ftp und Diskette verfügbar - die Anweisung, wie man sich die Dateien auf dem Server der TU Berlin holt, könnte aber die ganzen Anhänge ersetzen, die man dann als Benutzer auf den Files sowieso mit bekommt.

Die Lektoren und Herausgeber der Reihe und des Verlags haben auch eine Menge von z.T. sinnentstellenden und die Lektüre erschwerenden Tippfehlern übersehen - das müßte eigentlich im Zeitalter der automatischen Rechtschreibkorrektur nicht mehr sein:

S. 18: "Im KBMT-System wird zwar der Begriff Transfer vermeiden..." (das hätte allerdings auch das Korrekturprogramm nicht gefunden!)

S. 19: "die Übersetzung des Textes" etc. Ein weiterer Faktor, der bei der Lektüre zusätzlich ermüdet, sind die vielen Wiederholungen, sowohl stilistischer als auch argumentativer

Art:

S. 23: Angestrebte Architektur des MÜ-Systems. .

Die angestrebte Architektur ...

Das angestrebte MÜ-System

Schließlich scheint bei der Liste von Parametern zur MÜ-Architektur einiges mit der Kursivschrift nicht geklappt zu haben (S. 21/22): es ist z.B. aus der Liste nicht zu ersehen, ob KIT-FAST nun "Nur Satz-" oder "Satz- und Textrepräsentation" (oder beides?) auf einer Ebene oder auf jeder einzelnen Ebene bevorzugt, ob nun "Integration" oder "Trennung" von Satz- und Textrepräsentation gewählt worden ist. Das Ganze wird dann im folgenden Kapitel

noch einmal verbal entwickelt - ganz verstanden hat es der Rezensent jedoch immer noch nicht (kann an ihm liegen). Ähnliches passiert bei der kurzen Beschreibung von Regeln des SALAT-Systems (S. 35/36): sind die Regeln (1) und (2) nun Transfer- oder Syntheseregeln - oder beides?

An vielen wiederholten Stellen schließlich erfährt man (mit allen einschlägigen Zitaten) Details über das Berliner MÜ-System und seine Zusammenhänge mit BACK, z.B. Seite 25 und 68 und anderswo; einmal hätte genügt. Man gewinnt manchmal den Eindruck, das Buch sei aus diversen Artikeln zusammengeklebt worden, was durchaus ein gebräuchliches Vorgehen ist; man sollte nur etwas mehr Sorgfalt bei der Herstellung von Redundanzfreiheit anwenden.

Wenn man jedoch die kleinen Mängel in Kauf nimmt (von denen doch einige auch durch bessere Herausgebere tätigkeit hätten beseitigt werden können), hat man mindestens alle Informationen zu einem interessanten MÜ-Projekt in der EUROTRA-Begleitforschung in einem Band zusammen, die ja nach Meinung vieler Experten im Wissenschaftsbereich einen anerkannten Platz gefunden hat.

Johann Haller, IAI Saarbrücken

Wilhelm Weisweber:

Termersetzung als Basis für eine einheitliche Architektur in der maschinellen Sprachübersetzung. Zweite Rezension.

Weiswebers Arbeit ist in dem Rahmen der EUROTRA-D-Begleitforschung an der TU Berlin entstanden, also in einem Forschungsprojekt in dem die wissenschaftliche Fundierung der maschinellen Sprachübersetzung (MÜ) erweitert werden sollte. Das Projekt EUROTRA-D-Begleitforschung ist nun abgeschlossen und Ergebnisse werden vorgelegt.

Die EUROTRA-D-Begleitforschung ist insgesamt positiv zu bewerten. Es ist bei

den Gruppen, geleitet von Christian Rohrer an der Uni Stuttgart als auch Bernd Mahr an der TU Berlin gelungen das immanent anwendungsorientierte MÜ-Vorhaben mit der aktuellen linguistischen Forschung zu verbinden und mit eigenständigen Arbeiten zu fördern. Die EUROTRA-D-Begleitforschung vollbrachte vielleicht erstmalig in der Geschichte, daß MÜ nicht nur mit der aktuellen linguistischen verbunden worden ist, sondern auch für die Linguistik interessante, richtungsweisende Forschungsergebnisse vorlegte. Die Arbeiten der Berliner Gruppe über GPSG (sowie auch die parallelgeführten Arbeiten über LFG in Stuttgart) erweckten eine breite internationale Aufmerksamkeit. In diesem Rahmen ist die Arbeit von Wilhelm Weisweber eingebunden.

Termersetzung als mögliches linguistisches Modell ist durch Chomskys Präferenz für Produktionssysteme (rewriting systems) verdrängt worden. In der Linguistik, und insbesondere in der Syntax, sind stets axiomatische Symbolmanipulationssysteme benutzt worden, in welchen man aus einem abstrakten Ausgangssymbol \underline{S} ausgehend alle Sätze einer Sprache mit Hilfe von (rekursiven) Neuschreibe-Regeln (Generierungsregeln) aufzählen wollte. Dank der Theorie der formalen Sprachen gelang es schon recht früh, den Grammatiken (Aufzählmechanismen) systematisch den leistungsgleichen Automaten (Erkennungsmechanismen) zuzuordnen. Das BeschreibungsmodeLL wurde öfters modifiziert bzw. erweitert: Merkmale sind eingeführt worden. Es entstanden attributierte Grammatiken mit deren Hilfe die syntaktische Analyse verfeinert, die semantische Interpretation modelliert und auch die Textstrukturen erfaßt wurden.

Das inhärente Aufzählungskonzept der generativen Grammatiken kollidierte jedoch mit den MÜ-Modellen (und mit Sprachverstehenssystemen schlechthin), in welchen die normale Intuition die Befolgung eines Abbildungskonzepts erwarten läßt. Die Existenz von verschiedenen Repräsentationsebenen und die Stratifikation der MÜ-Systeme sind natürlich in der Literatur informell öfters behandelt worden (Pause

1986, Hauenschild und Umbach 1988, speziell auch in EUROTRA-Bereich durch Arnold et al. 1986:298). Die letzten Konsequenzen sind jedoch nicht gezogen worden.

Weisweber greift die EUROTRA-Modellierung auf, er verlagert aber dabei die linguistische Aufmerksamkeit von der Generierung (d.h. von der Bestimmungsverhältnis von Grammatik G_i zu Repräsentation RL_i) auf die Termersetzung (d.h. direkter Übergang von G_i zu G_{i+1}):

EUROTRA-Sicht:

$$TL_S \Rightarrow \begin{matrix} G_1 \\ \downarrow \\ RL_1 \end{matrix} - \begin{matrix} G_2 \\ \downarrow \\ RL_2 \end{matrix} - \dots - \begin{matrix} G_n \\ \downarrow \\ RL_n \end{matrix} \Rightarrow TL_T$$

Weiswebers-Sicht:

$$S_t \xRightarrow{sT_1} \begin{matrix} G_1 \\ \downarrow \\ R_1 \end{matrix} \xRightarrow{1T_2} \begin{matrix} G_2 \\ \downarrow \\ R_2 \end{matrix} \xRightarrow{2T_3 \dots n-2T_{n-1}} \dots \xRightarrow{n-1T_n} \begin{matrix} G_{n-1} \\ \downarrow \\ R_{n-1} \end{matrix} \xRightarrow{nT_t} \begin{matrix} G_n \\ \downarrow \\ R_n \end{matrix} \xRightarrow{nT_t} S_t$$

wobei die RL -s bzw. R -s die sprachlichen Repräsentationen die G -s die Grammatiken sind. Weisweber sieht zwischen allen Repräsentationsebenen Termersetzungen (doppel-indizierte T -s), d.h. direkte Übergänge vor.

Das Modell von Weisweber enthält zwei Verallgemeinerungen:

1. Für die Übersetzung (und implizit für die linguistische Modellierung überhaupt) ist die Abbildungsrelation zwischen den benachbarten Repräsentationen entscheidend. Die Grammatiken bestimmen lediglich die Wohlgeformtheit der Sätze auf der jeweiligen Ebene. Die Abbildung wird jedoch nicht zwischen den G -s, sondern zwischen den R -s definiert.
2. Die Regeln der Abbildungen von R_i zu R_{i+1} sind überall formal gleich, sie können als ein Termersetzungsverfahren verallgemeinert werden.

Wesentlich ist bei Weisweber, daß die linguistischen Operationen nicht als Generierung (bzw. ihre Umkehrungen) aufgefaßt

werden, sondern als Überführung (Abbildung) zwischen den Repräsentationsebenen im Sprachsystem, so zwischen der syntaktischen und der semantischen Repräsentation oder (funktional-)semantischen Repräsentation und Wissensrepräsentation usw. Die Regeln einer formalen Grammatik (die die Sprache konstituieren) sorgen lediglich für die Wohlgeformtheit der Eingabe und der Ausgabe. Die Bestimmung der Zielstrukturen RH geht aus der Eingabestruktur R_i als Ergebnis der Termersetzungen hervor.

Die Arbeit gliedert sich in acht Kapitel und Anhang mit Systemdokumentation.

In Kapitel 1 wird das Thema dargestellt und eingeordnet. Hier werden bereits die kritischen Wissensbereiche angesprochen, die in der Arbeit aufeinander bezogen werden, nämlich:

1. Die Modellierung der Maschinellen Sprachübersetzung,
2. Die Vorarbeiten an der TU Berlin und
3. Termersetzung und Termersetzungsverfahren

Kapitel 2 stellt die MÜ-Modelle dar. In dieser Darstellung handelt es sich nicht um die wissenschaftsgeschichtliche Genese der MÜ-Modelle, sondern um die Suche nach Ansätze der Vereinheitlichung der Systemarchitektur. Weisweber folgt verständlicherweise stark Christa Hauenschild und abstrahiert (über Analyse, Transfer und Synthese hinweg) ein einheitliches Grundmodell, das aus Paaren von Repräsentationen R und RH besteht, die durch zwei kontextfreien Grammatiken G_i und G_{i+1} definiert (generiert) werden (S. 17). Zwischen R_i und R_{i+1} werden Termersetzungen definiert.

Kapitel 3 stellt die Produktionssysteme (im Sinne von Davis & King 1977) vor. Insbesondere werden in diesem Rahmen drei MÜ-Systeme dargestellt: SALAT, CAT und TFS. Die Auswahl ist dadurch motiviert, weil diese Systeme eine einigermaßen homogene Architektur aufweisen, gegenüber etwa SUSY, TAUM-METEO oder METAL, die beim Übersetzen in den

verschiedenen Phasen (in der Analyse, in dem Transfer, in der Synthese) unterschiedliche Techniken verfolgen.

Kapitel 4 ist der abstrakteste Teil der Arbeit, hier werden die formalen Grundlagen der Termersetzung behandelt. Das Problem ist, zu gewährleisten, daß die Termersetzungen terminieren. Weisweber greift hier den Knuth-Bendix-Algorithmus auf, der Termersetzungen automatisch durchführt. Bei konfluenten Systemen kann auch die Terminierung der Termersetzungen garantiert werden.

Termersetzungen (=TE) und Termersetzungsverfahren (=TES) sind "bestimmte Ausprägungen" der Produktionssysteme (S. 41). Zwischen den in der Syntax üblichen "formalen Grammatiken" (=FG-en) und Termersetzungsverfahren gibt es zwei Unterscheidungs Momente:

1. Die TES bilden Terme in Terme ab, d.h. sie manipulieren Ketten (oder (Strukturen), während FG-en (Typ 3 und Typ 2) Einzelsymbole umsetzen. Erst bei Typ-O-Grammatiken operieren an Ketten oder Bäumen. TES sind also Operation mit transformationeller Mächtigkeit. Oder anders gesagt: Die Wirkung der TES ist vergleichbar mit der Wirkung von "höheren" (kontextsensitiven, Typ 1 und unrestringierten Typ 0) Grammatiken. Bekanntlich macht die Unentscheidbarkeit die unrestringierten (rekursivaufzählbaren) Sprachen für praktische Verwendung ungeeignet.
2. Die TES sind differenzierter als die einfachen Symbolersetzungssysteme, die lediglich terminale und nichtterminale Symbolen unterscheiden und innerhalb dieser zwei Klassen alle Symbole gleich behandeln. Terme werden durch Signaturen im Voraus (d.h. vor den TES) erzeugt. Sie unterscheiden stets Sorten und Operationen auf beiden Seiten der Gleichungen, insbesondere ist es von vornherein möglich, Verben als Funktoren und die Nominal- und Präpositionalphrasen als ihre Argumente unterschiedlich zu behandeln. Die linke

Seite einer TE erfüllt die Aufgabe der Strukturbedingungen (S. 42), wobei die Einbeziehung der Merkmalnotation möglich ist.

Die nicht-konfluenten Systeme (Kapitel 5) können u. U. auf konfluente zurückgeführt (d.h. lösbar gemacht) werden. Dies ist deshalb bedeutsam, weil die natürlichsprachlichen Systeme nicht immer konfluent sind.

Die Anwendung der Termersetzungsverfahren auf die natürliche Sprach - folgt im Kapitel 6. Vorgestellt wird das System KIT-FAST. Für die mathematisch weniger vorgebildeten bietet wahrscheinlich dieses Kapitel den besten Einstieg. Hier wird die Mächtigkeit der TES anschaulich vorgeführt und verständlich gemacht. Weisweber zeigt, daß die TES auf allen Beschreibungsebenen (Syntaktische Analyse, Funktor- Argument-Strukturen, Transfer und Generierung) eingesetzt werden können und die Leistung erbringen, für die sonst verschiedene Formalismen benötigt worden sind. Es wird hier klar, daß der aufwendige und u.U. schwerfällig wirkende TES-Formalismus für den Linguisten eine erstaunlich einheitliche Behandlung der Sprachverarbeitung ermöglichen wird.

Der eigentliche Textteil des Buches endet mit der Beschreibung der Implementierung des MÜ-Systems (Kapitel 7) und dem Kapitel 8 "Zusammenfassung und Ausblick".

In dem Anhang steckt eine vollständige professionell erstellte Systemdokumentation (A: Installationshandbuch, B: Benutzungshandbuch und C: Beschreibung der Systemdateien). Das System steht übrigens auf dem FTP-Server der TU Berlin zur Verfügung und kann "probiert werden"

ftp.cs.tu.berlin.de

Auch weitere Beschreibungsteile können von dem Autor aus Berlin geholt, bzw. angefordert werden.

Unabhängig davon, ob sie ihm voll zugeschrieben wird, oder als Teil des Forschungsprojektes EUROTRA-D-Begleitforschung angesehen wird, ist Weiswebers Dissertation eine

einmalige Leistung, sowohl aus der Perspektive der Computerlinguistik als auch aus der der theoretischen Linguistik. Denn die TES, kontrolliert durch den Knuth-Bendix-Algorithmus, können nicht nur in Übersetzungssystemen, sondern auch generell in beliebigen linguistischen Vorhaben eingesetzt werden und sie können Sprachen (natürliche Sprachen) mit transformationeller (Typ 0) Komplexität bearbeiten. Es wäre interessant, TES als universelle linguistische Beschreibungsmodelle zu betrachten und die Leistungsgrenzen des Modells genauer zu bestimmen.

Arnold, D.J. /Krauer, S. /Rosner, M. / de Tombe, L./ Varile, G.B.: The (C,A),T Framework in EUROTRA: A theoretically committed notation for MT, in: Proceedings 11th COLING86 Bonn, 1986, S. 297-303.

Chomsky, Noam: Formal Properties of Grammars, in: Luce /Bush and Galanter (eds.) Handbook of Mathematical Psychology 11. Wiley, New York, 1963, S. 323-418.

Hauenschild, Ch. /Umbach, C.: Funktor-Argument-Struktur, Die satzsemantische Repräsentations- und Transferebene im Projekt KIT-FAST, in: Schütz, R. (Hg.) Workshop Semantik und Transfer, EUROTRA-D Working Papers No 6. IAI Universität des Saarlandes, Saarbrücken 1988, S. 16-35.

Pause, P.E.: Zur Modellierung des Übersetzungsprozesses, in: B. Hori, I. /Weber, H.J. (Hgg.) Neue Ansätze in Maschinellem Sprachübersetzung: Wissensrepräsentation und Textbezug, Niemeyer, Tübingen 1988, S. 45-74.

Istvan Batori, Univ. Koblenz-Landau

Hans-Joachim Höll:
Computergestützte Analysen phonologischer Systeme. Exemplarisch am Beispiel einer historisch-vergleichenden Ortsgrammatik aus dem schwäbisch-fränkischen Übergangsbereich. Sprache und Information, Bd. 27, Niemeyer, Tübingen 1994, XVI, 319 S., DM 148,-.

Hans-Joachim Höll hat sich die schwierige Aufgabe gestellt, ein Programmsystem zur Phonemisierung laut schriftlich vorliegender

Corpora zu entwerfen, das so variabel ist, daß unterschiedliche praktische und theoretische Vorgaben gleichermaßen berücksichtigt werden können. Da gerade im Bereich der Phonologie viele konkurrierende Theorieansätze und ebenso viele unterschiedliche ‚Schulen‘ existieren, muß sich ein großer Teil der Arbeit der Darstellung und kritischen Wertung dieser z. T. stark divergierenden Theorien und Lösungsvorschläge widmen. Der Verfasser stellt schließlich mit dem Programm ‚DiaPhon‘ ein System vor, das viele der anvisierten Ziele erreicht und einem an Computerarbeit gewöhnten Phonologen eine wertvolle Hilfe sein kann!

In der Einleitung skizziert er die Schwerpunkte seiner Arbeit. Der erste besteht in der "Diskussion linguistischer und insbesondere phonologischer Fragestellungen" (S. 9), der zweite befaßt sich "mit bereits existenten phonologischen Analysesternen" (S. 9) und der eigenen, maschinenunterstützten Analyse einer Mundart des schwäbisch-fränkischen Übergangsbereiches, die von Ottmar Sexauer² beschrieben wurde. Diese Untersuchung dient als Folie für die Vorstellung des Programmsystems und seines Leistungsumfanges.

Der erste Schwerpunkt, der im wesentlichen in Kapitel 2 abgehandelt wird, ließe sich auch als Einführung in die Phonetik und Phonologie mit Rezensionsteil bezeichnen: Der Verfasser sammelt die einschlägigen phonologischen Theorien von Trubetzkoy bis Vennemann und sichtet sie kritisch. Gleichzeitig versäumt er es nicht, durch das Gegenüberstellen sich z.T. überschneidender oder widersprechender Theorien zu zeigen, wie stark die zu erwartenden Phonemisierungsergebnisse von dem jeweiligen Theorieansatz abhängen. Eine große Zahl von Fußnoten bringt hier jeweils noch detailliertere Informationen, die manchmal den spannenderen, weil kritischeren, Teil des Textes ausmachen. Insgesamt bündelt dieses Kapitel die zentralen Fragen der Phonologie mit besonderer Berücksichti-

gung des Deutschen und seiner latenten Probleme (z.B. Wertung des Schwa, Affrikaten/Diphthonge, silbische Konsonanten, Sonorität etc.), so daß jeder Student der Phonologie diesen Abschnitt als guten Einstieg in die Examensvorbereitung verwenden könnte. Jede der diskutierten Fragen wird mit einer Bewertung abgeschlossen, inwiefern das Problem für die Ausgestaltung des Programmes zur computergestützten Phonemisierung relevant ist.

Die folgenden Kapitel, die zunächst die Datenbasis der Arbeit beschreiben und schließlich die Verfahren, die der Verfasser für die Auswertung programmiert hat, sind naturgemäß stärker auf formale Aspekte konzentriert, wenngleich auch hier immer wieder der Rückverweis auf die theoretischen Differenzen erfolgt, die sich z.B. ergeben, wenn als Bezugseinheit für die Phonemanalyse einmal die Silbe, im anderen Fall aber eine morphologisch determinierte Einheit oder eine Mischung aus beidem gewählt wird. Der Verfasser verweist hier mehrfach auf die Offenheit des von ihm erstellten Programmes. Es ist seine Absicht, die Analysemodule so zu gestalten, daß der Benutzer die Möglichkeit hat, verschiedene Transkriptionssysteme ebenso zu bearbeiten, wie er z.B. die Auswahl zwischen monophonematischer oder biphonematischer Bewertung der Affrikaten bzw. Diphthonge hat.

Zunächst beschreibt der Verfasser in Kapitel 3 die Datengrundlage seiner Untersuchung: Die oben bereits erwähnte Ortsgrammatik von Sexauer aus dem Jahr 1927 ist wahrscheinlich für den Aufbau und das Austesten des Systems so gut geeignet wie es andere Corpora auch wären (S. 129); die moderate Transkription folgt dem in der deutschen Dialektologie vorherrschenden Teuthonista-Prinzip, das keineswegs "verstärkt" durch das IPA3 abgelöst wird, wie der Verfasser S. 53 behauptet (z.B. verwenden alle Teilprojekte des neuen Gesamtbayerischen Sprachatlas eine Teuthonista-Transkription). Obwohl viele Bemerkungen Sexauers darauf schließen lassen, daß er über ein hohes Probleme-

¹ Leider folgt die Arbeit dem Trend, bei immer geringerer Leistung des Verlages immer teurer zu werden.

² Sexauer, Ottmar: Die Mundart von Pforzheim. Leipzig 1927

wußtsein bezüglich der Unterschiede von Stadtmundart vs. Landmundart verfügt (vgl. Höll, S. 122), ist sein Blick doch sehr stark von der ‚Zerfallsmetaphorik‘ getrübt.⁴ In neueren Arbeiten/Datensammlungen ist das Problem der Varietäten besser durchdrungen und daher auch das Datenmaterial so stabilisiert und gekennzeichnet, daß Rückschlüsse auf koexistierende lautliche Systeme sicherer möglich wären (S. 129ff.; vgl. Bellmann 1994). Hier hat der Verfasser bei der Analyse der Sexauerschen Daten, deren Genese unbekannt ist (S. 124), möglicherweise Probleme zu bewältigen, die ein neueres, in dieser Hinsicht problemgeschärftes Corpus nicht enthalten hätte.

Kapitel 4 stellt wieder einen Rezensionsteil dar, in dem die älteren Arbeiten vorgestellt und kritisiert werden, die sich mit der teilautomatischen phonologischen Analyse von Corpora befassen. Es ist immer das Privileg jüngerer Arbeiten, die Fehler und Unschärfen älterer Ansätze schonungslos aufdecken zu dürfen: An dieser Stelle sei jedoch daran erinnert, daß es einen himmelweiten Unterschied ausmacht, ob ein Analysesystem dieser Art - wie hier - auf einem PC mit Turbo Pascal 5.0 konstruiert wird, oder auf Maschinen, wie z.B. der vom Verfasser zitierten, legendären "Sperry Univac 1100/60" (S. 172). Die langen Papierausdrucke waren damals nicht Selbstzweck, sondern die wesentliche Kommunikationsmöglichkeit mit der Maschine überhaupt; m.E. überschätzt der Verfasser in diesem Zusammenhang (vgl. S. 158) die Fähigkeit eines Benutzers, Ergebnisdaten von größerem Umfang auf dem Bildschirm zu erfassen und zu kontrollieren, wenngleich ihm natürlich zuzustimmen ist, daß das interaktive Arbeiten mehr "Kreativität und

Spontanität (sie!) eines linguistischen Bearbeiters" (S. 158) zuläßt.

Kapitel 5, das sich diesen Fragen widmet, folgt den heute üblichen Ansichten bezüglich der Mensch-Computer-Interaktion und skizziert die wesentlichen Merkmale, die das Programmsystem DiaPhon aus software-ergonomischer Sicht haben sollte.

Kapitel 6 beschreibt die Komponenten im einzelnen, aus denen DiaPhon besteht: Datenaufbereitungssystem, Datenanalysesystem und Peripheriesystem sind die Kernmodule.

Das Datenaufbereitungssystem ermöglicht das Editieren, Aufbereiten und Speichern eines Corpus. [. . .] Das Datenanalysesystem beinhaltet all die Systemkomponenten, die die wortphonologische Analyse unmittelbar unterstützen. [. . .] Das Peripheriesystem beinhaltet all die Systemkomponenten, die weder zum Datenaufbereitungssystem noch zum Datenanalysesystem gerechnet werden können. Hierzu gehören z.B. die Programmteile, die dem Benutzer die Konstruktion von Zeichensätzen oder die Anpassung der Tastatur erlauben. (S. 167)

Das Kapitel stellt diese Module vor und gibt - orientiert am Hauptmenü - die verschiedenen Analysemöglichkeiten des Programmes mit Beispielen wieder. Wichtige Menüpunkte sind die für eine phonologische Analyse grundlegenden Prozeduren wie

- > Inventaruntersuchungen, die entweder von einer phonemisierten oder nicht phonemisierten Datenbasis ausgehend die Anzahl der vorkommenden Segmente (Phone/Phoneme) auflisten;
- > Oppositionsuntersuchungen, bei denen die Wahl besteht zwischen ‚klassischen‘ Minimalpaaren und sog. subminimalen Paaren; weitere Menüunterpunkte geben u. a. die Möglichkeit, die Suche

⁴ Im allgemeinen ist jedoch die Lage derart, daß in den Dörfern des nähere Umkreises, ja auch in noch entfernteren, der Hauptteil wenigstens der männlichen Bevölkerung irgend wie in der Industrie tätig ist. [...] Mit dieser Entwicklung geht, wie gesagt, der Zerfall der Ma. [= Mundart, B.K.] Hand in Hand." (Sexauer 1927, S. 33)

⁵ Bellmann, Günther: Einführung in den Mittelrheinischen Sprachatlas (MRhSA). Tübingen 1994. Bellmann beschreibt dort die strengen Maßstäbe, welche an eine moderne "biserielle" Datenerhebung anzulegen sind, die zwei Varietäten (Basisdialekt vs. Regionaldialekt) untersucht.

auf bestimmte Paare zu beschränken. Der Verfasser verweist hier auf Probleme bei der Auswahl der Paare, die sich vor allem daraus ergeben, daß keine Bedeutungsüberprüfung vorgenommen werden kann und somit in den Ergebnislisten nicht nur bedeutungsverschiedene Ausdrücke kontrastiert werden, so daß der Bearbeiter kontrollierend eingreifen muß (S. 218);

- > Untersuchung der distributionellen Relationen, deren wohl wichtigste die Suche nach komplementärer Verteilung ist. Daneben werden Positionsstatistiken (für initiale, mediale und finale Position) erstellt, Listen für Vorgänger- und Nachfolgersegmente ausgegeben sowie Listen für gemeinsame Distributionen und die Silbenstruktur; den Abschluß der Abwählmöglichkeiten bildet eine phonotaktische Analyseeinheit für konsonantische und vokalische Cluster, die recht deutlich die Probleme der vom Verfasser breit diskutierten Frage zeigt, ob die Silbe und/oder morphologisch relevante Einheiten als Untersuchungsbasis gewählt werden sollen.
- > Zwei weitere Menüpunkte geben Listen für die Intensitätsdistribution (Haupt- und Nebentonverteilung) sowie die Möglichkeit zur Ausgabe von Merkmalsmatrizen und zu Suchaufträgen an das Material, die sich nach bestimmten Merkmalen ausrichten lassen.

Den Abschluß der Arbeit bilden ein alphabetisch nach einem Lemma sortiertes Wortverzeichnis der Sexauerschen Belege sowie ein ausführliches Literaturverzeichnis.

Im Resümee kommt der Verfasser nicht um die Feststellung herum, daß die "Analyseergebnisse [. . .] in starkem Maße von der Validität und Reliabilität der Transkription und der Repräsentativität der Materialbasis abhängig" sind (S. 260); dies ist die dialektologische Fassung der alten informatischen Grundweisheit *garbage in - garbage out* .

Offen bleibt die Frage, wie ein Einsatz des Programmsystems bei Daten aussehen muß, welche nicht in der Weise kodiert sind,

die der Verfasser präferiert. Hier verläßt ihn seine Methodentoleranz kurzfristig, wenn er z.B. Bürkle (1988) nachruft:

Ein hoffentlich letztes Credo für eine Codierung phonetischer Symbole als Prototypen findet sich bei BÜRKLE, der an dem traditionellen Verfahren ausdrücklich festhält, obwohl zu dieser Zeit (1988) längst graphikfähige Terminals mit frei programmierbaren Zeichensätzen im Handel waren. (S. 157)

Die geschmähte Kodierung von "Prototypen" (= Kodierung von Einzellaute durch Zeichenkombinationen, z.B. *TJ* als ‚N7‘ (S. 156)) ist kein "traditionelles Verfahren", sondern eine hardwareunabhängige, weltweit übertragbare Lösung, die es jedem Benutzer freistellt, daß - wie der Verfasser selbst sagt - "Letztlich [...] einige wenige, automatisiert durchführbare Substitutionsanweisungen [genügen], um eine Bearbeitung auf einem anderen Rechner [. . .] zu ermöglichen." (S. 157). Als kleiner Vorteil der Prototypen-Lösung ist m.E. anzusehen, daß das Material notfalls auch ohne technische Hilfsmittel lesbar bleibt, während bei der vorgeschlagenen Lösung beliebige Abfolgen von ASCII-Zeichen⁶ entstehen, die ohne passenden Bildschirm- oder Druckertreiber unlesbarer Zeichensalat sind. Im gleichen Zusammenhang ist stark zu bezweifeln, daß in bezug auf schwere Lernbarkeit und Fehleranfälligkeit zwischen einer ‚Prototypen‘-Kodierung, der dies vorgeworfen wird (S. 156), und einer beliebig im ASCII-Kode angesiedelten große Unterschiede bestehen: Auch bei dieser müssen ja die ca. 200 Zeichen (S. 181f.), mit denen eine Transkription auskommen soll, in irgendeiner Weise von einem menschlichen Bearbeiter **gemerkt** werden, zumal der Verfasser mittels einer Tastaturanpassung bis zu 20 (!)

⁶ Beliebige sind die Folgen deshalb, weil sich erst bei der lokalen Erstellung der Bildschirmzeichen (S. 183f.) entscheidet, welches Zeichen zu welchem ASCII-Wert gestellt wird. Bei der bekannten Erfindungsfreude der Dialektologen ist mit Sicherheit anzunehmen, daß sich jeder eine andere Zuordnung ausdenkt.

Tastaturebenen für die Eingabe der Transkriptionszeichen bereitstellt (S. 186); die ASCII-Zeichen werden durch "Umschalttaste,-Grundtaste" (S. 187) erreicht. Was nun leichter merkbar oder fehleranfälliger ist - N7 oder CTRL-@ ?

Trotz des etwas ernüchternden Eindrucks einer sehr großen Streubreite der phonologischen Theorien, die jeweils Einfluß auf die zugehörige Datenerhebung und -auswertung haben, bleibt zu erwarten, daß die vom Verfasser geleistete Problemdarbietung und Programmierung dem Linguisten eine gute Hilfe bei der Analyse großer Datenmengen sein kann. Aus der Schlußbetrachtung geht hervor, weshalb der Titel der Arbeit ("Computergestützte Analysen") so vorsichtig formuliert ist:

Eine computative phonologische Analyse ohne die aktive Kontrolle eines Linguisten ist aufgrund des Faktums, daß eine allgemein anerkannte Verfahrensweise zur Analyse phonologischer Systeme nicht explizit formuliert werden kann, nicht möglich. Es ist vor allem die *Komplexitätsbarriere* (natürliche Sprachen erlauben eine Vielzahl an Deutungsmöglichkeiten), die eine Vollautomatisierung einer phonologischen Sprachanalyse ausschließt. (S. 260)

Bernhard Kelle, Univ. Freiburg

Neuerscheinungen:

Stefan Eherer

Eine Software-Umgebung für die kooperative Erstellung von Hypertexten

1995. IX, 161 Seiten. Kart. DM96,-16s 749.-1 SFr 96,- ISBN 3-484-31929-1 (Sprache und Information. Band 29)

Der Einsatz von Textverarbeitungssystemen hat den Schreibprozeß stark verändert. Neben der Vereinfachung von Korrekturen ist vor allem das Einbinden der auf der Maschine möglichen Medien, wie z.B. Graphik, Bilder oder sogar Ton und Video, die entscheidende Neuerung. Hypertexte erweitern diese Möglichkeiten um maschinengestützte Verweise, die einzelne Texte oder Textfragmente miteinander verknüpfen. Gleichzeitig ist ein Trend zur Verteilung der Aufgaben auf mehrere Autoren zu beobachten, der nicht nur durch die Größe der Texte, sondern auch durch deren inhaltliche Komplexität motiviert ist.

Beide Aspekte werfen jeweils spezifische Probleme auf, die die hier vorgestellte Software-Umgebung lösen soll. Auf der einen Seite muß der Autor eines Hypertextes im Vergleich zu linearen Texten neue Organisationsformen von Texten erlernen, die den Leser bei der Erfassung des Gesamtzusammenhangs eines Hypertextes unterstützen. Auf der anderen Seite müssen die Aktivitäten der Gruppe von Autoren koordiniert und deren Ergebnisse zu einem einheitlichen Ganzen integriert werden.

Um die geforderte Unterstützung zu erreichen, wird auf der Basis von Erkenntnissen der Schreibforschung sowie des computerunterstützten kooperativen Arbeitens ein Modell des kooperativen Schreibens entwickelt und in einer deduktiven Objektbank implementiert. Das Objektmodell erweitert bestehende Modellierungsansätze um abstrakte Ideen, die die inhaltlichen Anforderungen an den zu erstellenden Text beschreiben. Das Prozeßmodell definiert die Aktivitäten der Autoren als hierarchisch geschachtelte, endliche Automaten, deren Zustände durch bestimmte Objektkonstellationen charakterisiert sind. Zustandsübergänge sind die von den Autoren ausgeführten Operationen, deren Anwendbarkeit durch Regeln und Integritätsbedingungen des Objektmodells festgelegt wird.

Future Perspectives of Dialogue Analysis

Edited by

FRANZ HUNDSNURSCHER and EDDA WEIGAND

1995. IX, 199 Seiten. Kart. DM 124.- 1 6s 967.- 1 SFr 124.-. ISBN 3-484-75008-1 (Beiträge zur Dialogforschung. Band 8)

Contents: Preface. - K. ADAMZIK, Dialoganalyse: eine Disziplin auf der Suche nach ihrer Identität. - F. HUNDSNURSCHER, Dialogue Analysis, the Proper Way. - E. WEIGAND, Looking for the Point of the Dialogic Turn. - E. Roulet, L'analyse du dialogue dans une approche modulaire des structures du discours: L'exemple du dialogue romanesque. - A. CAPONE, Dialogue Analysis and Inferential Pragmatics. - A.H. JUCKER, Discourse Analysis and Relevance. - B. Naumann, Mailbox Chats: Dialogues in Electronic Communication. - H. WESTHEIDE, Fremdsprachenvermittlung auf der Basis dialogischer Texte. Heutige Praxis und Zukunftsperspektiven.

Max Niemeyer Verlag GmbH & Co. KG
Postfach 2140. D-72011 Tübingen

GESELLSCHAFT FÜR LINGUISTISCHE DATENVERARBEITUNG e.V. GLDV

Die *Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV)* wurde 1975 - zunächst unter dem Namen "LDV - Fittings e.V." - als Verein zur Förderung der wissenschaftlichen linguistischen Datenverarbeitung gegründet. Sie ist der wissenschaftliche Fachverband für die maschinelle Sprachverarbeitung in Forschung, Lehre und Beruf. Sie vertritt die Interessen ihrer Mitglieder nach außen und fördert die Kooperation zwischen den Mitgliedern und ihren verschiedenen Arbeitsbereichen. Zu diesen Arbeitsbereichen gehören die computerlinguistischen Grundlagen- und Anwendungsforschung, die sprachorientierte Forschung zur Künstlichen Intelligenz, die sprachliche Informations- und Wissensverarbeitung sowie die philologische Datenverarbeitung. Die *GLDV* fördert diese Arbeitsbereiche durch Veranstaltungen, Arbeitskreise und Publikationen.

Die *GLDV* unterstützt die Zusammenarbeit mit Nachbardisziplinen (wie z.B. Linguistik und Semiotik, Informatik und Mathematik, Psychologie und Kognitionswissenschaft, Informations- und Dokumentationswissenschaft) und unterhält Kontakte zu den entsprechenden Fachverbänden. International kooperiert die *GLDV* mit Organisationen wie der *Association for Computational Linguistics (ACL)*, der *Association for Literary and Linguistic Computing (ALLC)* und der *Association for Terminology and Knowledge Engineering (J'KE)*.

Aufnahmeantrag zur GLDV - Mitgliedschaft

Name:- ----- Vorname:- -----
 Straße:- ----- PLZ:- ----- Ort:- -----
 Telefon:- ----- email:- -----
 Akad. Grade und/oder Berufsbezeichnung: Firma oder Hochschule:

Student: NEIN / JA Semesterzahl:- ----- Fächer:- -----
 Arbeitsschwerpunkt bzw. Interessengebiete:

Ort: ----- Datum: _____ Unterschrift: -----

REFERENZ - ERKLÄRUNG: Ich bin (nicht-studentisches) Mitglied der *GLDV* seit 19- und befürworte diesen Antrag.

Name:- ----- Vorname:- ----- email:- -----
 Ort:- ----- Datum:- ----- Unterschrift:- -----

Den Aufnahmeantrag senden Sie bitte an:

Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV) z.
 Hdn. Herrn Prof. Dr. Winfried Lenders
 Institut für Kommunikationsforschung und Phonetik Poppelsdorfer
 Allee 47
 53115 Bonn

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. März 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 Informationswissenschaft¹ war das passende Ambiente für eine erfolgreiche 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 Hitzberger, 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, Vorsit-

zender 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 Vormalig 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 abgelöst oder zu ergänzen 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 nicht-kommerzielle Impulse anzunehmen, 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 Informationssysteme 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-Präziserer) 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 müssen².

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 Kontextabhängigkeit 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 sinnvollen 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 Probanden- 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 Bedingungen 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 ZuhörerInnen 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

³ Normalerweise für n = 2 oder 3

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 Neudaption der Worterkennung an diesen Trainingsdaten auch die Worterkennungsrate von 59,5% auf 72% gesteigert. Eindrucksvoll war der Tonbandmitschnitt eines mißglückten Auskunftsdialogs, 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.

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 „*Spracheingabe 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 angebiert 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.

HYPertext - INFORMATION RETRIEVAL - MULTIMEDIA

HIM'95, 5.-7. APRIL 1995 IN KONSTANZ

Gerhard Knorz

1 HIM - ein rundum gelungenes Experiment

Anforderungen an moderne Informationssysteme lassen aus System- und Nutzer sieht eine isolierte Betrachtung von Problemen häufig nicht mehr zu 1. Deshalb unternahmen es die GI-Fachgruppen „*Hypertext*“, „*Information Retrieval*“ und „*Multimediale elektronische Dokumente*“, zusammen mit der *Österreichischen Computer Gesellschaft*, der *Schweizer Informatiker Gesellschaft* und dem *Hochschulverband Informationswissenschaft* eine gemeinsame Tagung durchzuführen. Die Universität Konstanz, für ihr wissenschaftliches Engagement in den Bereichen Hypertext und Information Retrieval bestens ausgewiesen, hat das Risiko möglicher Komplikationen nicht gescheut und hat die Tagung in jeder Hinsicht erfolgreich vorbereitet und durchgeführt. Dies ist vor allem ein Verdienst von *Rainer Kuhlen* als Programmkomitee-Vorsitzender, von *Marc Rittberger*, bei dem alle Fäden der Organisation zusammenliefen sowie von der erfreulichen Resonanz auf den Call for Papers, die eine Selektion von 50% sehr hochwertiger Beiträge erlaubte.

Ambiente

Die Tagung fand mit ca. 200 Teilnehmern in einem lichtdurchfluteten separaten Gebäudeteil auf mehreren Etagen statt, was den dort präsentierten Postern und Rechnerdemonstrationen viel Resonanz bescherte. Die grüne Umgebung der Universität, der Bodensee in Blickweite und das

¹ Zitat aus dem gedruckten Tagungsprogramm.

z.T. sonnige Wetter trug auf ihre Weise zum Tagungserfolg bei, der durch keinerlei technische oder organisatorische Pannen getrübt wurde.

2 Eingeladene Vorträge

Jede der veranstaltenden GI-Fachgruppen hatte ein *Invited Paper* beigesteuert. Die fachliche Eröffnung der Tagung hatte *Roy Rada* als Hauptredner für die Fachgruppe *Hypermedia* übernommen, mit einem Vortrag über „*publishing World Wide Web Material on CD-ROM: Coordination and CD-ROM*“. Ich will freimütig bekennen, daß ich im Hinblick auf persönliche Konstitution und Entfernung zur Projektionsfläche sehr ungünstige Voraussetzungen hatte, dem Vortrag zu folgen. So mag es an mir liegen, daß ich die recht kurze Plauderei von *Rada* (Universität Liverpool) über die Projekte seiner Studierenden als jene Art eingeladener Vorträge empfand, mit denen man nicht so recht glücklich ist - aus meiner Sicht der einzige Schwachpunkt des Tagungsprogramms.

2.1 Evaluierung von Information Retrieval Systemen

Am Donnerstag rettete *Donna Harmann* nach meiner Einschätzung das Ansehen eingeladener Prominenter mit einem Vortrag über „*The TREC Conferences*“. Sie erläuterte den Anspruch der TREC-Konferenzen, IR-Systeme an Fragestellungen und Daten realer Größenordnungen unter kontrollierten Wettbewerbsbedingungen zu evaluieren und kommentierte den Weg,

der mittlerweile mit 3 Konferenzen zurückgelegt wurde. Die Proceedings von TREC 3 werden im übrigen in den nächsten Wochen im Internet präsent sein:

<http://potomac.ncsl.nist.gov/TREC>

Aufbau und Struktur der Aufgaben sowie das Vorgehen bei der Evaluierung wurden besprochen. Als klares Ergebnis stellte Harmann heraus, daß sich vollautomatisch ermittelte Queries unzweifelhaft manuell entwickelten Suchfragen überlegen gezeigt haben, sofern Feedback verfügbar ist. Dies ist immer dann der Fall, wenn immer wiederkehrende Fragen bearbeitet werden sollen (Routing, Filtering, Profildienste). Sie machte an einem Beispiel auch einsichtig, warum dies so ist: Die statistische Optimierung an Feedback-Daten findet Suchterms, an die ein Mensch kaum jemals denken würde, wohingegen die inhaltlich orientierten Suchstrategien einfach nicht so effektiv sind wie erhofft.

2.2 Multimedia Trends

Das fachliche Programm der Tagung wurde mit dem Beitrag von *Simon Gibbs* (GMD) abgeschlossen: „*Directions in Multimedia Programming.II*“ Er diskutierte zunächst das Problem, zwischen Anwendung und Hardware eine für Multimedia-Entwicklungen geeignete Abstraktion zu finden und zu definieren, so wie dies etwa mit GKS (Graphisches Kernsystem) und anderen Modellen/Standards im Bereich des CAD bereits eingeführt sei. Dann widmete er sich seinem Schwerpunkt der Video-Konferenzen, deren Vor- und Nachteile er auf verschiedenen technischen Realisierungsniveaus verglich: von der gegenwärtig so teuren, vorbereitungs- und betreuungsintensiven Nutzung von Spezialräumen über die Versammlung einzelner Köpfe in Fenstern von Workstations mit integrierter Kamera bis zur Virtual Reality-Lösung, deren Technik mit der des virtuellen Filmstudio nahezu zusammenfällt: Ein kleiner Raum wird durch eine Projektionsfläche geteilt. Sensoren nehmen Erscheinungsbild, Stimme und aktuellen Standort des Diskussionsteilnehmers auf und sorgen über aufwendige Graphikrechner dafür, daß die anderen Diskussionsteilnehmer in der aktuell richtigen

Perspektive auf der Projektionsfläche dargestellt werden - so wie dies zeitgleich auch bei diesen geschieht. Das alles wurde anschaulich demonstriert anhand eines Videos. In diesem vom WDR moderierten und ausgestrahlten Film unter Verwendung von und über diese GMD-Technik, bei dem erst ein Graphikrechner den in einer Bluebox aufgenommenen Moderator mit einer computergenerierten Kulisse verrechnet und bei allen Kamerabewegungen perspektivenrichtig darstellt, konnte die bereits erreichbare "Illusionsqualität" unter Beweis gestellt werden.

2.3 Hypermedia in a Gambling Casino Setting

Neben den eingeladenen Vorträgen fiel ein weiterer bereits durch Thematik und Plazierung aus dem Rahmen: Zwischen Abendspaziergang auf der Insel Mainau mit anschließenden Cocktail im Palmenhaus und dem Abendessen im Comturey-Keller hielt *Hermann Maurer* (Graz) einen engagierten Vortrag über „*Hypermedia in a Gambling Casino Setting*“¹¹. Mit der Idee zu diesem Arrangement hatte sich das Programmkomitee über eine Schwierigkeit hinweggerettet: Was tun, wenn ein renommierter Wissenschaftler ein seriös bis futuristisch anmutendes Paper über eine nicht im traditionellen Wissenschaftsbereich verankerte Anwendungs idee einreicht - und man gar nicht sicher ist, ob auch alles so ganz ernst gemeint ist? Nun, Maurer lies keinen Zweifel, daß es ihm *sehr* ernst ist und daß für ihn Hypermedia eben grundsätzlich mehr und innovativer ist und sein muß als *Multimedia mit Hyper*. Klar, daß auch das WWW einige bange Minuten der begründeten Polemik zu überstehen hatte und letztlich gegen Maurers *Hyper-G* ohne Chancen blieb. Ja, und dann entwickelte Maurer seine Vision, die jedoch wohl ganz auf dem Boden erstaunlicher Tatsachen aufbaut: Es geht um ein Spielkasino, dessen Hotelräume samt und sonders mit hochwertigen Graphic-Workstations ausgestattet werden sollen, und für die innovative und creative Anwendungen zu schaffen sind. Wollten Sie nicht schon immer mal Videoconferencing in der Gestalt von

Spider (Wo) man mit ihrer - letztlich unbekanntem - Barbekanntschaft von gestern abend machen und anschließend nach geeigneter Verwandlung - inspiriert durch einen Teleshopping-Ausflug - als virtuelle MitspielerIn beim realen Roulette teilnehmen und gewinnen?

3 Vorträge in den einzelnen Sektionen

Die Vorträge profitierten davon, daß bis auf eine Ausnahme am Freitag vormittag das Programm ohne parallele Sektionen auskam. Der Tagungsband, herausgegeben von Rainer Kuhlen und Marc Rittberger, erschien als Band 20 der H12-Reihe Schriften zur Informationswissenschaft beim Universitätsverlag Konstanz lag rechtzeitig zu Tagungsbeginn vor. Von allen TeilnehmerInnen wurde sowohl bei "öffentlichen" Meinungsäußerungen wie auch im informellen Gespräch die Qualität der Beiträge hervorgehoben. Aus den ca. 20 angenommenen Beiträgen will ich - nach ganz persönlichen Gesichtspunkten - nur einige wenige herausgreifen.

3.1 Indexieren

Das Interesse am Thema der Dokumentenindexierung wurde durch zwei erfreuliche Beiträge von der ETH Zürich auch für solche belebt, denen dieser Bereich etwas ferner liegt. Bei beiden Papers geht es darum, Inhalte zu indexieren, die nicht in Text verpackt wurden und zwar, ohne daß man Voraussetzungen über den behandelten Gegenstandsbereich machen muß. Wenn also O. Lorenz und G. Monagan über *Automatisches Indexieren von Liniengraphiken* vortragen, dann meinen sie nicht, daß man vorab ein Inventar semantischer Primitive zu definieren hat, um etwa aus Symbolen für Wände und Tische Zeichnungen von Inneneinrichtungen zu interpretieren. Stattdessen bringen sie elementare Techniken aus Bilderkennung und Information Retrieval zusammen, um Strichzeichnungen, wie man sie etwa im Duden Bildwörterbuch findet, anwendungsneutral durch Indexierungsmerkmale zu beschreiben und auffinden zu können.

Die Query besteht dabei aus einem Bild bzw. Teilbild und die Retrievalantwort aus einem Ranking solcher Bilder, die sich auf der Ebene von Linienverläufen, Parallelen, Winkeln und auch Beschriftungen als ähnlich qualifizieren.

Die Besonderheit des Beitrags von M.

Wechsler: *Eine neue Indexierungsmethode für Information Retrieval auf Audiodokumenten* liegt in vergleichbarer Weise darin, daß die Verarbeitung ohne inhaltliche Vorarbeit greift: Es wird also kein allgemeines oder anwendungsspezifisches Wörterbuch vorausgesetzt, um Audiotexte zu indexieren. Vielmehr legt das auf Phonemerkennung nach Standardtechnik (Hidden Markov Modelle, Bi-gramme) basierte System selbst a posteriori fest, welche Phonemketten sich im Hinblick auf Länge, Abdeckung, Häufigkeit und Erkennungssicherheit eignen, um als Indexierungsmerkmale zu fungieren. Das System konnte in den Pausen auf der Basis eines Korpus mit gesprochenen Radionachrichten erprobt werden.

Beide Ansätze wurden intensiv diskutiert, und vermutlich haben die Vortragenden einige weitere anregende Ideen mit nach Hause genommen.

3.2 Modellierung von Hypermedia Retrieval mittels Datalog

Aus Dortmund kam der Vorschlag, die Sprache *Datalog* zur Modellierung komplexer Retrievalstrategien für strukturierte Hypermediadokumente einzusetzen: „*Modelling Hypermedia Retrieval in Datalog*“ (N. Fuhr). Unter Datalog kann man sich ein um Berechnungsprobleme bereinigtes Prolog zur Realisierung deduktiver Datenbanken vorstellen. Ziel war es, die grundsätzliche Eignung von Datalog, was die Ausdrucksmächtigkeit betrifft, nachzuweisen. Weiterführende Arbeiten werden sich mit der Implementierung von Datalog beschäftigen, um damit im Gigabyte-Bereich Information Retrieval zu erproben - und natürlich wird es ein probabilistisches Datalog werden müssen, wenn es in Dortmund Bestand haben will.

² Hochschulverband für Informationswissenschaft.

3.3 Informationssuche in Faktenbanken

Zwei recht verschiedene Beiträge beschäftigten sich mit dem Thema, die Informationssuche in Faktenbanken zu unterstützen. STN Karlsruhe stellte Überlegungen vor, die in ihren Faktenbasen implizit vorhandenen Links für eine komfortable Hypermedia-Oberfläche für Datenbanken wie DETHERM zu nutzen: *Concept of a Hypertext Interface for the Information Retrieval in Complex Factual Databases* (L. Helmes, M. Razum, A. Barth). Es ist zu hoffen, daß im STN-internen Abstimmungsprozeß diese Überlegungen nicht auf dem Weg zur Realisierung steckenbleiben.

Der Beitrag „Softcomputing“- Verfahren zur Behandlung von Ähnlichkeiten und Vagheit in objektorientierten Informationssystemen (Th. Mandl und Ch. Womser-Hacker) motivierte zunächst das Problem vager Anfragen und stellte die Ähnlichkeitssuche als Möglichkeit vor, das Problem von Datenlücken zu überwinden. Der Regensburger Prototyp für ein multimodales Werkstoff-Informationssystem integriert Lösungen auf der Basis von Fuzzy Sets und von neuronalen Netzen. Er stellt eine menügesteuerte Möglichkeit zur Verfügung, linguistische Variablen mit Werten zu belegen (z.B. Temperatur mit *hoch*, *niedrig*, etc.), wobei eine Kontextabhängigkeit 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.

Ein Detailproblem der Faktensuche behandelten U. Pfeifer, Th. Poersch, N. Fuhr: *Searching Proper Names in Databases*, in dem sie mehrere Verfahren daraufhin verglichen, mit den Problemen von Schreibvarianten von Namen fertigzuwerden. Ergebnis: Alles ist besser, als nur auf Identität zu prüfen.

3.4 Multimedia und Hypertext

Von den in dieser Besprechung aus Zeitgründen zu kurz gekommenen Beiträgen über Multimedia und Hypertext haben mir insbesondere die Beiträge *"Towards a Family of Formal Models for Hypermedia"* (K. Tochtermann, G. Dittrich), *„Navigation via Similarity in Hypermedia and Information Retrieval“* (D. Tudhope, C. Taylor, P. Beynon-Davie) sowie *„Ein graphischer Browser für das Konstanzer Hypertext-System“* (S. Arens, R. Hammwöhner) gut gefallen.

4 Digitale Bibliotheken

Eine sehr lebhaft entwickelte sich auf dem von Hans-Peter Frei moderierten Panel über *Digitale Bibliotheken*. Die Panel-Teilnehmer zeigten zunächst, daß sie durchaus Unterschiedliches im Sinn haben, wenn sie über digitale Bibliotheken sprechen: Herrmann Maurer sprach über Probleme des Internet, über die Integration von Internet und LAN, von Internet und CDROM und über die Integration von Anwendungen und Produkten (wie z.B. von elektronischen Wörterbüchern). Im Gedächtnis ist mir der zitierte De Kemp mit seiner Antwort auf die Frage geblieben: *Wie kann man mit elektronischen Zeitschriften Geld verdienen?* - „Drucken Sie sie!“ (online 94).

Norbert Fuhr sprach über Probleme der (mangelnden) Ergonomie von Bildschirmen, der Archivierung und der Kosten. Er zitierte seinerseits Czermak mit dem Prinzip der *„flächendeckenden Entgeltlichkeit“* und forderte einen weiterhin freien Zugang zu Information, wobei eine Steigerung der Kosten gegenwärtig sicher nicht in die Landschaft paßt.

Jean-Frederic Jauslin (Direktor der Schweizer Landesbibliothek) lieferte für die spätere Diskussion im Plenum sicher die meisten Stichwörter. Er baute folgende Kette verkürzt wiedergegebener Thesen auf: Bibliotheken hatten bisher eine Schlüsselrolle, gegenwärtig sind die Dinge in Bewegung, den Bibliotheken ist vielfach nicht bewußt, was auf dem Spiel steht, die Gefahr ist groß, daß Bibliotheken künf-

tig nur noch wenige Wissenschaftler bedienen bzw. ganz verschwinden. *Jauslin* stellt fest, daß das Mehrsprachenproblem zwischen Produzenten - Benutzern - Archiv auch in der EG noch nicht so recht adressiert wird und daß es insgesamt nicht um die Überwindung sondern um die Ergänzung des Buches durch elektronische Formen geht.

Rainer Kuhlen zitiert aus einer Studie, die das Problembewußtsein und die Aufgeschlossenheit der Bibliotheksprofession doch zu belegen scheint und gibt auf die Forderung, daß das Internet Mechanismen des Qualitätsmanagements braucht, zu bedenken, daß das *sky writing* auch als Basisdemokratisierung der Information empfunden werden könne.

In der weiteren Diskussion fallen die Stichwörter der Informations- Umweltverschmutzung und -Ökologie, wird gefragt, ob Bibliotheken noch Mauern brauchen, werden ernsthafte Überlegungen aus den U.S.A. zitiert, die die Notwendigkeit für eine Bibliothek für eine neue Universität grundsätzlich hinterfragen, werden Aufgaben- und Prioritätsverlagerungen in Bibliotheken dargestellt und das papierlose Büro mit Hohn und Spott überschüttet.

5 Fazit

HIM'95 war eine Tagung auf gutem bis sehr gutem Niveau, das Modell einer integrierenden Tagung taugt zur Nachahmung. Daß die Synthese zwischen *Hypertext* und *Information Retrieval* besser gelungen war als die Einbeziehung der Fachgruppe *Multimedia*, war erkennbar, doch - so darf man hoffen - ein überwindbares Problem. Die Rahmenbedingungen der Tagung, die Vorbereitung und die Organisation waren hervorragend. Daß eine konkrete Vorplanung für eine HIM'97 noch nicht existiert, liegt eigentlich nur am globalen Wissenschaftsdorf und den Schwierigkeiten, die die ACM mit ihrer Unsicherheit über die nächste europäische ECHT in die deutschsprachige Tagungsszene hineintransportiert. Aber es wird schon werden, zumindest bis - zur zweiten HIM!

EINLADUNG

ZUR
26. Jahrestagung
der Gesellschaft für
Angewandte Linguistik
et al.
28. - 30.09.1995
an der Universität
Gesamthochschule Kassel


Rahmenthema:
Schlagwort:

Kommunikations- gesellschaft

Auskunft über thematische bzw. organisatorische Einzelheiten:

Prof. Dr. Dieter Wolf
Bergische Universität

GH Wuppertal

FB 4: Anglistik/Amerikanistik

Gaußstr. 20

42097 Wuppertal

Telefon 02 02/4 39-22 55 (22 54)

Telefax 02 02/4 39-22 54

Prof. Dr. Manfred Raupach

GH Kassel

FB Romanistik

Georg-Forster-Str. 3

34109 Kassel

Telefon 05 61/8 04-33 59

Telefax 05 61/8 04-38 15

(örtliche Organisation)

Tagungsgebühren:

DM 45,- für GAL-Mitglieder

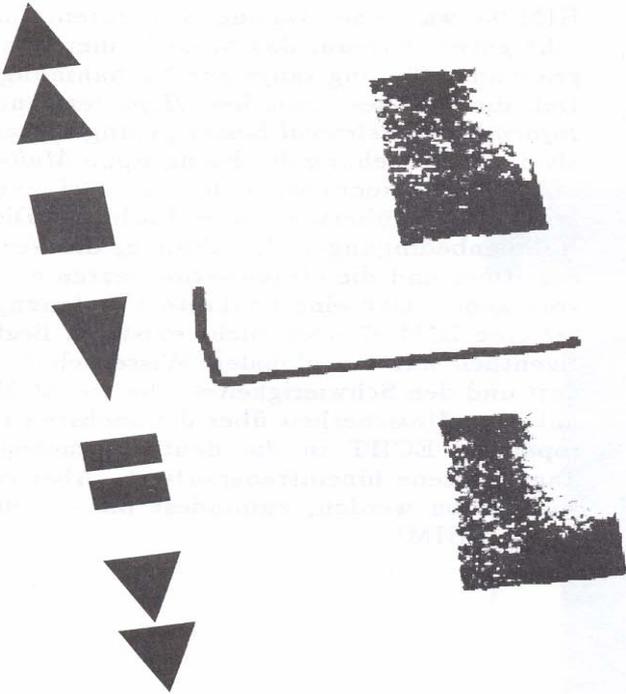
DM 80,- für Nichtmitglieder

DM 10,- für Studenten bei Vorlage des Studienausweises

Tagungsbesucher, die nur an den Arbeitskreisen teilnehmen
(Anmeldung am 28.09.1995), zahlen einen ermäßigten
Tagungsbeitrag von DM 20,-.

Bankverbindung:

Dresdner Bank Stuttgart (BLZ 600 800 00),
Kongelkonto 90 60 455/13



Veranstaltungen

GLDV-Herbstschule 1995

Moderne Methoden der Computeranalyse

11.-15. September 1995

Universität Bonn Psychologie
/Informatik - Zentrum, Römerstraße
164

Programm:

Kurs 1: Methoden der Standardisierung - Eine Einführung in SGML und TEI (Textual Encoding Initiative) (Dr. Peter Scherber, Göttingen)

Kurs 2: Morphologie und Tagging (Prof. Dr. Roland Hausser, Erlangen)

Kurs 3: Fachsprachliche Korpora (Prof. Dr. Bergenholtz, Aarhus)

Kurs 4: Statistischer Zugriff auf Korpora: Disambiguierung und Tagging (Dr. Robert Neumann, Cyril Belica und Doris al-Wadi, Institut für Deutsche Sprache, Mannheim)

Kurs 5: Textmodellbasierte Korpusanalyse (Dr. Karin Haenelt, GMD Darmstadt)

Kurs 6: Korpora gesprochener Sprache (Prof. Dr. Randall Jones, Brigham Young University, Provo).

Nähere Informationen zu den einzelnen Kursen wie Inhaltsangabe, Literatur, Zeiten etc. sowie zur gesamten Herbstschule werden an die angemeldeten Teilnehmer ab ca. 1. Juni 1995 versandt.

Exkursion:

Mittwoch, 13. September, Besichtigung von Forschungslabors der Gesellschaft für Mathematik und Datenverarbeitung (GMD) in St. Augustin-Birlinghoven, anschließend Wanderung im Siebengebirge und Weinabend in Oberdollendorf/Königswinter.

Plenarvorträge:

Prof. Dr. H. Schnelle, Bochum:
"Wortnetzungen in Computer und Gehirn"

Prof. Dr. Manfred Bierwisch, Berlin: Thema voraussichtlich.: "Universalien und Idiosynkrasien im Lexikon"

Gebühren:

Bei Anmeldung VOR dem 15. Juni 1995

	Studenten	Sonstige
Mitglieder der GLDV:	60,-	100,
Nichtmitglieder:	80,-	130,

Bei Anmeldung NACH dem 15. Juni 1995:

	Studenten	Sonstige
Mitglieder der GLDV:	100,-	140,
Nichtmitglieder:	120,-	170,

Die Teilnehmergebühr ist mit der Anmeldung auf das im Anmeldeformular (siehe unten) angegebene Konto zu überweisen.

Unterbringung:

Jugendgästehaus Bonn- Venusberg (Bus-Transfer) : 4-Bett-Zimmer, mit Frühstück, 31,50 DM p.P./Nacht .
Gästehaus d. Internationalen Jugendforum Bonn: 2-Bett-Zimmer, mit Frühstück, 63,00 DM p.P./Nacht
I-Bett-Zimmer, mit Frühstück, 95,00 DM p.P./Nacht
Studentenwohnheim: Sammelunterbringung (Küche, Bad, WC), 10,00 DM p.P./Nacht

Verpflegung:

Mensa, Cafeteria unmittelbar im Tagungsgebäude

Schwimmbad:

Römerbad, um die Ecke des Tagungsgebäudes

Anmeldeformular

Verbindliche Anmeldung zur Teilnahme an der GLDV-Herbstschule vom 11.-15. September 1995 in Bonn

Name: _____

Vorname: _____

Straße: _____

Wohnort: _____

Student an Universität/Hochschule: _____

Gewünschte Unterbringung (bitte ankreuzen):

- Jugendgästehaus Bonn- Venus berg (Bus-Transfer):
4-Bett-Zimmer, mit Frühstück, 31,50 DM p.P./Nacht
- Gästehaus d. Internationalen Jugendforum Bonn:
2-Bett-Zimmer, mit Frühstück, 63,00 DM p.P./Nacht
- 1-Bett-Zimmer, mit Frühstück, 95,00 DM p.P./Nacht
- Studentenwohnheim:
- Sammelunterbringung (Küche, Bad, WC), 10,00 DM p.P./Nacht
- besorge Unterbringung selbst

Die angebotenen Unterbringungsmöglichkeiten können nur bei einer Anmeldung bis zum 15.6.1995 gewährleistet werden. Später eingehende Unterbringungsünsche werden erfüllt, so gut es dann noch geht.

Zahlung der Gebühren:

Teilnehmergebühr:	DM
Unterbringung: Garantiebetrag für 1 Übernachtung:	DM
Den Gesamtbetrag von	DM

überweise ich auf das Konto Nr. 502199013
bei der Volksbank Bonn, BLZ 380 601 86
unter Angabe des Verwendungszwecks "GLDV-Herbstschule 1995".

Unterschrift:

Anmeldung bitte schicken an

GESELLSCHAFT für LINGUISTISCHE DATENVERARBEITUNG e.V. 1.
Vorsitzender
Prof. Dr. Winfried Lenders
Institut für Kommunikationsforschung und Phonetik
der Universität Bonn
Poppelsdorfer Allee 47
D-53115 Bonn

Tel. +49 (228) 735646
Fax +49 (228) 735639 e-mail:
Lenders@uni-bonn.de

Achter Internationaler Kongreß der Deutschen Gesellschaft für Semiotik

Vom 5. bis 9.8.1996 findet in Amsterdam in Zusammenarbeit mit niederländischen Institutionen der 8. Internationale Kongreß der Deutschen Gesellschaft für Semiotik (DGS) statt. Das allgemeine Thema des Kongresses lautet "Kultur - Zeichen Raum".

Auf diesem Kongreß wird es eine Sektion "Space and Computer" geben, die von Herrn Dr. Rene Jorna (Groningen) und Prof. Dr. Udo L. Figge (Bochum) vorbereitet wird. Gegenstand möglicher Sektionsbeiträge könnten beispielsweise Raum-Metaphern in der Mensch-Computer-Interaktion, Entwicklung und Funktion virtueller Räume, Repräsentation und Verarbeitung räumlichen Wissens oder automatische Verarbeitung lokaler Ausdrücke sein. Potentielle Teilnehmer sollten zunächst kurz ihr Interesse bekunden; sie erhalten dann weitere Informationen.

Prof. Dr. Udo L. Figge
Romanisches Seminar
Ruhr- Universität
D - 44780 Bochum

IWPT'95

Fourth International Workshop on Parsing Technologies

20 - 23 September 1995 Prague/Karlovy Vary
Czech Republic Sponsored by
ACL/SIGPARSE Association for
Computational Linguistics, Special Interest
Group on Parsing

The Fourth International Workshop on Parsing Technologies IWPT'95 will take place this year on September 20 through 23 in the Czech Republic in the heart of Europe. This workshop will continue the tradition, established by IWPT'89 and IWPT'93, of taking place partly on the premises of a university and partly in a rather secluded conference resort. The first part of IWPT'95

will thus be held at Charles University in Prague; the workshop then moves to a conference resort (Grand Hotel Pupp) in the famous old bathing resort of Karlovy Vary, also known as Karlsbad.

The papers will cover all areas of parsing technology. These areas include, but not limited to, theoretical and practical studies of parsing algorithms for natural language sentences, texts, fragments, dialogues, illformed sentences, and speech, as well as multidimensional (pictorial) language, and parsing issues arising or viewed in a multimodal context.

Program Committee: Robert Berwick, Makoto Nagao, Harry Bunt, Anton Nijholt, Bob Carpenter, Yves Schabes, Ken Church, Mark Steedman, Eva Hajicova, Henry Thompson, Arvind Joshi, Masaru Tomita, Ronald Kaplan, K. Vijay-Shanker, Martin Kay, Yorick Wilks, Bernard Lang, Kent Wittenburg.

The conference fee will be \$ 125. Details of a complete package including accommodation in Prague and Karlovy Vary, and other additional information will soon be available through WorldWideWeb:

Harry C. Bunt
ITK, Instituut voor Taalen
Kennistechnologie
Institute for Language Technology
and Artificial Intelligence Tilburg
University
P.G. Box 90153
5000 LE Tilburg, the Netherlands
Phone: +31 - 13 66.30.60
Fax: +31 - 13 66.25.37 bunt@kub.nl

Wahrscheinlichkeitstheoretische Ansätze für die linguistische Analyse geschriebener und gesprochener Sprache Workshop

Nachmittag des 11.9.95, Bielefeld
(KI/DAGM)

Organisatoren:

Günter Schukat-Talamazzini (Uni Erlangen),
Elmar Nöth (Uni Erlangen), Gerhard Niedermair
(Siemens AG) *Inhalt, Motivation und*

Diskussionsschwerpunkte:

Statistische Modellierungsverfahren haben in den achtziger Jahren auf dem Gebiet der Mustererkennung geschriebener und gesprochener Sprache für zwei entscheidende technologische Durchbrüche gesorgt: die mathematische Beschreibung akustisch - phonetischer Wortrealisierungen durch HMMs ('Hidden Markov models') und die wahrscheinlichkeits-theoretische Erfassung syntaktischer Struktureigenschaften natürlicher Sprache durch stochastische Grammatiken (z.B. N-Gramm-Verteilungen).

Die erfolgreiche Ausweitung des statistischen Paradigmas - oder allgemeiner: von Methoden maschinellen Lernens aus Sprachkorpora - auf weitere Fragestellungen der automatischen Sprachverarbeitung wie etwa der semantischen, prosodischen oder diskursbezogenen Analyse kündigt sich in der gegenwärtigen Forschungslandschaft bereits an und soll den inhaltlichen Schwerpunkt des vorgeschlagenen Workshops bilden.

Das zentrale Anliegen der Veranstaltung ist die Initiierung eines Austauschs zwischen industriellen und universitären Vertretern der Mustererkennung und KI über den aktuellen Stand der Forschung - bereits verfügbare Modellierungswerkzeuge, erfolgreiche Realisierungen, laufende Projekte, aber auch offene Probleme und Spekulationen über potentielle Anwendungen - auf diesem neuen Gebiet.

Die Beiträge werden vorzugsweise innovatorische Ansätze behandeln, die das Modellierungsinstrumentarium z.B. zur Bedeutungs- oder Dialoganalyse, Inhaltser-schließung oder Sprachenidentifikation einsetzen. Standardverfahren der akustischen oder grammatischen Modellierung - ohnehin in ausreichender Weise auf den einschlägigen Tagungen repräsentiert - sollen hingegen ausgespart bleiben.

Teilnehmer:

Zu dem halbtägigen Workshop sind neben den Referenten weitere interessierte Teilnehmer herzlich eingeladen sind.

Weitere Information:

Ernst G. Schukat-Talamazzini, Lehrstuhl für
Mustererkennung Universität
Erlangen Martensstraße 3
91058 Erlangen
E-Mail: schukat@informatik.uni-
erlangen.de
Tel. (09131) 85-7873
Fax (09131) 303811
Sekretariat (09131) 85-7775

Text Analysis and Computers

September 18 - 21, 1995 Mannheim,
Germany Preliminary Programme
ZUMA PO Box 122155 D-68072
Mannheim Germany

Aims

Text analysis has long been an interdisciplinary pursuit. The added potential of information technology and the growing availability of computer-readable corpora in fields as diverse as archaeology and discourse analysis have led to an increased activity and interest in the computer-assisted analysis of texts.

The ZUMA Text Analysis and Computers conference is of relevance to researchers interested or involved in computer assisted text analysis, irrespective of discipline. However, a special focus will be given to text analysis methods as implemented or implementable in the social sciences.

The goal of the conference is twofold: on the one hand to present and discuss text analysis procedures being used or developed in applied research, on the other to promote and coordinate the innovative development of text analysis tools. Invited keynote

speakers will open the conference with talks on progress and problems in the following three areas: computer-assisted content analysis, interpretative methods and computational linguistics.

The papers and demonstrations at the conference will reflect theoretical developments in a given field. During the podium discussion on the last day, the invited speakers and the other participants will try to sum up their conclusions from the meeting. On the basis of these, a concrete effort will be made to pin-point an area or areas in which a coordinated research initiative can be started. The conference language is English.

Organisation:

Center for Survey Research and Methodology (ZUMA) in cooperation with the Society for Conceptual and Content Analysis by Computer (SCACC)

Programme Committee:

Alfons Geis, ZUMA; Angelika Gloeckner-Rist, ZUMA; Janet Harkness, ZUMA and Albert-Ludwig University, Freiburg; Juergen H.P. Hoffmeyer-Zlotnik, ZUMA; Peter Ph. Mohler, ZUMA; Klaus M. Schmidt, Bowling Green State University, Ohio, USA; Peter Schrott, ZUMA; Cornelia Zuell, ZUMA

Conference Office:

Please address all correspondence to:

Jolantha Muellner
ZUMA
Post Box 122155
D-68072 Mannheim
Germany

Telephon: 0621 1246 174
Fax: 0621 1246 100
e-mail: cta@ZUMA-Mannheim.de
Gopher:
gopher .social-science-geis.de

General Information:
Conference Location

Hotel Wart burg
F4, 4-11
D-68159 Mannheim
Tel.: +49 621/28991
Fax: +49 621/101337

The hotel is located in Mannheim city centre.

A separate room has been reserved for software demonstrations. A conference office will be open for the duration of the conference, located in the foyer of the conference area in the hotel.

Conference Fee:

The conference fee is DM 185.
Reduced rate for students is DM 160.

Registration:

Registration includes payment according to the fee schedule listed above and a completed registration form with your full address. Payment inside Europe can be done by bank transfer to account no. 3421 88002 at the Commerzbank AG Mannheim (Giro clearing number 670 400 31) under the code word Textanalysis and the participants last name. Otherwise, a check payable to ZUMA must be sent along with the registration.

Final Programme:

The final programme including the abstracts will be available at the conference desk when you check in.

Further questions to:

ZUMA
Jolantha Muellner
PO Box 12 21 55 D-
68072 Mannheim
Germany

The Conference Programme includes the following Papers:

Monday, September 18, 1995

10.45 - 11.30 Text and Computers

Past, Present and some Future; Peter Ph. Mohler, Mannheim, Germany

11.30 - 12.30 Computer-assisted Content Analysis: an Overview; Erhard Mergenthaler, Ulm, Germany

13.30 - 15.00 Trend Analysis Using Computerized Text Analysis: A Case Study of Transportation News; Jane Torous, Irvine, USA

A Dictionary of Typical German Features; Horst-Alfred Heinrich, Giessen, Germany

Dynamics of Changes in Russia: Investigation with Open Questions and Content-Analysis (Longitudinal from 1989) Galina Saganenko, Eugeny Kanevsky, St. Petersburg, Russia

15.00 - 15.30 Coffee Break

15.30 - 16.30 Machine-readable Corpora and the Linguistic Description of Languages; Christian Mair, Freiburg, Germany

16.30 - 17.30 A Lexical Study as Dictionary Adjunct - An Analysis of YET in the Brown Corpus; Nina Devons, Jerusalem, Israel

Search Strategies with the on-line Conceptual Database for Medieval German Literature; Klaus M. Schmidt, Bowling Green, USA

Tuesday, September 19, 1995

9.00 - 10.00 Computer-Aided Qualitative Data Analysis: an Overview; Udo Kelle, Bremen, Germany

10.30 - 12.30 An Intelligent Multimedia Approach to Automating Qualitative Data Collection and Coding; Jacqueline Haynes, Gary Resnick, Rockville, USA

Writing in Computer-Assisted Qualitative Data Analysis; Seppo Roponen, Helsinki, Finland

Computer-Assisted Analysis in Qualitative Social Research - A Comparison; Jörg Struebing, Berlin, Germany

Automatic Language Recognition and Corpus Analysis; Marcus Spies, Heidelberg, Germany

13.30 - 15.00 Computer Environments and Expert Systems for

Content Analysis; William Evans, Atlanta, Germany

HyperJoseph: The Hypertextual Organization; Ephraim Nissan, London, U.K.

Transforming Linear Text to Hypertext. The WWW Experience; Werner Degenhardt, München, Germany

15.30 - 16.30 Information Processing and Texts; Jürgen Krause, Bonn, Germany

Wednesday, September 20, 1995

9.00 - 10.30 Multilingual Text Analysis: Contrastive Concordance between Leibniz's Monadologie and its Translations; A. Lamarra, A. Russo, Rome, Italy

SERAPHIN: A System for the Automatic Extraction of Main Sentences; Jawad Berri, Dominique Le Roux, Denis Malrieu, Jean-Luc Mine, Paris, France

Dictionary Layers Underlying Electronic Texts; Dusko Vitas, Cvetana Krstev, Gordana Pavlovic-Lazetic, Belgrade, Serbia

11.00 - 12.00 Global Issues and the New Geopolitics of Information; Anthony Giffard, Seattle, USA

Framing Effects of Strategy-Oriented New Coverage on Public Understanding: Semantic Network Analyses of Narrative Data; June W. Rhee, Philadelphia, USA

12.00 - 13.00 Computer-Assisted Rhythm Analysis of Mari Pagan Prayers; Natalia Glukhova, Mari EI, Russia

Prototype Effects vs. Rarity Effect in Literary Style; Paul A. Fortier, Winnipeg, Canada

14.00 - 15.00 Text Encoding Initiative; NN

Machine Translation; NN

15.30 - 17.00 Electronically Coding Corporate Justifications for Top Management Compensation Using the Method of Successive Filtrations; Joseph Porac, James B. Wade, Tim Pollock, Champaign, USA

Using Content Analysis in Support of Test Construction and Validation; Phyllis Kuehn, Fresno, USA

Management of Big Data Files, Disaster or Blessing?; R. Moelker, Breda, The Netherlands

Filtscor Joseph Porac, Champaign, USA

Hypertext Angelika Glöckner-Rist, Mannheim, Germany

INTEXT Harald Klein, Jena, Germany

TACT Janet Harkness, Mannheim, Germany

TEXTPACK Cornelia Zuell, Mannheim, Germany

Thursday, September 21, 1995

9.30 - 11.00 A Computer Content Analysis Approach to Measuring Social Distance in Residential Organizations for Older People; D. G. McTavish, Minneapolis, USA

Network Approaches to the Analysis of Texts; Roel Popping, Groningen, The Netherlands

Supporting Content Analysis of Software Development Data Gathered by Multiple Teams at Multiple Sites; Ira Monarch, Suresh Konda, Dave Gluch, Marvin Carr, Pittsburgh, USA

11.30 - 13.00 The Enemy Within: Auto-Correlation Bias in Content Analysis and in Historiometry and Scientometry; Robert Hogenraad, Dean McKenzie, Colin Martindale, Louvain, Belgium

Do the Words in Annual Reports tell us anything about Organisations' Performance? Boris Kabanoff, Kensington, Australia

Defining Key Words and Concepts through Computational Text Analysis; Renata Fox, Rijeka, Croatia

Exhibitions

Gesis (German Social Science Infrastructure Services)

iec Progamma (The Netherlands)

UM-96: Call for Papers and SIG Meeting Proposals

Fifth International Conference on User Modelling

Kailua-Kona on the Island of Hawaii

January 2-5, 1996

Research in the field of user modelling and user-adapted interaction provides crucial foundations for improving the usability of interactive software systems and technical devices in many application areas. UM-96, the Fifth International Conference on User Modelling, will provide a forum for presenting the results of academic and industrial research in this field to an international audience. UM-96 follows the very successful Fourth International Conference on User Modelling (UM-94) that was held in Cape Cod, Massachusetts in 1994. UM-96 will have the same format as UM-94 and will consist of tutorials, technical presentations, special interest group meetings, and system presentations, while encouraging close interaction among the participants. Tutorials will be held on Jan. 2, with the technical program on Jan. 3-5.

Submissions are invited on original and substantial academic or industrial research on all aspects of user modelling and user-adapted interaction, including (but not limited to) the following topic areas: user

Software Demonstrations

The following software demonstrations have been registered to date:

CAIR Eugeny Kanevsky, St.-Petersburg, Russia

Max.txt Udo Kukartz, Berlin, Germany

model acquisition, plan and intention recognition, user stereotypes, representation and inferencing techniques for user models, consistency of user models, user modelling shell systems, explanation generation, intelligent / adaptive user interfaces, adaptation to users with special needs, student modelling and intelligent tutoring systems, navigation aids, adaptive hypertext and hypermedia, information retrieval and information filtering, knowledge-based information presentation, natural language systems, and other applications.

Accepted submissions will be published in a proceedings volume. Kluwer is sponsoring a \$500 Best Paper Award. Extended versions of high-quality papers should be submitted to the international journal *User Modelling and User-Adapted Interaction* after the conference.

GENERAL CHAIR

David Chin

University of Hawaii
Honolulu, Hawaii, USA

LOCAL ARRANGEMENTS CHAIR Martha Cross by
University of Hawaii

Honolulu, Hawaii, USA

PROGRAM CHAIRPERSONS

Sandra Carberry

University of Delaware
Newark, Delaware, USA

Ingrid Zukerman

Monash University

Clayton, Victoria, Australia *INTERNATIONAL*

PROGRAM COMMITTEE

Mathias Bauer, DFKI, Germany

Alan Biermann, Duke Univ., USA

Mark Elsom-Cook, EBC, Great Britain
Gerhard Fischer, Univ. of Colorado, USA
Brad Goodman, MITRE, USA

Hitoshi Iida, ATR Labs, Japan

Anthony Jameson, Univ. of Saarbruecken,
Germany

Judy Kay, Sydney Univ., Australia
Alfred Kobsa, GMD, Germany

Kurt Konolige, SRI, USA

Diane Litman, AT&T Bell Labs, USA
Mark Maybury, MITRE, USA

Gordon McCalla, Univ. of Saskatchewan,

Canada

Kathy McCoy, Univ. of Delaware, USA

Cecile Paris, Univ. of Brighton, Great Britain

John Self, Lancaster Univ., Great Britain

Peter van Beek, Univ. of Alberta, Canada

Julita Vassileva, Univ. der Bundeswehr, Germany

Wolfgang Wahlster, DFKI, Germany

Marilyn Walker, Mitsubishi Electric, USA

Bonnie Webber, Univ. of Pennsylvania, USA "

Bev Woolf, Univ. of Massachusetts, USA

SUBMISSION OF PAPERS

Questions about submissions should be sent to

um96-pc@cis.udel.edu.

SUBMISSION OF PROPOSALS FOR SPECIAL

INTEREST GROUP MEETINGS

Proposals of not more than one page should be

submitted via email to um96sigs@cis.udel.edu.

The proposals should include the proposed topic,

an abstract, reasons for the usefulness of the

meeting, and the names of several other people

who have agreed to participate in the proposed

meeting.

REGISTRATION OF SYSTEM DEMONSTRATIONS

Various platforms will be available for unrefereed

system demonstrations. Prospective presenters

should contact David Chin

(chin@uhics.ics.hawaii.edu) and specify their

hardware and software requirements.

IMPORTANT DEADLINES

July 17, 1995: Electronic abstracts and
submissions received.

July 18, 1995: Authors notified by email
if hard-copy requirement is waived

July 21, 1995: Hardcopy of submissions must be
received by Sandra Carberry unless waived
by email on July 18.

Sept. 11, 1995: Proposals for special in-
terest group meetings received.

Sept. 25, 1995: Notification of authors
about acceptance or rejection

Nov. 15, 1995: Final camera-ready copy received by David Chin.

Nov. 30, 1995: System demonstrations registered with David Chin.

CONFERENCE SITE

Kona Surf Resort & Country Club, Kailua-Kona, Hawaii

FURTHER INFORMATION

Further information on the conference, including questions about hotel and travel information, can be obtained from David Chin (chin@uhics.ics.hawaii.edu) and Martha Crosby (crosby@uhics.ics.hawaii.edu). In addition, a Mosaic page (<http://www.ics.hawaii.edu/um-96>) will contain updated information on the conference and hotels as it becomes available.

**European Summer School
Information Retrieval (ES SIR)
University of Glasgow, 3. - 8.
September 1995**

Die zweite europäische Sommerschule für Information Retrieval (IR) wird im September am Department for Computing Science der Universität Glasgow stattfinden. Die Sommerschule wird für diejenigen abgehalten, die auf diesem Gebiet anfangen. Sie wird einen weiten Bereich von Fächern abdecken. Alle Kurse werden von angesehenen Experten auf dem entsprechenden Gebiet gehalten. Es stehen Studentenstipendien zur Verfügung.

Geplante Kurse:

- Introduction
- IR Models
- Architecture /Implementation Natural Language Processing for IR
- User Interfaces
- IR and Databases
- Evaluation
- Intelligent Retrieval
- IR and Hypermedia
- Multimedia and Retrieval
- IR and Wide Area Networks

Information

Informationen und Anmeldeformulare, sowie Stipendienanträge können über WWW bezogen werden:

<http://www.dcs.gla.ac.uk/essir/>

oder als Broschüre von

Jane Reid
Computing Science
University of Glasgow
Glasgow G1f 8QQ
UK

Tel: +44-141-330-5006 Fax:

+44-141-330-4913 Email:

essir@dcs.gla.ac.uk

**Informationsmanagement -
Informationsmärkte -
Informationsgesellschaft**

**10 Jahre Hochschulausbildung für einen
Schlüsselbereich der wirtschaftlichen
Entwicklung**

Veranstaltung vom 4. bis 6. Oktober 1995
am Fachbereich Information und
Dokumentation der Fachhochschule
Darmstadt

Anlässlich seines 10-jährigen Bestehens bereitet der Fachbereich Information und Dokumentation eine Veranstaltung vor, für die der Zeitpunkt nicht günstiger sein könnte.

Kaum jemals war die Bedeutung von Information für Wirtschaft und Industrie in gegenwärtigem Maße ein zentrales öffentliches Thema, das ganze Industriesparten in Aufbruch-, aber z. T. auch Alarmstimmung versetzt. Und wer hätte sich noch vor wenigen Jahren Personal Computing, Netzwerke, Multimedia, Online-Informationszugriff und CD-ROM als Themen des Konsumentenmarktes vorgestellt?

Gleichzeitig kann - allen Prognosen zum Trotz - gegenwärtig niemand verlässlich sagen, wie die höchst spektakuläre und

dynamische Entwicklung wirklich verlaufen wird. Und es herrscht vielerorts auch beträchtliche Unsicherheit darüber, welches Potential heute verfügbare Informationstechnologie für die Wettbewerbsfähigkeit und die eigenen Unternehmensziele tatsächlich besitzt.

Zu den wirklich erstaunlichen Tatsachen gehört, daß der gesellschaftliche und professionelle Bedarf an einer *Informationskompetenz* erst ganz allmählich erkannt wird. Denn so wie der Straßenverkehr flächendeckend nicht die Ausbildung von Autokonstruktoren verlangt, sondern Fahrschulen und Ausbildung für Berufsfahrer, so sind die technischen und informatischen Grundlagen zwar Voraussetzung, aber keineswegs ausreichend für einen professionellen effizienten Umgang mit Information. Ohne Marktkenntnisse, ohne methodische Kenntnisse, ohne anwendungsbezogene Fachkenntnisse wird Technikanwendung zum Strohfuder, ohne die notwendigen Ziele zu erreichen. Die erforderliche informationswissenschaftliche (Hochschul-) Ausbildung ist zwar erfolgreich, steht aber im deutschsprachigen Raum in ihrer Kapazität in krassem Gegensatz zum Bedarf. In Hessen ist es der Fachbereich Information und Dokumentation an der Fachhochschule Darmstadt, der Informationsspezialisten (mit dem akademischen Abschluß des Diplom-Informationswirtes) für den Bedarf vornehmlich von Industrie und Wirtschaft ausbildet.

Informationen über die Veranstaltung

ZIELGRUPPE: Sie wollen sich über Themen wie elektronisches Dokumentenmanagement, CD-ROM-Produktion, Online-Datenbankrecherchen, Informationsmanagement, Hypermedia oder Internet praxisnah informieren?

Sie planen selbst neue Lösungen ihrer Informationsprobleme und wollen sehen, ob Sie mit Ihren Einschätzungen und Kalkulationen richtig liegen?

Sie wollen den Fachbereich IuD und sein fachliches Spektrum kennenlernen?

Sie sind ehemaliger Absolvent oder Absolventin des Fachbereichs und wollen dabei sein?

Dann sollten Sie sich unbedingt den Termin vormerken!

PROGRAMM: Der Mittwoch Nachmittag ist den strategischen und politischen Aspekten des Themas gewidmet, der Donnerstag wird im Stil einer Anwendertagung Informationsprobleme und betriebliche Lösungen dazu vorstellen, der Freitag vormittag behandelt Aspekte und Trends der Ausbildung.

Begleitend werden Projektarbeiten und Diplomarbeiten ausgestellt.

Zu einem Social Event wird am Donnerstag abend eingeladen und der Freitag Abend bietet sich für eine Fete aller ehemaligen und gegenwärtigen IuD-Studierenden an.

ORT: Fachhochschule Darmstadt

INFORMATIONEN

ÜBER DIE VERANSTALTUNG: Die Veranstaltung wird gegenwärtig vorbereitet. Ein gedrucktes Programm mit allen notwendigen Informationen zur Anmeldung kann vom Fachbereich angefordert werden:

FH Darmstadt

Dekan des Fachbereichs IuD Prof.

Dr. G. Knorz Haardtring 100

D-64295 Darmstadt

Tel: (06151) 16-8491

Fax: -8980

e-mail: knorz@fh-darmstadt.de

Laufend aktuelle Informationen werden mittels e-mail weitergegeben bzw. sind auf dem WWW-Server des Fachbereichs spätestens Mitte Juli abrufbar.

Mitteilungen aus der GLDV

Neues zur Morpholympics 96

Wie bereits im letzten LDV-Forum angekündigt, laufen die Vorbereitungen für eine nächste Morpholympics planmäßig an. Als Austragungsort war schon von Anfang an Saarbrücken im Gespräch, in Verbindung mit der Wahl von Französisch als zusätzlicher Testsprache.

Inzwischen ist eine Reihe neuer Umstände hinzugekommen, die weitere Überlegungen notwendig machen: die 1. Morpholympics hatte bereits kurz nach ihrem Ende Aufsehen in Europa erregt. Kritische Stimmen ("schlecht vorbereitet", "keine gründliche Auswertung" etc.) mischten sich mit der Anerkennung der großen organisatorischen und technischen Leistung sowie der Bravour der Veranstalter, Teilnehmer und Preisrichter, die die gesamte Vorbereitung, Durchführung und Dokumentation der Morpholympics nur aus den Mitteln der GLDV und eigenen Mitteln, d.h. ohne staatliche und andere öffentliche Zuschüsse gewährleisteten.

Solche Zuschüsse werden seit einiger Zeit an Netzwerke, Ressourcen- und Standardisierungsprojekte vergeben, die das Problem erst einmal von der theoretischen Seite her zu bearbeiten begannen und dabei durchaus beachtenswerte grundlegende Ansätze zur Methodik der Evaluation von Softwaresystemen erarbeiteten, die Teilprobleme der automatischen Sprachverarbeitung zu lösen versuchen. Die entsprechenden Berichte sind über die DG XIII, R. Havenith (Fax +352-43013-4655) zugänglich.

In der nahen Zukunft will die EU-Kommission dem gesamten Gebiet der Validierung von Ressourcen und Systemen mehr Aufmerksamkeit widmen; hierbei sollen der oben geschilderte theoretische Ansatz, konkrete nationale Initiativen und das Interesse der Industrie in eine gemeinsame Aktion integriert werden, z.B. in der Form eines Projektes in der Aktionslinie "Linguistic Engineering" des 4. Rahmenprogrammes. Zu diesem Thema haben einige informelle Gespräche in Luxemburg stattgefunden, bei denen auch die praktischen Erfahrungen der Morpholympics präsentiert wurden. Eine weitere nationale Initiative kommt aus dem Nachbarland Frankreich, das durch sein Ministerium der Frankophonie auch dem Gebiet der automatischen Verarbeitung des Französischen öffentliche Mittel zukommen läßt, die die Mittel aus dem Sektor Forschung, Bildung und Technologie in idealer Weise ergänzen. Von einer ähnlichen Förderung kann die deutsche Sprache nur träumen. Da die französische Initiative GRACE ("Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation") eine Reihe von Gemeinsamkeiten mit den Ideen der LDV-Arbeitsgruppe "Morphologie und Parsing" aufweist (und die nächste Morpholympics ohnehin das Französische mit einbeziehen soll!), lag es nahe, den deutsch-französischen Dialog an dieser Stelle aufzunehmen. Der technische Koordinator von GRACE, P. Paroubek vom ehrwürdigen INALF ("Institut National de la Langue Francaise") besuchte J. Haller in Saarbrücken zu einem ersten Gespräch und sprach dann eine Einladung zum nächsten Koordinationstreffen GRACE am 23. Mai in Paris aus, das vom wissenschaftlichen GRACE-Koordinator J. Mariani (LIMSI-CNRS) geleitet wurde.

Bei diesem Treffen wurde die französische Gruppe kurz über die MORPHOLYMPICS-Erfahrungen informiert und J. Haller konnte an den Diskussionen über Grundkonzept und Zeitplan des französischen Evaluierungsprojekts teilnehmen.

In diesem Projekt sollen große Korpora des Französischen gesammelt, lemmatisiert, syn-

taktisch kategorisiert und oberflächennah analysiert werden. Auf der Basis zusammengeführter Lexika sollen verschiedene Instrumente (auch vergleichend) evaluiert werden, die diese Funktionalitäten erfüllen.

Die Zielsetzung geht also weit über die Bedingungen der MORPHOLYMPICS hinaus, sieht aber als ersten Schritt einen Vergleich von morphologischen Analyseverfahren und besonders von Taggern vor. Auch im geplanten Vorgehen gibt es Unterschiede: in mehreren Schritten sollen "Lernkorpora", "Versuchsdaten" und "Testdaten" verteilt und unter bestimmten zeitlichen Vorgaben bearbeitet werden. Der Schlußtest soll von Fachleuten überwacht und anschließend in einem Workshop zusammen mit der Präsentation der Systeme vorgestellt werden.

Die französische Gruppe zeigte sich jedoch offen für Diskussionen und manifestierte ein großes Interesse für ein experimentelles Zusammenführen der beiden Aktionen. Die Gespräche werden intensiv weitergeführt; wenn es hier zu einer guten Abstimmung kommt, müßte für eine gemeinsame Veranstaltung der Herbst 1996 ins Auge gefaßt werden.

In jedem Fall ist eine Teilnahme deutscher Institutionen an der französischen Evaluierung möglich; Interessenten, die über NLP-Komponenten für Französisch verfügen, können sich per Email an hans@iai.uni-sb.de wenden.

Johann Haller, Saarbrücken

Postrückläufer

Leider hatten wir wieder mal einige Postrückläufer beim Versand des letzten LDVForums. Wer die neue Anschrift eines der nachfolgend aufgeführten GLDV-Mitglieder kennt, möge sie bitte dem Vorstand bekanntgeben:

Diggelmann, Andi, Wettingen (Schweiz)
 Dr. phil. Kropac, Ingo Herbert, Graz
 Prof. Seeger, Thomas, Uni Potsdam
 Sigel, Alexander, Saarbrücken
 Dr. Ranieri, Filippo, Frankfurt
 Pape, Martin, Uni GHS Essen
 Argyriadis, Charoula, Max-Plank-Inst. ER. Frankfurt
 Langer, Hagen, Wert her
 Thiel, Manfred, Lauf
 Dezelski, Hans-Joachim, Uni Tübingen
 Mayer, Renate, Stuttgart
 Schykowski, Heinrich, Remagen
 Würzburger, Daniel, Garbsen
 Krewer, Monika, Trier
 Hasenknopf-Reknes, Adelheid, Olching
 Dikeoulia, Ionna, Saarbrücken
 Daiber, Jürgen, Trier
 Pruen, Claudia, Trier
 Dr. phil. Kahre, Annette, Gütersloh
 Stallwitz, Gabriele, TA, Nürnberg

ARBEITSKREISE DER GLDV

AK - Studienführer CL

Aufruf zur Mitarbeit

Nachdem seit dem Erscheinen des letzten Studienführers Computerlinguistik bereits einige Zeit verstrichen ist und sich in der deutschen CL-Landschaft auch institutionelle Veränderungen ergeben haben, hat der Vorstand der GLDV auf der diesjährigen Jahrestagung in Regensburg beschlossen, den Arbeitskreis Studienführer zu reaktivieren und eine neue Auflage des Studienführers in Angriff zu nehmen.

Folgender Arbeitsplan ist vorgesehen:

O. *Definition eines geeigneten Formates für den Studienführer*

Da anders als in der vorangegangenen Auflage die neue Version auch in elektronischer Form zugänglich sein soll, erscheint es ratsam, die Informationen in SGML bzw. HTML aufzuarbeiten. Der Entwurf einer zu diesem Zweck geeigneten DTD ist in Arbeit.

1. *Erfassung vorhandenen Materials*

Wenigstens teilweise können die Daten der alten Auflage übernommen werden, müssen allerdings in das neue Format überführt werden.

2. *Ermittlung neuer Daten*

Per Rundschreiben an die Hochschulen bzw. durch eine Telephonaktion sollen die neuen/ geänderten Informationen zur Computerlinguistik ermittelt werden.

3. *Evaluierung der Informationen*

Es scheint erforderlich, nach Abschluß

der Datensammlung eine gewisse Evaluierung durchzuführen, um dem Status und der tatsächlichen Bedeutung der ermittelten Veranstaltungen und Studiengänge gerecht zu werden (z.B. Studienschwerpunkt innerhalb eines Faches vs. eigener Studiengang; seltene Spezialvorlesung vs. regelmäßiger Vorlesungszyklus etc.).

4. *Aufarbeitung als Informationsangebot im WWW*

Um auf möglichst einfache Weise weiteres Feedback erhalten zu können und gleichzeitig die Information einem möglichst breiten Interessentenkreis zugänglich zu machen, werden alle Daten sowie die Datenstruktur für Neueinträge im WorldWideWeb verfügbar gemacht. Einstiegspunkt ist zunächst der WWW-Server des Instituts für Informatik der Universität Leipzig (<http://www.informatik.unileipzig.de>), weitere werden evtl. folgen. Je nach verfügbarer Zeit und bestehendem "Benutzerinteresse" sind natürlich unterschiedlich detaillierte (und elegante) Lösungen denkbar Anregungen sind jederzeit willkommen!

5. *Drucklegung einer neuen Auflage* Sobald die obigen Punkte abgearbeitet sind, kann die Drucklegung der neuen Auflage erfolgen.

Alle an der Erarbeitung und Ausgestaltung des neuen Studienführers Interessierten sind herzlich eingeladen (und ermuntert), am Arbeitskreis teilzunehmen - geplant ist ein Treffen möglichst noch vor der

Sommerpause, genauer Termin und Ort stehen allerdings noch nicht fest. Nähere Informationen finden sich demnächst unter der genannten WWW-Adresse in Leipzig.

Die Kontaktadresse für den Arbeitskreis ist (bevorzugt e-mail):

Dr. Christian Wolff
 Institut für Informatik
 Abt. Automatische
 Sprachverarbeitung Universität
 Leipzig Augustusplatz 10-11
 04109 Leipzig
 Tel. (0341) 97-32249 (0. -32230)
 Fax (0341) 97-32209
 e-mail:
 wolff@inform1.informatik.uni-
 leipzig.de

AK - Korpora

Treffen des Arbeitskreises KORPORA am 21. Februar 1995 im Institut für deutsche Sprache (IDS), Mannheim

Am 21. Februar 1995 tagte der Arbeitskreis "Korpora" der Gesellschaft für Linguistische Datenverarbeitung im Institut für deutsche Sprache, Mannheim, und beschäftigte sich mit der Thematik "Statistischer Zugang zu Korpora". Es nahmen teil:

Prof. Dr. Ursula Klenk (Universität Göttingen),
 Prof. Dr. Winfried Lenders (Universität Bonn),
 Kathrin Giesecking (Universität Trier),
 Alexander Mehler (Universität Trier),
 Markus Schulze (Universität Erlangen),
 Gerald Schüller (Universität Erlangen),
 Jochen Leidner (Universität Erlangen),
 Helmut Schmid (Institut für maschinelle Sprachverarbeitung,

Universität Stuttgart),
 Stefan Langer (Centrum für Informations- und Sprachverarbeitung, Universität München),
 Robert Naumann, Dons al-Waldi und Dr. Irmtraud Jüttner, jeweils IDS, Mannheim

Herr Robert Neumann (Leiter des Arbeitskreises) eröffnete die eintägige Zusammenkunft und umriß das Ziel der aktuellen Diskussion des Arbeitskreises, verschiedene Möglichkeiten statistischer Analysen vorzustellen. Die Arbeiten aus dem Bereich Computerlinguistik/LDV der Universität Trier wurden von Frau Kathrin Giesecking und Herrn Alexander Mehler vorgestellt.

Frau Giesecking sprach zum Thema "Empirische Überprüfung eines synergetischen Modells der Lexik anhand englischer Korpora", des Modells von Reinhard Köhler, das für die Betrachtung der Lexik die folgenden Systemgrößen "Frequenz", "Länge", "Polylexie", "Phonemanzahl" und "Lexikongröße" entwickelt hat.

Die synergetische Linguistik betrachtet die Sprache als ein sich selbst organisierendes System, das Mechanismen für eine Anpassung an die Systemumgebung entwickelt und gewisse Ansprüche an die Sprache - auch divergierende Systemansprüche - stellt. Und so sorgen Selbstorganisationsmechanismen für eine Anpassung an den Sollzustand, beispielsweise Mutations- und Selektionskriterien beim Silbenverschlucken. Neben dem Anpassungsbedürfnis erläuterte Frau Giesecking die Stabilitäts-, Kodierungs-, Spezifikations-, Anwendungs- und Ökonomiebedürfnisse, das Bedürfnis nach Sicherheit der Informationsübertragung und die speziellen Bedürfnisse nach Minimierung des Kodierungsaufwands. Die empirische Überprüfung des Basissystems von Köhler zeigt eine gerichtete Abhängigkeit zwischen der Frequenz, der Länge, der Polylexie und Polytextie. Zwischen Frequenz und Silbenlänge gibt es große Zusammenhänge; so wird bei häufigem Gebrauch des mehrsilbigen "Universität" ein Wort mit der Silbenlänge 2, nämlich "Uni". Herr Alexander Mehler (Universität Trier) beleuchtete in

seinem Vortrag "Fuzzy-Textbedeutungen" Probleme der Textbedeutungsrepräsentation und erarbeitete die Hypothese, daß die Zweistufigkeit der Bedeutungskonstitution, welcher die syntagmatischen und paradigmatischen Restriktionen der Sprachelemente untereinander entsprechen, auch auf der Ebene von Texten Geltung besitzt. Diese Zweistufigkeit führt auf der Textebene zur Konstitution der konnotativen Textbedeutung, deren Konstituenten auf der Wortebene durch die syntagmatischen und paradigmatischen Relationen der Wörter bestimmt werden.

Der von Herrn Mehler vorgeschlagene Repräsentationsformalismus für Textbedeutungen kann in einem weiteren Schritt zur Lösung des Problems der automatischen Textkonversion im Rahmen des Hypertextes herangezogen werden. Ein alternatives Verfahren zur bisherigen intellektuellen Textkonversion sieht er in einem automatischen Aufbau der Hypertextbasis durch textoberflächenorientierte Informationen, womit zumeist die Notwendigkeit einer Nachedition oder Überprüfung der resultierenden Hypertextstruktur durch den Informationswissenschaftler verbunden ist. Weil der Rekurs auf rein formale Textmerkmale - etwa die Untergliederung der Texte in Abschnitte - zu einer weitgehenden Ausklammerung des Textbedeutungsaspektes führt, schlägt Herr Mehler ein Verfahren zur automatischen Textkonversion vor, das primär auf Repräsentationen der Bedeutungen der zu verknüpfenden Wörter und Texte basiert und somit den Bedeutungsaspekt in den Mittelpunkt der Betrachtung rückt.

Herr Helmut Schmid (Universität Stuttgart) berichtete in seinem Referat über "Aktivitäten am Institut für Maschinelle Sprachverarbeitung (IMS) im Bereich des Taggens deutscher Texte". Unter Tag ging subsumiert er alle Annotationen von Textkorpora mit syntaktischen, morphologischen, semantischen und anderen Informationen. Beim Wortarten- Tagging treten die bekannten Probleme mit den syntaktischen ambigen Wortformen auf, und die einzige Lösung wäre eine Desambiguierung durch den Kontext - mit

den Möglichkeiten durch Markov-Ketten, durch neuronale Netze, durch Regelbasierung und durch einen Parser. Er diskutierte Vor- und Nachteile verschiedener Trainingsmodelle (Baum-Welch-Training, Viterbi-Training, Xerox-Tagger).

Herr Stefan Langer (Centrum für Informations- und Sprachverarbeitung, Universität München) sprach zu "Statistischen Untersuchungen der Semantik von Nominalkomposita im Rahmen der semantischen Kodierung der CISLEX". Das Kernlexikon enthält 38 000 einfache Nomina für die statistische Untersuchung der Semantik. Bei ersten Versuchen einer Einteilung ließen sich bisher ungefähr 250 semantische Klassen unterscheiden mit einer Grobgliederung in Lebewesen, andere Artefakte, Orts- und Zeitbezeichnungen. Herr Langer demonstrierte an Beispielen die statistische Analyse von Selektionspräferenzen in den Nominalkomposita.

Herr Neumann faßte den Vortrag "Statistische Analyse von Zeitstrukturen in Korpora" von Herrn Cyril Belica (IDS, Mannheim) kurz zusammen und beantwortete Fragen. Herr Belica, der wegen Krankheit kurzfristig seinen Vortrag absagen mußte, hat eine Methode zur Berechnung, Analyse und Auswertung zeitrelevanter Parameter von Sprachphänomenen in Textkorpora, die als Indikatoren des Terminus "Neologismus" verwendet werden, entwickelt. Diese mathematisch-statistische Analyse zeitlicher Verteilung von Häufigkeiten verschiedener Textmerkmale isoliert solche Sprachelemente, deren zeitliche Verteilung auf dem festgesetzten statistischen Signifikanzniveau bestimmte Auffälligkeiten aufweist. Herr Neumann erklärte die programmtechnische Implementierung der Methode, gab Beispiele für die statistische Berechnung und Findung der sogenannten Kandidaten für einen Neologismus im deutschen Wortschatz und für eine mögliche Kollokationsanalyse.

Frau Doris al-Wadi führte eine COSMAS-Präsentation vor und zeigte an relevanten Sprachbeispielen aus den Textkorpora des IDS - insbesondere am "Wendekorpus", das in sechs zeitlichen Phasen abfragbar ist - die statistischen Sor-

tierungen und Gruppenbildungen bei unterschiedlichem Signifikanzniveau sowie bei den veränderbaren Parametern der Granularität (Clusterbildung) und Klassenzuordnung; zur besseren Illustration dieser Methode gab sie den Teilnehmern eine vorbereitete Tischvorlage zur Kollokationsanalyse in die Hand.

Das nächste Treffen des Arbeitskreises "Korpora" ist für den 8. Juni 1995 an der Universität Stuttgart, im Institut für maschinelle Sprachverarbeitung - Computerlinguistik, geplant, wo über "Korpusbasierte Lexikographie" verhandelt werden soll.

Irmtraud Jüttner, Mannheim

AK - Fuzzy Linguistik

Ankündigung des neuen Arbeitskreises und Aufruf an Interessenten zur Mitarbeit

Während der beiden letzten GLDVJahrestagungen (Kiel 1993, Regensburg 1995) hat es jeweils eine Sektion *Fuzzy Linguistik* gegeben, deren Vorträge auf sehr erfreuliche Resonanz stießen. Da auch mein Vorschlag ein positives Echo fand, diesem Interesse ein zwischen den Jahrestagungen vorhandenes Forum zu bieten, möchte ich hiermit die - schon in Regensburg angekündigte - Einrichtung des neuen Arbeitskreises *Fuzzy Linguistik* in der GLDV bekanntgeben.

Fuzzy linguistische Modellbildungen erfassen zunehmend mehr und weitere Bereiche. Anfänglich nur im Rahmen einer traditionellen Semantiktheorie zur Analyse, Repräsentation und Verarbeitung natürlichsprachlicher Unschärfe eingesetzt, bilden heute die methodisch wie methodologisch (weiter-)entwickelten *fuzzy* Theorien ein neues Paradigma, das fast alle logischen und mathematischen, insbesondere aber die empirisch-numerischen Methoden und Verfahren beeinflusst, welche für die Modellbildungen in der Linguistik, der Computerlinguistik und der quantitativen Linguistik seit langem eingesetzt werden. Dieser

Einfluß scheint derzeit noch verstärkt zu werden durch eine merkliche Hinwendung linguistischer Forschung zu performativen Phänomenen des Sprachgebrauchs und deren funktionaler Erklärung in kognitivkommunikativen Zusammenhängen.

Welche Konsequenzen eine zunehmende Übertragung von *fuzzy* Modellierungstechniken für unsere Disziplin (wissenschaftstheoretisch wie praktisch) haben wird, ist wohlfundiert kaum abschätzbar und kann nur vage umrissen werden. Denkbar ist aber immerhin, daß ihr vermehrter Einsatz zu einer schrittweisen Revision des (wissensbasiert-kognitiven) Erkenntnisinteresses der Linguistik (zugunsten eines lernensbasiert-semiotischen) beitragen wird angesichts der schon jetzt beobachtbaren Erweiterungen, welche die (kompetenztheoretische) Bestimmung der linguistischen Forschungsgegenstände (in Richtung auf performative Sprachdaten) erfährt, und angesichts der *fuzzy* Modellierungen, deren integrative, kontinuierliche und dynamische Repräsentationen die (auf inkompatiblen Zwischenrepräsentationen operierenden) Modellbildungen der bisherigen Untersuchungsmethoden zunehmend ergänzen und sogar ablösen.

Die im Schnittbereich von kognitiver Linguistik und kognitiver Sprachverarbeitung mit Computerlinguistik und quantitativer Linguistik sich konstituierenden neuen (oder auch nur veränderten) *E-Interessen*, *F-Gegenstände* und *U-Methoden* könnten darüber hinaus eine Linguistik etablieren, die bei der Erforschung natürlicher und der Entwicklung künstlicher informationsverarbeitender Systeme auch anwendungsrelevante Problemlösungen zu bieten vermöchte. Beides, *fuzzy* Modellierungen als Vorbereiter einer semiotisch-linguistischen Sprachtheorie und als Basis einer linguistisch-simulativen Ingenieurwissenschaft markieren die Pole linguistischer Forschung, Entwicklung und Anwendung, zwischen denen sich die Aktivitäten des Arbeitskreises *Fuzzy Linguistik* bewegen könnten.

Interessenten und potentielle Mitglieder des AK mögen sich bitte per Email, Fax

oder auch brieflich mit mir in Verbindung setzen. Weil die Arbeitskreise der GLDV allen, insbesondere auch den jüngeren Mitgliedern offenstehen, bitte ich gleichzeitig auch um informative Angaben zu Hintergrund und Motivation für die Meldung zum AK Fuzzy Linguistik. Je nach Anzahl und Interessenlage bei diesen Anmeldungen möchte ich nach Möglichkeit noch während des laufenden Semesters (spätestens zu Beginn des kommenden Wintersemesters) ein erstes Arbeitstreffen hier in Trier organisieren. Es soll neben dem wechselseitigen Kennenlernen und der Vorstellung der jeweils eigenen Forschungsinteressen und Arbeitsansätzen vornehmlich dazu dienen, eventuell konvergierende Bereiche ausfindig zu machen und anhand erkennbarer (möglicher) Schwerpunkte über die künftige Arbeit und das Programm des AK zu diskutieren.

Burghard Rieger

Linguistische Datenverarbeitung
Lehrstuhl für Computerlinguistik FB 11:
Sprach- und Literaturwissenschaften,
Universität Trier D-54286 Trier
Email: fulinak~ldv35.Uni-Trier.de Fax:
+49-651-201-3946
Tel: +49-651-201-2270

AK - Hypermedia

Der Sommer kommt, ein Arbeitskreis erwacht: Der Arbeitskreis Hypermedia

Warum ein Arbeitskreis Hypermedia?

Im Zuge des Aufbruchs in die sog. Informationsgesellschaft arbeiten verschiedene Fachrichtungen gemeinsam an der Frage, wie elektronische Dokumente in neuartiger Weise organisiert und präsentiert werden können. Auch die Computerlinguistik kann zur Entwicklung qualitativ hochwertiger Hypermediasysteme und -anwendungen beitragen, z.B. durch Textparser zur (semi)automatischen Konversion konventioneller Texte in Hyperdokumente

oder durch intelligente, natürlichsprachliche Navigationshilfen. "Einfache" LDV-Verfahren wie morphologische Analyse oder Lemmatisierung sind wertvoll, wenn beispielsweise Termini in Hypertexten automatisch mit Glossaren oder Wörterbüchern verknüpft werden. Darüberhinaus bietet Hypermedia in verschiedenen Bereichen der angewandten Computerlinguistik völlig neuartige Möglichkeiten, Arbeitsabläufe zu organisieren und die dabei entstehenden Ergebnisse zu präsentieren.

Anlaß genug, das Arbeitskreisgeschehen der GLDV um einen Arbeitskreis zu bereichern, der sich speziell mit den computerlinguistischen Aspekten der Hypermedia-Technologie auseinandersetzt.

Wer sollte sich angesprochen fühlen?

>Entwicklerinnen und Entwickler von Hypermedia-Anwendungen in Bereichen der angewandten Computerlinguistik, z.B. computergestützte Lexikographie, Terminographie und Grammatikographie oder computergestützter Fremdsprachenunterricht .

> Personen, die Hypermediasysteme entwickeln und Interesse an der Integration computerlinguistischer Werkzeuge und/oder wissensbasierter natürlichsprachlicher Navigationshilfen haben.

> Personen, die mit der Konversion von Texten in Hyperdokumente bzw. der (Re) Konversion von Hypertexten in Texte befaßt sind.

Womit wird sich der AK Hypermedia befassen?

Was im einzelnen getan wird (werden kann), hängt von der Anzahl der AK-Mitglieder und deren Interessen ab. Denkbare Aktivitäten sind:

> Bestandaufnahme existierender und geplanter Hypermedia- Anwendungen im Bereich der angewandten Computerlinguistik

> Austausch von Erfahrungen, Tools, Programmen

- > Hypermediasysteme als neues Anwendungsfeld für computerlinguistische Werkzeuge und Methoden: Ideensammlung, Befragung von Hypermedia-Entwickler(innen)

- > Organisation themenspezifischer Workshops

Wer am Arbeitskreis mitarbeiten möchte, möge sich möglichst bald bei der unten angebenen Kontaktadresse melden. Für Mitte September ist ein erstes, konstituierendes Treffen in Mannheim geplant. Näheres ist ab Juli unter <http://www.ids-mannheim.de/projekt/grammis/ak-hypermedia.html> zu erfahren.

Kontakt:

Dr. Angelika Storrer Institut
für deutsche Sprache Abteilung
Grammatik
PO 101621
D-68016 Mannheim
Tel. 0621-1581207
Fax: 0621-1581200
e-mail: storrer@ids-mannheim.de

ERGEBNISSE DER GLDV-WAHLEN

Der GLDV-Wahlvorstand (I.S. Bátori, B. Schaeder, B. Schröder) zählte am 30.6.95 die eingegangenen Stimmen für die Wahlen zum Vorstand und zum Beirat der GLDV aus. Es waren 85 gültige Stimmzettel eingegangen.

Alle Mitglieder des Vorstands wurden wiedergewählt; im einzelnen entfielen

auf W. Lenders als Kandidaten für das Amt des 1. Vorsitzenden 78 Ja-Stimmen, 5 Nein-Stimmen und 2 Enthaltungen,

auf J. Haller als Kandidaten für das Amt des 2. Vorsitzenden 77 Ja-Stimmen, 5 Nein-Stimmen und 3 Enthaltungen,

auf R. Hausser als Kandidaten für das Amt des Schatzmeisters 76 Ja-Stimmen, 6 Nein-Stimmen und 3 Enthaltungen,

auf U. Seewald als Kandidatin für das Amt der Schriftführerin 81 Ja-Stimmen, 2 Nein-Stimmen und 2 Enthaltungen und

auf I. Hitzenberger als Kandidaten für das Amt des Informationsreferenten 77 Ja-Stimmen, 5 Nein-Stimmen und 3 Enthaltungen.

In den Beirat wurden gewählt: W. Hoepfner, G. Knorz, D. Rösner und U. Schmitz.