LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-Forum 10.1 (1993)
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

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 Schäder

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

Editorial

Sie haben es sicher bereits an einem neuen Element der Titelbildgestaltung bemerkt: Das zweite Heft des Jahres 1993 legt einen gewissen Schwerpunkt auf das Thema Morphologie. Und dies mit einem Bogen zwischen Vergangenheit und Zukunft, genauer mit einer ausführlichen Veranstaltungsbesprechung von U. Seewald und mit einer durchaus ungewöhnlichen Ausschreibung für einen Wettkampf computerlinguistischer Morphologieexperten, den Sie sich für den 7. und 8. März schon einmal vormerken sollten! Ich halte dieses Ereignis aus zweierlei Gründen für außerordentlich begrüßenswert: Es zeigt, daß der GLDV Initiative und Kreativität nicht abhanden gekommen sind, und es stellt einen Beitrag zum Thema Evaluierung dar, das in vielen Bereichen, aber gerade auch im Bereich der Computerlinguistik, oft sträflich vernachlässigt wird.

Auf ein zweites wichtiges Ereignis - ebenfalls ein Zeugnis für wachsende(7) Aktivitäten der GLDV - sei in diesem Zusammenhang hingewiesen: Die Herbstschule 1994! Auf die erste Ankündigung auf Seite 50 wird baldmöglichst eine ausführlichere Einladung folgen!

Wenn Sie sich schon ein wenig näher mit dem vorliegenden Heft beschäftigt haben, wird Ihnen aufgefallen sein, daß die Bitte um Mitarbeit bei den LeserInnen nicht ohne Echo geblieben ist. Insbesondere das aktive Interesse, Rezensionen zu erarbeiten, ist erfreulich. Wer das Beitragsspektrum noch eine Sekunde länger betrachtet, wird bemerken, daß der Schwerpunkt des Heftes besser noch als fachlich unter "Morphologie" geographisch unter "Erlangen/Nürnberg" einzuordnen wäre. Etwas direkter formuliert: Ohne die Aktivitäten von und um Hausser wäre es um das LDV-Forum schlecht bestellt gewesen. Es hat keinen Sinn, die mißliche Situation schamhaft zu tabuisieren: Es fehlt trotz gefüllter Seiten das, was bisher stets den fachlichen Kern des LDV-Forum ausgemacht hat: Es stand tatsächlich kein einziger Fachbeitrag zur Verfügung! Um es noch klarer auszudrücken: In dem ganzen Jahr, in dem ich nun das LDV-Forum verantworte, ist mir nicht ein Fachbeitrag angeboten worden (Der Beitrag im letzten Heft war eingeworben, für dieses Heft war es nicht gelungen). Verstehen Sie dieses Eingeständnis als dringende Bitte, das LDV-Forum bei Ihren Publikationsüberlegungen mit zu bedenken!

Zielen die letzten Feststellungen bereits auf die nächste und die weiteren Ausgaben, so auch die folgende Überlegung. Auf der letzten Beiratssitzung der GLDV war das Thema der hochschulseitigen Verantwortung für die berufliche Akzeptanz von Absolventen computerlinguistischer Studiengänge - in einem

für mich erstaunlichen Ausmaß - kontrovers gesehen worden. Zu diesem Thema ein Zitat aus dem vielbeachteten Titelbeitrag (Serie): "Welches Studium lohnt sich noch?" (Der Spiegel, 47 Jg., 1993, Nr. 42, 18. Okt. 93, S. 107/110):

> *Doch statt das Angebot in den traditionellen Studienfächern zu erweitern, konzipieren die Professoren lieber neue Spezialstudiengänge.*
>
> *Arbeitsvermittler Nettlau vom Hamburger Arbeitsamt mahnt seit Jahren die Abiturienten, nur ja nicht "auf alles reinzufallen, was die Universitäten anbieten ".*
> *Denn Flops gibt es zur Genüge. So ließ sich Petra Ricke,* 27, *vor acht Jahren zum Studium der Computerlinguistik verleiten. Den Sprachtechnikern, die Sprachübersetzungsprogramme entwickeln sollten, prophezeiten die Professoren damals eine große Zukunft, erinnert sich Ricke: Alle dachten: "Dieses Studium bringt's wirklich. "*
>
> *Nun sucht sie seit einem Jahr einen Job, wie die meisten ihrer Kommilitonen vergebens.*
> *Die Unternehmen haben kaum noch Interesse an den Exoten und stellen lieber Computerspezialisten ein. "Das macht keinen Spaß", sagt die Sprachexpertin, "wenn du erfährst, daß du überflüssig bist".*

Wenn Sie also auf das nächste Heft ungeduldig warten, dann vermutlich, weil Sie über den Ausgang der Morpholympics fundiert informiert sein wollen; weil Sie mit interessanten Fachbeiträgen sicher rechnen und weil Sie auf eine konstruktive Diskussion zum Thema berufliche Akzeptanz von Computerlinguisten und die Verantwortung von Hochschulen gespannt sind. Und wenn Sie zu diesem Thema glauben, etwas beitragen zu können, warten Sie nicht, daß man bei Ihnen anfragt: melden Sie sich bei mir!

G.K.

Titelgestaltung
Ute Hauck, 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!" bzw. 3!") als ASCII oder LATEX-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 11.1/94: 30. Apr. 1994; für Heft 11.2/94: 31. Okt. 1994

Herstellung
IAI, Saarbrücken

Druck
reha GmbH, Saarbrücken

Auflage
550 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.uni-sb.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

ZUR ENTWICKLUNG DER GLDV

Winfried Lenders
(1. Vorsitzender der GLDV)

Liebe Mitglieder der GLDV,

Mitte vorigen Jahres habe ich, nachdem Sie mich aufgrund des in der Satzung unserer Gesellschaft festgeschriebenen Verfahrens gewählt hatten, das Amt des ersten Vorsitzenden der GLDV von Burghard Rieger übernommen. Gleichzeitig wurden beinahe sämtliche anderen Vorstandsämter neu besetzt. Nachdem die berühmten 100 Tage längst vorbei sind, ist es jetzt sicher an der Zeit, daß Sie etwas über die Pläne und Vorstellungen dieses Ihres neuen Vorstandes erfahren.

Lassen Sie mich aber zunächst allen ausgeschiedenen Vorstandsmitgliedern, die viele Jahre, zum Teil über Jahrzehnte, für uns tätig waren, sehr herzlich für ihre erbrachte Arbeit danken. Es ist ein Dank für die tägliche Kleinarbeit, die ein Vorstandsamt mit sich bringt, für das 'Eintreiben' und Verwalten der Beiträge, das Führen der Kasse, das Protokollieren der Sitzungen, die Gänge zu Notaren, die Organisation von Tagungen, vielfältige Korrespondenz, die Leitung von Sitzungen, die Herausgabe des FORUMs, das Werben neuer Mitglieder, das Vertreten der Gesellschaft nach außen, den Versand von Schriften und für vieles andere mehr. Der alte hat dem neuen Vorstand ein 'wohlbestelltes' Feld hinterlassen, eine Gesellschaft von inzwischen ca. 320 Mitgliedern, mit zunehmender Tendenz, die auf beachtliche Erfolge in den letzten Jahren zurückblicken kann. Zu danken ist auch dem bisherigen Redakteur des LDV-FORUMs, Gerhard Knorz, der die letzten Hefte nur unter erschwerten Bedingungen betreut hat (vgl. das Editorial in Bd. 10, Nr. 1). Zu danken ist schließlich dem Beirat, den Arbeitskreisleitern, studentischen Hilfskräften und den vielen Mitgliedern, die sich z.B. in den Arbeitskreisen für die Belange unserer Gesellschaft eingesetzt haben.

Wenn oben von einem 'wohlbestellten Feld' die Rede war, so ist damit das gesamte Erscheinungsbild unserer Gesellschaft gemeint, aus dem einige Leistungen deutlich herausragen, und zwar das Erscheinen des Studienführers LDV ,Computerlinguistik, den wir wesentlich Frau Dr. Lutz-Hensel zu verdanken haben, die Vereinbarung mit anderen auf dem Gebiet der maschinellen Sprachverarbeitung tätigen Gesellschaften zur gemeinsamen Ausrichtung der Konvens-Tagung in zweijährigem Zyklus (die nächste Konvens findet im September in Wien statt, vgl. die Ankündigung in diesem Heft), die Konsolidierung unseres 'Vereinsorgans', des LDV-Forums, das nun als Zeitschrift mit eigenverantwortlichem Herausgeber erscheint, und die Idee der Herbstschule, die in diesem Jahr zum zweiten Mal durchgeführt wird (vgl. die Ankündigung in diesem Heft) .

Auf diesen und anderen Leistungen kann der neue Vorstand aufbauen; daß dabei neue und eigene Akzente gesetzt werden sollen, versteht sich von selbst.

Außer dem Vorstand wurde im vorigen Jahr auch der Beirat neu gebildet. Die Aufgabe des Beirats besteht darin, den Vorstand in seiner Tätigkeit zu beraten und Empfehlungen für geplante Aktivitäten zu geben. Wie es bisher schon Tradition ist, so sollen auch in Zukunft die Sitzungen des Beirats gemeinsam mit dem Vorstand

durchgeführt werden. Auf einer ersten gemeinsamen Sitzung konnten bereits zahlreiche Anregungen und Empfehlungen für das Arbeitsprogramm des Vorstandes besprochen werden.

Die wesentlichen Punkte dieses Arbeitsprogramms sind:

. Die wichtige und erfolgreiche Tätigkeit unserer Gesellschaft im Bereich Studienführer , Studienordnung, Berufsbild etc. soll unbedingt fortgesetzt werden. Der Arbeitskreis, der sich bisher mit diesen Fragen befaßt, soll neu konstituiert werden, insbesondere soll der Studienführer , der eine Übersicht über Studienorte, Studienordnungen zur Computerlinguistik in Deutschland enthält, fortgeschrieben und auf elektronischem Wege verfügbar gemacht werden. Vorstand und Beirat sind sich darüber im klaren, daß diese wichtige Aufgabe, wie bisher, auch in Zukunft auf den Schultern einiger weniger Personen liegen wird.

. Die Rolle der Arbeitskreise soll neu definiert bzw. belebt werden. Dabei wird von der Idee ausgegangen, daß sich ein gut Teil der Tätigkeit unserer Gesellschaft in den Arbeitskreisen abspielt. Hierzu waren zunächst alte Arbeitskreise zu reaktivieren und neue Arbeitskreise zu gründen. Die Arbeitskreise sollen sich mehrfach im Jahr treffen, sie sollen auch mit anderen Initiativen außerhalb der GLDV zusammenarbeiten und dadurch solche Initiativen in unsere Gesellschaft einbringen. Die Arbeitskreise sollen enger mit Vorstand und Beirat verbunden werden, indem die Arbeitskreisleiter regelmäßig an deren gemeinsamen Sitzungen teilnehmen. Ferner sollen die Arbeitskreise in Zukunft, dies hatte schon der vorige Vorstand beschlossen, aus der Vereinskasse in ihrer Arbeit bezuschusst werden können. Die Arbeitskreise sollen regelmäßig im FORUM oder im Newsletter berichten.

. Die Information der Mitglieder über aktuelle Ereignisse, Arbeitskreise, Tagungen, Veröffentlichungen etc. soll ausgebaut werden. Neben dem LDV-FORUM, das in alter, wenn auch neu definierter. Zuständigkeit regelmäßig erscheinen wird, gibt es wieder einen Newsletter, der in Kurzform die wichtigsten Neuigkeiten vermittelt. Dieser Newsletter, den der Informationsreferent betreut, soll vorerst noch per gelber Post, mittelfristig allerdings nur noch über e-mail verbreitet bzw. durch einen regelmäßigen elektronischen Nachrichtendienst ersetzt werden. Es sollen auch spezielle Nachrichten- und Datenserver (z.B. per anonymous-ftp zugreifbar) eingerichtet werden, die den Austausch von Information erleichtern, auf denen z.B. Bibliographien, Textdaten, Adresslisten abgelegt sind (natürlich unter Wahrung der Gebote des Datenschutzes).

. Die Zusammenarbeit mit anderen Gesellschaften im Bereich Linguistik und maschinelle Sprachverarbeitung, wie sie durch die gemeinsame Tagung KONVENS begonnen worden ist, soll intensiviert werden.

Bei der Bewältigung dieser Aufgaben ist der Vorstand auf die Mitwirkung aller Mitglieder angewiesen. Ich fordere Sie daher auf:

. Engagieren Sie sich in den GLDV-Arbeitskreisen. Eine Liste der bestehenden Arbeitskreise finden Sie im Newsletter; wenden Sie sich an den Leiter des Sie interessierenden Arbeitskreises und nehmen Sie an den Sitzungen der Arbeitskreise teil.

. Beteiligen Sie sich an den Tagungen der GLDV, nicht nur an unseren eigenen Jahrestagungen, sondern auch an der gemeinsamen Tagung "Konvens". Auf der nächsten Konvens in Wien sollte, das ist eine dringende Bitte, unsere Gesellschaft stärker vertreten sein, als auf der letzten in Nürnberg.

. Helfen Sie mit bei der Gestaltung unseres 'Vereinsorgans', des LDV-Forums. Senden Sie der Redaktion gute Beiträge, Nachrichten, die für alle interessant sind, Rezensionen etc. Fördern Sie den fachlichen Austausch durch Mitteilung Ihrer Erfahrungen in Lehre und Forschung im LDV-Forum (Rubrik Focus Computerlinguistik).

. Unterstützen Sie den Vorstand in seinem Bemühen um rasche Informierung aller Mitglieder. Teilen Sie uns, sobald Sie darüber verfügen, ihre e-mail- und fax-Adresse mit. Teilen Sie uns besondere Interessenschwerpunkte mit, damit Arbeitskreisleiter Sie ggf. gezielt anschreiben können. Und schließlich: Teilen Sie dem Vorsitzenden, Schatzmeister oder Schriftführer Adressen- und Kontoänderungen mit, damit wir die leider hohe Zahl der Irrläufer geringer halten können.

Von Ihrer Mitarbeit hängt, bei allem guten Willen des Vorstandes, der Erfolg unserer GLDV ab.

AUTOMATISCHE WORTFORMERKENNUNG IM DEUTSCHEN ANALYSEVERFAHREN UND SYSTEME
Ein Bericht über die auf dem 1. GLDV-Workshop zur automatischen Wortformerkennung in Erlangen präsentierten Entwicklungen

Uta Seewald

Universität Hannover

1 Der organisatorische Rahmen

Als Leiter des neu ins Leben gerufenen GLDV-Arbeitskreises "Parsing in Morphologie und Syntax" sowie als Initiator einer geplanten neuartigen Veranstaltungsreihe, MORPHOLYMPICS, hatte Roland Hausser, Leiter der Abteilung Computerlinguistik der Universität Erlangen- Nürnberg, für den 14. und 15. Oktober 1993 nach Erlangen eingeladen.. Die Initiative war aus der Beobachtung hervorgegangen, daß Programme zur automatischen Wortformerkennung sowohl in der Grundlagenforschung der Computerlinguistik als auch für praktische Anwendungen in der Textverarbeitung benötigt werden, eine Konsolidierung der in den vergangenen 20 Jahren entwickelten Teilergebnisse bisher jedoch nicht stattgefunden hat, nicht zuletzt auch aufgrund der schnellen Entwicklung auf dem Gebiet der Computerhard- und software sowie der linguistischen Theoriebildung.

Ziel der Veranstaltung war, die derzeit für das Deutsche in Entwicklung bzw. im Einsatz befindlichen Systeme zur automatischen Wortformerkennung vorzustellen und in einem aktuellen Überblick zu vergleichen. Ein besonderes Anliegen dieses Workshops lag auch darin, Bewertungskriterien für Systeme zur automatischen Wortformerkennung zu entwickeln. Zur Erarbeitung dieser Kriterien fand aus diesem Grunde eine gesonderte Diskussionsveranstaltung statt, deren Ergebnisse in einem Kriterienkatalog zur Bewertung automatischer Wortformerkennungssysteme gipfelte. Diese Kriterien sollen schließlich als Bewertungsmaßstab an jene Systeme angelegt werden, die sich dem Wettbewerb auf der 1. MORPHOLYMPICS stellen. Auf dieser für den Beginn des kommenden Jahres anberaumten Veranstaltung sollen verschiedene Systeme zur automatischen Wortformerkennung vorgestellt und *online* getestet werden.

2 Die vorgestellten Analyseverfahren und Systeme

Die zehn auf dem Workshop vorgestellten Systeme sind zum Teil im Hinblick auf bestimmte Anwendungsaspekte konzipiert worden, so daß ein unmittelbarer Vergleich der Systeme sowie der realisierten Analyseverfahren sich als schwierig oder zum Teil gar unmöglich erweist. Neben Systemen zur morphologischen Analyse und zum morphologischen Tagging finden sich denn auch Ansätze zur Inhaltserschließung von Lem-

mata sowie Hilfsmittel zur Verarbeitung von Nicht-ASCII-Zeichen.

Da die auf dem Workshop präsentierten Systeme alle *online* demonstriert wurden, war es den Teilnehmern dennoch möglich, bei Systemen mit ähnlicher Zielsetzung vergleichende Bewertungen anhand spontan zusammengetragener Testdaten vorzunehmen.

2.1 Word Manager (Mare Domenig, Universität Basel)

Der von Marc Domenig präsentierte Word Manager, der an der Universität Basel entwickelt wurde, ist ein System zur Erstellung und Nutzung morphologischer Wörterbücher. Das System beruht auf einer Netzwerkarchitektur, so daß verschiedene Benutzer bzw. natürlichsprachige Systeme auf ein Wörterbuch zugreifen können. Der Word Manager ist in der Lage, mehrere Sprachen zu bearbeiten und eine Vielzahl von Wörterbüchern zu verwalten. Die Idee, die diesem System zugrunde liegt, ist, - ein zentrales System bereitzustellen, daß das gesamte morphologische Wissen enthält, so daß verschiedene Anwendungen darüber verfügen können, ohne selbst mit einer morphologischen Komponente ausgestattet zu sein.

Ausgangspunkt der Konzeption des Word Manager ist das Wörterbuch. Diese Perspektive hat zur Entwicklung eines lexikalischen Datenbanksystems geführt, das im Vergleich zu verschiedenen regelbasierten Systemen den Vorzug hat, problemlos (von einem Linguisten) um große Datenmengen erweitert zu werden und für den menschlichen Benutzer doch transparent zu bleiben.

Nach der Vorstellung der Entwickler des Word Manager wird eine leere Datenbank zunächst von einem Linguisten bearbeitet, der Regeln über Flexion und Wortbildung einer bestimmten Sprache und entsprechende Beispiele formuliert sowie Ausnahmen von den aufgeführten Regeln angibt.

Die morphologische Regeldatenbank soll anschließend von einem Lexikographen, der lexikalische Einträge den formulierten Regeln zuordnet, weiter bearbeitet werden. Dabei kommt dem Lexikographen die Aufgabe zu, eine explizite Analyse einer Wortform in unmittelbare Konstituenten anzugeben und die bei jedem Schritt angewendete morphologische Regel zu nennen.

Die Sicht des Lexikographen auf das System wird durch ein Interface, das als CMKS (Conceptual Morphological Knowledge Specification environment) bezeichnet wird, geleistet. Bestandteil dieses Interface sind zahlreiche Fenster unterschiedlicher Funktionalität, auf die verschiedene Arten von Informationen verteilt sind. Zu diesen Fenstern gehören das Flexions- und das Wortbildungsfenster. Das Flexionsfenster zeigt Regeln und Formative an, die an der Flexion beteiligt sind, während das Wortbildungsfenster die entsprechenden an der Wortbildung beteiligten Regeln und Flexive visualisiert. Sowohl das Flexionsals auch das Wortbildungsfenster sind sogenannte *tree windows.* Bei der Implementierung von Wortbildungsmechanismen im Deutschen sind beispielsweise verzweigende Wortbildungsregeln erforderlich, die als Baumstruktur in einem Wortbildungsfenster angegeben werden.

Nachdem alle Regeln, Formative und Einträge einer Sprache angegeben sind, kann die betreffende Datenbank kompiliert werden. Stößt das Programm beim Kompilieren auf einen syntaktischen oder semantischen Fehler, wird eine Nachricht an das *message window* geschickt. Nach erfolgreicher Kompilierung stehen dem Benutzer zahlreiche *Browser* zur Verfügung. Zum Test bzw. zur Analyse einzelner Lexeme eignet sich vor allem der Lexembrowser, der ein Wort in seiner Zitierform zusammen mit kategorialen Angaben und möglichen davon abgeleiteten Wortformen ausgibt. Darüber hinaus enthält der Lexembrowser ein Teilfenster, das die am betreffenden Wort beteiligten Formative und deren syntaktische

Spezifikation auflistet. Spezifische morphologische Regeln, die sowohl bei der Wortbildung als auch bei der Flexion zum Tragen kommen, wie z.B. die Umlautung im Deutschen, werden ebenfalls in dafür vorgesehenen Fenstern dargestellt.

Durch die ausgebaute Fenster- und Browsertechnik sowie die menugesteuerten Funktionsaufrufe stellt der Word Manager dem Linguisten in einer offenen Client/Server-Architektur eine komfortable interaktive Entwicklungsumgebung zur Verfügung, die es erlaubt, große morphologische Wörterbücher effizient zu erstellen und zu verwalten.

2.2 Morphology Aid (Henriette Visser /Friederike Benjes, Universität Heidelberg)

Die von Henriette Visser und Friederike Benjes vorgestellte Morphology-Aid ist ein im Rahmen des Translater's Workbench-Projektes von Peter Hellwig entwickeltes Modul morphologischer Funktionen, das etwa von einem Texteditor aus aufgerufen werden kann, um morphologische Angaben einer Wortform abrufen zu können oder aber aufgrund der syntaktischen Angaben die korrekte Wortform eines Lemmas ermitteln zu können, um sie anschließend in den zu erstellenden Text zu übernehmen.

Der in der Morphology-Aid realisierte morphologische Ansatz geht auf das in den siebziger Jahren an der Universität Heidelberg entwickelte System PLAIN zurück. Das System greift auf zwei Lexika zu, wobei die morphologischen Phänomene einer Sprache in einem sogenannten morphosyntaktischen Lexikon beschrieben sind. Ein zweites Lexikon, als Valenzlexikon bezeichnet, enthält die Subkategorisierungsinformation, d.h. Angaben zur syntaktischen Kombinierbarkeit eines Lemmas. Das morphosyntaktische Lexikon ist als Übergangsnetzwerk aufgebaut und setzt sich aus mehreren Unternetzen zusammen. So existieren Unternetze für Stämme, solche für Derivationsmorpheme und andere für Flexionsendungen. Die morphosyntaktische Kombinierbarkeit der morphologischen Elemente wird durch die Übergänge im Netzwerk bzw. die Verbindungen zwischen den einzelnen Unternetzwerken beschrieben.

Die Eingabe eines Lexems in das morphosyntaktische Lexikon erfolgt anhand repräsentativer Wortformen des Lexems, in denen die möglichen Stämme des betreffenden Lexems enthalten sind. Im Fall des Verbs "sprechen" geschieht das anhand folgender Formen: *sprechen, sprichst, sprachst, sprächst, gesprochen,* das *gesprochene.* Anhand bestehender Muster erkennt das System die jeweils in den Wortformen enthaltenen Stämme und übernimmt sie in das Unternetzwerk, das Stämme enthält. Gleichzeitig werden die einzelnen Stämme mit Angaben über Netze bestimmter Flexionsformen verbunden, so daß entlang der autorisierten Netzwerkübergänge alle Formen der jeweils zu einem Lexem gehörenden Stämme erzeugt werden können. Aufgrund der in einem Wort enthaltenen Derivationsmorpheme und Flexionsendungen werden zunächst das Lexem sowie die morphosyntaktischen Angaben (Wortklasse, Kasus, Numerus, Genus, Person etc.) eines zu analysierenden Wortes ermittelt.

Die Anzeige der Flexionsparadigmen eines Lexems wird von sogenannten Tableaus gesteuert. Wird Information über eine bestimmte Wort form abgerufen, wird diese im Lexikon gesucht und die Stämme des zu dieser Wortform gehörenden Lexems ermittelt. Die möglichen Netzwerkübergänge und Verbindungen zu Unternetzen, die die Gesamtheit der Formen eines Lexems beschreiben, werden in einer sogenannten "section-table" gespeichert. Diese hat im Fall des Verbs "sprechen" folgende Form:

```
'sprech' Inf ,Prs1--t
     sprech    ' infen
     'sprech'    prs1-t
     'sprech'    prsK-est
     'sprech'    imp Sg-e
     'sprech'    imp Plt
```

```
     'sprech'      vBw-0
     'sprech' VN-en
          'Sprechen'       nSG-s
     'sprech' VN-er
          'Sprecher'       SMs-On
              'Sprecher'       mSg-s
              'Sprecher'       pI-On
     'sprech' VA-end
          'sprechend'      adjA
          'sprechend'      advO
       'sprich' Prs2Imp
     'sprich' prs2-st-t
     'sprich' impSg-O
'sprach' prtI-st
'spraech' prtK
'sproch' Ptz-ge-en
     'gesproch' ptzen
'sproch' VA-ge-en
     'gesproch' adjP-O
     'gesproch' adjA
```

Das System, mit dem menugesteuert verschiedene Optionen morphologischer Informationsgewinnung aufgerufen werden können, ist weniger für Zwecke der morphologischen Analyse großer Textmengen konzipiert, als für die interaktive Benutzung und den Zugriff von anderen Programmsystemen aus, wie es die Bezeichnung Morphology-Aid bereits nahelegt.

2.3 LA-Morph (Gerald Schüller/Oliver Lorenz, Universität Erlangen)

Das von Gerd Schüller und Oliver Lorenz vorgestellte System zur Wortformerkennung ist die Implementierung eines als LA-Morph bezeichneten Ansatzes zur morphologischen Analyse. LA-Morph basiert auf dem Algorithmus der Linksassoziativen Grammatik, der von R. Hausser (1992) mathematisch ausgewertet und (1989) im Hinblick auf morphologische Anwendungen beschrieben wurde.

Ein LA-Morph-System einer beliebigen Sprache setzt sich aus drei Komponenten zusammen: (1) einem Grundformenlexikon, (2) einer Regelmenge zur Ableitung von Allomorphen *(allo-rules)* aus den Einträgen des Grundformenlexikons und (3) einer Regelmenge, die die Kombination von Allomorphen *(combi-rules)* beschreibt. Eine der wesentlichen Maxime, die bei der Entwicklung von LA-Morph berücksichtigt wurde, ist die Speicherplatzeffizienz und die Verarbeitungsgeschwindigkeit des Systems. Um die Verarbeitungsgeschwindigkeit gering zu halten, werden beispielsweise die Regeln zur Erzeugung der Allomorphe vor der Laufzeit des Analysesystems angewendet, so daß das Lexikon zur Laufzeit bereits alle Allomorphe enthält. Enthält das Grundformenlexikon eines englischen Lexikons beispielsweise den Eintrag ("happy" (ADJ_GRAD) happy), dessen erste Position die Oberflächenform des Eintrags, dessen zweite Position die Kategorie und dessen dritte Position den Stamm bzw. die Semantik des Eintrags angibt, so sorgt eine entsprechende Allomorphieregel für die Erzeugung der zwei von dieser Grundform abgeleiteten Allomorphe ("happy" (ADJ) happy) und ("happi" (SR ADLGRAD) happy).

Allomorphieregeln bestehen aus einer Eingabebedingung und einer oder mehreren Allomorphdefinitionen. Die auf das englische Adjektiv *happy* angewendete Regel überprüft beispielsweise, ob die Eingabekette mit einem 'y' endet und ob das Lemma als (ADJ_GRAD) kategorisiert ist. Trifft das zu, so werden die beiden oben genannten Allomorphe erzeugt. Da bei einer zutreffenden Allomorphieregel die entsprechenden Allomorphe generiert und nachfolgende Allomorphieregeln ignoriert werden, ist die Abfolge der Regeln für die Erzeugung der Allomorphe von Bedeutung. Insbesondere ist wichtig, daß als letzte Regel eine Default-Regel zur Anwendung kommen kann, die solche Lemmata bearbeitet, deren einziges Allomorph mit dem Stamm identisch ist. Neben den aus dem Grundformenlexikon erzeugten Allomorphen enthält das Lexikon auch alle Affixe. Affixallomorphe werden - im Unterschied zu lexikalischen Morphemen - jedoch nicht über Regeln abgeleitet, sondern direkt spezifiziert, da es sich bei diesen um eine geschlossene Menge von Elementen handelt. Die Erzeugung der Allomorphe erfolgt in Abhängig-

keit vom Regularitätsgrad der jeweiligen Grundform. Für das Deutsche werden reguläre *(sagen -> sag),* semi-reguläre *(laecheln -> laechel) laech0,* semi-irreguläre *(hAus -> haus, haeus)* sowie irreguläre Lemmata *(gut -> gut, bess,* be) unterschieden.

Die morphologische Analyse wird zur Laufzeit von den Analyse- bzw. Kombinationsregeln gesteuert, die in editierbarer und in kompilierter Form vorliegen. Die Kombinationsregeln sind als links assoziative Regeln formuliert und enthalten im rechten Regelteil das Ergebnis aus der Konkatenation zweier im linken Regelteil enthaltenen Elemente sowie die Angabe eines Regelpaketes, das jene Regeln auflistet, die auf die nach Anwendung der Regel erzeugte Wortform angewendet werden können. Die erste Regel einer links assoziativen Ableitung beginnt mit einer Startregel, bei der als Startelement die leere Zeichenkette mit einem ersten Allomorph verbunden wird. Die morphosyntaktische Wohlgeformtheit von Wortformen wird mittels kategoriebasierter Restriktionen überprüft. Hierzu gehört beispielsweise die Angabe von Kategorien, die lexikalische Einträge charakterisieren, die unvollständig sind und die Anwendung einer Kombinationsregel erfordern, oder die Angabe von Elementen, die in einer Wortform niemals initial auftreten können.

Das Lexikon liegt bei der Analyse als *trie* vor, ebenfalls eine Maßnahme zur Effizienzsteigerung des Systems. Das deutsche Grundformenwörterbuch umfaßt ca. 13.000 Einträge, das Wörterbuch des englischen Systems ca. 8.000. Um eine Flexibilität des Systems zu gewährleisten, wurde LA-Morph auf verschiedene Plattformen portiert. Da sowohl das Lexikon als auch die Regeln deklarativ formuliert und als ASCII-Dateien gespeichert sind, ist das System praktisch sprachunabhängig.

2.4 Weiterentwicklungen von SADA W (Heinz Dieter Maas, IAI Saarbrücken)

Das von Heinz Dieter Maas vorgestellte Programmpaket *mpro,* das neben umfangreichen Lexikonfunktionen über die Möglichkeit verfügt, Wortformen zu generieren oder ein Texttagging durchzuführen, geht in seinen Anfängen auf das im Rahmen des SFB 100 entwickelte System SADAW zurück.

Das Programmpaket *mpro* analysiert Wörter des Deutschen morphologisch und liefert, sofern das entsprechende Ausgabeformat ('dima') gewählt wird, für Wörter bzw. Morpheme, die im Lexikon semantisch spezifiziert sind, Erklärungen, die den Inhalt der jeweiligen Wortbildung in Form einer Paraphrase darstellen. Neben dieser Option kann als Ausgabeformat der Analyse auch das CAT2-Format gewählt werden! Dabei werden sogenannte slex-Einträge erzeugt, die die Morphosyntax des zu analysierenden Wortes angeben, sowie sogenannte slex-Einträge, die die syntaktisch-semantische Information des betreffenden Wortes enthalten. Bei der Eingabe des Kompositums *Computerfreak* z.B., das in seine zwei nominalen Konstituenten zerlegt wird, enthält *mlex* als Wert des Attributs lex den Eintrag selbst, also 'Computerfreak' , der sich aus den lexikalischen Einheiten (Zu) 'Freak' und 'Computer' zusammensetzt. Die Angaben der syntaktisch-semantischen Information beziehen sich schließlich gesondert auf die beiden Konstituenten des Kompositums, wie das nachfolgende Analysebeispiel zeigt.

Eingabe: *Computerfreak*

```
mlex= {lex='Computerfreak', lu=freak,
       graphiks=eap, known=no, clu=
       {lu='eomputer', head={eat=n, ehead= {agr=
       {gen=masc}}, semf= {abstraet=eoner,
       anim=nil, temp=nil, bound=count,
       gran=nil}}},
```

l CAT2 ist ein maschinelles Übersetzungssytem, das als Seitenlinie von EUROTRA-D am IAI in Saarbrücken entwickelt wurde.

```
        head= {cat=n, deriv=nil, pref=_, ehead=
        {case=nom; dat; acc), agr= {num=sing,
        gen=masc}}}}.[].

slex= {lu='freak', head= {cat=n, ehead= {agr=
        {gen=masc}}, semf= {abstract=concr,
        temp=nil, gran=nil, anim= {'T'=hum,
        hum=male}}}}.[].

slex= {lu='computer', head= {cat=n, ehead= {agr =
        {gen=masc}}, semf= {abstract=concr,
        temp=nil, gran=nil, anim= nil,
        bound=count}}}.[].
```

Das Ausgabeformat mit den Angaben zur morphosemantischen Information des Kompositums enthält weitergehende Angaben, so z.B. daß es sich bei 'Computerfreak' um ein zählbares Substantiv handelt, das entweder im Nominativ, im Dativ oder im Akkusativ vorliegt und als menschliches Agens spezifiziert ist. Als Interpretation des obigen Wortes liefert *mpro* die Angabe "Ein Computerfreak ist jemand, der Computer gerne mag". In den Fällen; in denen das System eine Interpretation des zu analysierenden Wortes geben kann - bei dem Derivat 'Bäcker' lautet sie z.B. "Ein Bäcker ist jemand, der backt" - beruht diese auf einer semantischen Klassifikation der beteiligten Lexeme und Morpheme. So sind die im Lexikon enthaltenen Substantive in verschiedene Klassen untergliedert, wobei jede Klasse gleichsam als Hyperonym der unter sie fallenden Elemente aufgefaßt werden kann. Neben Klassen wie " abstract " , " agent" , "act" und "animal" sind z.B. Klassen für Gefäße ("box"), Krankheiten ("disease"), Instrumente ("instr"), Körperteile ("koerper") oder Materialien („material") vorgesehen. Das Substantiv *Computer* ist z.B. Element der Klasse ("instr"). Im Gegensatz zum Substantiv *Hammer,* das eben falls dieser Klasse angehört, ist *Computer* jedoch als komplexes Instrument markiert. Verben sind danach klassifiziert, welcher semantischen Klasse sie zuzuorden sind bzw. welche semantische Relation zwischen dem abgeleiteten Verb und seiner Basis vorliegt. Als Verb klassen finden sich beispielsweise "mitteilen" mit Verben wie *befehlen, diktieren, erzählen, melden, sagen,* "manipulate" mit Verben wie *ackern, boxen, drücken, reißen, stechen* oder "move_displace" mit Verben wie *fahren, fliegen, rollen* und *ziehen.* Eine Vielzahl von Verbklassen bezieht sich auf den jeweils vorliegenden Ableitungstyp. So gehören Ableitungen von Nomina des Typs *spionieren* (i.e. das tun, was ein Spion tut) zur mit "act_like" bezeichneten Klasse, während Verben, die ihr transitives Objekt mit über anschließen (Bsp.: *beherrschen* [über etw. herrschen], *bejammern* [über etw. jammern], *besiegen* [über etw. siegen], *bestaunen* [über etw. staunen]), als Elemente der Klasse "tra(ueber)" klassifiziert sind.

Bei der Segmentierung mit *mpro* werden alle in einer Wortbildung enthaltenen Stämme identifiziert, wobei auch Allomorphe berücksichtigt werden. Zu allen Teilketten der Zerlegung werden die Lexikoninformationen ermittelt, auf die dann je nach ausgewähltem Ausgabeformat zugegriffen wird. An zwei Beispielen seien die Angaben zum Inhalt der zu analysierenden Wörter abschließend noch einmal dargestellt:

kroatisch
Struktur des Wortes: [derived, a, n, isch, na,
[kroatien, [hyper=loc, loc=country]]]
Interpretation des Wortes:
kroatisch: bezieht sich auf Kroatien
Baecker Struktur des Wortes: [derived, n, v, er, er,
[backen, _31459]]] Interpretation des *Wortes:*
Ein Baecker ist jemand, der backt.

2.5 Korpusunterstützte Entwicklung lexikalischer Wissensbasen (Helmut Feldweg, Universität Tübingen)

Das von Helmut Feldweg beschriebene System zur Entwicklung lexikalischer Wissensbasen ist ein System ~ur automatischen Wortartenzuordnung für deutsche Texte (Taggingsystem), das nicht auf der Grundlage eines morphologischen Analysealgorithmus operiert, sondern auf einem sto-

chastischen Verfahren beruht. Das System LIKELY, um das es sich hier handelt, wurde im Rahmen des Tübinger Projektes ELWIS entwickelt, um umfangreiche annotierte Textkorpora des Deutschen zu erstellen. Das System baut auf einem Algorithmus zur stochastischen Wortartendisambiguierung auf, der von Marshall (1983, 1987) und .de Rose (1988) im Rahmen der Entwicklung von Taggingverfahren für das Englische beschrieben wurde.

Das System LIKELY sucht die einzelnen im Text auftretenden Wörter in einem Vollformenwörterbuch auf, das sowohl Angaben zu möglichen Wortarten des jeweiligen Wortes enthält als auch Angaben zu deren lexikalischen Häufigkeiten. Ist ein Wort in bezug auf seine Wortartenangaben ambig, so wird ein Netzwerk der verschiedenen Lesarten aufgebaut, das links und rechts durch ein eindeutig klassifiziertes Wort begrenzt wird. Die Übergangswahrscheinlichkeiten der aufeinander folgenden Wortarten werden einer Tabelle entnommen, und schließlich wird der durch- das Netzwerk führende optimale Pfad (mittels des Viterbi-Algorithmus) berechnet.

Das System unterscheidet 40 Wortarten, eine Teilmenge jener Wortarten, die im Saarbrücker Lemmatisierungsprogramm SALEM verwendet wurden, auf dessen Erfahrungen bei der Konzeption von LIKELY aufgebaut wurde. Die Wahrscheinlichkeitswerte, auf die bei der Analyse zugegriffen wird, werden einer sogenannten Übergangsmatrix entnommen, die im Falle von LIKELY zweidimensional ist. D.h. die in ihr angegebenen Werte basieren auf den Werten, die aus den absoluten Häufigkeiten der im Referenzkorpus auftretenden Zweierwortgruppen, i.e. Wortbigramme, errechnet wurden. Die Matrix der Übergangswahrscheinlichkeiten sowie ein Vollformenwörterbuch mit der Angabe der Wortarten und der jeweiligen relativen lexikalischen Wahrscheinlichkeiten wurden aus einem Trainingskorpus, das sich aus den ersten zwei Dritteln eines 239889 Wörter umfassenden Referenzkorpus zusammensetzte, ermittelt. Das Referenzkorpus seinerseits basiert auf drei in analysierter Form vorliegenden Segmenten des Mannheimer Korpus I, die mit dem Saarbrücker System SALEM bearbeitet worden waren, und hier einer maschinellen und intellektuellen Überarbeitung unterzogen wurden.

Auf einer Sun Sparcstation 10/20 annotiert das System ca. 6000 Wortformen pro Sekunde. Die ermittelte Fehlerquote von 7,38 % bei unvollständigem Wörterbuch, d.h. einem Wörterbuch, das nur die in einem vorgegebenen Trainingskorpus auftretenden Wörter, deren Wortarten und lexikalische Häufigkeiten berücksichtigt, ergibt sich u. a. aus den verschiedenen zur Analyse herangezogenen Textsorten. Das Ergebnis der Annotierung könnte sowohl durch ein umfangreicheres Wörterbuch als auch eine anschließende morphologische Analyse verbessert werden.

2.6 X2MORF - erweiterte 2-Ebenen-Morphologie (Harald Trost, Universität Wien)

Bei dem von Harald Trost vorgestellten System X2MORF handelt es sich um ein System, das ein erweitertes Zwei-Ebenen-Modell (Two-Level Morphology)2 mit einem merkmalbasierten Lexikon kombiniert. In der ursprünglichen Formulierung des Two-Level-Modells wurden Wortbildungsregeln als reguläre Grammatik.. in Form von Fortsetzungsklassen ausgedrückt. In X2MORF wurden diese Fortsetzungsklassen durch eine merkmalbasierte Wortbildungsgrammatik ersetzt. Darüber hinaus wurden die Zwei-Ebenen-Regeln mit einem Filter versehen, der die Anwendung der Regeln auf bestimmte morphologische Klassen beschränkt. Da der Filter als morphologischer Kontext in Form von Merkmalstrukturen ausgedrückt wird, können die Merkmalstrukturen der Two-Level-Regeln mit

2 Koskenniemi (1983)

den Merkmalstrukturen des jeweils konkret vorliegenden Morphs unifiziert werden.

Bei der Analyse greift das System auf ein Morphlexikon zu, das zu jedem Morph eine Merkmalstruktur enthält. Es handelt sich aufgrund der jedem einzelnen Morph zugeordneten Information also um ein morph- bzw. morphembasiertes System, im Gegensatz etwa zu einem paradigmenorientierten System. Dies bildet die Voraussetzung dafür, daß das System neben Flexion auch Derivation und Komposition behandeln kann. Dem Morphlexikon gewissermaßen übergeordnet ist ein Lexemlexikon, aus dem die syntaktisch-semantische Information eines Lexems entnommen wird. Unterhalb des Morphlexikons sind die Zwei-Ebenen-Regeln angesiedelt, die phonologische Phänomene behandeln. Durch die den Morphen zugeordneten Merkmale werden die Zwei-Ebenen-Regeln um einen morphologischen Kontext erweitert. Auf diese Art und Weise können mit dem System auch morphologische Phänomene wie nichtkonkatenative Morphologie, Schwa-Epenthese oder Ablaut und Umlaut, die in der ursprünglichen Form des Zwei-Ebenen-Modells problematisch waren, behandelt werden.

Die Kanten des dem Analysesystem zugrunde liegenden Übergangsnetzwerkes enthalten einen Filter. Anhand dieses Filters besteht die Möglichkeit, Tests durchzuführen, um bestimmte Morphkonkatenationen gegebenenfalls zu blockieren. Die Regel "A: ä {=}-; [MORPH: [HEAD: [UMLAUT: [VALUE: +]]]]" enthält bspw. in dem dem Semikolon folgenden Klammerausdruck den Kontext für die Umlautregel, Dieser Kontext wird hergestellt durch Suffixe, die Umlaut erfordern, d.h. deren entsprechender Wert mit '+' angegeben ist, so daß bei der Morphemkonkatenation in der Merkmalstruktur des Stamms das Attribut Umlaut schließlich in Abhängigkeit von den jeweiligen Suffixen bzw. deren Umlautwerten gesetzt wird. Auf diese Weise ist es möglich, kränklich und handlich zuzulassen, nicht aber händlich. Diese Bildung wird durch Unifikation ausgefiltert.

Insgesamt werden unnötige Unifikationen bei der Analyse vermieden, um die Effizienz des Systems zu steigern.

2.7 Deutsche Flexions- und Kompositionsmorphologie auf 2-Ebenen-Basis (Anne Schiller, Universität Stuttgart)

Auch das von Anne Schiller vorgestellte morphologische Analysesystem basiert auf dem Zwei-Ebenen-Modell, wobei es gegenüber dem ursprünglichen Modell von Koskenniemi nur geringfügige Veränderungen aufweist, was auch durch die Verwendung des PC-Kimmo-Systems als Grundlage der Implementierung deutlich wird.
Ziel dieser Arbeit ist die Erstellung eines. großen Lexikons mittels der hier beschriebenen morphologischen Analyse.

Entsprechend der in der Zwei-Ebenen-Morphologie eingeführten Konvention werden Morphemgrenzen oder auf lexikalischer Ebene unterspezifizierte Merkmale durch diakritische Zeichen dargestellt. Je nachdem ob ein morphologischer Stamm umlautet oder nicht, wird die Morphemgrenze zum nachfolgenden Suffixmorphem unterschiedlich dargestellt.

Bsp: jung+er →jünger
jung(+)er →junger

Die Einträge des aus bestehenden Wörtern zu kreierenden Lexikonteils werden hergeleitet aus Wortlisten, die kategoriale Angaben zu den einzelnen Wörtern enthalten. Da das System in erster Linie für die Analyse verwendet wird, enthält es auch Regeln, die zu Übergenerierung führen.

Diskontinuierliche Affixe, wie sie bei der Flexion von Verben im Deutschen beispielsweise im Fall des Partizips Perfekt auftreten können, werden im Lexikon ebenfalls mittels diakritischer Zeichen angegeben. Der

Stamm des Infinitivs *bauen* wird dementsprechend mit dem Merkmal [-Part] versehen, so daß aufgrund der lexikalischen Form „[-Part]bau" bei der Bildung des Partizips die Form *eingebaut* erkannt, *eingebaut* jedoch blockiert wird. Regeln für die Umlautung werden ebenfalls durch Diakritika an den Morphemgrenzen angegeben (Bsp: *Haus $er* oder: *groß $er).* Über solche Regeln hinaus verfügt das System über Kontrollregeln, die beispielsweise Fälle separierbarer präfixaler Elemente kontrollieren. So sorgt die Regel " [-Imp] :=::} [-Sep]_*" dafür, daß aufgrund des. Lexikoneintrags von *steh,* der als "steh[-Imp]" angegeben ist, und desjenigen von auf, der mit "auf[Sep]" angegeben ist, zwar *steht,* aber nicht *aufsteht* akzeptiert wird. Komposita des Typs N N (Substantiv+Substantiv), N Adj (Substantiv+Adjektiv) oder Adj N (Adjektiv+Substantiv) können im vorliegenden System auch unter Zuhilfenahme von Kontrollregeln bearbeitet werden, erfordern jedoch die Einführung spezieller zusätzlicher Kontrollregeln.

Das System in seiner derzeitigen Form verfügt über 15 Regeln. Die übrigen morphosyntaktischen Zusammenhänge werden unter Verwendung von Fortsetzungsklassen ausgedrückt. Für Nomina bestehen insgesamt 60 Fortsetzungsklassen, für Adjektive 20 und für Verben 40 solcher Klassen.

2.8 MORPH - ein modulares und robustes Morphologieprogramm für das Deutsche (Gerhard Hanrieder, FORWISS Erlangen)

Das von Gerhard Hanrieder entwickelte System MORPH ist ein morphologisches Programm, das sich aus drei Analysemodulen zusammensetzt. Ein Modul ist für die Flexionsanalyse konzipiert, ein weiteres Modul bearbeitet Wortbildungen, und das dritte Modul bearbeitet Formen mit unbekannten morphologischen Kernen und erzeugt auf der Grundlage der jeweils vorliegenden Endungen Hypothesen über die morphosyntaktischen Kategorien. Sofern eine Wortform also nicht mittels der ersten beiden Module analysiert werden kann, werden im dritten Modul Hypothesen über seine morphosyntaktische Kategorie angegeben. Dieses dritte Modul verleiht dem System den Aspekt der Robustheit.

Bei der Analyse eines Wortes werden alle möglichen Lesarten ausgegeben. Die Lexikoneinträge enthalten keine Angaben über Valenzen. Alle Einträge enthalten ausschließlich Angaben zu morphosyntaktischen Merkmalen wie Wortklasse, Numerus oder Kasus. Das Lexikon ist ein gemischtes Voll- und Stammformenlexikon, das als Liste von Listen angelegt ist. Kürzel zur Bezeichnung von Morphemklassen stellen die Verbindung einzelner Morpheme zu Klassen von Morphemen her. Mehrdeutigkeiten sind im Lexikon als verschiedene Lesarten kodiert, Allomorphe sind als gesonderte Einträge verzeichnet, wie das Beispiel (HAUS (MSING (ALLO HAEUS))) zeigt.

Die Lexikonlisten sind aus Effizienzgründen als Buchstabenbäume kompiliert. Die Suche eines Eintrags im Lexikon kann somit auf der Grundlage einer Baumtraversierungsfunktion erfolgen. Die Wortbildungsanalyse des zweiten Moduls erfolgt als *left-to-right-Analyse.* Die Kombination von Morphemen ist jeweils als Übergangsnetzwerk implementiert. Die möglichen Nachfolger eines Substantivs (N) sind z.B. angegeben mit (N), (Adj), (Präf) und (Suff).

Bei der " Endungshypothesenanalyse" wird nach dem Verfahren des *longest matching* die Endung eines Wortes mit unbekannter Basis isoliert. Die einer Endung zugeordneten Merkmale erlauben schließlich, Hypothesen über die vorliegende Wortform anzugeben.

2.9 Informationsgewinnung aus der Struktur lexikalischer Lemmata (Nico Weber, Universität Bonn)

Die von Nico Weber vorgestellte Informationsgewinnung aus der Struktur lexika

lischer Lemmata basiert auf der Untersuchung von Definitionstexten eines maschinenlesbaren Wörterbuchausschnitts des Deutschen3, in denen die morphologische Struktur der Lemmata thematisiert bzw. durch geeignete Lexeme variiert wird. Besonders ergiebig für die Informationsgewinnung aus der Struktur lexikalischer Lemmata haben sich dabei 1- bis 3-WortDefinitionen erwiesen, ein Definitionstyp, der - so die Schätzungen des Autors ca. 90 % der Definitionen im betrachteten Wörterbuchausschnitt umfaßt. Definitionen dieses Typs enthalten als Definiens des in Frage stehenden Lemmas meistens ein Wort, das entweder das Lemma, die Bedeutung von Ableitungsaffixen am Lemma oder die Wortbildungsbedeutung eines abgeleiteten bzw. zusammengesetzten Lemmas variiert.

Außer den Satzbanddaten des DUW zur Sichtung und anschließenden Klassifikation von Definitionstexten nach phrasalen Strukturen wurden zu Testanalysen 300.000 Einträge der Bonner. Wortdatenbank (Bonniex), 230.000 Einträge des Wortanalytischen Wörterbuchs [WAW]4, die sowohl in segmentierter als auch in unsegmentierter Form vorliegen, sowie eine Morphemliste mit 7677 Elementen herangezogen:

Bei den verwendeten Tools zur Extraktion der Daten aus den Definitionstexten handelt es sich um einen Satzbandparser mit SGML-konformer Ausgabe sowie. um ein Selektionsprogramm für Lemmata, Lesartenangaben und die eigentlichen Definitionstexte.

Der morphologische Parser, der die vorbereitende Segmentierung der Lemmata und Einwortdefinitionen vornimmt, ist lexikonorientiert und wurde auf der Grundlage der Daten des Wortanalytischen Wörterbuchs (WAW) erstellt. Die Segmentierung erfolgt von links nach rechts nach dem Verfahren des *longest matching.* Das Segmentierungsergebnis wird zunächst daraufhin überprüft, ob es in der vom Segmentierungsprogramm gelieferten Form im W A W, das "Einträge in segmentierter Form verzeichnet, vorhanden. ist. Wird im WAW keine entsprechende Segmentierung des in Frage stehenden Wortes gefunden, wird eine Kompositaanalyse durchgeführt. Die dabei ermittelten Lexeme werden mit den Wörterbucheinträgen verglichen. Ist das betreffende Lexem im WAW enthalten, wird es auf die in den Definitionstexten auftretenden Variationen seiner Wortbildungsbedeutung hin überprüf~. Verschiedene Typen von Variationen treten hier auf:5

(1) Kombinatorische Varianten: Lexeme variieren]

Bahn	verbindung
Zug	verbindung
Abend	mahl
Abend	essen
Ausguß	becken
Ausguß	0

(2) Diasystematische Ersetzung: [Lexemsubstitution]

achtern
hinten

(3) Affixexplikation mit Stammwiederholung:

[Präfixe variieren]

Ab	-	lauf
Ver	-	lauf

[Suffixe variieren]

abänder - lich
abänder - bar

Die in den Lexemdefinitionen auftretenden Alternanzen lassen sich semantisch in-

3 Es handelt sich bei dem betrachteten Wörterbuchausschnitt um die Einträge <A> bis <Band> des in Form maschinenlesbarer Satzbanddaten vorliegenden Buden - Deutsches Universalwörterbuch [DUW]..

4 Wortanalytisches Wörterbuch. Deutscher Wortschatz nach Sinn-Elementen (192ff.). Kandler, Günther/Winter Stefan. München: Wilhelm Fink Vlg.

5 Die Unterscheidung Webers nimmt die Klassifikation morphosemantischer Definitionen von J. Rey-Debove (1971) auf. Vlg. Weber (1992:10).

terpretieren auf der Grundlage der klassischen Relationen wie Synonymie, Hyponymie oder Antonymie. Der Parser, der die Analyse der Lexemdefinitionen vornimmt, sucht nun mit regulären Ausdrücken nach formalen Variationsmustern. Die regulären Ausdrücke berücksichtigen die Forderung, daß in zwei variierenden Wörtern mindestens ein Element an derselben Position materiell identisch auftreten muß. Der Abgleich mit dem Referenzmuster Ax Lx ist z.B. möglich in Fällen wie

Ax Lx:	*Ab-zug*	*Ab-guß*	*Ab-tausch*
"-Lx	*Abdruck*		
Ax-"		*Aus-guß*	
0-"			0- *Tausch*

Dabei variiert jeweils die im Suchpattern explizit aufgeführte Form (Lx = Lexem; Ax = Affix; 0 = Nullmorphem), während das mit dem Referenzmuster identische Element jeweils durch <<“>> angegeben wird. Eine Interpretation kann sowohl bei Komposita als auch bei Derivaten erfolgen. Die Derivationsbedeutung wird auf der Grundlage der Derivationsaffixe und den jeweils vorliegenden Definitionstexten ermittelt. So werden beispielsweise Ableitungen mit dem Suffix *ung* wie *Programmierung* in der Regel durch eine entsprechende substantivierte Infinitivform *(das Programmieren)* wiedergegeben, wodurch das Suffix eine prozessuale Interpretation erhält, und aufgrund des Definitionstextes des Eintrags *andünsten* ("kurz dünsten") kann dem Präfix *an* z.B. die Bedeutung "kurz" zugeordnet werden.

2.10 Schnittstellen zu Nicht-ASCII-Zeichen (S.Y. Cho, Universität Saarbrücken)

Den Abschluß der Präsentationen des Erlanger Workshops bildete die Vorstellung der Schnittstellen zu Nicht-ASCII-Zeichen von S.- Y. Cho, der die Möglichkeiten der Darstellung koreanischer, japanischer oder chinesischer Schriftzeichen in einem koreanischen Textsystem auf der einen bzw. einem englischen Textsystem auf der anderen Seite demonstrierte. Im Vordergrund dieses Beitrags stand kein morphologisches Analyseverfahren, sondern die von Cho und Lew 1992 entwickelte Lösung des Problems, in Textverarbeitungs- oder natürlichsprachlichen Systemen auf Schriftzeichen, die nicht auf dem lateinischen Alphabet beruhen bzw. nicht im Umfang der ASCII-Codierung enthalten sind, zugreifen zu können.

Die vorgelegte Lösung basiert auf der internen Darstellung eines Hangul-Zeichens durch zwei ASCII-Codes. Im koreanischen Textverarbeitungsprogramm lassen sich die Zeichen so in ihrer koreanischen Form anzeigen. In einer englischen Programmumgebung werden die jeweiligen Zeichen den zwei ASCII-Codes entsprechend durch zwei miteinander verbunden ASCIIZeichen (etwa: ea für ASCII: 136 97) wiedergegeben.

3 Ausblick

Die verschiedenen auf dem Workshop präsentierten Ansätze und Systeme sowie die sich an die Systemvorführungen anschließende engagierte Diskussion der Workshopteilnehmer über Bewertungskriterien für automatische Wortformerkennungssysteme deuten sicher schon jetzt auf einen spannenden Verlauf der 1. MORPHOLYMPICS, auf der sich im März nächsten Jahres konkurrierende Systeme zur Wortformerkennung des Deutschen einer Jury stellen sowie um den ersten Platz und ein Preisgeld <computerlinguistisch kämpfen> werden.

Literatur

De Rose, S.J. (1988): Grammatical category disambiguation by statistical optimization. In: Computational Linguistics, 14/1, S. 31-39.

Domenig, Marc/Ten Hacken, Pius (1992): Word Manager: A System for Morphological Dictionaries. Hildesheim: Olms.

Feldweg, Helmut (1993):
Stochastische Wortartendisambiguierung für das Deutsche. Untersuchungen mit dem robusten System LIKELY. SfS-Report-08-93, Universität Tübingen.

Hausser, Roland (1989): Principles of Computational Morphology. Laboratory of Computational Linguistics, Carnegie Mellon University.

Hausser, Roland/Schüller, Gerald/Zierl , Marco (1993): MAGIC. A Tutorial in Computational Morphology. Univ. Erlangen-Nürnberg.

Hellwig, Peter (1992): The Morphology-Aid Function. Workpackage 2.4., T2, Universität Heidelberg.

Koskenniemi, Kimmo (1983): Two-level morphology. A general computational theory for word-form recognition and production. Department of General Linguistics, University of Helsinki, Publications no. 11.

Marshall. I. (1983): Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus. In: Computers in the Humanities, 17, S. 139-150.

Marshall. I. (1987): Tag selection using probabilistic methods. In: Garside, R./Leech, G./Sampson, G. (Hrsg.): The computational analysis of English. London/New York: Longman, S. 4256.

Trost, Harald (1993): Coping with Derivation in a Morphological Component. Erscheint
in: Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL '93), Utrecht.

Weber, Nico (1992):
Morphosemantische Wörterbuchdefinitionen. In: Sprache und Datenverarbeitung 16/2, S. 45"::63.

Uta Seewald, Hannover

AUSSCHREIBUNG DER 1. MORPHOLYMPICS FÜR AUTOMATISCHE WORTFORMERKENNUNG

Roland Hausser
7. und 8. März 1994, Universität Erlangen-Nürnberg

Die *Morpholympics* ist ein neuartiger Typ von Tagung, auf dem sich unterschiedliche Systeme zur automatischen Wortformerkennung in einem öffentlichen Wettlauf messen. Der Wettbewerb wird von einer unparteilichen Jury aus 5 Preisrichtern beobachtet, die die Sieger bestimmt.

Im Folgenden werden Ablauf der Morpholympics und Teilnahmemodalitäten genauer erläutert. Eine analoge Veranstaltung für syntaktische Parser ist von der GLDV unter dem Namen *Parsolympics* für das Jahr 1995 geplant.

1 Organisation und Ablauf

1.1 Allgemeine organisatorische Struktur der Morpholympics

1.1.1 Zielsetzung

Ziel der Morpholympics ist ein objektiver, theorieunabhängiger Vergleich existierender Systeme zur automatischen Wortformerkennung. Dabei sollen linguistische Motivation, technische Konzeption, Datenabdeckung und Geschwindigkeit bewertet werden. Für den praktischen Vergleich der Systeme ist es sinnvoll, Testdaten aus einer natürlichen Sprache in den Mittelpunkt zu stellen. Dies ist die sogenannte *Haupttestsprache.*

1.1.2 Turnusmäßige Gestaltung

Die Morpholympics werden, je nach Bedarf, alle zwei bis vier Jahre abgehalten. Dabei sollen die Systeme an einer jeweils neuen natürlichen Sprache (*Haupttestsprache)* getestet werden. Tagungssprachen sind Englisch und die jeweilige Haupttestsprache.

Um die Weiterentwicklung bereits getesteter Systeme und Sprachen zu fördern, werden auf einer gegebenen Morpholympics auch die Sprachen früherer Wettkämpfe als *Nebentestsprachen* zugelassen. Auf diese Weise kann der *defending champion* einer ehemaligen Haupttestsprache herausgefordert werden. Voraussetzung für die Aufnahme einer Sprache als Nebentestsprache ist, daß sich mindestens 3 Teilnehmer für diese Sprache melden.

Wenn nach mehreren Morpholympics das Interesse an einer ehemaligen Haupttestsprache wieder genügend angewachsen ist, kann diese zur neuen Haupttestsprache bestimmt werden.

1.1.3 Koordination

Die Verantwortung für die allgemeine Planung, Koordination und anschließende Publikation einer Morpholympics liegt in den Händen eines KOORDINATORS. Der Koordinator soll ein in Computerlinguistik aus

gewiesener Wissenschaftler sein, der gleichzeitig Repräsentant einer nationalen oder internationalen computerlinguistischen Organisation (Trägerorganisation) ist. Der Koordinator sorgt für die Plazierung von Austragungsort und -zeitpunkt, gewinnt die Zusage von fünf kompetenten und neutralen Preisrichtern und trägt die Gesamtverantwortung für den reibungslosen Ablauf. Der Koordinator kann nicht gleichzeitig als Preisrichter fungieren.

1.1.4 Initiierung der Nachfolge Morpholympics

Der Arbeitskreis *Parsing in Morphologie und Syntax* der GLDV bestimmt in Absprache mit dem Koordinator und der Trägerorganisation der letzten Morpholympics den Koordinator der nächsten Morpholympics.

1.1.5 Gestaltung der Preise

Der Preis für den Sieger in der Haupttestsprache einer Morpholympics besteht in einer Urkunde, einer Trophäe und einem Preisgeld. Die Trophäe wird zu Beginn der nächsten Morpholympics dem dann aktiven Koordinator zugestellt und erneut vergeben (Wanderpreis). Abhängig von der Zahl der Teilnehmer ist es möglich, einen zweiten und dritten Preis in der Haupttestsprache, sowie jeweils einen Preis für jede der N ebentestsprachen in Form von Urkunden zu vergeben. Die Preisgelder werden vom Koordinator von industriellen Sponsoren eingeworben.

1.1.6 Zeitlicher Ablauf

Termin und Ort einer Morpholympics, sowie die Haupt- und Nebentestsprachen, werden in der Regel mindestens drei Monate vor der Veranstaltung bekannt gegeben. Zwei Monate vor der Veranstaltung haben potentielle Teilnehmer Zugang zu einem Fragebogen (siehe 2, 'Standardisierte Darstellung der Systeme'), den vorläufigen Testdaten (siehe 3.1) und dem Anmeldeformular (siehe 4). Um an der Tagung teilzunehmen, muß das ausgefüllte Anmeldeformular bis zu dem vom Koordinator genannten Anmeldeschluß eingereicht werden.

1.1. 7 Form der 'Wettläufe'

Vor Beginn der Tagung können Teilnehmer auf einem geeigneten Unix-Rechner am Veranstaltungsort einen 10gin erhalten und haben damit Gelegenheit, ihre Systeme per rlogin über das Netz zu installieren. Für die Installation von MAC- und PC- Versionen haben Teilnehmer zwei Tage vor der Veranstaltung Zugang zu den Rechnern am Veranstaltungsort (können aber auch ihre eigenen Geräte mitbringen).

Die offiziell angemeldeten Teilnehmer treffen sich zwischen 9:00 und 9:50 am Morgen des ersten Konferenztages. Bei dieser Gelegenheit wird der ausgefüllte Fragebogen in 6-facher Ausführung bei den Veranstaltern abgegeben (fünf für die Preisrichter, einer für die abschließende Publikation) und die Reihenfolge der 'Wettläufe' durch das Los bestimmt. Jedes System wird durch maximal zwei offizielle Teilnehmer repräsentiert.

Der 'Wettlauf' eines Systems besteht aus

I> einer PRÄSENTATION MIT PRIMÄREM TESTLAUF und

I> zwei SEKUNDÄREN TESTLÄUFEN

Die Präsentation eines Systems dauert 50 Minuten. Zu Beginn der Präsentation erhalten die Teilnehmer die endgültigen Testdaten (siehe 3.3) *on line* und starten ihr System auf einem Unix-Rechner. Dann stellen sie ihr System auf der Grundlage des eingereichten Fragebogens in Form eines Vortrags mit Folien dar. Die Ergebnisse der Tests auf dem Unix-Rechner (primärer Testlauf mit Zahl der erkannten Wortformen, Geschwindigkeit, Speicher bedarf und Dauer des *turn around,* siehe 3.2) werden per LCD-Display während der Präsentation bekannt gegeben und von den Veranstaltern als *hardcopy an* die Preisrichter verteilt.

Nach der Präsentation eines Systems, und parallel zur fortlaufenden Hauptveranstaltung, werden in einem separaten Raum sekundäre Tests unter Aufsicht unabhängiger Protokollführer durchgeführt. In den sekundären Tests werden die endgültigen Testdaten noch einmal auf PC und MAC analysiert. Der Sinn der sekundären Tests ist, die Portierbarkeit des jeweiligen Systems und die Verwendbarkeit unter kommerziellen Bedingungen zu demonstrieren. Die Ergebnisse der sekundären Tests werden den Preisrichtern von den Protokollführern als *hardcopy* nachgereicht.

1.1.8 Grundlagen der Bewertung

Die Entscheidung der Preisrichter basiert auf

1. der standardisierten schriftlichen Darstellung des Systems (2) und - falls vorhanden - Exemplaren der Dokumentation,
2. dem mündlichen Vortrag und
3. den Ergebnissen der endgültigen (primären und sekundären) Tests.

1.1.9 Bestimmung des Siegers, Preisvergabe und Publikation

Nach Abschluß der letzten Präsentation und des letzten sekundären Tests ziehen sich die Preisrichter mit ihren Unterlagen einzeln für ca. 90 Minuten zurück, um ihre Entscheidungen zu treffen, bzw. noch einmal zu überdenken. Anschließend treffen sich die Preisrichter, der Koordinator, Vorstandsmitglieder seiner Organisation und die Vorstandsmitglieder der GLDV für etwa 30 Minuten. Bei dieser Sitzung teilen die Preisrichter ihre Entscheidungen mit und bestimmen die Sieger.

Anschließend werden die Ergebnisse den Teilnehmern und dem Publikum mitgeteilt und die Preise vergeben. Mit ihrer Anmeldung zur Morpholympics haben die Teilnehmer ihr Einverständnis für eine Publikation ihrer Systembeschreibung und der aktuellen Testergebnisse gegeben (siehe 4.2). Es ist die Aufgabe des Koordinators, daß diese baldmöglichst als Sammelband (oder Sonderheft einer Zeitschrift) einem breiteren Publikum bekannt gemacht werden.

1.2 Spezifika der 1. Morpholympics

1.2.1 Koordinator

Koordinator der 1. Morpholympics ist Roland Hausser als Vorstandsmitglied der GLDV und Urheber dieses Veranstaltungstyps.

1.2.2 Termine

Anmeldeschluß: 20. Februar 1994
Veranstaltungsdatum der 1. Morpholympics (1994):
Montag, den 7. März und Dienstag, den 8. März, 1994

1.2.3 Veranstaltungsort:

Abteilung Computerlinguistik,
Universität Erlangen-Nürnberg,
Bismarckstr. 12, 91054 Erlangen.

Kommentare, Fragen, etc. bitte E-Mail an: rrh@linguistik.uni-erlangen.de (Roland Hausser)
Für remote logins und technische Hilfe bitte E-Mail an: rolf@linguistik.uni-erlangen.de (Rolf Haberrecker)

1.2.4 Testsprache(n)

Die Haupttestsprache der 1. Morpholympics ist Deutsch. Nebentestsprachen können beim Koordinator bis zum 6. Januar 1994 beantragt werden.

1.2.5 Preisrichter der 1. Morpholympics 1994

Prof. Dr. I. S. Batori
Universität Koblenz-Landau

Prof. Dr. Grzegorz Dogil
Universität Stuttgart
Prof. Dr. Günther Görz
Universität Erlangen-Nürnberg
Prof. Dr. Winfried Lenders
Universität Bonn
Prof. Dr. Wolfgang Wahlster
Universität Saarbrücken

2 Standardisierte Selbstdarstellung

Die folgenden Fragen zur linguistischen Motivation, technischen Konzeption, Datenabdeckung und Geschwindigkeit der teilnehmenden Systeme soll in einer standardisierten Selbstbeschreibung der zu testenden Systeme resultieren. Sie dient den Preisrichtern als schriftliche Unterlage, die bei der Bewertung mit in Betracht gezogen wird, und ist Grundlage der mündlichen 'Präsentation' (siehe 1.1. 7).

2.1 Name und Herkunft des angemeldeten Systems:

2.2 Konzeptuelle Kriterien:

2.2.1 Deklarative Spezifikation lexikalischer Einträge und Regeln

Stellen Sie bitte die Struktur der lexikalischen Einträge und der Regel(type)n der Morphologie-Grammatik schematisch dar. Geben Sie bitte zu jedem Schema ein realistisches (unediertes) Beispiel.

2.2.2 Bezug zwischen lexikalischen Einträgen und Wort formen

Zeigen Sie bitte anhand einer schematischen Ableitungsskizze, wie in Ihrem System das Verhältnis von Allomorphie und Konkatenation behandelt wird. Skizzieren Sie bitte den Bezug zwischen lexikalischem Eintrag und den abgeleiteten Wortformen *anlächle/anlächeln* in Ihrem System.

2.2.3 Verständlichkeit und linguistische Motivation der Regeln

Beschreiben Sie bitte, wie in Ihrem System die morphologischen Prozesse der Flexion, Derivation und Komposition behandelt werden. Illustrieren Sie die Handhabung dieser Prozesse an der Ableitung der Formen

Tisch, Tisches, Tischen
vorbeischwammst,
vorbeischwämme,
vorbeigeschwommenen
Hausdächern, Häusermeers
Unabhängigkeitserklärung
unlesbares
durchdachte
gut, besser, besten

aus den Lexikoneinträgen Ihres Systems. Zitieren Sie die hierbei verwendeten Regeln und erläutern Sie ihre Funktionsweise.

2.2.4 Morpho-syntaktische Analyse (Kategorisierung)

Erläutern Sie bitte anhand der oben analysierten Beispiele, daß Ihr System Wortformen nach traditionellen Kriterien wie Genus, Numerus, Kasus, (Verb-)Modus, Tempus, Person, Komparation etc. charakterisiert. Werden Morphemgrenzen dargestellt (Segmentierung) und/oder syntaktisch relevante Eigenschaften wie Valenzrahmen berücksichtigt? Wie schätzen Sie die Möglichkeit ein, Ihr System im Rahmen unterschiedlicher Syntaxsysteme zu verwenden (mithilfe einer automatischen Transduktion Ihrer morpho-syntaktischen Analysen in die entsprechenden Formate)?

2.2.5 Behandlung der Generierung

Werden in Ihrem System dieselben Regeln für die Generierung verwendet wie für die Analyse? Illustrieren Sie Ihr System bitte an der Generierung der Paradigmen von *geben* und *Küchentisch.*

Welche Formen dienen bei der Generierung als Eingabe (z.B. Stamm oder Infinitiv beim Verb)? Wie wird die Generierung spezifischer Zielformen (z.B. 3. Person Singular Präsens von *lernen)* gehandhabt?

2.2.6 Übertragbarkeit auf andere Sprachen

Auf welche anderen Sprachen wurde Ihr System bisher angewendet? Gab es dort spezifische morphologische Phänomene (z.B. Vokalharmonie oder *intercalation* von Morphemen), für die sich Ihr System als besonders geeignet erwies?

2.3 Technische Konzeption und Einsatzfähigkeit

2.3.1 Zielsetzung der Konzeption

Gibt es Anwendungsaspekte, auf die bei Konzeption und Implementation Ihres Systems besonderer Wert gelegt wurde? Mögliche Aspekte sind z.B. Geschwindigkeit, Vollständigkeit der Datenabdeckung, Modellierung einer spezifischen linguistischen Theorie, Verwendbarkeit bei den verschiedensten Sprachen, geringer Bedarf an primärem oder sekundärem' Speicher, Lauffähigkeit auf bestimmten Maschinen (z.B. PC) oder unter bestimmten Betriebssystemen, Unterstützung von akustischer Spracherkennung, Verwendung durch einen Syntaxparser und die Indizierung textueller Datenbanken. Wie haben sich Ihre speziellen Einsatzziele auf die Konzeption ausgewirkt, und inwiefern ist Ihr System für diese Einsatzziele besonders geeignet?

2.3.2 Portabilität der Software und der Daten

In welcher Programmiersprache/Version ist Ihr System bisher implementiert worden? Auf welchen Maschinen und mit welchen Betriebssystemen/Version haben Sie Ihr System bisher zum Laufen gebracht? Bei welchen Umgebungen gab es Probleme?

2.3.3 Schnittstellen zur Syntax und zur Semantik

Gibt es in Ihrem System Schnittstellen zur Syntax und zur Semantik? Wie sind sie spezifiziert und werden sie tatsächlich verwendet?

2.3.4 Hilfestellung bei Benutzerfehlern

Demonstrieren Sie bitte die Güte der Fehlermeldungen und des Debuggers in Ihrem System anhand der *trace* dreier aussagekräftiger Beispiele.

2.3.5 Größenbeschränkung des Systems

Welche Größenbeschränkungen gibt es in Ihrem System bzgl. der Speicherbelegung, des Lexikons oder der Länge der Eingabe?

2.3.6 Schnittstelle zu Nicht-ASCII Zeichen

Kann Ihr System Nicht-ASCII Zeichen (z.B. Sonderzeichen wie Umlaute, Akzente etc., oder nicht-lateinische Sprachzeichen wie Hangul) in der Eingabe verarbeiten und in der Ausgabe darstellen? Wie ist dies technisch gelöst?

2.3.7 Benutzerfreundlichkeit des *turn around*

Wieviele Schritte sind erforderlich, um an Ihrem System linguistisch-empirische Modifikationen vorzunehmen, und wie langwierig ist die damit einhergehendene Kompilation? Demonstrieren Sie bitte eine solchen Vorgang mit einer *getimeten trace.* Fügen Sie bitte (im Druckbild deutlich markierte) erläuternde Kommentare ein, um dem Leser einen schnellen Überblick zu verschaffen.

Stellen Sie bitte auch dar, wie robust Ihr System auf unbekannte Wortformen reagiert und welche Hilfestellungen es bei der Eingabe neuer Wörter ins Lexikon gibt.

2.3.8 Transparenz und Vollständigkeit der Dokumentation

Welche Dokumentation gibt es zu Ihrem System? Umfaßt sie (a) eine Darstellung der zugrundeliegenden linguistischen Theorie mit algebraischer Definition und Komplexitätsanalyse, (b) ein Bedienungsmanual, daß es neuen Benutzern ermöglicht, Ihr System schrittweise zu erlernen und auf neue Sprachen anzuwenden und/oder (c) den kommentierten Quellkode? Wann wurde die Dokumentation das letzte Mal überarbeitet? Falls aktuelle Teile Ihrer Dokumentation veröffentlicht wurden, geben Sie bitte die bibliographischen Daten an. Bitte legen Sie Exemplare Ihrer Dokumentation zu den Fragebogen bei.

2.3.9 Verfügbarkeit und Wartung

Unter welchen Voraussetzungen und in welcher Form kann ein potentieller Benutzer Ihr System als laufendes Softwarepaket bekommen? An welchen Institutionen wurde Ihr System von wann bis wann für welche Sprachen benutzt? Wird das System von Ihnen z. Zt. aktiv gepflegt, bzw. weiterentwickelt? In 'welchen Anwendungen wird das System z. Zt. von Ihnen genutzt? An welche Personen (email-adresse) in Ihrem Team kann sich ein Benutzer im Fall von *bugs* und anderen Problemen wenden?

3 Testdaten und Fragebogen

3.1 Vorläufige Tests des Systems an bekannten Daten

Zwei Monate vor den Morpholympics stellen die Preisrichter zwei Sorten von Testdaten *on line* zur Verfügung:

I> eine Wortliste mit besonders interessanten morphologischen Phänomenen (inzwischen als 'Schmankerlnliste' bekannt)

I> einen repräsentativen Text ausreichen der Länge.

Diese Testdaten sollen den teilnehmenden Systemen zur Vorbereitung auf die Morpholympics und zur Berechnung vorläufiger *bench marks* dienen.

3.2 Portabilität

Um die Portabilität Ihres Systems zu zeigen, testen Sie bitte Ihr System an den in 3.1 genannten Daten, und zwar auf einer Unix-Workstation, einem PC und einem Apple-Macintosh. Für jeden der drei Tests beantworten Sie bitte im Rahmen dieses Fragebogens folgende Fragen:

Typ der Maschine, des Operating Systems und der Programmiersprache, sowie

I> Geschwindigkeit (Wortformen pro Sekunde)

I> Speicherbelegung

I> Zeit, die für einen *turn around* benötigt wird (siehe 2.2.7)

Bitte fügen Sie die Liste analysierter und nicht-analysierter Wortformen als Anlage bei, wobei die Zahl der analysierten und der nicht-analysierten Wortformen sowohl hier als auch am Anfang der Anlage genannt wird. Geben Sie hier bitte auch Geschwindigkeit, benötigten Speicherplatz und 'turn around'-Zeit für jeden der drei Maschinentests an.

3.3 Endgültige Tests des Systems an neuen Daten

Die in 3.2 beschriebenen Testläufe werden während der Morpholympics an neuen, aber ähnlichen Testdaten wiederholt. Dabei wird der Testlauf auf der Workstation während der Präsentation vorgeführt (primärer Testlauf). Die Testläufe auf dem PC und dem Mac (sekundärer Testläufe) werden nach der Präsentation durchgeführt (siehe 1.1. 7).

3.4 Auswahl und Verteilung der vorläufigen Testdaten

Potentielle Teilnehmer und Zuschauer bei der Morpholympics werden hiermit höflich gebeten, Wortformen oder Texte, die sie für relevant halten, an einen der Preisrichter, Herrn Prof. Lenders, unter lenders@uni-bonn.de zu schicken. Die Preisrichter werden diese Daten bei der Zusammenstellung der vorläufigen und endgültigen Testdaten in Betracht ziehen.

Die vorläufigen Testdaten können ab 6. Januar 1994 via ftp von

server name:
sol.linguistik.uni-erlangen.de
login name: anonymous
password: your user name
directory: morpholympics

abgerufen werden. Die Auswahl und Weitergabe der vorläufigen und endgültigen Testdaten liegt in der Verantwortlichkeit der Preisrichter.

3.5 *on line-Abruf* von Fragebogen und Anmeldeformular

Die vorliegende Darstellung von Ablauf und Teilnahmemodalitäten kann bereits jetzt in den Tagungssprachen Englisch und Deutsch von dem in 3.4 genannten Server abgerufen werden.

4 Anmeldeformular für die 1. MORPHOLYMPICS

4.1 Anmeldung des Systems

Das folgende System zur automatischen Wortformerkennung für *on line* Texte im Deutschen nimmt an dem Wettlaufen der 1. Morpholympics teil, das am 7. und 8. März 1994 in Erlangen von der Gesellschaft für Linguistische Datenverarbeitung (GLDV) veranstaltet wird.
Name und Herkunft des Systems:

Name und Adresse der Personen, die das System bei den Morpholympics vorstellen:

4.2 Einverständniserklärung für die Publikation

Wir werden 6 Kopien der standardisierten Beschreibung unseres Systems zu Beginn der Morpholympics beim Veranstalter einreichen. Wir sind damit einverstanden, daß diese Beschreibung vom Koordinator in der offiziellen Publikation über die Ergebnisse der Morpholympics veröffentlicht wird.

Unterschriften, Ort und Datum

Rezensionen

Sven Naumann:
Generalisierte Phrasenstrukturgrammatik: Parsingstrategien, Regelorganisation und Unifikation. (Linguistische Arbeiten 212) Tübingen: Niemeyer, 1988. 180 Seiten; kart. 70,- DM

Naumanns Buch ist eine der wenigen deutschsprachigen Abhandlungen über GPSG und dennoch ist es kaum bekannt geworden und wird selbst in späteren verwandten Arbeiten nicht zitiert (Fisher 89, Weisweber & Preuß 92). Motivation genug fünf Jahre nach dem Erscheinen nachzufragen, wer dieses Buch lesen (oder auch gelesen haben) sollte.

Das Buch gliedert sich in drei Teile: eine Einführung in die GPSG, die Vorstellung des von Naumann entwickelten GPSG-Parsers und seine Implementation.

Die GPSG-Einführung behandelt systematisch alle Komponenten einer GPSG, die verschiedenen Regeltypen, die Instantiierungsprinzipien sowie die Merkmalprinzipien. Die Darstellung ist knapp und beschränkt sich teilweise so stark auf formale Definitionen, daß es schwer ist zu folgen. Aber wo sie ausführlicher ist, da ist sie auch kritisch (hinterfragt den Sinn der Vorschläge aus Gazdar et al. 1985 (bekannt als GKPS)), abwägend (zwischen verschiedenen GPSG Versionen) und einordnend (vergleicht GPSG mit TG). An einer Stelle überrascht Naumann damit, daß er zeigt, daß die ECPO-Eigenschaft von kontextfreien Grammatiken nicht die Voraussetzung dafür ist, daß solche Grammatiken in ID- LP Grammatiken umgewandelt werden können, wie dies in vielen anderen Publikationen behauptet wird (z.B. in GKPS
1986:49 oder in Evans 1987:38). Naumanns Argumentation basiert jedoch darauf, daß neue Konstituentennamen beliebig eingeführt werden können. Das belegt, daß er die GPSG-Implementation aus informatischer und weniger aus linguistischer Sicht betreibt.

Im Zusammenhang mit der Erläuterung seines Parsers stellt Naumann klar heraus, welche Probleme bei der Abbildung der reinen GPSG-Lehre in ein ablauffähiges System entstehen. Er wählt dann ein Drei-Phasen-Modell, das zunächst aus ID-, LP-, und Metaregeln eine kontextfreie Grammatik erzeugt, die die Eingabe für den Parser bildet. Dieser generiert für einen Eingabesatz aufgrund der Grammatik und unter Anwendung der Merkmalprinzipien (FFP, HFC, CAP) eine Menge von Strukturbeschreibungen, aus der in der dritten Phase durch Merkmaldefaults (FSD) und Kookkurenzbeschränkungen (FCR) die gültigen Strukturen ausgefiltert werden. Auffällig bei diesem Vorgehen ist vor allem der frühe Einsatz der LP-Regeln, die in ihrer Anwendung auf die Merkmale der ID- und Metaregeln beschränkt bleiben. Ein anderer, zur gleichen Zeit entstandener Ansatz der Berliner Gruppe wendet beispielsweise die LP-Regeln ganz zum Schluß an (Hauenschild & Busemann 1988:14). Dieses Vorgehen bedingt jedoch den Einsatz eines ID-LP Parsers, während Naumann sich auf einen Parser für kontextfreie Sprachen beschränkt.

Naumanns System ist in LISP geschrieben. Die Seiten 99 bis 126 des Buches enthalten den gut annotierten Code.
Das System war seinerzeit auf einem MSDos Rechner entwickelt worden und bot entsprechend unbefriedigende Antwortzeiten. Naumann hat das System getestet mit einigen englischen Beispielsätzen aus

dem Bereich der "unbounded dependencies" . Der Test basierte auf einem Lexikon mit 38 Einträgen und einer "ausmultiplizierten" Grammatik mit 169 kontextfreien Regeln. Im Kapitel 10 werden die berechneten Strukturen präsentiert.

Naumann versucht nirgendwo vorzugeben, daß sein Buch mehr darstellt als es ist, nämlich der Bericht über einen getreuen Ansatz zur Implementation des GPSG-Ansatzes, wie er in GKPS vorgeschlagen wurde. Als solcher (und als gute Einführung in die Theorie) ist er durchaus lesenswert. Es bleibt dennoch die Frage, warum das Buch so unbeachtet geblieben ist. Vielleicht liegt es daran, daß hier ein deutschsprachiges Buch (von einem deutschen Verlag) vorliegt, das jedoch nur englische Syntax behandelt. Das erschwert sicherlich die Akzeptanz in der angelsächsischen Forschungsgemeinde. Der von Uszkoreit (1986) vorgestellte GPSG-Ansatz für das Deutsche wird nur am Rande erwähnt, was wiederum das geringe Interesse im deutschsprachigen Raum erklären mag. Vielleicht liegt es aber auch daran, daß hier ein System in LISP vorgestellt wird, während Prolog-Implementationen seit Mitte der 80er Jahre mehr im Blick
punkt standen. Schließlich mag es auch daran liegen, daß dieses Buch erst erschien, als sich vielerorts das Interesse bereits auf HPSG zu verlagern begann.

Literatur

Evans, Roger: *Theoretical and Computational Interpretations of Generalized Phrase Structure Grammar.* (Cognitive Science Research Paper 085) Brighton, GB: The University of Sussex. August 1987.

Fisher, Anthony: *Practical Parsing of Generalized Phrase Structure Grammars.* Computational Linguistics 15, 3 (1989), 139-148.

Gazdar, Gerald; Klein, Ewan;
Pullum, Geoffrey; Sag, I van: *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press, 1985.

Hauenschild, Christa; Busemann, Stephan:
A constructive version of GPSG for machine translation. (KIT-Report 59) Berlin: TU-Berlin, Projektgruppe KIT. Februar 1988.

Uszkoreit, Hans: *Word order and constituent structure in German.* (CSLI Lecture Series) Chicago University Press, 1987.

Weisweber, Wilhelm; Preuß, Susanne: *Direct Parsing with Metarules.* (KIT-Report 102) Berlin: TU-Berlin, Projektgruppe KIT. Dezember 1992.

Martin Volk, Univ. Koblenz-Landau

Ulrich Schmitz: Computerlinguistik. Eine Einführung. Westdeutscher Verlag, Opladen, 1992.

Es handele sich um ein Denkbuch, so der Autor. Dies bedeutet Distanz zum Inhalt und Reflexion darüber, was man tut, wenn man Sprache und Sprachforschung mit Computern zusammenbringt. Es ist eher gedacht als Standortbestimmung der Computerlinguistik als ein Lehrbuch oder eine Einführung ins Studium der Computerlinguistik im engeren Sinne. Hierzu wird ausführlich auf aktuelle Literatur verwiesen. Die Adressaten sind Studienanfänger oder Laien mit geisteswissenschaftlichem Hintergrund, die sich für die Denkweise computerlinguistischen Vorgehens interessieren. Es behandelt kritisch und in der Bewertung ambivalent eine wissenschaftliche Haltung gegenüber dem Gegenstand Sprache, in der der Zweck im Vordergrund steht. Vor allem geht es bei der Instrumentalisierung von Sprache um solche Zwecke, die wiederholbar und technisch in Form von Maschinen nachbaubar sind. Die Frage ist, wie sich diese Mensch-Maschine-Beziehung auswirkt, wenn Sprache im Spiel ist. Ein markantes Beispiel dafür ist die Annahme von Regeln und/oder Schemata, mit denen Wiederholbares im Sprachgebrauch formuliert werden kann. Um Distanz und gleichzeitig eine begründete Technikfreundlichkeit herzustellen, stellt der Autor diese

Perspektive den Möglichkeiten des "Nicht-Geregelten" gegenüber. Insofern geht es um mehr als Computerlinguistik, es geht um "einen gewachsenen Teil unserer Kultur" (S.13). Diese Sichtweise wird an den wesentlichen Themenbereichen der Computerlinguistik aufgezeigt.

Das einleitende Kapitel ("Einladung zur Computerlinguistik") beschränkt sich auf die Nennung des anvisierten Ziels des Buchs sowie auf allgemeine Angaben zum Fach Computerlinguistik. Plastisch dargestellt sind die Anwendungsbeispiele, die den Zugang zu den sprachlichen Problemen und damit zur linguistischen Herangehensweise eröffnen. Neben den traditionellen Bereichen, die Form und Struktur in den Vordergrund stellen, werden prozedurale Ansätze skizziert, die die Verarbeitung von Information thematisieren. Im weiteren werden statistische Methoden und die Arbeit an Wörterbuchdatenbanken aufgezeigt. Nicht vergessen bleibt die gegenwärtig (noch) vorherrschende Tendenz der Entwicklung von Grammatikformalismen. Die Skizze von Nachbarfächern, Forschungsaufgaben und Studiengängen rundet den Überblick ab. Die Bemerkungen zur Berufsperspektive bleiben situationsgemäß sehr vage.

Im 2. Kapitel geht es dann bereits um grundsätzliche Fragen (" Wissenschaft und Technik: das Leben im Griff"). Besprochen werden formale und maschinenorientierte Verfahren (z.B. Abbildung und Simulation oder die Verwendung von Schemata). Im weiteren gibt es eine Kurzeinführung in die Programmiersprache PROLOG, die den bereits Kundigen die Prinzipien klarer macht, zum Erlernen aber wirklich zu kurz ist. Deutlich werden soll die Eingeschränktheit solcher Ansätze im Vergleich zu Fragestellungen, die sonst noch möglich waren.

Konkret auf die traditionellen linguistischen Teilbereiche geht das 3. Kapitel ein ("Menschliche Sprache und mathematische Form"): Phonetik, Morphologie, Syntax, Semantik, Pragmatik, Textlinguistik. Zugleich handelt es sich hier um die Kerngebiete der Computerlinguistik. Sie werden unter dem Gesichtspunkt des analytischen Denkens, des Meßbaren und der Regel diskutiert. Immer wieder erscheint der Hinweis auf das Gegenteil, unterstützt durch Zitate abendländischer Denker (u. a. Derrida, Schleiermacher, Wittgenstein), daneben Zitate von Sprachwissenschaftlern wie Chomsky, Fillmore, Paul, Saussure.

Die beiden folgenden Kapitel setzen sich mit einigen der Hauptanwendungsgebiete der Computerlinguistik auseinander.

Die Mensch-Maschine-Kommunikation steht im 4. Kapitel im Vordergrund und damit grundsätzliche Unterscheidungen zwischen Menschen und Maschinen ("Menschen handeln, Maschinen operieren"), zwischen natürlicher Sprache und Programmiersprachen. Nichtsdestoweniger wird der Gesichtspunkt eingebracht, daß neue Werkzeuge das Handeln der Werkzeuggebraucher beeinflussen. Andersherum gesehen, werden natürlich nur solche Werkzeuge erfunden, die einer bestimmten Geisteshaltung ihrer Gebraucher entsprechen und vor allem: sie werden "anthropomorphisiert". Der Abschnitt" Über Computer sprechen" behandelt den geläufigen metaphorischen Sprachgebrauch. "Mit Computern sprechen" geht auf Dialogmodellierungen ein, wobei (selbstironisch) ebenfalls Metaphorisierungen eingebracht werden ("die Maschine als Lehrer" und "die Maschine als Wissens-Diener"). Schließlich aber wird ganz dezidiert Stellung bezogen: Mensch-Maschine-Kommunikation ist tot (und daher unter Umständen auch attraktiv) und etwas ganz anderes als Mensch-Mensch-Kommunikation, die lebendig und durch Gefühle geprägt ist. Gleichwohl ist eine Vielzahl kommunikativer Vorgänge zwischen Menschen auf der Ebene von Operationen anzusiedeln. In diesem Sinne ist das 5. Kapitel der automatischen Textgenerierung gewidmet (" Texte von Menschen und Maschinen"). Hier bringt der Autor einen eigenen Ansatz ein, was sich posi-

tiv durch stärkere Konkretheit des Beispiels bemerkbar macht. Es geht um die Generierung von Nachrichtentexten. Das Modell ist entwickelt auf der Grundlage von authentischen Fernsehnachrichten (Textmaschine für die Tagesschau). Es wird versucht zu zeigen, daß hier ein Bereich vorliegt, in dem Regeln und Schemata, Wiederholbares und Voraussehbares, die Produktion von Nachrichtentexten bestimmen. Ein Katalog von semantischen Stereotypen (z.B. Besuch, Treffen, Vereinbarung) wird in einer Datenbank gespeichert. Die aktuelle Version wird nach bestimmten Regeln gemischt und durch Referenzangaben (Spezifikationen von Personen, Zeit und Ort) ergänzt. Fazit der Geschichte: Menschen machen Texte in der Art, wie sie für Maschinen modelliert werden (Informieren anstelle von Argumentieren). Insofern sind Werkzeugkonstruktion und Werkzeuggebraucher nicht so weit voneinander entfernt! Zum Schluß (6. Kapitel: "Computer als Werkzeug") nochmal die Gegenüberstellung von Reduktion und Grenzen einerseits und Nützlichkeit andererseits. Und dann schließlich noch ein kleines Fragezeichen in bezug auf zukünftige Entwicklungen. Der Ausgang der Geschichte ist offen.

Generell zum Buch läßt sich folgendes sagen. Beim Lesen sind Personen- und Sachregister hilfreich. Das Personenregister skizziert überdies einen weit gespannten Rahmen, innerhalb dessen eine Vielzahl von Sprachphilosophen und Sprachwissenschaftlern zu Wort kommen. Das Sachregister gibt einen guten Überblick über die behandelten linguistischen Themen. Zum Weiterarbeiten hilft ein umfangreiches Literaturverzeichnis. In Einführungskursen ist das Buch m.E. vor allem begleitend nützlich. Gut ist, daß alle Probleme durch Beispiele klar gemacht werden. Dies betrifft sowohl die sprachliche und linguistische Ebene als auch die Implementierungsebene, wiewohl die oft sehr knappe und vereinfachende Darstellung für tatsächliche Anwendungen nicht ausreicht. Will man allerdings der komplexen Sichtweise und der argumentativ aufgebauten Darstellung tatsächlich gerecht werden, so ist etliches an Vorwissen und intensivem Mit-Denken Voraussetzung. Den größten Gewinn wird die Lektüre daher für die LeserInnen bringen, die die Ansätze, in die eingeführt wird, bereits kennen. Für sie jedoch ist es ein Buch, in das hineinzusehen sich auch zum wiederholten Male lohnt.

Annely Rothkegel, Univ. Saarbrücken

G. Leitner (Hsg.):
Theorie und Praxis der Korpus-Analyse: New Directions in English Language Corpora. (Topics in English Linguistics; Bd. 9). 368 Seiten, Berlin, New York: Mouton de Gruyter, 1992

Einleitung

Die Untersuchung von Korpora hat gegenwärtig Konjunktur. Dies erklärt sich aus zwei Faktoren:

1. Der erste ist das weitgehende Mißvergnügen an bisherigen Bemühungen im Bereich der formalen Linguistik. Angetreten mit dem Anspruch, Sprache (und das hieß damals: beliebige Texte) formal beschreiben zu können, hat die Zunft sich mittlerweile eines schlechteren besonnen. Die Diskussion um die verschiedenen Forma!ismen hat deren Eleganz mit der weitgehenden Abstinenz erkauft, die Daten zu betrachten: Es konnten immer weniger Phänomene immer eleganter erklärt werden, und die Beispiele sind immer weniger nachvollziehbar geworden (in meinem Dialekt ist dieser Satz möglich). Der Versuch, Grammatiken mit größerem Abdeckungsgrad zu er-

stellen, wenn er denn überhaupt unternommen worden ist, hat zu dem Eindruck geführt, die Aufgabe sei erstens langwierig (was zutrifft), zweitens mit den gegenwärtigen Formalismen nicht zu machen (was weitestgehend ebenfalls zutrifft), und drittens prinzipiell endlos, weswegen man in Richtung auf probabilistische Grammatiken und eben Korpora ausgewichen ist. Auch bei den Sponsoren von NLP-Projekten (wie der EG) ist die Bereitschaft geschwunden, Projekte mit nicht einsetzbaren Ergebnissen zu fördern (der BMFT ist hierin eine Ausnahme). Dies hat speziell in den USA dazu geführt, über die Kriterien der Abnahme von Projekt-Ergebnissen nachzudenken und die Frage, was ein System eigentlich leisten muß, mit dem Hinweis auf eine Sammlung von Texten zu beantworten, die ein System analysieren können muß.

2. Der zweite Faktor des Interesses an Korpus-Arbeiten liegt in den überraschenden Möglichkeiten, die eine robuste und schnelle, meistens statistisch basierte Analyse großer Textmengen mit sich bringt.

   t> Eine Vorreiterrolle haben hier die Sprachmodelle in den Speech-Understanding-Projekten gespielt, die in den praktischen Einsätzen klare Dominanz über linguistische Ansätze erzielen konnten. Kaum ein Projekt arbeitet ohne solche Techniken.

   t> Eine Ausweitung der Techniken auf den Bereich der Übersetzung (über die Terminologie-Zuordnung hinaus) ist zwar bisher nicht erfolgreich, hat aber den Effekt gehabt, daß die Thematik des Alignments von bilingualen Korpora studiert und mit überraschend guten Ergebnissen angewandt worden ist.

   t> Basierend darauf konnte man sich der Identifizierung von mehrsprachiger Terminologie, der Zuordnung von Textteilen und der Erstellung entsprechender Referenz-Datenbanken zuwenden; erste Produkte dieser Art sind bereits erhältlich.

   t> Im Lexikon-Bereich wurden Korpus-Techniken angewandt bei der Identifizierung von Verb-Argumenten, einem notorisch komplexen Thema (Arbeiten von Shieber), bei der Ermittlung von semantischen Definitionen aus Lexika (Projekt Aquilex), bei der Disambiguierung von Lesarten mithilfe bilingualer Korpora, oder bei der Ermittlung semantischer Ähnlichkeiten (Arbeiten von G. Ruge).

   t> Im Syntax-Bereich sind Fragen der Frequenz bestimmter Strukturen von Interesse, die das Parsing leiten können; im Zusammenhang damit Fragen von fachsprachlichen Varianzen syntaktischer Patterns; und die Fragen nach der Repräsentation solcher Information.

All diese Arbeiten zeigen, daß die Analyse von Korpora ein Feld ist, das Erfolge zumindest im Language Engineering verspricht. Die Frage dabei ist, wie sich statistische und linguistische Techniken kombinieren lassen, um linguistisches Wissen durch die Untersuchung der praktischen Sprach-Benutzung optimal anwenden zu können.

Das vorliegende Buch stammt aus einer anderen Tradition. Es handelt sich um die Proceedings der 11. Konferenz des ICAME 1990 in Berlin, des *International Computer Archive* of *Modern English.* Diese Institution hat seit jeher starke Wurzeln in der

corpus-analytischen Tradition, die sich umgekehrt soziolinguistischer Argumente bedient. Dort hat sich ja von Anfang an eine gewisse Skepsis gegenüber dem Konzept des idealen Sprecher-Hörers gehalten, und man ist den Fragen, was ein Sprachsystem konstituiert, mithilfe der Empirie, und das heißt, mithilfe von Sprachkorpora, nähergetreten. Jetzt finden diese Arbeiten Interesse bei der NLP-Zunft; und aus diesem Gesichtspunkt, und aus dem Gesichtspunkt eines praktischen Interesses an der Entwicklung nicht-trivialer NLP-Systeme, ist die folgende Besprechung geschrieben.

1. Corpus Design and Text Encoding

Das Buch beginnt enttäuschend. Der erste Abschnitt behandelt methodische Fragen beim Design von Korpora. Es finden sich Beiträge von

t> Pieter de Haan: The optimum corpus sampie size. Er stellt fest, daß die Größe von Korpora vom Untersuchungszweck abhängt, weil die zu bearbeitenden Phänomene in entsprechender Signifikanz vorkommen müssen.

t> J. Clear: Corpus Sampling. Er gibt einige Prinzipien an, die man beim Anlegen von Korpora beachten sollte ("core" language, Texttypen-Einteilung, Unterscheidung zwischen Sample und Monitoring Korpus) und definiert Zielgruppen des Oxford Korpus.

t> G. Leitner: International Corpus of English (ICE): Corpus design - problems and suggested solutions. Er präsentiert die Taxonomie des ICE und die wesentlichen Design-Prinzipien und schlägt einige Änderungen vor: In einem Top-Down Approach sollen zunächst "large-scale socio-communicative environments that (co-)determine their own specific socio-communicative norms and the choice of language" (48) betrachtet werden, darunter die jeweiligen "major communicative events", die zeit- und ortsgebunden sind, und dann die Einteilung nach "textual and linguistic structure in text linguistic categories".

t> J. Kirk: The Northern Ireland Transcribed Corpus of Speech. Er präsentiert diese Sammlung gesprochener Sprache (240 K Wörter) und einige Probleme bei ihrer Erfassung.

t> Chr. Mair: Problems in the compilation of a corpus of Standard Caribbean English: A pilot study. Der Beitrag stellt sich die Frage, ob Caribbean English ein eigenes Sprachsystem sei.

t> L. Burnard: The Text Encoding Initiative: A progress report. Der Beitrag berichtet über den damaligen Stand der TEI, kurz nach Erscheinen der ersten Version der Guidelines. Hier hat sich mittlerweile ja einiges getan, so daß der Beitrag etwas zu spät kommt, um noch völlig aktuell zu sein.

Wenn man den Beiträgen des ersten Abschnitts nun Glauben schenken soll, so ist das Erstellen von Korpora - das ist das Enttäuschende - eine aussichtslose Sache:

t> Man weiß nicht, was man sammeln soll: "the phenomenon to be sampled is poorly defined" (Clear, 21).

t> Die Sammlung folgt bestimmten Prinzipien und theoretischen Ambitionen, die doch andererseits sich erst erweisen sollen: "The sampling problem is precisely that a corpus is inevitably biassed in some respects" (Clear, 23).

t> Man weiß nicht, wieviel man sammeln soll: "the suitability of the sample depends on the specific study that is undertaken, and there is no such thing as the best, or optimum, sample size as such". (de Haan, 4; dagegen Clear, 40: "more is definitely better").

t> Man weiß nicht, wer sammeln soll, weil das Ergebnis vom Befrager abhängt: "where the fieldworker used dialect forms, the informant cooperatively followed, but in some cases not until then" (Kirk, 68); außerdem sollte er vom Torfstechen und vom Guinness-Trinken etwas verstehen.

t> Man weiß nicht, was man mit den einmal gesammelten Texten anfangen soll: Jeder interpretiert die Daten, wie sie ihm wichtig scheinen (Mair, 78), und übersieht vielleicht das Wesentliche.

t> Wenn man schon sachlich nichts weiß, sollte man sich zumindest auf eine einheitliche Darstellung einigen (TEI mit SGML), aber auch hier gibt es berechtigte Sorgen: "The obvious risk that, in endeavoring to please all , the TEl may end by pleasing no-one" (Burnard, 106).

Das Problem an einem Beispiel:

1. Der Vorteil der Korpus-Arbeit soll sein, daß eine vieldiskutierte Problematik des Englischen, nämlich" whether it is one system or a set of overlapping, differing codes would be treated as an empirical issue" (Leitner, 33).

2. Mair führt dies am Problem des Caribbean English durch: Unterscheidet sich Caribbean English vom Englischen? Vorsicht: Gibt es das Caribbean English seinerseits überhaupt? Müßte man nicht annehmen, daß auf Jamaica, Trinidad, Tobago vielleicht ein jeweils eigenes Sprachsystem herrscht? "A wrong assessment of the situation at the time of the collection of the texts for the corpus will have fatal results" (Mair, 77). Dasselbe Argument für die verschiedenen Regionen, Schichten usw. von Trinidad ...

3. Selbst wenn man dann Unterschiede entdeckt: Begründen sie ein eigenes Sprachsystem, oder handelt es sich nur um einen Ausrutscher? "In evaluating the evidence, one faces the usual problems of having to decide whether an instance of unusual verb complementation is due to a typographical error, e.g. omission of a preposition, or whether one is dealing with a genuine linguistic datum" (Mair, 87).

4. Sodaß sich die ursprüngliche Fragestellung mit Hinweis auf Korpus-Arbeiten gerade nicht beantwortet: "a computerised corpus will show whether or not these are systematic enough to posit an independent local norm" (Mair, 85), weil man Normen als solche nicht beobachten kann, sondern allenfalls ihre Manifestationen; und weil sich immer darüber streiten läßt, ob es sich wirklich um Manifestationen der Norm handelt.
   Alles hängt eben von der Bewertung durch den Interpreten ab, sodaß sich die postulierte Objektivität der Korpus-Arbeit zuletzt wieder in den Standpunkt des Forschers verflüchtigt
   …

Tröstlich ist nun freilich, daß die meisten Autoren nicht konsequent sind und ihre eigenen Beteuerungen der Unmöglichkeit der Korpus-Arbeit nicht allzu wichtig nehmen; denn faktisch sammeln sie ja die Daten, teilen sie ein, finden auch interessante Phänomene; und das sind die Teile des Buches, die wirklich lesenswert sind.

2. Automated Text Analysis

Der zweite Teil des Buches behandelt Probleme bei der praktischen Arbeit mit Korpora, untersucht Fragen der Text-Aufbereitung, der Annotationen und des Text-Retrievals.

Der erste Beitrag (N. Belmore, Pinpointing problematic tagging decisions), diskutiert Unterschiede im Tagging zwischen Brown und LOB corpus; sie sind vor allem

auf verschiedene Arten der Tokenisierung zurückzuführen (ist can't als ein oder zwei Tokens zu behandeln?), und der Beitrag berichtet von Arbeiten, die eine Synchronisation der beiden Systeme gestatten sollen; das eigentliche Tagging-Thema (Zuweisung von Kategorien zu Tokens) wird aber noch gar nicht behandelt.

Zwei Beiträge berichten von niederländischen Projekten (DEVIL und LINKS). Der Beitrag von W. Meijs zu "Inferences and lexical relations" ist ein Irrläufer in diesem Buch; er versucht, Wort bedeutungen mit atomaren Prädikaten des Typs *baby → child (x) and recently_born (x)* zu behandeln, was nun oft genug zu nichts geführt hat. Interessant ist, daß die Bedeutungsdefinitionen in LDOCE untersucht worden sind, und Meijs stellt fest, daß sie mitnichten so hierarchisch klassifiziert sind, wie die AI-schemata suggerieren, sondern daß sie oft zirkulär sind, Lücken enthalten, und daß Verweise fehlen, die man intuitiv machen würde *(person* und *animal* z.B.). Das weist darauf hin, daß Korpus-Arbeiten immer noch der verständigen Aufbereitung bedürfen, um sinnvoll benutzbar zu sein, und die lexikalische Arbeit zwar erleichtern, aber nicht ersparen. Der zweite Beitrag (S. Janssen: Tracing cohesive relations in corpus sampies using dictionary data) ist interessant zu lesen: Sie entwickelt ein Verfahren zur Disambiguierung, in dem zunächst lexikalische Bedeutungsdefinitionen analysiert werden; daraus werden Genus-Terme gefunden, zwischen denen dann Hyponym/Hyperonym-Relationen bestimmt werden. Es gelingt der Autorin, sechs Bedeutungen von rich in LDOCE korrekt kontextuell zu disambiguieren. Die Technik ist erwägenswert, indem man nämlich entsprechende Kontext-Termini identifiziert, die in den Lexikon-Definitionen verwendet werden, und zur Laufzeit versucht, die besten Matches solcher Termini mit dem aktuellen Text zu finden.

Die Beiträge von St. Fligelstone (Developing a scheme for annotating text to show anaphoric relations) und G. Sampson (SUSANNE - A deeply analysed corpus of American English) zeigen Schwierigkeiten, die auftreten, wenn man komplexe linguistische Phänomene mit Korpora behandeln will. Fligelstone berichtet vom Versuch, Anaphora zu taggen; dazu ist in der Lancaster Grammar factory ein eigener Editor für die entsprechenden Markups erstellt worden. Der Versuch ist nicht erfolgreich gewesen, da das Phänomen theoretisch nicht klar abgegrenzt war: Trotz eines Manuals von über 100 Seiten war es nicht möglich, das Corpus konsistent zu behandeln ("consistency could not be achieved" (161); und das Ziel, ein Phänomen theorie-neutral zu beschreiben, wird als "unattainable objective" (165) gesehen. Ähnliche Ergebnisse sind aus dem Beitrag von G. Sampson zu folgern, der das Göteborg Korpus mit syntaktischer und semantischer Tiefeninformation anreichern wollte, ebenfalls theorie-neutral. Ergebnis ist die Feststellung, daß bereits der Versuch, semantische Tiefenkasus (diesmal auf Stockwell basierend) für 150 englische Verben konsistent zu markieren, bezogen auf das gegebene Korpus, nicht gelungen ist (187), sodaß man sich wieder auf die "klassischen" Rollen (Subjekt, direktes j indirektes Objekt), und die "klassischen" Adverbialen Angaben (time, place usw.) zurückgeworfen sah (und sieht).

Diese Ergebnisse sind erwägenswert, erstens weil sich zeigt, daß konsistente Korpus-Arbeiten nur gemacht werden können, wenn die zu markierenden Phänomene zuvor bereits klar durchdrungen sind. Dann stellt sich aber die Frage, was die Korpus-Arbeit noch leisten kann, außer Häufigkeitsanalysen: Der Anspruch, daß die Korpora solche Phänomene erst erweisen, konterkariert sich, wenn diese zuvor bereits gewußt sein müssen. Zweitens sind gerade solche komplexeren Themen sicherlich nicht theorieneutral zu beschreiben; und das beeinträchtigt die Wiederverwendbar-

keit und damit den Wert der gesamten Markierungs- Arbeiten.

Der letzte Beitrag in dieser Gruppe (S.Sutton, A. McEnery, Information Retrieval and Corpora), stellt fest, daß "the field of IR is clearly relevant to the tasks of accessing and using machine readable corpora" (208). Diese Auskunft ist billig zu haben, verpaßt aber die Chance, die Vorteile der IR-Techniken in Datenorganisation, Zugriff usw. genauer zu diskutieren (wie überhaupt den Fragen der Organisation und des technischen Managements von großen Korpus-Daten kein gesteigertes Interesse zu gelten scheint: Man hat den Eindruck, die Welt-Festplatten-Kapazität ist unerschöpflich). Die Verweise auf Hypertext- und wissensbasierte Systeme sind dagegen nutzlos, weil auch im Feld der IR unklar ist, wie solche Techniken in größeren Applikationen funktionieren sollen.

Insgesamt zeigt der zweite Teil doch einige Grenzen der Korpus-Arbeiten auf und stellt klar, daß es den Königsweg für die computerlinguistischen Arbeiten (viel Preis ohne Schweiß) auch hier nicht gibt. Die automatische Extraktion von Bedeutungen aus Lexikon-Definitionen erfordert ein erhebliches Maß an Korrektur- und Nachbearbeitungs-Aufwand, wenn man wirklich gute Ergebnisse haben will; und die Hoffnung, komplexere Phänomene aus Korpus-Annotationen extrahieren zu können, wird man wohl deutlich dämpfen müssen. Vorläufig scheint es leidlich möglich, Wortklassen-Angaben und darüber errichtete syntaktische Patterns zu markieren und auch zu benutzen. Komplexere Phänomene scheinen aber vorerst außerhalb der Zugriffsmöglichkeiten der Korpuslinguistik zu liegen.

3. Corpora in Language Description

Der dritte Abschnitt versammelt eine ganze Palette von Arbeiten und Themen aus der praktischen Arbeit und dem Umgang mit Korpora. Zwei Beiträge befassen sich empirisch mit der Frage, ob English ein oder eine Menge von Sprachsystemen sei.

E. W. Schneider (Who(m)? Case marking of wh-pronouns in written British and American English) kann diesbezüglich keine signifikanten Unterschiede erkennen, lediglich textsortenabhängige (whom ist häufiger in US-religiösen und UKregierungs amtlichen Texten): "In many, in fact most respects of this fairly ,~ubtle area of grammar, British and America:Q. English behave identically" (242). Das ist zwar ein gutes Resultat, stellt den Autor aber nicht zufrieden: Vielleicht gibt es noch Nischen, Dialekte, subtilere Phänomene? Weitermachen!: "it can be stated that the diversity of the two major varieties of English appears to be further-reaching in scope but subtler in kind than might have been suspected" (243). Hier ist der Wille der Vater der Forschung.

Auch S. V. Shastri (Opaque and transparent features of Indian English) kommt zu dem Ergebnis, "that the most obvious transparent feature of IE is code-mixing" (273), opaque features "turn out to be statistically nonsignificant" (274). Die Hoffnung, im indischen Englisch ein eigenes Sprachsystem vor sich zu haben, gibt er allerdings ebenfalls nicht auf: "all this points towards the need for systematic quantitative studies based on large databases before we can say anything decisive about variety features" (274).

Zwei Beiträge bestätigen bereits anderweitig bekannte Probleme und Lösungsansätze. Dazu gehören einmal G. Barnbrook (Computer Analysis of spelling variants in Chaucers *Canterbury* T*ales),* er untersucht Schreibvarianten im Mittelenglischen mit bekannten Techniken von Spell Checkern (add final letter / add initial letter / change two letters usw.), kommt dabei mit seinem Problem ganz gut zurecht und berichtet die bekannten Probleme (inkorrektes Matching usw.).

G. Knowles (Pitch Contours and tones in

the Lancaster /IBM spoken English corpus) berichtet über Versuche, Akzente akustisch zu ermitteln; er stellt fest, daß die FO-Analyse allein dafür nicht ausreichend ist, und weist auf Kontextabhängigkeiten hin: fallende oder steigende Konturen begründen Bedeutungsunterschiede und hängen gleichzeitig von ihnen ab. Dieses Ergebnis bestätigt entsprechende deutsche Untersuchungen.

Kay Wikberg (Discourse category and text type classification: Procedural discourse in the Brown and LOB corpora) diskutiert die Klassifikationen der Texte (Love Story, detective fiction usw.) und weist korrekt darauf hin daß »discourse topic hardly qualifies as a serious candidate for scientific classmcation" (249). Dies führt er aus an der Kategorie E von LOB, der Texte versammelt wie homecraft, food, trade, professional journals, travel, pets, hobbies usw. Der Hinweis, daß die Klassifikation ein bißchen linguistisches Wissen (z.B. im Bereich des procedural discourse) berücksichtigen sollte, ist berechtigt, einige interessante Details (daß Gebrauchstexte viele *if/when* Konstruktionen, viele Aufzählungen und Bullet-Elemente enthalten) fallen auch ab.

Der Beitrag von A. Renouf (What do you think of that: A pilot study of the phraseology of the core words of English) ist ein Beitrag für das "Handbuch des unnützen Wissens". Sie sucht im Birmingham Korpus Ketten, die nur aus high-frequency words bestehen (Birmingham corpus). Die längsten Ketten sind:

t> *he* was and *what he* was and *he* was *not* to be

t> *it* not to be *there but it* was *there* and *when the*

t> *it* was for you *it* was all for you you *said* 1.

Daß gewisse Patterns auch Sinn ergeben, läßt sich gar nicht vermeiden (all you *have to* do *is, in such* a way as to, *it* was one of *the most). What* do you *think* of *that?*

Auch der Beitrag von Chr. Geisler (Relative Infinitives in spoken and written English) tendiert in diese Richtung; er findet, daß wissenschaftliche Texte mehr PassivKonstruktionen enthalten, daß Face-to-Face- Texte mehr Personalpronomina enthalten (222), und daß Pronomina, die auf *thing, -one,* und –body enden, in fiktionalen Texten häufiger sind als in nichtfiktionalen. Erweitert das jetzt unseren Kenntnisstand über Infinitive?

Die letzten drei Beiträge des Buches sind fast die interessantesten vom Standpunkt der NLP-Anhängerschaft.

H. Hasselgard (Sequences of spatial and temporal adverbials in spoken and written English) untersucht Stellungen von Adverb-Phrasen. Sie unterscheidet *clusters* (freie Adverbiale) und *combinations* (wohl valenzgebundene). In ihrem Korpus (736 Sätze) überwiegen die *clusters.* Wenn dabei mehrere Adverb-Phrasen auftreten, stehen sie zusammen, und zwar Orts- vor Zeit-Angaben. Dies gilt nicht für *combinations.* Temporale Adverbiale sind dabei mobiler als lokale; speziell im gesprochenen Englisch wandern temporale Adverbien gern in die initiale Position (was zu 50% am Adverb *then* liegt, das zu Zwecken der Textkohäsion benutzt wird). Offen bleiben zwei Fragen: Wie verhalten sich die Adverbien vs. Adverbiale in dieser Hinsicht (mobil sind wohl v. a. die kurzen Adverbien, Adverbiale aus mehreren Wörtern müßten wohl eher zur Cluster-Bildung neigen)? Und, interessant für unsereinen: Wie stellt man solche Informationen in einer formalen Grammatik dar?

Der Aufsatz von G. Kjellmer (Grammatical or nativelike?) ist deswegen interessant, weil er genau das Dilemma der gegenwärtigen NLP-Grammatik-Arbeiten bespricht: Nicht alles, was grammatisch möglich ist, ist auch akzeptabel, oder wird auch gesprochen. Kjellmer spricht von" the existence of lexically-based restrictions that operate wi-

thin the rule-defined field" (329) und gibt dafür drei Beispiele (aus dem Brown Corpus):

1. Verbal tenses: Manche Verben kommen fast nur im Präsens vor (amount illustrate *derive denne distinguish),* andere fast nur in Past Tense *(exdaim pause smile mutter peer).*

2. Den Infinitiv nach TO benützen manche Verben oft *(TO purchase minimize dear free recover),* manche selten *(amount differ intend deduct range).*

3. Passivity: manche Verben stehen gern im Passiv *(baptize subject situate tempt retrieve institute),* manche fast nie *(come get want wish like love talk).*

Diese Beobachtungen werfen die Frage auf, wie solche Phänomene in einer formalen NLP-Grammatik darzustellen sind: Man kann ja nicht behaupten, daß ein Verb wie *baptize* nie im Aktiv auftritt; es tritt eben nur sehr selten im Aktiv auf. Eine formale Sprachbeschreibung muß solche Phänomene ausdrücken können, d.h. sie muß Schemata von Präferenzen, Gewichtungen usw. viel mehr in ihre Überlegungen einbeziehen, als das bisher getan worden ist. Und der Ausgangspunkt dieser Gewichtungen muß im lexikalischen Bereich liegen.

Der letzte Aufsatz (J. Noel: - Collocation and bilingual text) "explores a computerized procedure for uncovering collocation data in bilingual texts" (346). Die Arbeit variiert die Church-Methode des Alignment, indem sie es auf Zeilenbasis berechnet (Fr-En mit 30:35 Zeichen pro Zeile); damit werden 2700 Zeilen eines bilingualen Korpus behandelt. Die Markierung der Kollokationen, und das ist das Enttäuschende, wird aber' per Hand vorgenommen, und die Aussage, daß man manches findet, was nicht im Lexikon steht, verliert an Gewicht, da man eben nur das findet, was man vorher im Text markiert hat. Die Frage wäre gewesen, den Versuch einer automatischen Analyse zu wagen und zu sehen, ob es relevante Patterns gibt, und ob diese im Lexikon zu finden sind oder nicht. Ansonsten hat der Ansatz seine Schranken bei Sprachpaaren, in denen Kollokationen stark im Satz (und d.h. über die Zeilengrenzen hinaus) verschoben werden können; hier scheinen die "klassischen" satzbezogenen Techniken robuster zu sein.

Insgesamt fällt im dritten Teil auf, daß einige Beiträge unter dem Fehlen einer klaren Aufgabenstellung leiden, und eben Wissenswertes und weniger Wissenswertes in einerlei Gewand versammeln.

Ein Interesse der Beiträge liegt darin, herauszufinden, ob das Englische *ein* Sprachsystem ist, oder aus vielen Subsystemen besteht (d.h. ob indisches und US-Englisch verschiedene Sprachen sind); die Beiträge im vorliegenden Band bestätigen diese Annahme eher nicht, ihre Verfechter geben aber noch nicht auf.

Fragen, die von der praktischen Seite der computerlinguistischen Anwendungen sich ergeben, sind z.B.:

[> Häufigkeiten bestimmter syntaktischer Strukturen

[> Einfluß der verschiedenen Textsorten /Fachsprachen auf diese Häufigkeiten

[> praktische Relevanz und Häufigkeit bestimmter grammatischer Phänomene (syntaktische Strukturen, in Abhängigkeit vom lexikalischen Material)

Die Aufgabe ist dann, diese Informationen darzustellen und in die entsprechendene Analyse-Verfahren zu integrieren.

Einige Phänomene, die dabei beachtet werden müssen, beschreibt das vorliegende Buch. Insofern ist es eine interessante und empfehlenswerte Lektüre, auch wenn man sich manches Kopfzerbrechen, das die Autoren anregen wollen, nun wirklich nicht

machen muß.

Gregor Thurmair, Sietec München

Marc Domenig, Pius ten Hacken: Word Manager: A System for Morphological Dictionaries. Hildesheim/Zürich/New York: Olms. 211 Seiten, 39,80 DM, 1992.

Marc Domenig ist Professor für Informatik an der Universität Basel und hat sich schon in seiner Dissertation und seiner Habilitation mit computerlinguistischen Themen beschäftigt, die zu dem hier beschriebenen System hinführen. Pius ten Hacken war, bevor er nach Basel wechselte, als Computerlinguist an der Universität Utrecht u. a. in der niederländischen Eurotra-Gruppe tätig. Das System Word Manager ist in der Grundanlage Domenigs Werk; ten Hacken dürfte vor allem gründliche Kenntnisse einer weiteren Sprache (Niederländisch) und einer überaus anspruchsvollen Anwendung, der automatischen Übersetzung (Eurotra), beigetragen haben.

Word Manager ist ein aufwendiges Softwaresystem zum Aufbau und zur Pflege wortgrammatischer Daten und Regeln. Der Ausdruck morphology umfaßt hier die ganze Wortgrammatik, d.h. Morphologie und Wortbildung (vgl. Schubert 1993). Die innere Grammatik des Wortes wird in der Computerlinguistik nur zu oft auf die leichte Schulter genommen, was sicherlich nicht zuletzt daran liegt, daß viele Wissenschaftler sich an Modellen orientieren, die für das Englische entwickelt worden sind, das bekanntlich auf der Wortebene weniger formenreich ist . Versucht man wie Domenig und ten Hacken, das Problem der Wortgrammatik umfassend anzugehen, wird schnell klar, daß weder schnelle noch linguistisch anspruchslose Lösungen der Aufgabe gewachsen sind, komplexe Wörter in Stämme, Flexions- und Ableitungsmorpheme zu zerlegen und umgekehrt zu einem bestimmten Stamm alle flektierten Formen und Ableitungen zu bilden, insbesondere dann nicht, wenn mehrdeutige Analysen und Übergenerierung ausgeschlossen werden sollen. Angesichts der Komplexität des Problems versuchen Domenig und ten Hacken daher eine großangelegte Lösung, die Domenig (1990) schon früher als "computationally expensive" bezeichnet hat. Umfang und Kosten führen die Autoren zu dem Konzept eines zentralen, viele Anwendungen bedienenden Servers.

Die grundlegende Architektur ist modular. Die leere, sprachunabhängige Datenbank ist von den Autoren entwickelt worden und gilt für die darauf aufbauenden Ebenen als vorgegeben. Sie bietet Arbeitsoberflächen für Grammatiker, die für eine bestimmte Sprache ein wortgrammatisches Regelsystem, und für Lexikografen, die das dazugehörige wortgrammatische Wörterbuch aufbauen. Hinzu kommen Schnittstellen für die Clientanwendungen, die die Datenbank befragen. Word Manager ist damit eine Arbeitsumgebung für beliebige Sprachen und nicht das fertige Morphologiepaket für Sprache X, das gelegentlich Anbieter sprachtechnologischer Anwendungen schnell irgendwo zu kaufen versuchen. Domenig und ten Hacken nennen den Kern ihres Systems daher auch Conceptual Morphological Knowledge Specification environment.

Die Autoren haben ihr System gründlich getestet. Sie referieren Teile ihrer Implementierung für Deutsch, Italienisch, Niederländisch und Englisch. Dies sind Sprachen, deren Wortgrammatik sich nicht auf "einfache" Agglutination beschränkt, sondern Vor- und Nachsilben, Ablaut, Fugenlaute und Trennbuchstaben sowie viele andere Morphemveränderungen kennt. Daß Domenig und ten Hacken ihren Word Manager an diesen wortgrammatisch

komplexen Sprachen mit einem Maß an Vollständigkeit ausprobiert haben, das auch anspruchsvollen Anwendungen gerecht wird, läßt hoffen, daß sich das System als praxisrobust erweisen wird.

Die Arbeit bewegt sich zwischen einer begrifflichen Definition des gewählten Modells und der in ihm abgebildeten sprachlichen Regelmäßigkeiten einerseits und einer detaillierten technischen Beschreibung der Bedienoberflächen, Zeichensätze und Regelnotationen andererseits hin und her. Sie verzichtet jedoch nicht darauf, auch konkurrierende Ansätze zu besprechen. Insbesondere betrachten die Autoren das System der niederländischen Lexikologiedatenbank CELEX, Koskenniemis Zweiebenenmorphologie und DATR nach Evans und Gazdar.

Eine interessante Arbeit, die mit erfreulicher praktischer Untermauerung einen Be
reich darstellt, der gern vergessen wird, auf dem aber der ganze Rest des computergrammatischen Gebäudes ruht.

Literatur

Domenig, Mare (1990): *Lexeme-based morphology: a computationally expensive approach intended for a server architecture.* In: Hans Karlgren (Hg.): Papers presented to the 13th International Conference on Computational Linguistics. Helsinki: Universitas. Bd. 2: 77-82

Schubert, Klaus (1993): *Semantic compositionality: Esperanto word formation for language technology.* Linguistics 31: 311-365

Klaus Schubert} FH Flensburg

Hooshang Mehrjerdian:
Automatische Übersetzung englischer Fachtexte ins Persische. Sprache und Information Bd. 25, Niemeyer, Tübingen 1993. 171 Seiten.

Bei diesem Band der Reihe Sprache und Information handelt es sich um einen interessanten Versuch, die im EUROTRA-Projekt erarbeiteten Ergebnisse auf eine nicht-westeuropäische Sprache anzuwenden. EUROTRA war als Forschungs- und Entwicklungsprogramm der Kommission der Europäischen Gemeinschaft konzipiert, das einen Prototyp eines MÜ-Systems für alle 9 Amtssprachen zum Ziel hatte: Dieses ehrgeizige Ziel wurde innerhalb der Projektlaufzeit von 1985 bis 1992 nur teilweise erreicht; die Bewerter stimmen jedoch darin überein, daß neben dem Aufbau einer konkreten Zusammenarbeit und der Schaffung einer Infrastruktur in weniger entwickelten Ländern auch eine Reihe von interessanten linguistischen Ergebnissen erzielt worden sind. Auf diesen Ergebnissen setzen in der Folge einige nationale Forschungsprojekte sowie neue industrielle Entwicklungen wie EUROLANG auf.

Auch in der akademischen Welt, wo teilweise kontroverse Diskussionen um die von EUROTRA verfolgten Konzepte geführt und oft alternative Konzepte entworfen wurden, hat dieses große Projekt seine Ausstrahlung gezeigt; an einzelnen Stellen sind Anstrengungen unternommen worden, die in langen Diskussionen für die 9 westeuropäischen Sprachen entwickelten linguistischen Modelle an weiteren Sprachen zu erproben.

Hierzu gehört auch die vorliegende Arbeit von H. Mehrjerdian, die als Dissertation von W. Lenders (Bonn) betreut worden ist.
Sie gibt (nach einem kurzen Überblick über den Aufbau der Arbeit) zunächst einen Einblick in die speziellen Belange der persischen Sprache (Farsi), die für Phonologie, Morphologie und feste Wortklassen relativ detailliert und auch für den Nicht-Spezialisten (z.B. den Rezensenten) verständlich ist; die Angaben über die Syntax sind dagegen eher etwas knapp ausgefallen. Hier müßte ein Farsi-Kenner detailliertere Urteile anbringen können.

Im weiteren wird ein Überblick über die

Prinzipien des EUROTRA-Systems sowie die Implementierung der englischen Analyse gegeben, die der Autor als Basis für den von ihm zu erstellenden Transfer Englisch-Persisch sowie die persische Generierung verwendet.
Die dabei entstehende Kurzdokumentation der in der englischen Analyse verwendeten Attribut-Wert-Paare ist von großem Nutzen für alle, die sich mit einer automatischen Syntaxanalyse des Englischen beschäftigen müssen; EUROTRA selbst hat sich die Mühe einer solchen (leicht verständlichen) Darstellung nur selten gemacht! Erst in den jetzt erscheinenden Heften der Reihe "Studies in Machine Translation and Natural Language Processing" (zu beziehen über das Veröffentlichungsbüro der EG), die dem Verfasser noch nicht zur Verfügung standen, ist eine solche Richtung zu erkennen.

Weniger von allgemeinem Interesse ist dagegen die (sehr mit technischen Einzelheiten belastete) Beschreibung von Transfer, für die Beschreibung der Synthese sind wiederum detaillierte Farsi-Kenntnisse notwendig. Hier wäre fast zu überlegen gewesen, die ganze Dissertation in englischer Sprache abzufassen - sie hätte sicherlich mehr Publikum gefunden.

Das letzte Kapitel" Wörterbücher" bringt in etwas ungleich gewichteter Verteilung nunmehr die detaillierten Informationen zum Persischen (die für das Englische im Analysekapitel auftauchten); Transfer- und Fachwörterbücher werden nur noch kurz gestreift, was etwas im Widerspruch zur expliziten Erwähnung der Fachsprache im Titel steht. Ähnliches gilt für das Schlußkapitel "Zusammenfassung und Ausblick", das Verbesserungsideen etwas zusammenhanglos präsentiert: neben technischen Problemen der vom Autor verwendeten EUROTRA-Version steht die inhärente Problematik der Satz-für-Satz-Übersetzung sowie die Notwendigkeit eines vorgeschalteten Grammar-Checkers.
Man hat den Eindruck, als sei der Autor froh, daß er sich endlich durch den großen Brocken EUROTRA durchgearbeitet habe und nun ein bißchen erschöpft sei.

Darin besteht eigentlich das größte Verdienst dieser Arbeit: in einem verhältnismäßig frühen Stadium mit einem äußerst komplexen und noch nicht ausgereiften System in einer exotischen Sprache experimentiert zu haben; zumindest liegen damit erstmals Elemente' einer (moderneren) formalen Beschreibung des Persischen vor.

Johann Haller, Univ. Saarbrücken

INFORMATION RETRIEVAL IN REGENSBURG

Mark Rittberger
Universität Konstanz

Information Retrieval '93 Von der Modellierung zur Anwendung

1. Fachtagung der Gesellschaft für Informatik (GI) in Regensburg vom 13.-15. September 1993

Zwei Jahre nach ihrer Gründung veranstaltete die Fachgruppe Information Retrieval der GI in Zusammenarbeit mit der Linguistischen Informationswissenschaft der Universität Regensburg ihre erste Fachtagung im oberpfälzischen Regensburg.

Etwa 100 Teilnehmer, vornehmlich aus dem deutschsprachigen Raum, fanden sich im mittelalterlichen Regensburg zusammen, um 17 Vorträge zu verfolgen. Gleich im voraus sei gesagt, daß sowohl die 3 eingeladenen Referenten als auch die 14 begutachteten Vortragenden überaus interessante Entwicklungen aus dem Bereich des Information Retrieval und angrenzender Gebiete, wie etwa Datenbanken, Benutzerschnittstellen oder Hypertext vorstellten. Karen Sparck Jones von der Cambridge University (GB) eröffnete die Tagung mit ihrem Vortrag "What might be in a summary?". Nachdem sie einen ausführlichen Bericht über Methoden zur Unterstützung von automatischen Zusammenfassungen gegeben hatte, beschrieb sie ihr eigenes Projektvorhaben. Sie teilte die für die Generierung von Zusammenfassungen notwendige Information in die drei Typen linguistische Information, Information über das Weltwissen und kommunikative Information ein. Diese werden zusammen mit zwei Repräsentationsformaten zur Textrepräsentation für die Generierung von Zusammenfassungen verwendet.

Im zweiten eingeladenen Vortrag sprach Nicholas J. Belkin von der Rutgers University (USA) über "Interaction with Texts: Information Retrieval as Information Seeking Behaviour". Ausgehend vom Standard Information Retrieval Modell zeigte er, daß beim Information Retrieval als Suchvorgang das gesamte Verhalten eines Benutzers bei der Informationssuche berücksichtigt werden muß. Daraus folgt notwendigerweise, daß ein Information Retrieval System den Benutzer in allen Stufen des Information Retrieval Prozesses unterstützen muß. Der zentrale Prozeß dieser Unterstützung ist die Interaktion mit dem Nutzer. Die Interaktion spielt aber auch gegenüber anderen Prozessen, wie etwa Repräsentation oder Vergleich eine wichtige Rolle.

Als letzten eingeladenen Gast begrüßten die Teilnehmer Edward A. Fox von dem Virginia Tech (USA). Sein Vortrag "From Information Retrieval to Networked Multimedia Information Access" bezog sich vornehmlich auf die Chancen, die neue Informationsdienste, insbesondere solche, die über das Internet angeboten werden, für die Forschung im Information Retrieval bieten. Insbesondere beklagte er den Mangel an Aktivität im Bereich Multimedia bzw. Netzwerke und Information Retrieval. Er

rief die Forschungsgruppen im Bereich Information Retrieval auf, sich mehr in diesen Gebieten zu engagieren, da dort die Notwendigkeit für praktisch einsetzbare Information Retrieval Verfahren besonders hoch sei.

Parallel zu den in fünf Sitzungen aufgeteilten Vorträgen wurden verschiedene der in den Vorträgen diskutierten Systeme vorgeführt.

A. Burghardt, N. Fuhr, K. Grossjohann, U. Pfeifer, H. Spielmann und O. Stach begannen die erste Sitzung mit "NOSFERATU - ein integriertes Datenbank- und Information-Retrieval-System basierend auf dem Datenstrom-Paradigma". Sie stellten ein System vor, welches vage Anfragen und unsichere Repräsentationen in Datensammlungen berücksichtigt. Mit einem probabilistischen NF2-Modell versuchen sie bei großen Datenmengen den Zugriff auf Fakten- und Textinformation mit gerichteten, azyklischen Graphen zu verbessern.

H.-P. Frei und Y. Qiu berichteten über ein Retrievalexperiment mit der Online-Datenbank INSPEC ("Effectiveness of Weighted searching in an Operational IR Environment"), in dem boole'sche und gewichtete Anfragen auf ihre Genauigkeit und Brauchbarkeit hin verglichen wurden.

Die zweite Sitzung wurde von J. Gu, U. Thiel und J. Zhao mit "Efficient Retrieval of Complex Objects: Query Processing in a hybrid DB and IR System" eröffnet. Dabei wurde die Verbindung von relationalen Datenbanksystemen und auf invertierten Dateien beruhenden Information Retrieval Systemen untersucht, um komplexe Objekte leichter suchbar zu machen.

Im Anschluß daran erklärte M. Hemmje "eine inhaltsorientierte, intuitive 3D Benutzerschnittstelle für Information-Retrieval-Systeme", bei der der Benutzer seine Suchterme auf der Oberfläche einer Kugel entsprechend ihrer Gewichtung zueinander bewegt. Abhängig von der Bewegung werden dann die Antworten als räumlich den Suchtermen am nächsten liegende Objekte präsentiert.

R. Kuhlen und M. Hess beschreiben das "Passagen-Retrieval- auch eine Möglichkeit der automatischen Verknüpfung in Hypertexten" . Die Autoren verglichen verschiedene Ähnlichkeitsmaße in Bezug auf ihre Effektivität bei der Verknüpfung von Textpassagen innerhalb eines aus dem Buchformat in Hypertextformat transformierten Buches.

Am zweiten Tag begannen M. Schmidt und U. Pfeifer die Vorträge zur dritten Sitzung mit einer Untersuchung zur "Eignung von Signaturbäumen für Best-Match Anfragen". Dabei zeigte sich, daß ein höhenbalancierter Signaturbaum als Zugriffspfad zur Beantwortung von Best-Match-Anfragen bei größeren Kollektionen nicht geeignet erscheint.

Im Beitrag von S. Meienberg ("Relevance Feedback by Relative Relevance Assessments") ging es um eine neue Relevance-Feedback-Methode zur Verbesserung der Frageformulierung. Deren Effektivität erwies sich im Vergleich zu anderen Verfahren als sehr hoch.

S. Roppel, C. Wolff und C. Womser-Hacker ("Intelligentes Faktenretrieval am Beispiel der Werkstoffinformation") stellten die im Projekt WING-IIR entwickelten Retrievalkomponenten für das graphische Retrieval von numerischer Information vor, welches Antworten über vages Wissen zu Materialeigenschaften präsentiert. Sie beschrieben ein Benutzermodell, mit dem die Benutzerschnittstelle in Abhängigkeit von den Benutzerinteressen vereinfacht wird.

Zu Beginn der vierten Sitzung knüpften J. Marx und M. Schudnagis an den vorangegangenen Vortrag an, indem sie die Entwicklung der Benutzerschnittstelle im Projekt WING-IIR erläuterten. Dabei wurden verschiedene Schnittstellenkonzepte in eine multimodale Datenbankoberfläche eingearbeitet.

A. Glöckner-Rist bezog sich in ihrem Vortrag "Suche und Du wirst finden: Die Formulierung von Suchproblemen und ihre

Transformation in Suchfragen " auf eine Versuchsanordnung, in der die Abhängigkeit der Suchfrageformulierung von den Retrievalkenntnissen der Benutzer bei Recherchen in Online-Datenbanken untersucht wurde. Die Effektivität wurde dabei sowohl durch die Inhalte der Problembeschreibung, als auch die Information Retrieval- und Problemkenntnisse der Rechercheure beeinflußt.

M. Herfurth, P. Mutschke und H.P. Ohly stellten "AKCESS: Konzept-orientiertes Retrieval mit bibliographischem Kontextwissen" vor. Sie nutzen wissenschaftsstrukturelle Information über Literatur, Forschungsprojekte und -institute sowie über Lehrveranstaltungen, um einschlägige Wissenschaftler in Bezug auf eine Fragestellung zu identifizieren.

Die Schlußsitzung wurde von R. Hammwöhner und M. Rittberger mit "KHS-ein offenes Hypertextsystem" eröffnet. Sie gaben zunächst einen Überblick über das Konstanzer-Hypertext-System (KHS) und die Aspekte der Offenheit, die das System auszeichnen. Anschließend beschrieben sie die Integration von e-mail und Online-Datenbanken in das KHS.

U. Kampffmeyer stellte. "Multilinguale Dokumenten- Retrievalsysteme. Implementierung und Beispiele" vor. Er erläuterte die Notwendigkeit solcher multilingualer Retrievalsysteme, ehe er ihre Konzeption und die Arbeitsweise anhand eines Archivierungssystems für Faksimile-Dokumente und Dateien, einer Pressedokumentation in der Schweiz und einer Metadatenbank für die Vereinten Nationen beschrieb.

Ein weiterer Beitrag im Schnittfeld Hypertext und Information Retrieval von den Autoren M. Hofmann, E. Bartsch und S. Schmezko ("Klassendefinition und Anfrageformulierung in Informationssystemen durch graphische Interaktion") beschäftigte sich mit der graphischen Schnittstelle im Hypertextsystem CONCORDE, die zur Definition von Knoten- und Linktypen dient und die Abfrage von typisierten Strukturen in einem Hypertext erlaubt.

Begleitet wurde die Tagung von einem Tutorial zum Thema Information Retrieval, welches von N. Fuhr, U. Pfeifer und G. Ruge angeboten wurde. Für Auflockerung sorgte ein kulturell und kulinarisch gelungener Abend, der mit einer Besichtigung der Handels- und Reichstagsstadt Regensburg begann und einem gemeinsamen bayerischen Abendessen in der Altstadt schloß.

Zusammenfassend läßt sich sagen, daß sowohl die Vielfalt und Breite der Themen, das große Interesse von Information Retrieval-Spezialisten aus Wissenschaft und Praxis, als auch die hervorragende Organisation zum Gelingen der Tagung beitrugen. N ach diesem Erfolg wird es daher auch eine zweite Information Retrieval Tagung im Mai 1995 in Konstanz geben.

Alle Beiträge sind im Tagungsband (Information Retrieval '93. Von der Anwendung zur Modellierung) herausgegeben von G. Knorz, J. Krause und C. Womser-Hacker im Universitätsverlag Konstanz erschienen.

REALISTISCHE BEWERTUNG VON RETRIEVALVERFAHREN

Ulrich Pfeifer
Universität Dortmund

Realistische Bewertung von Retrievalverfahren

Bericht von der zweiten
TREC-Konferenz
(30.8-2.9.93 in Was hingt on)

Ziele der Initiative

Ziel der US-amerikanischen TREC (Text Retrieval Conference) Initiative ist es, zum einen Standard-Kollektionen von Texten zu erstellen, auf deren Basis verschiedene Retrieval-Ansätze verglichen werden können. Zum anderen sollen Kollektionen einer Größe aufgebaut werden, die realistischen Anwendungen nahe kommen. Damit soll das alte Vorurteil der Untauglichkeit neuerer IR-Methoden (bzgl. Effizienz und Effektivität) für große IR-Datenbasen widerlegt werden.

Inhalt

Im Rahmen der Initiative wurde eine Kollektion von mittlerweile 3 GB Text (Zeitungsmeldungen, Patentschriften, Referate wissenschaftlicher Artikel, ...) aufgebaut, die auf drei CD-ROMs an die Teilnehmer verteilt wurden.

Die Teilnehmer hatten zwei Arten von Aufgaben zu lösen:

ad-hoc: Eine Liste von 50 Fragen im Umfang von je etwa einer A4-Seite Text, die gegen eine Teilkollektion (1 bzw. 2 GB) von Dokumenten getestet werden sollten.

routing: Weitere 50 Fragen, für die für eine CD Feedback-Daten zur Verfügung standen, und die dann für eine weitere CD prozessiert werden sollten.

Die Teilnehmer (etwa 30 Systeme) sandten für jede Frage eine Rangliste von Dokumenten als Antwort ein, für die dann Relevanzurteile erstellt wurden. Darauf aufbauend konnte die Retrievalqualität der Systeme verglichen werden.

Kategorien für die Teilnahme

Die Teilnehmer konnten verschiedene Kategorien von Ansätzen verfolgen:

1. Vollautomatische (initiale) Frage-Erstellung
2. Manuelle (initiale) Frage-Erstellung
3. Manuelle (initiale) Frage-Erstellung mit Feedback. *Das heißt eigenes Feedback bei den ad-hoc Fragen*

Weiter wurde zwischen voller (Kategorie A) und eingeschränkter Teilnahme (Kategorie B) unterschieden. Nur Kategorie-A-Teilnehmer mußten mit der vollen Datenmenge arbeiten.

Ergebnisse

Ohne der genaueren Auswertung vorgreifen zu wollen, lassen sich tendenziell folgende Aussagen machen.

=> Bei den ad-hoc Fragen schneiden voll-automatische Systeme im Mittel besser ab als Systeme, die auf manueller Konstruktion der Fragen beruhen. Andererseits gibt es bei automatischen Systemen meist eine breitere Streuung der Ergebnisse zwischen einzelnen Fragen.

Die sechs besten automatischen Systeme basierten auf dem Vektorraummodell bzw. verschiedenen probabilistischen Modellen. Das System von Cornell, Siemens und Carnegie Mellon basierten auf dem Vektor-raummodell, wobei die beiden letzteren Thesauri zur Frageerweiterung verwendeten. Die Systeme von Berkley, Dortmund und University of Massachusettes basieren auf probabilistischen Modellen, wobei letzteres ein probabilistisches Netzwerk als Retrie-valmaschine verwendet.

Bei den manuellen Systemenschnitten das System der UMASS bzw. Carnegie Mellon am besten ab, die sich auf Korrekturen der automatisch erstellten Fragen beschränkten. N ur diese beiden Systeme konnten sich mit den automatischen Systemen messen. Die rein manuellen Systeme waren signfikant schlechter.

=> Bei den routing Fragen entstand bezüglich des Vergleichs von automatischen mit manuellen Systemen das gleiche Bild. Hier sind die Unterschiede nur noch signifikanter als bei den ad-hoc Fragen.

Bei den automatischen Systemen war das System aus Cornell dem aus Dortmund überlegen, das wiederum signifikant besser als der Rest der Mitbewerber abschnitt.

Neben den Systemen der University of Massachusettes und Carnegie Mellon schnitten die Systeme von General Electrics und TRW gut ab, die konventionelle boolesche Systeme einsetzten.

=> Die von vielen Anbietern kommerzieller Systeme unterstellten Effizienzprobleme gibt es tatsächlich: bei den kommerziellen Systemen! Während man bei einem Firmenprodukt 8 Minuten auf eine Antwort warten muß, liefert beispielsweise das SMART-System (Cornell) auf einer SUN Workstation bei einer Frage (mit 50 ... 100 Wörtern) bezüglich einer 2GB Kollektion die Antwort innerhalb von 10 Sekunden ab.

Ausblick

Die TREC-Initiative wird fortgesetzt und geht damit schon in das dritte Jahr. In den neuen Kollektionen werden wohl zum einen fremdsprachige Teile (als Option, wahrscheinlich Spanisch) zum anderen *noisy*-Teile wie e-mail aufgenommen werden.

Informationen

Die Proceedings zur ersten TREC-Konferenz wurden vom National Institute of Standards and Technology (NIST) als Special Publication No. 500-207 veröffentlicht.

Redaktionsschluß für den zweiten Tagungsband ist im November diesen Jahres, der kurz darauf von gleicher Quelle zu beziehen sein wird.

Ankündigung

Begriffliche Wissensverarbeitung

Technische Hochschule Darmstadt
Fachbereich Mathematik
Schloßgartenstraße 7 Raum 201
(Forschungsgruppe Begriffsanalyse)

Beginn: 23.02.94, 16.00 Uhr Ende: 26.02.94, 13.00 Uhr

Mit der Tagung wird eine Tagungsreihe fortgesetzt, die mit der Arbeitstagung Begriffsanalyse im Januar 1986 begonnen wurde (s. Ganter, Wille, Wolff: Beiträge zur Begriffsanalyse, BI-Wissenschaftsverlag 1987) und die in größerer Breite Wissenschaftler aus dem Bereich der Daten-, Informations- und Kognitionswissenschaften zusammenführt.

Die angekündigte Tagung wird von der Darmstädter Forschungsgruppe Begriffsanalyse in Zusammenarbeit mit dem Ernst Schröder Zentrum für Begriffliche Wissensverarbeitung und der International Society of Knowledge Organization veranstaltet. Mit der Tagung soll die Forschung, Entwicklung und Anwendung auf dem Gebiet der Begrifflichen Wissensverarbeitung gefördert werden. Ein zentrales Anliegen ist dabei, Menschen im rationalen Denken, Urteilen und Handeln durch geeignete Werkzeuge und Methoden zu unterstützen und damit einem drohenden Abbau kognitiver Autonomie durch nicht mehr beherrschbare Wissens- und Informationssysteme entgegenzuwirken. Wie schon Tradition ist, wird die Tagung so geplant, daß viel Zeit zu Diskussion und Informationsaustausch bleibt. Unmittelbar vor der Tagung (am 22. und 23. Februar 1994) finden in der TH Darmstadt Kurse zur Begrifflichen Datenanalyse und Begrifflichen Wissensverarbeitung statt.
Anmeldeunterlagen sind anzufordern bei:

Prof. Dr. Rudolf Wille, FB Mathematik, Technische Hochschule, 64289 Darmstadt, Schloßgartenstr. 7 (Tel.: 06151 163415, FAX: 06151 164011, Internet: ag01@mathematik.th-darmstadt.de)

Anmeldeschluß: 31.01.94

Einführung in die Begriffliche Datenanalyse

Kurs im Rahmen des Aus- und Weiterbildungsprogramms des Ernst Schröder Zentrums TH Darmstadt (22.-23.2.94)

In vielen Bereichen werden Daten in Tabellen festgehalten, um den Zusammenhang zwischen Objekten und ihren Eigenschaften darzustellen. Die Daten enthalten Strukturen, die in der Begrifflichen Datenanalyse erschlossen und als Begriffshierarchien graphisch dargestellt werden. Dadurch lassen sich komplexe Strukturen analysieren und Abhängigkeiten erkennen. Das ermöglicht eine Interpretation unmittelbar anhand der Originaldaten.

Anwendungen

. Auswertung von Befragungen

. Analysen in Soziologie und Psychologie

. Auswertung medizinischer Daten

. Optimierung von Produktionsprozessen

. Untersuchung sprachlicher Zusammenhänge

Ziele

Die Teilnehmer sollen mit den Methoden der *Begrifflichen Datenanalyse* vertraut gemacht werden. In praktischen Übungen wird das Erstellen von *Begriffssystemen* und *Liniendiagrammen* geübt, so daß die Teilnehmer in der Lage sind, eigene Daten zu analysieren und zu interpretieren.

Kursinhalte

. Tabellen und Kontexte

. Begriffshierarchien

. Liniendiagramme

. Implikationen in Daten

. Begriffsanalyse realer Daten

. Übungen mit den Programmen CON IMP und DIAGRAM

Zielgruppen

Datenanalytiker in Industrie, Verwaltung und Wissenschaft.
Linguisten, Soziologen, Psychologen, Mediziner, Biologen, ...

Voraussetzungen

Der Kurs ist elementar und erfordert keine speziellen mathematischen Kenntnisse. Erfahrungen in der Datenanalyse sind hilfreich, aber nicht notwendig.

Kursleitung

Prof. Dr. K. E. Wolff

Anfragen (Preise) und Anmeldungen an das Ernst Schröder Zentrum (Adresse siehe Seite 54)

Einführung in die Begriffliche Wissensverarbeitung

Kurs im Rahmen des Aus- und Weiterbildungsprogramms des Ernst Schröder Zentrums TH Darmstadt (22.-23.2.94)

Die mathematisch fundierte Methode der *Begrifflichen Wissensverarbeitung* befaßt sich mit der Repräsentation, der Inferenz, der Akquisition sowie der Kommunikation begrifflichen Wissens. Ein zentrales Analyse- und Auswertungsinstrument der *Begrifflichen Wissensverarbeitung* sind *Begriffliche Datensysteme,* die sich in zahlreichen Anwendungsgebieten bewährt haben. Ihre Aufgabe besteht in der Repräsentation und in der Kommunikation von Wissen. Durch geeignete graphische Aufbereitung lassen sich durch *Begriffliche Datensysteme* die Strukturen kaum überschaubarer Datenbestände klar und übersichtlich darstellen. Hierbei sind die Ausgangsdaten jederzeit rekonstruierbar, so daß die Interpretationen der Daten überprüfbar bleiben. *Begriffliche Wissensverarbeitung* unterstützt somit die rationale Kommunikation und macht Entscheidungen effizienter.

Anwendungen

. Klassifikation von Dokumenten

. Diagnose von Situationen

. Retrieval von Informationen

. Unterstützung von Entscheidungen

Ziele

Die Teilnehmer sollen vor allem mit *Begrifflichen Datensystemen* sowohl in theoretischer wie auch praktischer Hinsicht vertraut werden, so daß sie in der Lage sind, selbständig Wissensbestände begrifflich zu untersuchen. Um Erfahrung mit diesem Auswertungsinstrument zu sammeln, werden praktische Übungen an vorgegebenen Datenbeständen durchgeführt. Dabei wird auch die Handhabung des rechnergestützten Verwaltungssystems TOSCANA für Begriffliche Datensysteme erklärt und eingeübt.

Kursinhalte

. Einführung in die Begriffliche Wissensverarbeitung

. Tabellen und Kontexte

. Entwicklung von Fragekomplexen (Skalen)

. Liniendiagramme von Begriffshierarchien

. Struktur Begrifflicher Datensysteme

. Anwendung des Programms TOS CANA

Zielgruppen

Mitarbeiter der Wirtschaft wie des öffentlich-rechtlichen Bereichs, die sich mit Daten- bzw. Wissensauswertung beschäftigen. Personen, die Sitzungen, Besprechungen und Entscheidungen informationell vorbereiten. Wissenschaftler der verschiedenen Fachrichtungen.

Voraussetzungen

Erfahrungen in der Datenanalyse sind hilfreich, aber nicht notwendig.

Kursleitung

Dr. W. Kollewe, Dipl.-Math. F. Vogt

Anfragen (Preise) und Anmeldungen an das Ernst Schröder Zentrum (Adresse siehe Seite 54)

TAGUNGSKALENDER

07.02.–11.02.1994 Hamburg, BRD 17. Europäischer Congreß für Technische Kommunikation **Online'94**
Information: Online GmbH, Postfach 100866, 42508 Velbert Tel.: 02051 23071 Fax: 02051 21993 BTX: 20353

21.02.–25.02.1994 Darmstadt, BRD
International Conference on **Multimedia Computing and Systems**
Information: Erich J. Neuhold, GMD–IPSI/Technische Hochschule Darmstadt, Dolivostr. 15, 64293 Darmstadt

28.02–02.03.1994 Hamburg, BRD 3. Workshop Informationssysteme und Künstliche Intelligenz **IS-KI'94**
Information: Dr. Heinz Marburger, Mikroelektronik Anwendungszentrum Hamburg GmbH, Karnapp 20, 21079 Hamburg

01.03.–02.03.1994 Ilmenau, BRD 3. Fachtagung Softwaretechnik in Automatisierung und Kommunikation. Datenbanken unter Realzeit- und technischen Entwicklungsanforderungen **STAK'94**
Information: Prof. Dr. U. Engmann, TU Ilmenau, Fakultät für Informatik und Automatisierung, Postfach 327, 98684 Ilmenau

09.03.–11.03.1994 Oldenburg, BRD 18. Jahrestagung Gesellschaft für Klassifikation in Oldenburg e.V.
„Von Daten zu Wissen" Information: Prof. Dr. D. Pfeifer Fachbereich Mathematik, Universität Oldenburg, Postfach 2503, 26015 Oldenburg, Email: 206150@doluni1.bitnet, Tel.: 0441 798–3243,–3237, FAX: 0441 7983004

10.03.–11.03.1994 Marburg, BRD Workshop **Semantikgestützte Analyse, Entwicklung und Generierung von Programmen**
Information: Dr. J. Uhl, IBM Labor, Schönaicherstr. 22, 73560 Böblingen

16.03.–18.03.1994 Hamburg, BRD
Arbeitstagung **Erfahrung und Abstraktion – Frauensichten auf die Informatik**
Information: Heidi Schelhowe, Universität Hamburg, Fachbereich Informatik, Vogt–Kölln–Str. 30, 22527 Hamburg

16.03.–23.03.1994 Hannover, BRD
CeBIT'94
Information: Deutsche Messe AG, Messegelände, 30521 Hannover, Tel.: 0511 8931014, FAX: 0511 8932630

17.03.–18.03.1994 Kassel, BRD
Workshop **Benutzungsschnittstellen für Datenbanken** Information: Prof. Dr. Lutz Wegner, Universität Gesamthochschule Kassel, Fachbereich Mathematik/Informatik, Heinrich–Plett–Str. 40, 34109 Kassel, Email: wegner@db.informatik.unikassel.de

28.03.–31.03.1994 Cambridge, Großbritanien
4th International Conference on Extending Database Technology **EDBT94**
Information: Dr. Matthias Jarke, Informatik V. RWTH Aachen, Ahornstr. 55, 52074 Aachen

11.04.–13.04.1994 Las Vegas, USA Third Annual Symposium on Document Analysis and Information Retrieval
Information: Mary C. Guirsch, c/o Information Science Research Institute of Nevada, Las Vegas, NV 89154–4021, Email: mary@isri.univ.edu

13.04.–15.04.1994 Sankt Augustin, BRD
Conference on Electronic Publishing, Document Manipulation and Typography **Electronic Publishing '93**
Information: EP 94, Frau Harms, GMD Birlinghoven, Postfach 1316, 5205 Sankt Augustin 1. Email: ep94@gmd.de, Tel.: 02241 142473, FAX: 02241 142472

21.04.–23.04.1994 Oberhof/Thüringen, BRD
18. Oberhofer Kolloquium über Information und Dokumentation **„Informationsvermittlungsstellen als Kern des internen Informationsmanagement"**
Information: TU Ilmenau, FG Informationsmanagement, Prof. H.–J. Manecke, Postfach 327, 98684 Ilmenau, Tel.: 03677 694041 Fax: 03677 694204

24.04.–28.04.1994 Boston, USA CHI'94
Information: Prof. Dr. Horst Oberquelle, Universität Hamburg, Fachbereich Informatik, Vogt–Kölln–Str. 30, 33527 Hamburg

02.05.–04.05.1994 Hong Kong, China
Internationale Konferenz **Data and Knowledge Systems for Manufactoring and Engineering**
Information: Prof. Dr. F. Wahl, TU Braunschweig, Institut für Robotik und Prozeßinformatik, Hamburger Str. 267, 38114 Braunschweig

03.05.–04.05.1994 Pensacola, USA
International Symposium on **Integrating Knowledge and Neutral Heuristics**
Information: Prof. Dr. F. Belli, Pohlweg 47–49, 33098 Paderborn

14.05.–19.05.1994 Boston, USA
1994 International Conference on Multimedia

Computing and SystemsICMCS'94
Information: Scott M. Stevens, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA 15313, USA, Email: sms@sei.cmu.edu

16.05.–20.05.1994 Banff, Kanada
Canadian Artificial Intelligence Conference **AI/GI/VI'94**
Information: Renée Elio, Computing Science Department, University of Alberta, Edmonton AB T6G, Canada

17.05.–19.05.1994 Frankfurt/Main, BRD
DGD–Online–Tagung **„Information & Medienvielfalt"**
Information: Deutsche Gesellschaft für Dokumentation, Hanauer Landstraße 126–128, 60314 Frankfurt/Main, Tel.: 069 430313, FAX: 069 49090996

17.05.–19.05.1994 Frankfurt/Main, BRD
INFOBASE
Information: Messe Frankfurt GmbH, Postfach 150210, 60001 Frankfurt/Main

24.05.–26.05.1994 Tutzing, BRD
Workshop über **Formale Grundlagen für den Entwurf von Informationssystemen**
Information: Prof. Dr. G. Vossen, Universität Münster, Institut für Wirtschaftsinformatik, Grevenerstr. 91, 48159 Münster

24.05.–27.05.1994 Bonn, BRD 4th International Conference on Principles of Knowledge Representation and Reasoning **KR'94**
Information: Werner Horn, Austrian Research Institute for Artifical Intelligence, Schottengasse 3, A–1010 Austria

25.05.–27.05.1994 Dortmund, BRD
Bibliotheka
Information: Westfalenhalle Dortmund GmbH, Rheinlanddamm 200, 44139 Dortmund, Tel.: 0231 1204521

31.05.–02.05.1994 Paris, Frankreich
Information management for company excellence **IDT'94**
Information: Marie–Louise Zeutzem, IDT/ANRT, 101, Avenue Raymond Poincaré, F–75116 Paris, Tel.: ++33 145017227, FAX: ++33 145018529

21.06.–24.06.1994 Kopenhagen, Dänemark
Third International **ISKO** Conference on Knowledge Organization and Quality Management.
Information: Hanne Albrechtsen, The Royal School of Librarianship, Birketinget 6, 2300 Copenhagen S. Tel.: +45 31586066, Fax: +45 32840201

03.07.–06.07.1994 Dublin, Irland 17th International Conference on Research and Developement in Information Retrieval **SIGIR'94**
Information: Alan F. Smeaton, School of Computer Applications, Dublin City University, Glasnevin, Dublin 9, Ireland. Email: sigirinfo@ca.dcu.ie

25.06.–29.06.1994 Vancouver, Canada
World Conference on Educational Multimedia and Hypermedia **ED–MEDIA–94**
Information: Gary Marks AACE, P.O.Box 2966, Charlottesville, VA 22902, USA

11.07.–14.07.1994 Bonn, BRD 1st International Conference on **Temporal Logic**
Information: Hans–Jürgen Ohlbach, Max–Planck–Institut für Informatik, Im Stadtwald, 66123 Saarbrücken

08.08–12.08.1994 Amsterdam, Niederlande
11th European Conference on Artificial Intelligence **ECAI94**
Information: Mirijam de Leeuw, Erasmus University Rotterdam, P.O.Box 1738, NL–3000 DR Rotterdam, Tel.: +31 104082302, FAX: +31 104530784, Email: M.M.deLeeuw@apv.oos.eur.nl

28.08.–02.09.1994 Hamburg, BRD
13th World Computer Congress **IFIP Congress'94**
Information: Conference Secretariat IFIP'94, c/o Congress Centrum Hamburg, Congress Organisation, 20308 Hamburg

04.09.–08.09.1994 Hamburg, BRD Tagung **Intelligent Systems Engineering**
Information: Prof. Dr. J. Lunze, TU Hamburg–Harburg, Elektronik XVII, Eißendorfer Str. 40, 21073 Hamburg–Harburg

27.09.–29.09.1994 Trier, BRD
Dokumentary 94
Information: DGD, Ostbahnhofstr. 13, 60314 Frankfurt/Main

05.10.–10.10.1994 Frankfurt/Main, BRD
46. Franfurter Buchmesse
Information: Austellungs– und Messe–GmbH des Börsenvereins des Deutschen Buchhandels, Postfach 1001, 60001 Frankfurt/Main, Tel.: 069 2102257, FAX: 069 2102227

02.11.–04.11.1994 Graz, Österreich 4. Internationales Symposium für Informationswissenschaft **ISI'94**
Information: Institut für Informationswissenschaft, Karl–Franzens–Universität Graz, Strassoldogasse 10, A–8010 Graz, Tel.: ++43 3163803560, FAX: ++43 316381413, Email: isi@edvz.kfunigraz.ac.at

29.03.–31.03.1995 Konstanz, BRD 8. Internationale Fachkonferenz der Deutschen Gesellschaft für Dokumentation e.V. (DGD) **Informationscontrolling**
Information: Werner Schwuchow, Gesellschaft für Mathematik und Datenverarbeitung mbH, Postfach 1316, 53731 Sankt Augustin. Tel.: 02241 143320 Fax: 02241 143003

26.09.–28.09.1995 Trier, BRD
Dokumentary 95
Information: DGD, Ostbahnhofstr. 13, 60314 Frankfurt/Main

2. GLDV–Herbstschule

Lexikon & Morphologie

19.9. — 24.9.1994

Universität Leipzig

Hauptvorträge: H. Schnelle, Bochum
M. Bierwisch, Berlin

Referenten

W. Lenders (Bonn)/*M. Volk* (Koblenz): Maschinenlesbare Lexika (mit einer Einführung in SGML)

W. Paprotté (Münster): MLexD; ein Standard für die (korpusbasierte) Lexikographie

G. Heyer/K. Waldhoer (Leipzig): Produkte der Sprachtechnologie

R. Hausser (Erlangen): Automatische Wortform–Erkennung und morphologische Algorithmen

J. Haller (Saarbrücken): Transfer–Lexika in der Maschinellen Übersetzung

J. Krause (Regensburg): Ergonomie und Evaluation von lexikonbasierten Sprachprodukten

Anmeldegebühren:

			Studenten
bis	1.7.94:	DM 130.-	DM 65.-
ab	2.7.94:	DM 180.-	DM 90.-

Gemeinsame Exkursion
Unterbringungsmöglichkeit in Studentenheimen

Die Teilnehmerzahl ist auf max. 100 begrenzt!

---

2. GLDV–Herbstschule 1994
c/o Prof. Dr. G. Heyer
Institut für Informatik
Automatische Sprachverarbeitung
Universität Leipzig
Augustusplatz 9
04109 Leipzig

Nachruf für Gerhard Lustig

Im Januar 1992 wurde Prof. Dr. Gerhard Lustig mit einer akademischen Feier, zu der die Technische Hochschule Darmstadt und. der Hochschulverband für Informationswissenschaft (HI) eingeladen hatten, wegen schon lange andauernder Krankheit aus seinem Hochschulamt verabschiedet. Anläßlich dieser vorzeitigen Verabschiedung wurde ihm die Ehrenmitgliedschaft des HI verliehen und eine Festschrift herausgegeben. Keine zwei Jahre später, am 6. Oktober 1993, ist er nach wechselhaftem, aber wohl doch ausichtslosem Kampf gegen die Krankheit und die schwächenden Folgen der Therapie gestorben.

Die wissenschaftliche Laufbahn von Gerhard Lustig - von seiner Ausbildung her Mathematiker - war bereits seit den 60er Jahren unmittelbar mit den Themen Computerlinguistik (automatische Sprachverarbeitung) und Information Retrieval verbunden: In ISPRA, einem internationalen EURATOM-Forschungszentrum, wo man u.a. automatische Übersetzung vom Russischen ins Englische betrieb und wo Lustig erste grundlegende Experimente mit "einer neuen Art von Assoziationfaktoren" zur automatischen Indexierung durchführte und publizierte. Als Leiter der Forschungsabteilung der Zentralstelle für maschinelle Dokumentation (ZMD) in Frankfurt, die mit ihrer 2-jährigen nachuniversitären Ausbildung zum Informationswissenschaftler die experimentell arbeitende Informationswissenschaft in Deutschland begründet hat (Diese Formulierung verdanke ich Rainer Kuhlen, der seine Entwicklung zum Informationswissenschaftler eben dieser Ausbildung verdankt). An der TH Darmstadt, im Fachgebiet Datenverwaltungssysteme II (treffender eigentlich als "Information Retrieval" zu benennen) des Fachbereichs Informatik, wo Lustig die Themen Information Retrieval, automatische Indexierung und Computerlinguistik (letzteres auch als Anwendungswahlfach für Informatiker) einbrachte. Und wo Lustig 1978 daran ging, die selbstgestellte Aufgabe im Rahmen mehrerer BMFT -geförderter Projekte konstruktiv anzugehen und bis zum Abschluß zu bringen: Die automatische Indexierung in wissenschaftlich befriedigender und praktisch überzeugender Weise als anwendungsreif nachzuweisen. Daß dieser Nachweis in eine seit 1985 laufende praktische Anwendung seines Systems AIR/PHYS für eine große Datenbasis des Fachinformationszentrums (ca. 10.000 Dokumente/Monat) eingemündet ist, steht in direkter Kontinuität zu den ersten ISPRA-Experimenten mit den sogenannten "z-Werten" und kann mit Recht als Lebenswerk von Lustig angesehen werden.

Gerhard Lustig hat mich als Wissenschaftler und Mensch zutiefst beeindruckt. Ich glaube in erster Linie durch die Verantwortung, von der sein gesamtes Handeln und Arbeiten bestimmt war: Gegenüber den StudentInnen, wenn er etwa für n> 10 keinen Moment zögerte, auch noch ein (n+1)-stes Lehrbuch durchzuarbeiten, auch wenn es "nur" darum ging, ein Konzept für eine Einführungsvorlesung "Grundzüge der Informatik I" zu erarbeiten. Gegenüber seinen Mitarbeitern, wenn z.B. seine konstruktiven Anmerkungen zu einem vorgelegten Manuskript vom Textvolumen her in die Größenordnung des Originaltextes kamen (genauso wie die dafür auf

gewendete Arbeit). Gegenüber der Fachwelt, wenn ihm nicht nur jede gefällige spekulative Formulierung fremd war und er die Grenzen seiner Kompetenz erheblich enger zog, als es der "allgemeine" Maßstab erwarten ließe, sondern er etwa die Rechtfertigung jeder einzelnen möglichen Publikation sorgfältig hinterfragte. Gegenüber der Forschungsförderung, die ihn nie zu einem lockeren Umgang mit der Formulierung von Projektzielen und Arbeitsinhalten verführen konnte. Gegenüber Umwelt und Gesellschaft, für die er sich in Bürgerinitiativen und im Rahmen der evangelischen Kirche engagierte. Lustigs Arbeitspensum war unglaublich und in einer selten gewordenen Weise frei von jeder Eitelkeit, frei von jedem Bemühen, in den Augen anderer zu glänzen - "nur" den eigenen strengen Maßstäben verpflichtet.

Und dann gab es noch einen Gerhard Lustig, den nur wenige, die ihn aus dem Hochschul- und Wissenschaftsbetrieb her kennen, in ihm vermutet haben dürften: Einen Gerhard Lustig, der bei Geburtstagen seiner Mitarbeiter höchst kreativ und witzig mit Sprache umging und seine Einfälle in abenteuerliche Limericks kleidete; der die Zeit fand, für einen Workshop-Abschlußabend ganze Balladen und Nachrichtensendungen zu erfinden und der bei solchen Gelegenheiten als Pantomime, Schauspieler und Regisseur zu glänzen verstand. Und dem Musik nicht nur als Konsument wichtig und vertraut war.

Bei allem körperlichen Verfall hat Gerhard Lustig auch noch in seinen letzten Monaten sich stets eine gewisse Arbeitskapazität erhalten, die er nach zeitweiser erzwungener Deaktivierung oft unter schwierigsten Bedingungen wieder aufleben lassen konnte. So hat er die laufenden Arbeiten, die sein wissenschaftliches Werk weiterführen, mitverfolgt und auch noch zwei Doktoranden betreut. Daß seine Arbeiten weiterentwickelt werden - wenngleich bedauerlicherweise gerade an der Stelle seines hauptsächlichen Wirkens nicht - war ihm sicher wichtig. Und davon kann im Rahmen dessen, was im Wissenschaftsbereich vorhersehbar ist, auch ausgegangen werden. Gerhard Lustig gehörte ganz sicher zu den Menschen, denen das Alter so schnell keine Grenze für ein aktives und positiv auf ihre Umgebung ausstrahlendes Leben setzt. Die Fachwelt trauert um den viel zu frühen Tod von Gerhard Lustig und manchen fällt es schwer, die Todesnachricht ganz zu begreifen. Beim Schreiben dieses Nachrufs wird mir nochmals klar, daß ich zu letzteren gehöre.

Gerhard Knorz

Don Walker 1986 mit Ehefrau Betty

Don Walker

Die GLDV betrauert, zusammen mit den amerikanischen Kollegen der ACL und mit der internationalen Gemeinde der Computerlinguisten, den Tod von Don Walker. Don Walker hat über 3 Jahrzehnte hinweg die Entwicklung der Computerlinguistik maßgeblich beeinflußt, sowohl durch seine wissenschaftlichen Beiträge, als auch durch sein großes Engagement in der Durchführung und Betreuung wissenschaftlicher Tagungen, der Herausgabe von Zeitschriften und Sammelbänden, die Tätigkeit als Sekretär der ACL und als Mitglied des ICCL. Er hat sich auch unter den

Mitgliedern der GLDV viele Freunde erworben. Wir alle werden uns immer dankbar an ihn erinnern.

Zur Information über seinen Tod fügen wir im folgenden die über e-mail verbreiteten Benachrichtigungen durch Judith Klavans und Fernando Pereira bei.

W. Lenders

Date: Sat, 27 Nov 1993 14:36:04
Betty has just informed me that Don past away, peacefully at home, surrounded by his family last night at 10:15 pm. For the past few days, he had been failing rapidly, as some of you may know. Don has chosen to be cremated, and so there will be a memorial service for family and friends in California at a time Betty will decide on in the future.

Betty wants to thank the many of you who have had Don and her in their thoughts and prayers over the past period of illness. This has given them both great strength in dealing with the trials, pain, and also joys of the past years. Your help has been appreciated.

An fuller obituary will be published in the New York Times on Tuesday, November 30, 1993. This is the issue of the Times that includes the Science Times section. Betty feels Don would want this.

Betty has asked me to please spread the word that donations to charities be sent instead to the Don and Betty Walker Student Fund. You can send them

Association for Computational Linguistics
c/o Judith Klavans
Box 105
Hastings-on-Hudson, New York 10706.

Credit card payment will be accepted. Be sure to include the expiration date of your card as well as the number (Betty has warned me that many people forget the expiration date.) You will receive an acknowledgement, and Betty will be informed as well.

Personal notes for Betty can be sent to:

Betty Walker
36 Oak Place
Bernardsville, New Jersey 07924

In this time of sorrow, all personal notes are well-appreciated, especially with personal anecdotes. My own experience is that this comforts by keeping the memories alive.

Betty will continue to stay active with us at the ACL for as long as she wants. Her presence will, as always, be much valued.
Please feel free to forward this message.

Peace, Judith

Subject: Memorial Service for Donald E. Walker
A memorial service for Donald E. Walker, Director of Language and Knowledge Resources Research at Bellcore, Secretary- Treasurer of the Association for Computational Linguistics and Secretary-Treasurer of JJCAll, will be held on Sunday December 19th from two to four PM at the Unitarian Fellowship Church in Summit, 4 Waldron Avenue, Summit, New Jersey. Don passed away peacefully with his wife Betty and all three daughters at his side on Friday November 26, 1993, after a long battle with cancer.

Friends and colleagues of Don's who are unable to attend the service might want to send written reminiscences, photographs or other materials remembering him, for a booklet to be presented to his family. Please send these to:

Kathy McKeown
20 Prospect Rd.
Wayne, New Jersey
07470 USA

In addition, any charitable contributions in his memory may be sent to:

ACL Don and Betty Walker
International Student Fund
Association for Computational
Linguistics

c/o Judith Klavans
Box 105
Hastings-on-Hudson, New York 10706

Please forward this message. Thank you,

Fernando Pereira, ACL President

Ernst Schröder Zentrum für Begriffliche Wissensverarbeitung begründet

Am 3. Dezember 1993 stellte sich das Ernst Schröder Zentrum für Begriffliche Wissensverarbeitung an der TH Darmstadt im Rahmen eines Kolloquiums erstmals der Öffentlichkeit vor.

Den Kolloquiumsvortrag im bis auf den letzten Platz gefüllten Hörsaal hielt Prof. Dr. Karl-Otto Apel über "Diskursethik und Semiotik".

Ziele und Aufgaben des Zentrums wurden vom Präsidenten der TH Darmstadt, Prof. Dr. H. Böhme und dem Vorsitzenden des Zentrums, Prof. Dr. R. Wille vorgestellt.

Das *Ernst Schröder Zentrum für Begriffliche Wissensverarbeitung e.V.* fördert Ausbildung, Forschung, Entwicklung und Anwendung auf dem Gebiet der Begrifflichen Wissensverarbeitung. Dazu werden vom Zentrum Seminare, Tagungen, sowie Aus- und Fortbildungsseminare veranstaltet. Grundsätzlich geht es dem Zentrum um kritische Bestandsaufnahme, Entwicklung und Vermittlung von Ergebnissen, Methoden, Verfahren und Programmen der Begrifflichen Wissensverarbeitung.

Im Ernst Schröder Zentrum haben sich Human- und Sozialwissenschaftler, Mathematiker, Informatiker und Informationswissenschaftler zusammengefunden. Sie wollen einem drohenden Abbau kognitiver Autonomie durch Wissens- und Informationssysteme, die vom Menschen nicht mehr kontrollierbar sind, entgegenwirken. Sie befürworten deshalb Methoden und Instrumente Begrifflicher Wissensverarbeitung, die Menschen im rationalen Denken, Urteilen und Handeln unterstützen und den kritischen Diskurs fördern.

Der LDV-Forum-Leserln ist das Thema "Begriffsanalyse" (zumindest noch) aus LDV-Forum 198711 bekannt:

> Kipke, U.; Wille, R.: "Formale Begriffsanalyse - erläutert an einem Wortfeld", S. 31-36.

Anschrift des Zentrums:
Ernst Schröder Zentrum
für Begriffliche Wissensverarbeitung e.V.
(THD, FB 4)
Schloßgartenstraße 7
64289 Darmstadt

"KODIERUNG UND NORMUNG MASCHINENLESBARER TEXTE" - BERICHT AUS DEM GLDV-ARBEITSKREIS

Peter Scherber
Göttingen

Zum Arbeitskreis

Dieser Arbeitskreis wurde im Herbst 1991 auf der GLDV-Tagung in Trier gegründet. Er verstand sich als ein lokales deutschsprachiges Forum für die weltweite Text Encoding Initiative (TEI), die damals gerade den ersten Teil ihrer Arbeit mit der Herausgabe von "Guidelines" abgeschlossen hatte1. Da die zweite Phase des TEI-Projekts zeitlich sehr knapp kalkuliert worden war, - man plante, im Sommer 1992 sowohl die 2. Projektphase (P2) zum Abschluß zu bringen und direkt anschließend daran die endgültigen Guidelines als P3 zu veröffentlichen - glaubten wir damals, der Arbeitskreis werde schon sehr bald vor allem mit der Evaluierung und praktischen Ausfüllung dieses neuen Normwerkes vollauf zu tun haben. Doch es ist dann anders gekommen.

Zwar haben wir von Anfang an die einzelnen Publikationen der P2-Phase mitgehalten und kurz nach Erscheinen auch auf dem Göttinger Listserver bereitgestellt, doch schwand auch mit dem geduldigen Warten auf die einzelnen Teile des Normenwerks die Bereitschaft zu ungeduldiger Aktivität im Arbeitskreis. Hinzu kam, daß auch die Motivation der AK - Teilnehmer, das angebotene elektronische Diskussionsforum zu nutzen, äußerst gering war. Man kann das Interesse der Teilnehmer am Arbeitskreis sicher ganz zutreffend als überwiegend rezeptiv charakterisieren, d. h. es besteht zwar ein breites und intensives Interesse daran, diese Entwicklung im Auge zu behalten, aber bevor die Phase P2 abgeschlossen ist, wartet man ab bzw. widmet man sich den Dingen des wissenschaftlichen Alltags, die keinen Aufschub gestatten.

Der Zeitverzug um mittlerweile eineinhalb Jahre und auch das Anwachsen der ursprünglich noch übersichtlichen Guidelines auf das mittlerweile zu erwartende Achtfache ihres Umfanges in der Phase P2 hat uns bewogen, Zielsetzung und Aufgaben des Arbeitskreises neu zu überdenken.

Nachdem auf der GLDV-Tagung in Kiel im März 1993 das Interesse am Arbeitskreis erneut bestätigt worden ist, trafen wir uns am 22. Oktober in Göttingen zu einem relativ kleinen Treffen, bei dem sich zeigte, daß das Vorhaben der TEI mittlerweile auch durchaus skeptisch gesehen werden muß. Hierzu soll weiter unten noch einiges berichtet werden.

Was ist die Text Encoding Initiative?

Seit 1987 haben sich in der Text Encoding Initiative, an der auch mehrere

1 Guidelines For the Encoding and Interchange of Machine-readable Texts TEI PI, hg. von C. M. Sperberg-McQueen u. Lou Burnard, Chicago u. Oxford 1990/91,289 S.; Träger der 1987 gegründeten Initiative waren die drei Organisationen ACH, ACL und ALLC.

deutsche Wissenschaftler beteiligt waren, zahlreiche Kommissionen mit der Erstellung von Kodierrichtlinien (Guidelines) beschäftigt. Dabei ist schon in einer sehr frühen Phase die Entscheidung gefallen, die bestehende ISO-Norm 8879 (SGMLY zur Grundlage der Verhandlungen zu machen. Daraus resultierte eine zweigeteilte Aufgabenstellung. Es mußten einerseits Kodierrichtlinien entwickelt werden, die es gestatteten, möglichst jede vorhandene Textsorte kontrolliert in maschinenlesbare Form zu überführen, und es waren andererseits SGML-konforme Dokument-Typdefinitionen (die sogenannten DTDs) zu entwickeln, die einen weltweiten Austausch derartiger Dokumente zu garantieren im Stande waren.

Zweckbestimmung dieser Richtlinien waren neben dem Austausch und damit der möglichen Wiederverwertung einmal erfaßter Ressourcen außerdem die Unterstützung von applikations-unabhängiger Dokumenterstellung und eine Auszeichnung der Texte nach standardisierten Regeln, die es ermöglichen sollten, so gut wie alle relevanten Textmerkmale der Analyse zugänglich zu machen 3.

Es hat sich gezeigt, daß die Guidelines von 1990/91 auch quantitativ nur ein erster Schritt waren. Die Phase P2, die noch andauert, hat sowohl den ersten Entwurf der Guidelines, als auch die für die erste Phase gültigen DTDs einer gründlichen Revision unterzogen. Der Umfang aller 42 Kapitel von P2 wird voraussichtlich einen Umfang von über 2000 Seiten besitzen.

2 International Organization for Standardization, ISO 8879: Information processing - Text and office systems - Standard Generalized Markup Language (SGML), 1986.

3 ausführlicher dazu vgl. Winfried Lenders: Fragen der Standardisierung, in: Computereinsatz in der Angewandten Linguistik, hg. v. W. Lenders, Frankfurt u. a. 1993, S. 63-74.

Derzeitiger Stand des P2-Projekts

In der definitiven und expliziten Form eines P2-Dokuments4 sind bislang nur die ersten beiden (von acht) Teilen erschienen. Der einleitende Teil I enthält eine kompakte Einführung in die SGML-Philosophie und eine Beschreibung der TEI-DTDs. Teil II (Core Tags and General Rules) äußert sich zu Zeichen und Zeichensätzen, beschreibt den TEI-Reader, in dem Informationen über das erfaßte Dokument mitgeführt werden können und enthält diejenigen Kodierelemente (Tags), die allen TEI-konformen Texten gemeinsam sein sollen. In diesem Zusammenhang wird eine Standard- Textstruktur konstituiert, die sozusagen die gemeinsamen Merkmale aller zu erfassenden Textsorten enthält.

Die Teile III und VI, die aus den Kodierrichtlinien (Tag sets) für alle textsortenspezifischen Merkmale bestehen, sind noch immer lückenhaft. Bei den Teilen V bis VIII sind bislang erst vier von 14 Kapiteln erschienen.

Lou Burnard, der Mitherausgeber der Guidelines hat im Juli 1993 festgestellt, daß damals ca. 25 % des für P2 vorgesehenen Materials erschienen sei, dies läßt den Schluß zu, daß wir heute, d. h. Ende 1993 ca. 40 - 50 % erreicht haben, und uns wohl noch auf eine längere Durststrecke einzustellen haben.

Das Treffen des Arbeitskreises am 22. Oktober

Am 22. Oktober 1993 trafen sich in Göttingen 9 Teilnehmer (incl. der beiden Moderatoren) und diskutierten die aktuelle Situation des Arbeitskreises. Nach Berichten der beiden Moderatoren zu Erfahrungen mit zwei SGML-Parsern *(G.* Koch) und zum Stand des P2-Projekts (P. Scherber)

4 Die Phase 3 (P3) läßt, im Gegensatz zu P2, nur mehr geringfügige redaktionelle Veränderungen des Normenwerkes zu, so daß man davon ausgehen kann, daß P2 tatsächlich den definitiven Status der Guidelines widerspiegelt.

referierte aus seiner Arbeit Martin Volk (Universität Koblenz-Landau), der ein Korpus von deutschen Sätzen (zum Einsatz für Zwecke sowohl in der Forschung als auch in der Lehre) mit SGML-Markierung aufgebaut hat. Bruno Schulze aus Stuttgart (Institut für maschinelle Sprachverarbeitung, Projekt Textkorpora und Erschließungswerkzeuge) berichtete von den dortigen Arbeiten zu Feature-Strukturen. Bei beiden Referenten wurde offensichtlich, daß die Arbeit der TEI, wenn sie überhaupt "ankommen" will und nicht schon zu spät kommt, noch eine geraume Weile als konkurrierendes System gesehen werden muß, das sich im Wettbewerb mit anderen Taggingsystemen erst noch bewähren muß. Insbesondere der bisherige erfolgreiche Einsatz SGML-konformer, speziell auf eine bestimmte Anwendung optimierter Systeme, die sich in der Praxis bewährt haben, wird eine skeptische und abwartende Haltung gegenüber dem TEI-Normenwerk zur Folge haben.

Dies war auch das Fazit der anschließenden Diskussion, die sich mit der Thematik eines nach dem Abschluß der P2-Phase vorgesehenen Status-Workshops auseinandersetzte. Es wurde festgestellt, daß ein derartiges Unternehmen sich unter dem Arbeitstitel: "Standardisierung bei der Erfassung maschinenlesbarer Texte" vor allem drei Aufgaben widmen sollte

1. dem Für und Wider der TEI-Guidelines als Kodiervorschrift

2. Erfahrungsberichten von Anwendern

3. den Softwareinstrumenten (Tools, Parser, Editoren).

Was ist zu tun, wie geht es weiter

Auch in den mehr als zwei Monaten seit Oktober ist der Stand des P2-Projekts noch fast unverändert geblieben[5]. Hinzu traten organisatorische Probleme, mit denen die beiden Moderatoren des Arbeitskreises konfrontiert wurden und die es erforderlich machten, den auf dem Treffen diskutierten Termin eines Status-Workshop zu den TEI-Guidelines im Mai 1994 vorerst fallen zu lassen.

In den nächsten Wochen werden wir zu allererst die Distribution der TEI-Dokumente auf eine neue Basis stellen. Dem Trend der Zeit folgend werden wir die Diskussionsliste auf anderer Plattform, aber mit demselben bewährten Namen , 'MARKUP-L" weiterführen und die Verteilung der Dokumente und DTD-Dateien über FTP vornehmen. Wir hoffen, daß diese Arbeiten Ende Januar 1994 abgeschlossen sein werden.

Erfahrungen und Kenntnisse über Softwareprodukte (Parser, Editoren, Tools usw.), die die Arbeit mit SGML unterstützen, sind immer noch sehr dünn gesät, dies liegt nicht zuletzt auch daran, daß für die meisten dieser Produkte nur kommerziell kalkulierte Preise verlangt werden. Aus diesem Grunde möchten wir an dieser Stelle an diejenigen appellieren, die praktische Erfahrung mit derartigen SGML-Produkten haben, sich uns anzuschließen und/oder auf einem der nächsten AK- Treffen zu berichten.

Einige Ideen, wie man das umfangreiche Material, das uns durch die P2Dokumentation geboten wird, für den praktischen Gebrauch des "Corpusarbeiters vor Ort" erschließen kann, wurden auf dem vergangenen Treffen diskutiert. Dazu gehört das Projekt eines Leitfadens (dies ist vorerst nur der Arbeitstitel: ein Leitfaden für die Guidelines!), der es dem Wissenschaftler entbehrlich machen soll, vor der Erfassung eines Corpus mittlerer Reichweite und Größe zuerst das gesamte Normwerk von

[5]im Dezember 1993 erschienen endlich die Kodierrichtlinien für Verstexte, so daß wenigstens die TEI-konforme Erfassung von allen drei großen belletristischen Textsorten abgeschlossen ist.

dann vielleicht 2000 Seiten durchzustudieren.

Diese eben genannten Themen und die letzthin aufgeworfene "Gretchenfrage": sollte man das ganze TEI-Projekt bereits jetzt schon als vorerst gescheitert betrachten und nach anderen Lösungswegen (natürlich auf der Basis der rundum akzeptierten SGML-Norm) suchen, werden uns auch weiter beschäftigen, vielleicht sogar Zündstoff bieten für die weitere Diskussion. Wir beabsichtigen deshalb, für 1994 mindestens zwei Treffen vorzusehen, eines im April oder Mai und eines im Zusammenhang mit der GLDV-Tagung auf der KONYENS in Wien (28.-30.9.1994)

Zum Abschluß möchte ich diesen Bericht auch verbinden mit dem Aufruf zur tätigen Mitarbeit an unserem Arbeitskreis, besonders wertvoll wären uns natürlich Kollegen, die bereits mit den TEI-Richtlinien oder anderen SGML-konformen Kodierungen arbeiten und darüber berichten könnten.

Bitte wenden Sie sich an:

Peter Scherber
bzw. Günter Koch GWDG
Am Faßberg
D-37077 Göttingen Tel.
0551/201559 bzw.
0551/201550
FAX: 0551/21119
e-mail: pscherb@gwdg.de
bzw. gkoch@gwdg.de

Das KONTEXT–Textanalysesystem

Karin Haenelt
IPSI, Darmstadt

Am Institut für Integrierte Publikations- und Informationssysteme (IPSI) der GMD, Darmstadt wurde unter Leitung von Frau Dr. Karin Haenelt das KONTEXT–Textanalysesystem konzipiert und entwickelt, das u.a. auch auf der Buchmesse 93 (Frankfurt) vorgestellt wurde.

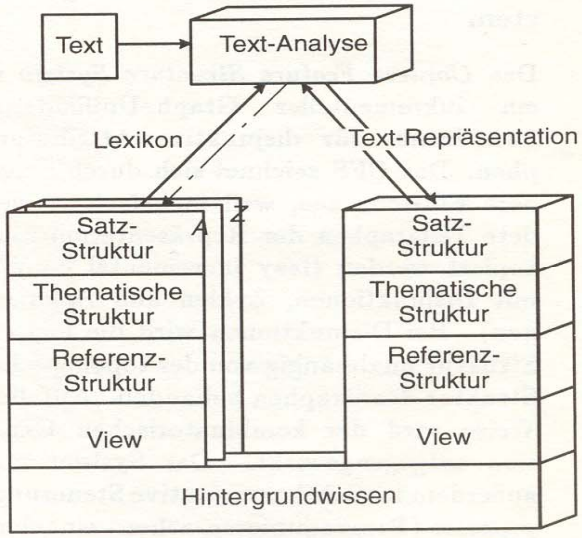
Das KONTEXT–System ist ein erster Prototyp einer neuen Generation von Textanalysesystemen. Grundlage des Systems ist das im IPSI neu entwickelte KONTEXT–Textmodell, das es ermöglicht, Texte in völlig neuartiger Weise textorientiert (nicht auf der Basis isolierter Wörter oder einzelner Sätze) unter Berücksichtigung von Textstruktur und Kontext zu verarbeiten. Dank dieses Modells können in der Textanalyse auch in Texten vorkommende neue Informationen erkannt und integriert werden.

Wissenschaftlicher Hintergrund

a) Das KONTEXT–Modell

Das KONTEXT–Textmodell beschreibt, wie mit Mitteln der natürlichen Sprache (neues) Wissen aufgebaut und mitgeteilt wird. Die Grundannahmen des Modells sind

- die Information, die in einem Text vermittelt wird, sowie die Information, die die kontextuelle Organisation dieser Information beschreibt, kann in fünf Ebenen strukturiert werden: Satzstruktur, Thematische Struktur, Referenzstruktur, View (textspezifische Sicht auf Wissen), Hintergrundwissen.
- der Grundmechanismus der Konzeptkonstruktion ist der Diskurs. Diskurs ist definiert als eine Sequenz von Zustandsübergängen zwischen Diskursstates, und Diskursstates sind definiert durch die Information, die in den fünf Ebenen repräsentiert ist.



Der Textanalyseprozeß konstruiert und durchläuft die Ebenen unter der Kon-

trolle der Diskursentwicklung. Auf diese Weise kann er inkrementell die text spezifische Sicht (View) auf Hintergrundwissen aufbauen.

b) Die Textgrammatik

Eine lexikalisierte Textgrammatik beschreibt den Beitrag, den sprachliche Ausdrücke zum Aufbau der Schichten des KONTEXT-Modells leisten. In der Entwicklung dieser Textgrammatik wird u. a. auch der Zusammenhang zwischen Tiefenkasus (semantischen Rollen) und syntaktischen Funktionen (Satzteile, Kasus) systematisiert. Alle Schichten (also nicht nur die Syntax, sondern auch die Wissensrepräsentation) werden in einem einheitlichen Formalismus, dem *Context Feature Structure System* (CFS) repräsentiert. Das Analyseverfahren erzeugt schrittweise lesend eine linguistische und konzeptuell angereicherte Textrepräsentation. Im Sinne des Integrierten Parsing werden alle Beschreibungsebenen des Modells zum Zwecke der wechselseitigen Ergänzung und Einschränkung gleichzeitig konstruiert.

c) Das Context Feature Structure System

Das *Context Feature Structure System* ist ein inkrementeller Graph-Unifikationsformalismus für disjunktive Attributgraphen. Das CFS zeichnet sich durch besondere Effizienz aus, weil mehrfach verwendete Teilgraphen der Repräsentation nicht kopiert werden (lazy incremental copying mit Disjunktionen, Zyklen und Rekursionen). Bei Disjunktionen wird die logische Struktur unabhängig von der topologischen Struktur des Graphen behandelt. Auf diese Weise wird der kombinatorischen Explosion entgegengewirkt. Das System kann außerdem in nichtkommutative Steuerungsprozesse (Programmiersprachen) eingebunden werden, ohne daß die jeweiligen Datenstrukturen vollständig kopiert werden müssen.

Eine Lexikon- und Grammatikentwicklungsumgebung steht zur Verfügung.

Anwendungsgebiete

Der Vorteil der textmodellbasierten Verarbeitung natürlichsprachiger Texte (wie sie in KONTEXT realisiert ist) besteht darin, daß *ein und dasselbe Prinzip* zu einer *Vielzahl unterschiedlicher Anwendungen* führt. Kern ist die Textanalysekomponente. Sie bleibt für verschiedene Anwendungskontexte in ihrer Funktionalität jeweils die gleiche.

Die Textanalysekomponente baut eine mehrschichtige Textrepräsentation auf. Eine dieser Schichten enthält die aus Texten gewonnenen Fakten, die anderen enthalten die Textstruktur. Die Textrepräsentation erfüllt mehrere Funktionen:

=> sie ermöglicht einen direkten inhaltsorientierten Zugriff auf Volltexte, wobei der Zugriff durch Textinhalt und Textstruktur bestimmt wird (Textretrieval++);

=> sie kann mit Methoden des Faktenretrieval ausgewertet werden (Information Retrieval++);

=> sie kann als Thesaurus ausgewertet werden;

=> sie dient der automatischen Indizierung und Verschlagwortung von Texten;

=> sie dient als Grundlage Syntax- und Konzept-basierter Rechtschreibprüfungen;

=> sie bildet eine abstrakte Schnittstelle zu Texten und liefert so die Grundlage für weitere textuelle Operationen, wie automatische Paraphrasengenerierung (Lexikonaufbau), Maschinelle Übersetzung, Kondensierungen (Inhaltsangaben, Zusammenfassungen), Dialog-Anwendungen oder Aufbau, Analyse und Linearisierung von Hypertexten.

Ein weiterer Vorteil der textmodellbasierten Verarbeitung natürlichsprachiger Texte in KONTEXT besteht darin, daß auch *neues Wissen* systematisch behandelt werden kann. So lassen sich solche Systemkomponenten auch als „lernende Wörterbücher" und „lernende Faktendatenbanken" einsetzen, die neue Veröffentlichungen lesen, das in Texten mitgeteilte neue Wissen erkennen, es in ihre Textrepräsentation übernehmen und sich so selbst auf dem laufenden halten.

Kontaktadresse

Dr. Karin Haenelt
Institut für Integrierte Publikations- und Informationssysteme (IPSI)
Dolivostraße 15
6100 Darmstadt, Tel.: 869–828, e–mail: haenelt@darmstadt.gmd.de

AUTOMATISCHE INHALTSERSCHLIESSUNG STRUKTURIERTER DOKUMENTE

Projekt an der FH Darmstadt
Fachbereich lud und Informatik
unter Beteiligung der G MD (IPSI) und der THD

Es wird zunehmend erkannt, daß es für Unternehmen von strategischer Bedeutung sein kann, Dokumente nicht nur als fortlaufenden Text zu betrachten, sondern als Objekte mit einer wohldefinierten logischen Struktur. Darüber hinaus ist es häufig für eine qualitativ befriedigende Lösung von Informationsproblemen notwendig oder wünschenswert, Dokumente auch inhaltlich aufzubereiten und im Hinblick auf ein gezielt es Auffinden mit Informationen anzureichern. Traditionell geht es dabei um die manuelle Zuordnung von Klassifikationscodes oder Deskriptoren - eine Lösung, die vielfach aus wirtschaftlichen oder grundsätzlichen Gründen nicht ernsthaft in Frage kommt. Darüber hinaus stellt sich im Kontext neuerer Systemarchitekturen (Hypertext) das Problem der inhaltlichen Erschließung in weitaus komplexerer Weise. 1

Im Rahmen des Projektes der Fachhochschule Darmstadt (93 - 95) unter Beteiligung des IPSI-Institutes der GMD und der TH Darmstadt geht es darum, ein Werkzeug zu konzipieren, zu implementieren und anhand zweier konkreter Anwendungsfälle zu erproben, mit dem sich Aufgaben der inhaltlichen Erschließung spezifizieren und automatisieren lassen.

Die Arbeiten basieren auf der Konzeption des AIR/X-Indexierungssystems, den in der GMD entwickelten Werk zeugen SFK (Smalltalk Frame Kit) und Dream (Parser zur Konvertierung nach SGML).

Der Schritt vom "reinen" Textdokument zu einer strukturell reichhaltigen Repräsentation als SGML-Text über die Zwischenschritte *Document Type Definition (DTD)* und *Document Structure Definition* unter Verwendung des Dream-Parsers ist mittlerweile gut untersucht und im Griff. Gegenwärtig wird damit begonnen, die entwickelten Werkzeuge zur morphologischen und syntaktischen Analyse an die Randbedingungen des SGML-Formates anzupassen. Über die Effektivität spezieller Regelklassen für eine weitergehende inhaltliche Klassifikation liegen erste Ergebnisse vor. Es ist mittlerweile eine PC-Datenbank aufgebaut (und dokumentiert), in der manuell klassifizierte Textstellen mit ihrem linken und rechten Kontext abgelegt sind. In dieser Datenbank werden die Ergebnisse der Anwendung inhaltlicher Regeln abgespeichert und ausgewertet.

Die entscheidenden Implementierungsarbeiten, deren Ergebnis ein vielseitig einsetzbares Werkzeug zur automatischen inhaltlichen Erschließung sein wird, sollen 1994 abgeschlossen werden.

1 Hier geht es darum, innerhalb eines Dokumentes oder zwischen Dokumenten einzelne Dokumentteile inhaltlich zu verknüpfen. Neu ist das Problem zu entscheiden, wie denn überhaupt Quelle und Ziel einer solchen Verknüpfung im Einzelfall zu definieren sind: als Nominalphrase, als Satz, als Abschnitt, **...**

Kontakt: *G. Knorz,* FHD

AUFGABEN DER COMPUTERLINGUISTIK

Roland Hausser
Computerlinguistik Universität
Erlangen-Nürnberg

Seit der ersten Entwicklung von elektronischen Rechenanlagen in den vierziger Jahren des 20. Jahrhunderts unterscheidet man zwischen der *numerischen* und der *nicht-numerischen* Informatik. Die numerische Informatik befaßt sich mit der Berechnung von Zahlen und hat in der Physik, der Chemie, der Wirtschaftswissenschaft, der Soziologie etc., zu einer explosionsartigen Wissensexpansion geführt. Aber auch in vielen praktischen Bereichen, wie dem Bankwesen, dem Flugverkehr, der' Lagerhaltung, etc., sind numerische Anwendungen heute nicht mehr wegzudenken. Ohne Computer und ihre Software würde die Funktionsfähigkeit dieser Bereiche zusammenbrechen.

Die nicht-numerische Informatik, andererseits, befaßt sich mit den Phänomenen der Wahrnehmung und der Kognition. Die theoretische und praktische Entwicklung der nicht-numerischen Informatik blieb, trotz hoffnungsvoller Anfänge, weit hinter denen der numerischen Informatik zurück. In neuerer Zeit findet die nichtnumerische Informatik jedoch als KOGNITIONSWISSENSCHAFT und KÜNSTLICHE INTELLIGENZ wieder starkes Interesse. Die *"cognitive science,* " und *"artificial intelligenceJ"* wie sie im Amerikanischen genannt werden, untersuchen die menschliche Informationsverarbeitung unter Einbeziehung der Informatik, der Psychologie, der Linguistik, der Philosophie und der mathematischen Logik.

1 Methoden und Anwendungen

In der Computerlinguistik werden die sehr unterschiedlichen Methoden der theoretischen Linguistik, der Psychologie, der Philosophie und der mathematischen Logik dadurch auf einen Nenner gebracht, daß man theoretische Hypothesen systematisch als Computerprogramme implementiert. Dieses Entwickeln und Testen von theoretischen Hypothesen auf dem Computer bietet die Möglichkeit einer neuen, vereinheitlichten Methodologie und Theoriebildung, die sich deutlich von der der traditionellen Linguistik, Psychologie, Philosophie und mathematischen Logik unterscheidet.

Theoretisch wird die Entwicklung in der Computerlinguistik durch eine neuartige Methodologie (Überprüfung formaler Grammatiksysteme durch den systematischen Einsatz von Computern) und prak-

tisch durch einen ungeheuren Bedarf an einer effizienten automatischen Sprachverarbeitung angetrieben. Im wissenschaftlichen Alltag der modernen Sprachwissenschaft wirkt sich der systematische Einsatz von Computern folgendermaßen aus:

1.1 Methodische Auswirkungen des Programmierens

=> Die Programmierung, z.B. einer Grammatik als Parser, erfordert eine viel detailliertere Analyse des zu behandelnden Phänomens als vorher üblich.

=> Die unterschiedliche Eignung grammatikalischer Formalismen für die Programmierung wird zu einem neuen, wichtigen Faktor im Wettbewerb konkurrierender Theorien.

=> Effizient implementierbare (und implementierte) Formalismen der automatischen Sprachanalyse haben praktische Anwendungen, die eine völlig neue Eigendynamik in die geisteswissenschaftliche Empirie bringen.

In folgenden Bereichen der praktischen Anwendung werden die Methoden der Computerlinguistik in stark zunehmendem Maße eingesetzt:

1.2 Praktische Aufgaben der Computerlinguistik

=> Indexierung von und Abruf aus textuellen Datenbanken

Textuelle Datenbanken speichern Texte in elektronischer Form, zum Beispiel die Jahrgänge einer Tageszeitung, die Publikationen einer Medizinischen Zeitschrift oder sämtliche Gerichtsurteile in den USA seit 1960. Der Benutzer einer solchen Datenbank muß in der Lage sein, all diejenigen Dokumente und Textstellen zu finden, die für seine spezifische Fragestellung relevant sind.

=> Automatisierte Textproduktion

Große Firmen, die ständig neue Produkte wie Motoren, Pumpen, Fernseher etc., herausbringen, müssen hierfür immer wieder neue Produktbeschreibungen und Wartungsmanuale herstellen. Ähnliches gilt für Rechtsanwalt- und Steuerkanzleien, Personalabteilungen, etc., die ein sehr hohes Korrespondenzvolumen haben, wobei sich die Briefe nur an klar definierten Stellen unterscheiden. Die Methoden der automatisierten Textproduktion reichen von einfachen Schablonen zu hochflexiblen und interaktiven Systemen, die auf linguistischem Wissen basieren.

=> Automatische Textüberprüfung

Auch auf diesem Gebiet reichen die Anwendungen von einfachen Orthographie-Checkern auf der Grundlage von Wortformlisten über Morphologiesysteme, die die sprachliche Vielfalt im Bereich der Wortbildung systematisch behandeln, bis zu Syntax-Checkern, die Fehler in Wortstellung, Kongruenz etc. finden können.

=> Automatische Inhaltsanalyse

Es heißt, daß sich die gedruckte Information auf der Erde alle 10 Jahre verdoppelt. Auch auf wissenschaftlichen, rechtlichen, wirtschaftlichen etc., Spezialgebieten ist die Flut der relevanten Literatur so groß, daß die Lebenszeit der Mitarbeiter einfach nicht mehr ausreicht, um ständig auf dem neusten Stand zu sein. Eine zuverlässige automatische Inhaltsanalyse mit kurzen Zusammenfassungen wäre hier von größtem Nutzen. Die automatische Inhaltsanalyse ist auch die Voraussetzung einer *konzeptbasierten Indizierung*, wie sie für den optimalen Abruf aus textuellen Datenbanken notwendig ist, sowie die Vorraussetzung für eine wirklich leistungsfähige maschinelle Übersetzung (s. u.).

=> Maschinelle Übersetzung

Die maschinelle Übersetzung war eine der Hauptanwendungen in den Anfängen der nicht-numerischen Informatik. In der Dekade von 1955 bis 1965 wurde auf diesem Gebiet intensiv geforscht, wobei es in der Öffentlichkeit große Aufmerksamkeit fand. Diese Erwartungen erfüllten sich jedoch nicht, und die Hoffnungen auf kommerzielle Erfolge zerschlugen sich.

Inzwischen ist das Interesse wieder stark gestiegen. HUTCHINS 1986 nennt folgende Gründe für die fortgesetzten Bemühungen um eine maschinelle Übersetzung:

- Wissenschaftler, Techniker, Ingenieure, Manager und viele andere Geschäftsleute müssen täglich viele Briefe und Dokumente in Sprachen lesen und schreiben, die sie nicht beherrschen... Es gibt einfach nicht genug Übersetzer, um mit dieser ständig wachsenden Menge an Material fertig zu werden.

- Viele Forscher betreiben die Entwicklung der maschinellen Übersetzung aus Idealismus. Sie wollen die internationale Kooperation und den Frieden fördern, indem sie Sprachbarrieren überwinden und die Verbreitung technischen, landwirtschaftlichen und medizinischen Wissens in die Entwicklungsländer fördern.

- Die maschinelle Übersetzung wird aber auch von Institutionen gefördert, die Anwendungen im Militärbereich sehen, also z.B. die schnelle Übersetzung gegnerischer Dokumente.

- Als Problem der reinen Forschung stellt maschinelle Übersetzung ein schwieriges Problem dar, dessen Lösung von manchen Forschern als Test ihrer linguistischen Arbeit betrachtet wird.

- Schließlich konstituieren leistungsfähige Systeme der automatischen Übersetzung wertvolle Software-Produkte mit vielfältigen Anwendungen, die entwickelt werden, weil man damit viel Geld verdienen kann.

Gerade in der Europäischen Gemeinschaft, wo derzeit in 12 Mitgliedsländern 9 verschiedene Sprachen gesprochen werden, ist der potentielle Nutzen von automatischen oder auch nur halb automatischen Übersetzungssystemen unübersehbar.

> Automatisierter Unterricht

Es gibt zahlreiche Unterrichtsfächer, in denen viel Zeit für sogenannte Drillübungen verwendet wird, z.B. die mehr oder weniger mechanische Erlernung von regelmäßigen und unregelmäßigen Paradigmen im Sprachunterricht. Diese können mindestens ebenso gut am Computer durchgeführt werden, wodurch dem Lehrer mehr Zeit für andere Aktivitäten, z.B. Konversation, bleibt. In der neueren Forschung beschäftigt man sich intensiv mit weitergehenden Automatisierungen des Unterrichts, zum Beispiel der automatischen Fehleranalyse bei Übersetzungsübungen.

Automatisierte Unterrichtssysteme haben den zusätzlichen Vorteil, daß automatisch über die Interaktionen zwischen Schüler und Computer buchgeführt werden kann. Durch das Wissen darüber, wo die Schüler am meisten Fehler machen und wo die meiste Zeit verbracht wird, erhält man dann eine wertvolle Heuristik für die Verbesserung der Ergonometrie des automatisierten Unterrichts. Dies hat eine Entwicklung vom 'elektronischen Textbuch' alter Prägung zu

neuartigen Lehrprogrammen mit einer eigenständigen, medium-gerechten Pädagogik initiiert.

=> Dialogsysteme und automatische Auskunft

Ein wesentlicher Engpaß bei der Interaktion mit Computern ist die Tatsache, daß die Interaktion entweder aufwendig (z.B. selbstgeschriebene Programme) oder unflexibel (z.B. Menuegesteuerte Interaktion) ist. Deshalb besteht großes Interesse an der Entwicklung robuster, natürlichsprachlicher Systeme, die bei praktisch allen Interaktionen zwischen Mensch und Maschine zum Einsatz gebracht werden können.

Die Zahl der möglichen Anwendungen der Computerlinguistik ist damit keineswegs abgeschlossen. Sie umfaßt vielmehr ganz allgemein sämtliche Bereiche, in denen Menschen (heute und in Zukunft) mit Computern umgehen. Bei all diesen Anwendungen dienen die Erkenntnisse der Sprachwissenschaften, zumindest potentiell, der Optimierung der automatischen Sprachverarbeitung. Umgekehrt spielen die Computer als Hilfsmittel der linguistischen Analyse und Theoriebildung eine immer größere Rolle.

2 Übertragung in das elektronische Medium

Damit natürliche Sprache auf dem Computer automatisch analysiert und verarbeitet werden kann, muß sie als elektronisch repräsentierte Buchstabenfolge gespeichert sein. Texte, die in dieser Form auf dem Computer gespeichert sind, nennt man auch *on-line* Texte.

Texte und Ausdrücke natürlicher Sprachen existieren jedoch meist in verschiedenen nicht-elektronischen Medien: als *Lautzeichen* der gesprochenen Sprache, als *Buchstaben* der geschriebenen oder gedruckten Sprache oder als die *Gesten* einer Taubstummensprache. Während Lautzeichen und Gesten normalerweise nur eine sehr kurze Lebensdauer haben (von Tonoder Videoaufnahmen einmal abgesehen), zeichnet sich die Schrift, konventionell fixiert auf Papier, Pergament oder Stein, durch ihre überdurchschnittlich lange Haltbarkeit aus.

Im Gegensatz zur konventionellen Speicherung von Schriftsprache, werden im elektronischen Medium Magnetband, Diskette oder CD-ROM als Datenträger verwendet. Die Übertragung nicht-elektronisch gespeicherter Sprache in das elektronische Medium ist aufwendig und kann in verschiedener Weise erfolgen.

Eine Möglichkeit ist das Eintippen gesprochener oder geschriebener Sprache (etwa durch eine Sekretärin) in den Computer. Dies ist heute noch eine weit verbreitete Methode, z.B. das Tippen vom Diktiergerät in Büros, das Transskribieren von Tonbandaufnahmen in der Psychologie, oder das Eingeben von Büchern, die bisher nur in gedruckter Form vorlagen.

Hinzu kommen heute Technik-basierte Methoden. Die automatische Überführung von (auf Papier) gedruckter Sprache in das elektronische Medium fällt in den Bereich der *optischen Mustererkennung* und wird mit Hilfe sogenannter *Scanner* vorgenommen. Diese Maschinen machen nicht nur ein Abbild der Seite, wie es auch eine Fotografie tun würde, sondern sie tasten die einzelnen Buchstaben zeilenweise ab und vergleichen sie mit gespeicherten Mustern. Auf diese Weise wird das Druckbild nicht nur in den Computer abgebildet (als sogenannte *bitmap),* sondern buchstabenweise erkannt. Dies ist der Prozess der *optical character recognition* oder OCR.

Nun kann sich das Druckbild von einem Buch zum nächsten sehr stark unterscheiden. Hinzu kommen unterschiedliche Buchstabengrößen und Formatierungen wie bei Überschriften, Fußnoten, Bildunterschriften oder Tabellen. Dies bewältigen moderne Scanner mit Hilfe einer initialen

Lernphase, in der der Benutzer Fehlklassifikationen korrigieren kann, indem er dem Programm eingibt, ob es sich bei einem bestimmten Buchstaben z.B. um ein 'd' oder um ein 'a' handelt.

Zusätzlich verwenden Hochleistungs-Scanner große Lexika, mit deren Hilfe sie in Zweifelsfällen entscheiden, welche von zwei Möglichkeiten eine sinnvolle Wort form darstellt. Auf diese Weise kann man, abhängig vom verwendeten Schrifttyp und der Qualität des Schriftbilds, eine Erkennungsrate von bis zu 99% erreichen, wobei das Gerät für eine Seite zwischen 50 Sekunden und mehreren Minuten benötigt!

Im Vergleich zu dem Abtippen einer Buchseite durch einen Menschen ist die Geschwindigkeit heutiger Scanner bereits durchaus konkurrenzfähig, besonders wenn man bedenkt, daß die Maschine nicht ermüdet und die Bedienung eines Scanners von einer ungelernten Kraft geleistet werden kann. Der wichtigste Faktor ist jedoch die Fehlerfreiheit, und hier erfordern beide Übertragungsformen, daß bei wichtigen Dokumenten nachträglich Korrektur gelesen werden muß.

Die Leistungsfähigkeit von Scannern und ihrer OCR-Software hat sich seit den 80-er Jahren enorm verbessert, bei einem gleichzeitigen Preisverfall, wie er für die Computerbranche charakteristisch ist. Deshalb kann man seit 1991 eine stark steigende Verbreitung von Scannern in Büros beobachten.

Die Übertragung *gesprochener Sprache* in das elektronische Medium gestaltet sich dagegen wesentlich schwieriger. Während das Druckbild klar getrennte Wörter mit verhältnismäßig gleichförmigen Buchstaben aufweist, muß die sogenannte Spracherkennung *(speech recognition)* einen kontinuierlichen Lautstrom analysieren und zudem mit unterschiedlichen Dialekten, Stimmhöhen und Hintergrundsgeräuschen fertigwerden.

Der Qualitätsmaßstab für die automatischen Spracherkennung ist die Spracherkennung des Menschen. Somit ergeben sich folgende Ansprüche an Systeme der automatischen Spracherkennung:

2.1 Desiderata der automatischen Spracherkennung:

=> Sprecher-Unabhängigkeit

Das System soll spontane Sprache verschiedener Sprecher bewältigen, auch wenn deren Aussprache sich in Tonhöhe, Dialekt, Geschwindigkeit etc. unterscheidet.

=> Domänen-Unabhängigkeit

Das System soll in der Lage sein, gesprochene Sprache in geschriebene Sprache zu übertragen, und zwar unabhängig vom Inhalt.

=> Realistischer Wortschatz

Die Zahl der erkennbaren Wortformen soll der eines normalen Sprechers entsprechen.

=> Robustheit

Auch bei Abbrüchen, Kontraktionen und Verschleifungen der gesprochenen Sprache soll das System in der Lage sein, die intendierten Wortformen zu erschließen.

Heutige Systeme der Spracherkennung erreichen eine gewisse Sprecherunabhängigkeit, indem eine Domäne vorgegeben wird (z.B. Zugauskunft), in deren Rahmen nur ganz beschränkte Dialoge sinnvoll sind. Das Wissen über diese inhaltlichen Einschränkungen der verwendeten Domäne wird - in Kombination mit grammatischem Wissen - dazu genutzt, die wahrscheinlichsten Wortfolgen zu erschließen.

Der Wortschatz dieser Spracherkennungssysteme liegt jedoch nach wie vor bei unter 1000 *Wortformen.* Ein normaler Sprecher verwendet dagegen etwa 10000 Wörter, was im Deutschen etwa 100.000

1 Siehe hierzu D. McClelland 1991.

Wort formen entspricht. Das passive Vokabular eines durchschnittlichen Sprechers ist noch einmal drei bis vier mal so groß.

Trotz dieser Schwierigkeiten wird an der automatischen Spracherkennung z. Zt. weltweit intensiv und mit großem finanziellen Aufwand gearbeitet. Der Grund ist, daß das Diktieren wesentlich einfacher (benutzerfreundlicher) ist als das Eintippen. Die praktischen Ziele reichen von der elektronischen Sekretärin über die automatische Zugauskunft per Telephon zum 'Verbmobil' , einem tragbaren Computer, in den man auf deutsch oder japanisch hineinspricht und der dann (über einen kleinen Lautsprecher) eine englische2 Übersetzung ausgibt.

Daß die heutigen Systeme der akustischen Spracherkennung bei der Interpretation der Schallwellen grammatisches Wissen und Domänenwissen sehr stark mit einbeziehen, ist keineswegs als eine Notmaßnahme anzusehen, um mit deren Hilfe überhaupt zu einem Ergebnis zukommen. Vielmehr entspricht diese Strategie der Situation beim Menschen, der ja auch alle ihm zu Verfügung stehenden Informationen bei der Interpretation von gesprochener Sprache mit zum Einsatz bringt.

Dies ändert jedoch nichts an der Tatsache, daß die Aufgabe der Spracherkennung in nicht mehr und nicht weniger als der Übertragung von gesprochener Sprache in das elektronische Medium besteht. Das elektronische Medium ist naturgemäß das eigentliche Medium der computerlinguistischen Analyse von Lexikon, Morphologie, Syntax, Semantik und Pragmatik.

Mit anderen Worten, die computerlinguistische Analyse elektronisch gespeicherter Sprache erfolgt unabhängig von den anderen Sprachmedien. Je höher aber die Qualität und Effizienz dieser allgemeinen, abstrakten Analyse von Sprache auf dem Computer ist, desto leistungsfähiger ist sie als Grundlage der optischen und akustischen Signalerkennung.

3 Technische Vorteile des elektronischen Mediums

Auch ohne den Einsatz sprachwissenschaftlich-basierter Methoden bietet das elektronische Medium den anderen Medien gegenüber ganz wesentliche Vorteile. Die Möglichkeiten der elektronischen Verarbeitung auf dem Computer sind der Grund, warum Texte, die ursprünglich nur im Druckmedium vorhanden waren, systematisch in das elektronische Medium übertragen werden und nun auf CD gekauft werden können. Beispiele sind:

=> sämtliche Texte des klassischen Griechisch
sämtliche Texte des klassischen Latein

die Shakesspeare Gesamtausgabe die
Encyclopedia Britannica der
Brockhaus/Wahrig

Vergleichen wir z.B. die Benutzung der gedruckten Version eines 10 bändigen Lexikon mit der elektronischen Version auf einer CD-ROM. Der Vorteil liegt in der Geschwindigkeit und Bequemlichkeit beim Finden von relevanten Textstellen auf der CD-ROM. Statt mehrere Bände aus dem Regal zu wuchten und nach den richtigen Seiten zu suchen, genügt bei der CD-ROM die Eingabe der Schlüsselwörter.

Mit der geeigneten Software kann man nicht nur nach den Haupteinträgen suchen, sondern sämtliche Vorkommnisse eines Schlüsselwortes in den Einträgen in Sekundenschnelle finden. Und schließlich kann man Kombinationen von Wörtern suchen, etwa alle Stellen, an denen die Wörter 'Maler', 'Venedig' und '16. Jahrhundert' innerhalb einer Länge von 40 Wörtern vorkommen.

2 Als Voraussetzung für die Benutzung wird angenommen, daß die deutschen und japanischen Gesprächspartner eine passive Kenntnis des Englischen besitzen. Auf diese Weise soll nicht nur der Hörer den Sprecher verstehen können, sondern der Sprecher soll auch in der Lage sein, die automatische Übersetzung des Geräts zu überprüfen.

Diese elektronisch-basierten Suchmethoden sind nicht nur bei Benutzung eines Lexikons oder der wissenschaftlichen Arbeit über Shakespeare, die klassischen Texte der Griechen und Römer, etc. von praktischem Nutzen. Auch bei der Vorbereitung auf einen Prozeß mit Hilfe einer juristischen Datenbank, der computergestützten Diagnose einer seltenen Krankheit oder der Wahl eines spezifischen Medikaments sind elektronische Datenbanken traditionellen Schriftstücken und Zettelkästen haushoch überlegen.

Gegenüber der Schreibmaschine bieten Computer zudem die Möglichkeit, Texte elektronisch zu korrigieren, in andere Dateien zu kopieren, zu edieren und formatieren. Aus diesem Grund entstehen die meisten Texte, die heute publiziert werden, schon primär in elektronischer Form und werden erst ganz am Schluß in das sekundäre Medium des Buch- oder Zeitungsdrucks übertragen.

Betrachten wir z.B. eine Tageszeitung. Früher wurden die einzelnen Artikel mit einer mechanischen Setzmaschine im Bleisatz aus einzelnen Buchstaben zusammengesetzt. Der Inhalt der Tageszeitung existierte nur in Form der Druckplatten, die nach dem Druckvorgang wieder zerlegt, bzw. eingeschmolzen wurden, und in Form der Zeitungsexemplare, die auf Papier gedruckt wurden.

Wenn die Redaktion einen Beitrag eines Informationsdienstes übernehmen wollte, mußte der Beitrag Buchstabe für Buchstabe von der Vorlage nachgesetzt werden. Wenn es vor Druckbeginn eine Sensationsmeldung gab, die man unbedingt aufnehmen wollte, mußte man die Druckplatten mit der Hand umbauen, um Platz für den neuen Beitrag zu schaffen.

Heute existiert der Inhalt der Zeitungen in elektronischer Form. Beiträge der Informationsdienste werden nicht mehr auf Papier geliefert, sondern kommen über das Telephon in einer Form, die mit Hilfe eines Modems in das elektronische Medium rekonvertiert werden kann. Eine Zeitungsausgabe in elektronischer Form kann beliebig umformatiert, kopiert und ediert werden. Jede dieser Versionen kann dann automatisch gedruckt werden.

In 3.1 findet sich ein kurzer Artikel aus einer Tageszeitung, wie er in einem Verlagsrechner gespeichert wurde. Dieser Text enthält die charakteristischen Steuerzeichen für die Setzmaschine.

3.1 Zeitungstext mit Steuerzeichen

```
0509636
      / otagD22801P1008501271738otagotag
<01001> <SB15.HOSO.HX2,3.D42.S451.SGS> <ef> po1itik - panorama
vinde1en - vd++ - otag<OO1,OOO3> <01002> <002,0006> <01003>
<sb14> Heinrich Vinde1en, ++ <mp> <S450> <SGS> <sv> <SK> <DZ>
<ef> Bundesminister f}r <01004> Innerdeutsche++ Beziehungen,
sieht 4nzeichen <01005> f}r eine Einigung zwischen Bonn und++
<01006> Ostber1in in der umstrittenen Frage der <01007> DDR-
Staatsbürgerschaft++ . Im Hessischen <0100S> Rundfunk meinte der CDU-
Po1itiker, ohne <01009++> auf Einze1heiten einzugehen, es
verdichteten <01010> sich die Indizien daf}r++ , da' die SED-
F}hrung <01011> offenbar nicht mehr auf einer "vo11en 4ner++<->
<01012> kennung" bestehe, sondern sich mit einer <01013>
"Respektierung++" durch Bonn zufrieden geben <01014>
klnnte.<014,0042> <014,0042>
```

Wie ist nun dieser Text an die Stelle von Beispiel 3.1 gekommen? Linguisten interessieren sich für Tageszeitungen nicht wegen der aktuellsten Informationen, sondern weil sie Zeitdokumente der Sprache sind. Da Tageszeitungen heute primär in elektronischer Form vorliegen, ist es für Computerlinguisten eigentlich nur ein rechtliches Problem (Copyright), beliebige Mengen von elektronisch gespeicherten Zeitungsausgaben zu besorgen.

Sobald die Genehmigung vorliegt, muß man sich nur noch eine Kopie der Druckerbänder besorgen und in den eigenen Computer einspielen. Danach kann man die Information beliebig verarbeiten. So kann man z.B. eine bestimmte Textstelle herauskopieren und in einem anderen Text unterbringen, wie in 3.1. Eine weitere Möglichkeit ist, die Steuerzeichen zu entfernen bzw. zu interpretieren.

3.2 'gereinigter' Zeitungstext

```
05.09.86
politik - panorama vindelen
Heinrich Windelen, Bundesminister fuer Innerdeutsche
Beziehungen, 5ieht Anzeichen fuer eine Einigung zwischen
Bonn und Ostberlin in der umstrittenen Frage der DDR-
Staatsbuergerschaft. Im Hessischen Rundfunk meinte der
CDU-Politiker, ohne auf Einzelheiten einzugehen, es
verdichteten
sich die Indizien dafuer, dass die SEDFuehrung offenbar
nicht mehr auf einer "vollen Anerkennung" bestehe,
sondern sich mit einer "Respektierung" durch Bonn
zufrieden geben koennte.
```

Falls der Text nicht auch als gedruckte Zeitung vorliegt, kann die Interpretation der Steuerzeichen über den textuellen Kontext erfolgen: z.B. soll 'Staatsb}rgerschaft´ in 3.1 offenbar *Staatsbürgerschaft* heißen und' 'kInnte', offenbar *könnte.*

Weltweit existiert heute noch die Schwierigkeit, daß nicht nur jedes Land seine eigenen Konventionen für Steuerzeichen hat, sondern praktisch jede Druckerei. Da es umständlich und zeitraubend ist, ständig wechselnde Steuerkonventionen zu interpretieren, wurde von der INTERNATIONAL STANDARDS ORGANIZATION (ISO) der sogenannte SGML-Standard entwickelt:3

3.3 SGML: *standard generalized markup language.*

> A family of ISO standards for labeling electronic versions of text, enabling both sender and receiver of the text to identify its structure (e.g. title, author, header, paragraph, etc.)
>
> Dictionary of Computing, S. 416
> (Illingworth et al. 1990)

Der SGML-Standard wird auch in Europa, und damit Deutschland, anerkannt und findet mit der Zeit immer weitere Verbreitung.

3 Eine ausführliche Darstellung zum Thema SGML findet sich in Herwijnen 1990.

Denn elektronische Texte, die die Konventionen dieses Standards einhalten, haben den Vorteil, daß ihre Steuerzeichen von allen anderen SGML-Benutzern automatisch interpretiert werden können.4

Ein im Computer gespeicherter Text kann nach den individuellen Absichten und Bedürfnissen des Benutzers elektronisch verändert werden. So kann z.B. der "gereinigte" Zeitungsartikel 3.2 mit Hilfe eines Editors für die Verarbeitung in LATEX wie folgt aufbereitet werden.

3.4 19\TEX-Forma.tierung eines Texts

```
\documentstyle{artikel}
\begin{document}

\noindent
05.09.86\\
{\bf Politik:} - {\it panorama windelen}\\ Heinrich Windelen,
Bundesminister f\"{u}r Innerdeutsche Beziehungen,
sieht Anzeichen f\"{u}r eine Einigung zwischen Bonn
und Ostberlin in der umstrittenen Frage der DDR-
Staatsb\"{u}rgerschaft. Im Hessischen Rundfunk meinte
der CDU-Politiker, ohne auf Einzelheiten einzugehen,
es verdichteten sich die Indizien daf\"{u}r, da{\ss}
die SED-F\"{u}hrung offenbar nicht mehr auf einer
"vollen Anerkennung" bestehe, sondern sich mit einer
"Respektierung" durch Bonn zufrieden geben k\"{o}nnte.
\end{document}
```

LATEX ist eine vereinfachte Version von TEX, welches von D. KNUTH als Programmiersprache für das Schriftsetzen auf dem Computer entwickelt wurde. Nachdem 3.4 durch das LATEX-Programm geschickt worden ist, gibt der Computer folgendes Schrift bild aus:

3.5 GeTEXte Version des Texts

> 05.09.86
> Politik: - *panorama windelen*
> Heinrich Windelen, Bundesminister für Innerdeutsche Beziehungen, sieht Anzeichen für eine Einigung zwischen Bonn und Ostberlin in der umstrittenen Frage

4 Als Vereinfachung der sehr mächtigen SGML wurde inzwischen der TEI Standard vorgeschlagen. TEI ist eine Spezialisierung (Untermenge) der SGML und steht für *text encoding initiative.*

der DDR- Staatsbürgerschaft. Im Hessischen Rundfunk meinte der CDU-Politiker, ohne auf Einzelheiten einzugehen, es verdichteten sich die Indizien dafür, daß die SED- Führung offenbar nicht mehr auf einer "vollen Anerkennung" bestehe, sondern sich mit einer "Respektierung" durch Bonn zufrieden geben könnte.

3.4 und 3.5 illustrieren nur ganz einfache D-TEX Befehle, z.B. für das Fettdrucken ({\bf }, für *bald face)* und das Kursivdrucken ({\it }, für *italic).* Weiterhin finden wir eine Behandlung von Umlauten und scharfem 's', sowie ein rechtsbündiges Druckbild, wobei das Programm Worttrennungen am Zeilenende automatisch vornimmt.

Hinzu kommt die automatische Behandlung der Kapitel- und Sektionsüberschriften, die automatische Erstellung von Inhaltsverzeichnissen und Indices, und vieles mehr. Vor allem bei der Darstellung mathematischer Formeln sind TEX und LATEX außerordentlich leistungsfähig.

Seit ihrer Einführung im Jahre 1984 werden TEX und LATEX immer mehr zur Publikation von wissenschaftlichen Büchern und Zeitschriften verwendet, wobei Wissenschaftler ihre Aufsätze und Bücher nicht nur auf dem Computer schreiben, sondern auch selbst formatieren und in druckfertiger Form beim Verlag abliefern. Die Publikation über das elektronische Medium ist nicht nur wesentlich kostengünstiger als ein konventionell gesetztes Buch, sondern hat auch viele praktische Vorteile, insbesondere die direkte Einflußnahme des Autors auf die Gestaltung und die Tatsache, daß das Korrekturlesen der Setzer arbeit entfällt.

Neben den Möglichkeiten der schnellen, computergestützten Suche von Textstellen und dem *desktop publishing* (DTP) bieten elektronisch gespeicherte Texte auch leistungsfähige Möglichkeiten der linguistischen Analyse. So kann man den Text 3.2 z.B. in wenigen Schritten in eine alphabetische Wortliste verwandeln.

3.6 Alphabetische Wortformenliste des Texts

05.09.86	Windelen	koennte
Anerkennung	auf	mehr
Anzeichen	auf	meinte
Beziehungen	bestehe	mit
Bonn	dafuer	nicht
Bonn	dass	offenbar
Bundesminister	der	ohne
CDU-Politiker	der	panorama
DDR-Staats-		
buergerschaft	der	politik
Einigung	die	sich
Einzelheiten	die	ohne
Frage	durch	sieht
Heinrich	eme	sondern
Hessischen	emer	umstrittenen
hn	einer	und
Indizien	einzugehen	verdichteten
Innerdeutsche	es	vollen
Ostberlin	fuer	windelen
Respektierung	fuer	zufrieden
Rundfunk	geben	zwischen
SED- Fuehrung	m	

Eine Wortliste wie 3.5 zählt jedes einzelne Vorkommnis einer Wortform und bietet somit die Grundlage für statistische Untersuchungen zur Worthäufigkeit in Texten. Man kann aber auch ebenso einfach eine Wortliste erstellen, in der jede Wortform nur einmal vorkommt (und wo kein Unterschied zwischen Groß- und Kleinbuchstaben gemacht wird). Dieser zweite Typ ist dann z.B. für eine lexikalische Kategorisierung das Geeignete.

Die bisher genannten Verfahren der automatischen Suche von Wörtern oder Wortfolgen in großen textuellen Datenbanken, der automatischen Fehlersuche ("spelling checker" über Vergleiche mit Wortlisten), das Formatieren mit Hilfe von Steuerzeichen, die Umformung von Texten in alphabetische Wortformenlisten etc. sind rein technologische Verfahren der Zeichenverarbeitung im elektronischen Medium. Sie basieren in keiner Weise auf linguistischen

Konzepten, Theorien oder Methoden.5

Im Vergleich zu nicht-elektronischen Verfahren (Bleisatz, Zettelkästen, Suche in großen Dokumenten in Form von Durchblättern, bzw. Durchlesen etc.) sind diese elektronischen Verfahren enorm schnell, präzise und bequem zu handhaben. Sie erleichtern nicht nur die praktische Arbeit mit Texten, sondern sie liefern auch wertvolle Daten (alphabetische Listen von Wortformen, statistische Aussagen über die Häufigkeit von Wortformen, Paaren von Wortformen, Tripeln von Wortformen- sogenanten *trigrams* - etc. in großen Texten) für die linguistische Analyse.

Gleichzeitig zeigen sich aber auch deutliche Grenzen. Sie bestehen darin, daß die Technologie-basierten Verfahren rein Buchstaben-orientiert sind. Eine grammatikalische Analyse der Wortformen, der syntaktischen Struktur und, darauf aufbauend, des Inhalts, liegt außerhalb dieser Technologie in der Domäne der Sprachwissenschaft.

4 Komponenten der Grammatik

In welchen Bereichen kann mit den Methoden der Sprachwissenschaft eine substantielle Verbesserung der elektronischen Textverarbeitung erreicht werden? Als erste Grundlage für eine Antwort auf diese Fragen werden im Folgenden die Komponenten der Grammatik und ihre Funktionen beschrieben.

Dabei muß berücksichtigt werden, daß sich innerhalb der Sprachwissenschaft drei unterschiedliche Ansätze der grammatischen Analyse herausgebildet haben, nämlich (a) die TRADITIONELLE GRAMMATIK, (b) die THEORETISCHE LINGUISTIK und (c) die COMPUTERLINGUISTIK. Diese drei Ansätze unterscheiden sich bezüglich ihrer

1. Methoden,
2. Fragestellungen
   (also den deskriptiven bzw.
   explanatorischen Zielen) und
3. Anwendungen.

Bevor wir die Komponenten der Grammatik beschreiben, beginnen wir mit einem schematischen Vergleich der drei verschiedenen Ansätze innerhalb der Sprachwissenschaft.

4.1 Drei unterschiedliche Ansätze der Sprachanalyse

=> Traditionelle Grammatik

Die traditionelle Grammatik ist von der Methode her taxonomisch (deskriptiv-klassifikatorisch) orientiert.

Ihr Ziel ist ein möglichst vollständiges Sammeln und Klassifizieren der sprachlichen Einzelphänomene, insbesondere die Darstellung der sprachlichen Regelmäßigkeiten und der Ausnahmen.

Von der Anwendung her kommt sie aus dem Sprachunterricht.

Für die Computerlinguistik ist die traditionelle Grammatik wegen ihrer empirischen Datenfülle von großem Interesse.

=> Theoretische Linguistik

Die Methode der theoretischen Linguistik ist logisch-mathematisch: es werden formale Regelsysteme formuliert, aus denen alle und nur die wohlgeformten sprachlichen Strukturen ableitbar sein sollen. Dies hat der traditionellen Grammatik gegenüber den methodologischen Vorteil der *expliziten Hypothesenbildung* - allerdings nur theoretisch, denn eine Überprüfung der formalen Regelsysteme an realistischen Datenmengen ist mit Papier und Bleistift praktisch unmöglich.

5 Es zeigt sich aber schon an einem so einfachen Problem wie der automatischen Worttrennung am Zeilenende, etwa im Rahmen des *desk top publishing*, daß Technologie und Linguistik bei der automatischen Textverarbeitung eng zusammenliegen.

0bwohl die theoretische Linguistik nach wie vor in viele verschiedene Schulen zersplittert ist, gilt als gemeinsames Erklärungsziel die formale Charakterisierung des menschlichen Sprachvermögens, und zwar unter Ausgrenzung der Sprachwendung *(Performance)* .

Anwendungsversuche reichen von Erklärungsmodellen in der Psychologie bis zum Sprachunterricht in der Schule.

Für die Computerlinguistik sind vor allem die Untersuchungen zu formalen Sprachklassen und Komplexität relevant.

=> **Computerlinguistik**

Methodisch verbindet die Computerlinguistik das Ziel der traditionellen Grammatik einer möglichst vollständigen Klassifikation natürsprachlicher Phänomene mit dem logischmathematischen Ansatz der theoretischen Linguistik. Hinzu kommt allerdings die wichtige Neuerung, daß die expliziten Hypothesen, repräsentiert durch als Parser implementierte formale Grammatiken, .*automatisch* an großen Datenmengen *überprüft* werden können.

Das deskriptive und explanatorische FERNZIEL der Computerlinguistik ist eine *Modellierung der Informationsübertragung mit Hilfe natürlicher Sprachen.* Auf dem Weg zu diesem Ziel muß eine vollständige morphologische, lexikalische, syntaktische, semantische und pragmatische Erfassung der natürlichen Sprachen in einem funktionalen Rahmen geleistet werden.

Mit dem Erreichen dieses Ziels ergeben sich weitreichende Möglichkeiten in der Anwendung der automatischen Sprachverarbeitung. 6

Trotz ihrer unterschiedlichen Methoden, Zielen und Anwendungen liegt den genannten Varianten der Sprachwissenschaft eine gemeinsame Aufteilung der Grammatik in die Komponenten *Phonologie, Morphologie, Lexikon, Syntax, Semantik* und das zusätzliche Gebiet der *Pragmatik* zugrunde. Allerdings variieren Stellenwert und wissenschaftliche Behandlung dieser Komponenten in den verschiedenen Ansätzen der Sprachwissenschaft:

4.2 Die Komponenten der Grammatik

- **Phonologie**

Wissenschaft von den Sprachlauten.

In der theoretischen Linguistik spielt die Phonologie eine zentrale Rolle als eine Art Grundlagendisziplin, in der universale Prinzipien der Sprachanalyse (distinktive Merkmale, formale Regelapparate) exemplarisch vorgeführt werden. Das Ziel ist eine möglichst allgemeine und elegante Darstellung in Form von Regelsystemen, die (a) historische Veränderungen (Lautwandel) oder (b) synchrone Alternationen in der Aussprache (z.B. die sogenannte "Auslautverhärtung" im Deutschen) beschreiben.

In der Computerlinguistik spielt die Phonologie dagegen, wenn überhaupt, eine untergeordnete Rolle. Der einzige Bereich, wo sie möglicherweise zum Einsatz kommen könnte, ist die automatische Spracherkennung. Allerdings wird dieser Bereich heute mit Hilfe der *Phonetik* (und nicht der Phonologie) bearbeitet. Die Phonetik untersucht die Struktur der (a) artikulatorischen, (b) akustischen und (c) auditiven Abläufe. Im Gegensatz zur Phonologie wird die Phonetik nicht zu den Komponenten der Grammatik gerechnet.

- **Morphologie**

Lehre von den Wortformen einer Sprache. Die Morphologie ist der Hauptbereich der

6 Ein Grammatikformalismus, der von vornherein mit dem Ziel einer effizienten Verarbeitung entwickelt wurde, ist die Linksassoziative Grammatik (LAG). Die Komplexitaetseigenschaften der LAG sind in Hausser 1992 beschrieben.

traditionellen Grammatik, wie sie etwa in Schulgrammatiken (z.B. für Latein) zu finden ist. Sie klassifiziert die Wörter einer Sprache nach ihren Wortarten und beschreibt die Wortformen in bezug auf *Flexion, Derivation* und *Komposition.*

In der Computerlinguistik ist die sogenannte *Computermorphologie* ein zentraler Bereich mit der Aufgabe der automatischen Wortformerkennung. Dies geschieht auf Grundlage eines *on-line* Lexikons und eines morphologischen Analyseprogramms. Die automatische Wortformerkennung ist die praktische Voraussetzung für alle anderen linguistisch basierten Verfahren der automatischen Textanalyse.

. Lexikon

Auflistung der Wörter einer Sprache.

Das möglichst vollständige Sammeln und Einordnen der Wörter einer Sprache fällt in die Lexikographie und Lexikologie. Die Lexikographie beschäftigt sich mit den Prinzipien der lexikographischen Kodierung und der Struktur lexikalischer Einträge und ist ein praktisch orientiertes Randgebiet der Sprachwissenschaften. Die Lexikologie untersucht den Wortschatz einer Sprache in Hinblick auf ihre interne Bedeutungsstruktur und ist in der traditionellen Sprachwissenschaft beheimatet.

Was die theoretische Linguistik betrifft, ist seit Mitte der 60-er ein ständig wachsendes Interesse am Lexikon zu verzeichnen. Die Tendenz dieser Arbeiten ist es, immer mehr syntaktische und semantische Eigenschaften komplexer Ausdrücke aus den lexikalischen Eigenschaften der Teilwörter abzuleiten. Das Ergebnis sind umfangreiche formale Darstellungen einzelner Wörter, die der Erklärung syntaktischer Phänomene dienen sollen.

In der Computerlinguistik fungieren *online* Lexika in Kombination mit Morphologieprogrammen bei der automatischen Wortformerkennung. Das Ziel ist eine größtmögliche Vollständigkeit bei möglichst kompakter Speicherung und schnellem Zugriff. Neben der Erstellung neuer Lexika im Rahmen der automatischen Wortformerkennung besteht großes Interesse dar an, das Wissen traditioneller Lexika wie dem OXFORD ENGLISH DICTIONARY (die inzwischen in elektronischer Form existieren) für die automatische Textanalyse nutzbar zu machen (*"mining of dictionaries").*

. Syntax

Beschreibung der grammatisch legalen Kompositionen von Wortformen.

In der theoretischen Linguistik (generativen Grammatik) ist das Ziel der syntaktischen Analyse die Darstellung der grammatischen Wohlgeformtheit in einer Sprache, und zwar mit Hilfe formaler Regeln, die alle, und nur die wohlgeformten, Ausdrücke einer Sprache generieren (erzeugen) bzw. erkennen. Um aus der Fülle der formalen Möglichkeiten die langfristig richtige Beschreibung zu finden, bemüht man sich dabei primär um eine Charakterisierung des menschlichen Sprachvermögens auf der Grundlage von sogenannten Universalien.

Die Computerlinguistik liefert einerseits die technische Voraussetzung, die generative Kapazität formaler Grammatiken für natürliche Sprachen wirklich an der ganzen Fülle der Daten effektiv zu überprüfen. Andererseits haben sich angesichts des großen praktischen Bedarfs an leistungsfähiger automatischer Syntaxanalyse in den letzten dreißig Jahren immer wieder eigenständige, computer-orientierte Systeme entwickelt, die die Syntaxsysteme der theoretischen Linguistik mehr oder weniger direkt beeinflußt haben.

. Semantik

Analyse der wörtlichen Bedeutung sprachlicher Ausdrücke.

In der theoretischen Linguistik reichen die Aufgaben der Semantik von der Charakterisierung syntaktischer Ambiguität und Paraphrase mit Hilfe von (verschiedenen bzw. gleichen) 'Tiefenstrukturen' zu einer logisch-semantischen Darstellung der Wahrheitsbedingungen mit Hilfe logischer Formeln (z.B. MONTAGUE GRAMMATIK).

Dabei befaßt sich das Teilgebiet der Wortsemantik mit der Bedeutungsanalyse von Wörtern bzw. Wortformen, während die Satzsemantik beschreibt, wie sich die Bedeutung komplexer Ausdrücke aus der Bedeutung ihrer Teile und der Art ihrer Zusammensetzung ableitet (FREGE'SCHES PRINZIP).

Die semantische Analyse sprachlicher Ausdrücke umfaßt u. a. die logische Charakterisierung von Einzahl und Mehrzahl (Quantoren), Konjunktion *(und)* und Disjunktion *(oder),* die Verbergänzung durch Subjekt und Objekt (Kasus und Valenz), die Modifikation von Nomina und Verbalkomplex durch Adjektive und Adverbien, die Neben- und Unterordnung von Teilsätzen und vieles mehr. In der Computerlinguistik reicht die Problematik der Bedeutungsanalyse von der Interpretation von Programmiersprachen über die Konsistenz von Datenbanken zu Verfahren der konzeptbasierten Indexierung und der Desambiguierung in der maschinellen Übersetzung.

. Pragmatik

Theorie von der Verwendung sprachlicher Ausdrücke.

Während sich die bisherigen: Komponenten (Phonologie, Morphologie, Lexikon, Syntax und Semantik) mit den strukturellen Eigenschaften sprachlicher Ausdrücke (Wortformen und Sätzen) beschäftigen, untersucht die Pragmatik, wie sich diese strukturellen Eigenschaften bei der Verwendung der Ausdrücke in einem Äußerungskontext auswirken. Deshalb gehört die Pragmatik streng genommen nicht zu den Komponenten der Grammatik, sondern umfaßt (i) die Strukturanalyse der sprachlichen Ausdrücke (Grammatik), (ii) die Beschreibung des (Äußerungs- und Interpretations-) Kontexts und (iii) die Analyse der Interaktion zwischen Sprache und Kontext.

Wenn es darum geht, die natursprachliche Bedeutungsübertragung zwischen Menschen, bzw. zwischen Mensch und Maschine, theoretisch und praktisch zu modellieren, dann darf die Pragmatik mit ihren drei Teilkomponenten keinesfalls fehlen. Die Verwendung von sprachlichen Ausdrücken umfaßt die Referenz (also den Bezug sprachlicher Ausdrücke auf die vom Sprecher intendierten Objekte), die Interpretation von indexikalischen Ausdrücken (Pronomina, temporale und lokale Adverbien, Verbalflektion), den rhetorisch korrekten Einsatz von Pronomina und die Wortstellung bei der Generierung, sowie die Interpretation nicht-wörtlicher Verwendungen (z.B. Metaphern).

In der theoretischen Linguistik wird die Pragmatik meist als Teil der modelltheoretischen (logischen) Semantik oder im Rahmen der sogenannten Sprechakttheorie behandelt. In der Computerlinguistik findet die Pragmatik wachsendes Interesse aufgrund von praktischen Problemen bei der rhetorisch korrekten Implementierung des *Generierungsaspekts* (etwa bei Dialogsystemen oder Systemen der maschinellen Übersetzung).

Die beschriebene Aufteilung der grammatikalischen Komponenten ist für traditionelle Linguistik, theoretische Linguistik, und Computerlinguistik deshalb gleichermaßen gültig, weil sie sich an unterschiedlichen strukturellen Aspekten natürlicher Sprachen orientiert, nämlich den *Lauten* (Phonologie), den *Wortformen* (Morphologie), den *Wörtern* (Lexikon), den *Sätzen* (Syntax), den *Bedeutungen* (Semantik) und den *Verwendungen* (Pragmatik).

5 Fernziel der Computerlinguistik

Bis heute wurden (und werden) die Grammatiken der theoretischen Linguistik allein nach ihrer Eleganz, Plausibilität, mathematischen Mächtigkeit, Umfang der Datenerfassung oder Kompatibilität mit psychologischen Tests bewertet. Die Methodik der Computerlinguistik erfordert dagegen zusätzlich, daß die formale Struk-

tur der morphologischen, syntaktischen und semantischen Ableitungen eine gute Basis für einen klaren Programmfluß bietet, der eine einfache, schnelle Fehlersuche und Erweiterung erlaubt. Außerdem ist es für die Modellierung des menschlichen Sprachverständnisses auf dem Computer unerläßlich, daß die grammatischen Analysen für eine funktionsfähige pragmatische Interpretation sowohl bei der *Analyse* als auch bei der *Generierung* optimal geeignet sind.

Deshalb liegt aus wissenschaftlicher Sicht das Wesentliche am computerlinguistischen Ansatz nicht so sehr in den möglichen Anwendungen (obwohl die sicherlich wichtig und interessant sind), sondern vielmehr in seiner neuartigen Methodologie. Das Fernziel der Computerlinguistik, nämlich die Modellierung der menschlichen Sprachverwendung auf dem Computer (siehe 4.1), ist von größter methodologischer Bedeutung, weil es eine funktional-holistische Sichtweise erzwingt.

Die neuere Geschichte der theoretischen Linguistik zeigt, daß ihre Vertreter häufig dem Fehler verfallen, vorhandene Komponenten und Formalismen auf Phänomene anzuwenden, für die sie überhaupt nicht entwickelt wurden, und dies nur, weil die Beschreibungsapparate eine solche Ausweitung oberflächlich zuzulassen scheinen. Es gibt hierfür in der Literatur massenhaft Beispiele, etwa die Behandlung semantischer Phänomene in der Syntax, pragmatischer Phänomene in der Semantik, morphologischer Phänomene in der Syntax, usw. Dies hat immer wieder in Sackgassen geführt, bei denen es meist Jahrzehnte dauerte (und dauert), wieder herauszukommen. Diese Gefahr kann nur durch ein einheitliches und funktionstüchtiges Gesamtkonzept wirklich gebannt werden.

Das Fernziel der Computerlinguistik führt zu der Frage, wieweit es prinzipiell überhaupt möglich ist, ein realistisches Modell natursprachlicher Kommunikation
zu entwickeln und zu implementieren. Ich möchte diese Frage anhand einer Analogie beantworten.

Die heutige Situation in der Computerlinguistik entspricht in vielem der Entwicklung des mechanischen Fluges. Viele hundert Jahre lang hat der Mensch die Spatzen und andere Vögel beobachtet, um zu verstehen, wie sie fliegen. Und er hat versucht, sich auf möglichst ähnliche Weise in die Lüfte zu erheben.

Es hat sich dann herausgestellt, daß es mit Flattern nicht geht. Dies wurde gerne zum Anlaß genommen, den menschlichen Flugverkehr für prinzipiell unmöglich zu erklären, häufig mit dem frommen Spruch:

Wenn Gott gewollt hätte, daß die Menschen fliegen könnten, hätte er ihnen Flügel verliehen.

Heute ist das Fliegen für die Menschen selbstverständlich geworden. Außerdem weiß man inzwischen, daß ein Spatz aufgrund desselben theoretischen Prinzips in der Luft bleibt wie ein Jumbojet, nämlich dem Prinzip der "*air foir'*, der Tragflächen.
Es gibt also eine Ebene der Abstraktion, auf der der Flug des Spatzen und der Flug des Jumbojets nach demselben Prinzip ablaufen.

Bei der Modellerierung natursprachlicher Kommunikation in der Computerlinguistik geht es ebenfalls um das korrekte Prinzip auf der korrekten Abstraktionsebene. Dabei besteht naturgemäß die Gefahr, die Ebene entweder zu niedrig oder zu hoch anzusetzen. Zum Beispiel wären geschlossene Signalsysteme, wie man sie etwa bei einem Fahrkartenautomaten findet, sicherlich als Modell ungeeignet.7

Genauso unsinnig wäre es aber auch, die Modellierung natursprachlicher Kommunikation von vornherein mit naiven antropomorphen Vorstellungen *ad absurdum* zu führen. Wer z.B. mit einem Begriff

7 Der entscheidende Punkt der in Frage stehenden Modellierung ist, daß die charakteristische Vielseitigkeit natursprachlicher Kommunikation erhalten bleibt - also die Tatsache, daß dieselben Ausdrücke in den verschiedensten Äußerungskontexten sinnvoll eingesetzt werden können.

von "Verstehen" antritt, wonach sich das System bei der Analyse von FINNEGAN'S WAKE subtil amüsieren muß, liegt ebenso daneben, wie jemand, der Paarungsverhalten und Brutpflege von einem Jumbojet erwartet.

Zum Schluß noch eine zweite Analogie aus der Geschichte des Flugzeugbaus: nachdem man an Doppeldeckern, Propellerflugzeugen und Jets ein immer besseres Verständnis der Flugprinzipien entwickelt hat, analysiert man heute wieder verstärkt natürliche Flugvorgänge, um ihre wunderbare Leistungsfähigkeit zu begreifen und erfolgreich in den Bau leiserer und effizienterer Flugzeuge einzubringen.

Dieses Beispiel zeigt, daß theoretisch-technologische Lösungsversuche in der Computerlinguistik keinesfalls ein fehlendes Interesse an der Analyse menschlicher Sprachfähigkeiten implizieren. Vielmehr ist es so, daß eine Untersuchung des speziell Menschlichen am sprachlichen Kommunikationsprozeß erst dann wirklich sinnvoll wird, nachdem eine prinzipielle Modellierung natursprachlicher Kommunikation geleistet worden ist und sich in massiven Anwendungen bewährt hat.

Bibliographie

Illingworth et al. (Hrsg.) (1990) *Dictionary oi Computing,* Oxford University Press, Oxford.

Hausser, R. (1989) *Computation oi Language,* Springer-Verlag, Symbolic Computation: Artificial Intelligence, Berlin-New York.

Hausser, R. (1992) *"Complexity in Left Associative Grammar, Theoretical Computer Science,* Vol. 103:283-308, Elsevier.

Herwijnen, E. van (1990) *Practical SGML,* Kluwer Academic Publishers.

Hutchins, W.J. (1986) *Machine Translation: Past, Present, Future,* Ellis Horwood Lmt., Chichester .

McClelland, D. (1991) *"OCR: Teaching Your Mac to Read,"* MACWORLD, November 1991:169-175.