LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-Forum 9.1 (1992)
Forum der Gesellschaft für Linguistische Datenverarbeitung e.V.

Herausgeber
Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)
Anschrift: Prof. Dr. Burghard Rieger, Universität Trier, FB II: LDV/CL, D-5500 Trier, Postfach 3825, Tel.: (0651)201-2272/2270; Fax (0651)201-3946; Email: rieger%utrurt@unido.uucp.de *oder* unido!utrurt!rieger

Redaktion
Burghard Rieger, Roland Fraese, Amancio Kolompár

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

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

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

Editorial

Wie angekündigt, erfüllt das *LDV-Forum* mit dem vorliegende Heft zum Teil eine Aufgabe, die in den zurückliegenden Jahren durch die Publikation von Tagungsbänden abgedeckt wurde: schon die früheren 'LDV-Kolloquien' unserer Gesellschaft wie auch die jüngeren 'GLDV-Jahrestagungen' wurden während der letzten 10 Jahre durchweg[a] durch Buchveröffentlichungen der Tagungsbeiträge dokumentiert, was auch für diejenigen Jahrestagungen der *GLDV* zutrifft, die zusammen mit anderen Organisationen veranstaltet wurden (wie etwa 1985 mit der Fachgruppe 3 *Natürlichsprachliche Systeme* der Gesellschaft für Informatik, oder 1988 wiederum mit dieser Fachgruppe der GI und der Sektion *Computerlinguistik* der Deutschen Gesellschaft für Sprachwissenschaft). Die unterschiedlichen thematischen Schwerpunkte der GLDV-Tagungen seit 1982 haben so zu der beachtlichen Reihe der bisherigen Tagungsbände geführt:

Koblenz 1982: *Linguistische Datenverarbeitung und Nachbarn*, hrsg. von Bátori, I.S./ Krause, J./ Lutz, H.D., Tübingen (Niemeyer) 1982 [Sprache und Information 4]

Trier 1983: *Mikrokomputer und Textverarbeitung*, hrsg. von Krause, J./ Niederehe, H.J., Hamburg (Buske) 1984 [Romanistik in Geschichte und Gegenwart 16]

Heidelberg 1984: *Trends in der Linguistischen Datenverarbeitung*, hrsg. von Hellwig, P./ Lehmann, H., Hildesheim/Zürich (Olms) 1986 [Linguistische Datenverarbeitung 1]

Hannover 1985: *Sprachverarbeitung in Information und Dokumentation* hrsg. von Endres-Niggemeyer, B./ Krause, J., Berlin/Heidelberg (Springer) 1985 [Informatik Fachberichte 114]

Göttingen 1986: *Computerlinguistik und philologische Datenverarbeitung*, hrsg. von Klenk, U./ Scherber, P./ Thaller, M., Hildesheim/Zürich (Olms) 1987 [Linguistische Datenverarbeitung 7]

Bonn 1987: *Analyse und Synthese gesprochener Sprache*, hrsg. von Tillmann,H.G./ Willée, G., Hildesheim/Zürich (Olms) 1987 [Linguistische Datenverarbeitung 9]

Saarbrücken 1988: *Computerlinguistik und ihre theoretischen Grundlagen*, hrsg. von Bátori, I.S./ Hahn, U./ Pinkal, M./ Wahlster, W., Berlin/ Heidelberg (Springer) 1988 [Informatik Fachberichte 195]

Ulm 1989: *Interaktion und Kommunikation mit dem Computer*, hrsg. von Endres-Niggemeyer, B./ Hermann, T./ Kobsa, A./ Rösner, D., Berlin/ Heidelberg (Springer) 1990 [Informatik Fachberichte 238]

Siegen 1990: *Lexikon und Lexikographie*, hrsg. von Schaeder, B./ Rieger, B., Hildesheim/Zürich (Olms) 1990 [Linguistische Datenverarbeitung 11]

Trier 1991: *Quantitative Linguistics*, hrsg. von Rieger, B./ Köhler, R., Amsterdam/NewYork (Elsevier Science) 1992 [in Vorbereitung]

[a] Ausnahme: Tagungsbeiträge der Sektion Inhaltsanalyse, Jahrestagung 1983, hrsg. von Rostek, L./ Schulz, G.F., *LDV-Forum*, Beiheft 1(1985)

Die gemeinsame Veranstaltung der *1st Quantitative Linguistics Conference (QUALICO)* zusammen mit der GLDVJahrestagung 1991 im letzten Jahr in Trief hat nun erstmals wegen der vornehmlich internationalen Beteiligung - einen englischsprachigen Tagungsband angezeigt sein lassen, so daß die fünf angenommenen Tagungsbeiträge aus den beiden "nicht-quantitativen" GLDV-Sektionen *Unification* Based Grammars and *Models* und *Natural* Language *Processing* and *Tools* der *QUALICO* mit Zustimmung der Autoren in diesem *LDV-Forum* veröffentlicht werden.

In der letzten Nummer des *LD V-* Forum hatten wir keine Rezensionen veröffentlicht. Ich möchte daher hier nochmals einige Titel auflisten, die der Redaktion als Rezensionsexemplare zugingen. Weil aber die Verlage zunehmend dazu übergehen, Rezensionsexemplare ihrer Neuerscheinungen nur dann zu versenden, wenn begründete Aussicht einer baldigen Besprechung und deren Veröffentlichung besteht, möchten ich hiermit dazu einladen und Sie auffordern, uns weitere Rezensionsanregungen und thematischen Vorschläge (möglichst mit der Bereitschaft zur Besprechung der vorgeschlagenen Veröffentlichungen) zukommen zu lassen, damit wir die zusätzlichen Titel mit der Besprechungszusage des *LDV-Forum* bei den betreffenden Verlagen anfordern können.

~ *Steiner, Erich/ Schmidt, Paul/ Zelinsky- Wibbelt, Cornelia:* From Syntax to Semantics. London (Pinter Publishers) 1988 (262 S.)

~ *Batori, Istvan S./* Lenders, *Winfried/ Putschke, Wolfgang:* Computational Linguistics - Computerlinguistik. Berlin/New York (Walter de Gruyter Verlag) 1989 (933 S.)

~ *Rieger, Burghard:* Unscharfe Semantik. Frankfurt/Bern/New York (Peter Lang Verlag) 1989 (369 S.)

~ *Köhler, Reinhard/ Janßen, Andreas:* PASCAL. Programmieren für Sprach- und Textwissenschaften. (Uni-Taschenbücher 1638) Tübingen (Francke Verlag) 1991 (250 S.)

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

~ Johansson, *Stig/ Stenström, Anna-Brita* (Eds.): English Computer Corpora. Selected Papers and Research Guide. (Topics in English Linguistics 3), Berlin/New York (Mouton de Gruyter) 1991 (402 S.)

Kompetenten und sachkundigen Rezensenten unter unseren Lesern, die eines (oder mehrere) Bücher beurteilen können
und möchten, senden wir gern die entsprechenden Exemplare zu, welche nach Erscheinen ihrer Besprechung - wie üblich Eigentum der Rezensenten werden.

Das nächste Heft 9.2 (1992) des *LDV-Forum* wird Ende des Jahres (im Dezember) erscheinen; Themenschwerpunkt (zu dem noch Beiträge eingereicht werden können): LDV/CL und empirische Sprachdaten.

Trier, im März 1992 B.R.

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

Titelgestaltung
Werbestudio Zimmermann, D-6083 Biebesheim

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 auch auf Diskette (5 1/4“ bzw. 3 ½“) oder elektronisch (Email: ldvforum%utrurt@unido.uucp.de *oder* unido!utrurt!ldvforum) als ASCII oder J9.TEXDatei *(LDVforum.sty* wird zugesandt) übermittelt werden.

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 9.2/92: 31. Oktober 1992; für Heft 10.1/93: 30. April 1993

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

Auflage
550 Exemplare

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

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

Syntaxbasierte Satzgenerierung mit PLNLP

ANDREA BEURER

Zusammenfassung

In der Programmiersprache PLNLP (wird "Penelope" ausgesprochen und steht für Programming Language for Natural Language Processing) ist ein gleichnamiges System implementiert, das sich zur Analyse und Generierung von Sätzen eignet. Die durch die syntaktische Analyse eines Eingabesatzes ermittelten Knoten und Records liefern die Information zum Aufbau einer dependenzorientierten Prädikat-Argument-Struktur. Zusammen mit Attributen, die der Benutzer interaktiv zuweist, können aus der Prädikat-Argument-Struktur deutsche Sätze generiert werden. Im Gegensatz zu Generierungssystemen wie Appelts KAMP 1 oder McKeowns TEXT 2 , die auf den Mensch-:Maschine-Dialog ausgerichtet sind, dient PLNLP dazu, morphosyntaktische Fehlerkorrekturen vorzunehmen und auf stilistische Schwächen in Sätzen hinzuweisen.

1 PLNLP - Programmiersprache und System

Die Programmiersprache PLNLP wurde speziell für Anwendungen in der maschinellen Sprachverarbeitung entworfen und wird in LISP, FORTRAN und C übersetzt. Sie läuft in einer VM-Umgebung auf dem Großrechner, wobei die C-Version auch unter dem Betriebssystem OS/2 auf dem PS/2 eingesetzt werden kann. In dieser Sprache ist ein gleichnamiges System implementiert, das aus Heidorns NLP- System (1972) hervorging und in diversen IBM-Projekten weiterentwickelt wurde. Das PLNLP-System eignet sich zur Analyse und Generierung von natürlichsprachlichen Sätzen und dient dazu, morphosyntaktische Fehlerkorrekturen vorzunehmen und auf stilistische Schwächen in Sätzen hinzuweisen, um so den späteren Einsatz in einer Komponente zur Textkritik zu ermöglichen. Eine solche Komponente könnte im Zweitspracherwerb, aber auch in der maschinellen Übersetzung Anwendung finden.

Abb. 1 veranschaulicht, wie PLNLP auf einen Kongruenzfehler zwischen dem Genitivattribut der Präpositionalphrase (" Brasiliens") und der zugehörigen Apposition (" dem größten Land des Subkontinentes") aufmerksam macht und wie der Verbesserungsvorschlag aussehen könnte.

Es läßt sich am Beispiel Brasiliens, dem größten Land des Subkontinentes, zeigen.

Zuerst analysiert der bottom-up und parallel arbeitende PLNLP- Parser den Eingabesatz. Der Analysebaum ist nicht von oben nach unten, sondern von links nach rechts zu lesen: Oben links wird der Startknoten aufgeführt, der den Satzmodus - in diesem Fall DECL für deklarativ - anzeigt. Rechts davon erscheinen die einzelnen Konstituentenebenen in Form von Spalten. Das Sternchen markiert den Head auf jeder Ebene, d.h. in jeder Spalte. In der letzten Spalte stehen die lexikalischen Elemente und die Satzzeichen, die die Terminalknoten des Syntaxbaumes bilden. Nach der Analyse gibt PLNLP eine Fehlermeldung aus, führt den Eingabesatz noch einmal auf, macht einen Verbesserungsvorschlag und nimmt am Ende eine Fehleranalyse vor - in diesem Fall Case Disagreement.

1 Basierend auf einem hierarchischen Planer, generiert "der Experte" KAMP Äußerungen, die dem laienhaften Benutzer bei der Montage und Reparatur von technischen Geräten helfen (Appelt 1985).

2 TEXT beantwortet drei Arten von Meta-Level-Fragen zur Struktur der zugrundeliegenden Datenbasis, die Informationen über Militärfahrzeuge und Waffen enthält. Die Antworten umfassen mehrere kohärente Sätze und haben die Länge eines Textabschnittes. Zur Organisation des Diskurses verwendet TEXT Schemata (McKeown 1985).

DECL1	NP1	PRON1*	"Es"		
	VP1	VERB1*	"läßt"		
	NP2	PRON2*	"sich"		
	PP1	PREP1	"am"		
		NOUN1*	"Beispiel"		
		NP3	NOUN2*	"Brasiliens"	
		PUNC1	","		
		NP4	PRON3*	"dem"	
		NP5	AP1	ADJ1*	"größten"
			NOUN3*	"Land"	
		NP6	ART1	ADJ2*	"des"
			NOUN4*	"Subkontinentes"	
			PUNC2	","	
	VERB2	"zeigen"			
	PUNC3	"."			

GRAMMATICAL ERROR IN SENTENCE 1.

Es läßt sich am Beispiel Brasiliens, dem größten Land des Subkontinentes zeigen.
CONSIDER: Es läßt sich am Beispiel Brasiliens, des größten Landes des Subkontinentes zeigen.
CASE DISAGREEMENT

Abb. 1: Eine von PLNLP vorgenommene Korrektur.

1.1 Der PLNLP-Formalismus im Hinblick auf die Generierung

PLNLP stellt einen Formalismus bereit, um Prozeduren und Analyse- und Generierungsregeln zu entwickeln. Als Datenstruktur werden Segment Records benutzt. Dabei handelt es sich um Bündel von Attribut-Wert-Paaren, die beliebig lange Zeichenketten (= Segmente eines Textes) beschreiben. Analog zu anderen Programmiersprachen können die Attribute als Werte Zahlen, Zeichenketten und Listen annehmen. Darüber hinaus gibt es *Pointer-Werte*, die auf andere Segment Records verweisen. Auf diese Art lassen sich Records ineinander einbetten, so daß ganze Record-Strukturen entstehen können. Die Anwendung einer Generierungsregel führt nun dazu, daß neue Records kreiert oder alte modifiziert werden.

2 Generierung als encoding process

Generierung in PLNLP bedeutet, daß die Information, die in den Records vorliegt, in Information in Form von linearen Zeichenketten umgewandelt wird. Dazu müssen zunächst mit Hilfe der Generierungsregeln Segment Records auf höherer Ebene in Segment Records auf niedriger Ebene transformiert werden. Segment Records auf höherer Ebene definieren längere Textsegmente, Segment Records auf niedriger Ebene entsprechend kürzere. Insgesamt gibt es sechs Ebenen: Satz, Phrase, Wort, Morpheme, Wortstamm, Zeichen. Auf der untersten Ebene, d.h. der Zeichenebene, erfolgt dann die Ausgabe der Zeichenketten. Diesen Vorgang bezeichnet Heidorn (1972, 111 et passim) als "encoding process". Da das PLNLP-System nicht auf die Generierung von Antworten im Rahmen einer Dialogkomponente ausgelegt ist, wird Generierung hier nicht als Planungsprozeß aufgefaßt, mit dem bestimmte Diskursziele erreicht werden sollen, sondern als Umwandlung von Segment Records in Textsegmente. So lassen sich Wortformen und ganze Sätze in einer Komponente substituieren, die Fehler und stilistische Schwächen in Sätzen aufzeigt. Die Festlegung des Inhalts eines zu generierenden Textes, die bei Systemen zur Antwortgenerierung (vgl. TEXT und KAMP in der Zusammenfassung) von großer Bedeutung ist, tritt bei PLNLP daher zugunsten der linguistischen Realisierung in den Hintergrund.

2.1 Aufbau der Generierungsregeln

Die Generierungsregeln setzen sich aus *erweiterten Phrasenstrukturregeln* zusammen, die ohne die Erweiterungen kontextfreien Phrasenstrukturregeln gleichen:

VP→ VP NP
VP→ VERB VP

Abb. 2: Kontextfreie Phrasenstrukturregeln

Da mehrere Regeln dasselbe nicht-terminale Symbol auf der linken Regelseite enthalten können, fehlt ein Entscheidungskriterium, das in jeder Phase des Generierungsprozesses die passende Regel auswählt. Die in Abb. 3 aufgeführten Erweiterungen liefern ein solches Kriterium. So würde z.B. Regel (2) ausgewählt, wenn ein Satz im Passiv generiert werden soll.

```
(1) VP(DOBJ,...)      → VP(...,-DOBJ)
                        NP(%DOBJ(VP),...)
(2) VP(PASSIVE,...)   → VERB(%VP,PRED=
                        'WERDEN',...)
                        VP(-PASSIVE,...)
```

Abb. 3: Erweiterte Phrasenstrukturregeln in vereinfachter und modifizierter Form[3]

Die Objekte NP, VP etc., auf die die Regeln Bezug nehmen, repräsentieren nicht nur Namen, sondern die zuvor erwähnten Segment Records, d.h. komplexe Attribut-Wert-Strukturen. Bei den in Klammern erscheinenden *Spezifizierungselementen* handelt es sich nicht nur um die Namen von Prozeduren, sondern um die Prozeduren selbst, die während der Generierung ausgeführt werden und in der Programmiersprache PLNLP formuliert sind. Die Spezifizierungselemente auf der linken Regelseite bestimmen, wann eine Regel angewandt wird, und die Spezifizierungselemente auf der rechten Regelseite kreieren oder modifizieren Records.

2.2 Der Generierungsalgorithmus

Die Generierungsregeln operieren auf einem Algorithmus, der top- down, seriell und von links nach rechts mit einem Stack arbeitet. Er läßt sich folgendermaßen beschreiben (Heidorn 1972, 118):

1. Setze am Anfang das Startsymbol auf den Stack.
2. Nimm das oberste Symbol vom Stack und prüfe:
   (a) ist es ein Terminalsymbol, so gib es aus;
   (b) ansonsten wähle eine Regel, deren linke Seite das Symbol enthält, und setze die Symbole, die sich auf der rechten Regelseite befinden, auf den Stack. 3. Wiederhole Schritt (2), bis der Stack leer ist.

In Wirklichkeit ist der Algorithmus komplexer als hier angegeben, weil er nicht nur Symbole, sondern Segment Records verwendet. Auf diese Segment Records verweisen Pointer, die der Generierungsalgorithmus auf dem Stack verwalten muß. Er überprüft die Bedingungen in den Segment Records, um zu bestimmen, welche Regel angewandt werden kann. Wenn eine Regel ausgewählt wird, kreiert er neue Records.

3 Generierung von deutschen Sätzen

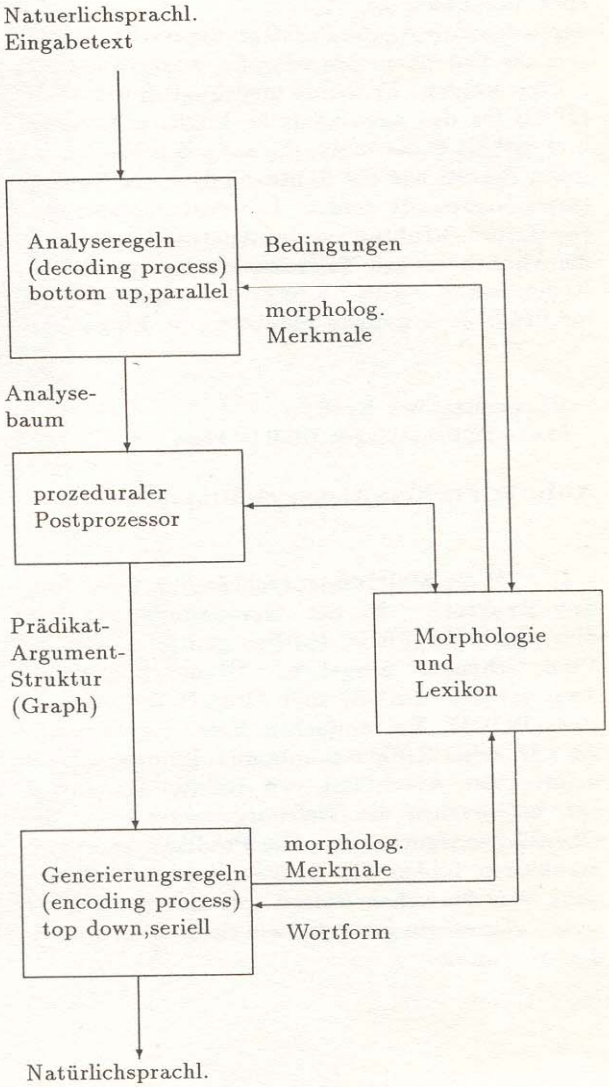


Abb.4: Das PLNLP-System

[3]Erklärung der Symbole:

```
DOBJ = DEEP OBJECT
-    = Löschen eines Attributes
       aus dem Record
%    = Kopieren eines Record
PRED = PREDICATE
=    = Pointer
```

Die Systemübersicht zeigt, daß PLNLP aufgrund seiner Zielsetzung über Resourcen verfügt, die zur linguistischen Realisierung, aber kaum zur Inhaltsfestlegung eines zu generierenden Satzes beitragen. Das System besteht aus einem Parser, der den Eingabesatz analysiert und die dazugehörigen Records erstellt, einem prozeduralen Postprozessor, der den Generierungsprozeß initialisiert, und der Generierung selbst, deren Funktionsweise in Abschnitt 2 erläutert wurde. Alle Komponenten haben Zugriff auf eine komplexe Lexikon- und Morphologiekomponente.

3.1 Prädikat-Argument-Strukturen als Ausgangsbasis für den Generierungsprozeß

Als Eingabe für den Generierungsprozeß dienen sprachunabhängige,
dependenzorientierte *Prädikat-Argument- Strukturen*, die von einem *prozeduralen Postprozessor* ermittelt werden. Er wurde ursprünglich von Jensen (1990) für das amerikanische Englisch entwickelt und enthält Prozeduren, die auf jeden Knoten und jeden Record aus der Syntaxanalyse des Eingabesatzes angewandt werden. Der Postprozessor weist der Record-Struktur aus der Analyse Attribute zu, die Ähnlichkeit mit Tiefenkasus oder funktionalen Rollen haben. Abb. 5 zeigt in Graphennotation die Prädikat-Argument-Struktur zum Eingabesatz

Hans schlägt den Hund.
hans←DSUB-schlagen-DOBJ→hund

Abb. 5: Prädikat-Argument-Struktur

Die VP des Matrixsatzes repräsentiert den Head der Struktur. In der Graphennotation wird dies durch die Pfeile deutlich gemacht, die vom Verb "schlagen" ausgehen. "Hans" fungiert als deep subject (DSUB) und "Hund" als deep object (DOBJ). Bei einfachen Satzkonstruktionen, die z.B. keine Reflexivpronomina, Infinitivkomplemente oder Anschlüsse mit Relativsatz aufweisen, entsprechen die Tiefenargumente exakt den Oberflächenargumenten. Die Prädikat-Argument-Strukturen bilden die Ausgangsbasis zur Erzeugung von deutschen Sätzen mit Hilfe von erweiterten Generierungsregeln, wie sie in Abb. 3 angedeutet wurden.

3.2 Generierungsregeln und die zugrundeliegenden Record-Strukturen

Welche Veränderungen die Generierungsregeln an den Records hervorrufen, die ihnen zugrundeliegen, soll an der Generierung der Passivversion zu dem aktiven Eingabesatz aus Abb. 5 demonstriert werden. Nach Eingabe des Satzes erfolgt automatisch die Syntaxanalyse.

Hans schlägt den Hund.

```
DECL1 NP1     NOUN1* "Hans"
      VERB1*  "schlägt
      NP2     ART1   ADJ1  "den"
              NOUN2* "Hund"
      PUNC1   "."
```

Die Berechnung der Prädikat-Argument-Struktur wird vom Benutzer durch den folgenden Befehl ausgelöst.

.graph < 1 >
hans ← DSUB − schlagen − DOBJ → hund
Um die Erzeugung des Passivsatzes zu aktivieren, weist der Benutzer interaktiv das Merkmal PASSIVE zu.

.+passive("schlagen")
Der Generierungsprozeß selbst wird durch den Aufruf

.encodeg< "schlagen" >
initialisiert und produziert den Satz

Der Hund wird von Hans geschlagen.

Abb. 6: Interaktion zwischen PLNLP und dem Benutzer

Abb. 7 zeigt einen Ausschnitt aus einer Regel, die bei der Generierung des obigen Satzes involviert ist, indem sie zu Verbformen im Präsens die passivierten Formen erzeugt.

```
(6185) VP(PASSIVE,^NONFINIT,...)
→ VERB(%VP,PRED='WERDEN',...)
VP(-PERSNUMB,+PSTPRT,-PASSIVE,...)
```

Abb. 7: Vereinfachte und modifizierte Generierungsregel

Regel Nr. 6185 ist folgendermaßen zu lesen: Wenn es eine VP gibt, die das Attribut PASSIVE besitzt und nicht infinit ist (^NONFINIT), transformiere diese VP in ein VERB und eine

VP. Alle Merkmale des VP-Record auf der linken Regelseite werden in den VERB-Record kopiert (%VP). Da zur Bildung des Präsens Passiv das Hilfsverb "werden" benötigt wird, wird dem Attribut PRED (predicate) über den Pointer "=" der entsprechende Wert zugewiesen. Die neue VP ist eine Kopie der Eingabe-VP. Die Kopierfunktion % muß hier nicht explizit angegeben werden, weil die Namen der Segment Records identisch sind. Das Attribut PERSNUMB, das Informationen zu Person und Numerus enthält, muß gelöscht werden, weil eine infinite Verbform gebraucht wird. Daher wird das Merkmal PSTPRT (past participle) hinzugefügt. Eine andere Regel steuert den Zugriff auf die Morphologiekomponente, die dann die endgültige Partizipform liefert. Um Endlosschleifen zu verhindern, muß noch das Passivmerkmal getilgt werden. Was die obige Regel bewirkt, zeigen die Veränderungen in dem Record, der ihr zugrundeliegt(siehe unten)

Eine ausführliche Erläuterung aller Attribut-Wert-Paare des Record ist an dieser Stelle nicht möglich. Zu erwähnen ist jedoch folgendes: Der Record zur Eingabe-VP weist ein Merkmal ATTRS (attributes) auf, das die beiden Werte DSUB und DOBJ enthält. Wie die Prädikat-Argument-Struktur aus Abb. 5 veranschaulicht, war das DOBJ des aktiven Eingabesatzes "Hund". Eine andere Regel (vgl. hierzu Beurer & Harriehausen-Mühlbauer 1991) hat aus diesem deep object bereits das Oberflächensubjekt des zu generierenden Passivsatzes erstellt. Das wird an dem Merkmal SUBJECT deutlich, das in allen drei Records vorhanden ist. Im VERB-Record wird der ursprüngliche Wert des Attributes PRED ("schlagen") durch "werden" ersetzt. Die endgültige Verbform ergibt sich später durch Zugriff auf die Morphologiekomponente. Das Attribut INDIC (indicators) besitzt u.a. den Wert P3, d.h. 3. Person Singular. Er gehört zu einer Reihe von Werten, die unter PERSNUMB zusammengefaßt werden. Da Regel (6185) die Tilgung von PERSNUMB vorsieht, wird P3 in dem neuen VP-Record gelöscht und durch PSTPRT ersetzt. Das Passivmerkmal am Ende der Records ist ein *boolesches Attribut*, das nur die Werte 1 (wahr) oder 0 (falsch) annimmt. Wie es zeigt, lassen sich Attribute in PLNLP nicht nur innerhalb der Regeln löschen oder zuweisen, sondern auch interaktiv durch den Benutzer. Diese Eigenschaft ist nützlich für die Generierung von Alternativen. Generierungsalternativen werden z.B. zur Erzeugung diverser Negationsformen und topikalisierter Konstituenten in Sprachen mit freier Wortstellung gebraucht, aber auch zur Generierung von verschiedenen Flexionsformen aus stilistischen Gründen. So kann z.B. in PLNLP über die Zuweisung des booleschen Attributes ALTVERBFORM in daß-Sätzen, die Bestandteil einer indirekten Rede sind, die indikative oder konjunktive Verbform generiert werden.

Er sagte ihr, daß er komme.

```
DECL1 NP1       PRON1*Er"
      VERB1*    "sagte"
      NP2       PRON2*ihr"
                PUNC1 ","
      DASSCL1   CONJ1 "daß"
                NP3   PRON3*  "er"
                VERB2*"komme"
      PUNC2     "."
```

Berechnung der Prädikat-Argument-Struktur[4]:

.graph< 1 >

er← DSUB-sagen-DIND → ihr

er2 ← DSUB-kommen ← DOBJ

Defaultmäßig wird die indikative Verbform generiert.

.encodeg< "sagen">

ER SAGTE IHR, DASS ER KOMMT.

Dem VP-Record des daß-Satzes wird das Attribut ALTVERBFORM zugewiesen.

.+altverbform("kommen")

.encodeg<"sagen">

ER SAGTE IHR, DASS ER KOMME.

Abb. 9: Interaktion zwischen PLNLP und dem Benutzer

4 Abschließende Bemerkungen

Der PLNLP-Formalismus und die bislang existierenden Systemkomponenten bieten einen ersten, rudimentären Ansatz für die stilistische Analyse von Texten, die innerhalb des didaktischen Fremdsprachenerwerbs zur Anwendung kommen kann. Damit unterscheidet sich PLNLP von Systemen wie TEXT und KAMP (vgl. Zusammenfassung), die Antworten im Rahmen von Dialogen generieren, und rückt im Einsatzbereich in die Nähe von Systemen wie z.B. PAULINE (Hovy 1988), das unter Berücksichtigung diverser Stilparameter verschiedene Beschreibungen derselben Situation liefert. Der PLNLP-Formalismus selbst be-

[4] DIND = Deep indirect object

(6185) VP→ VERB VP

VP:

COPYOF	schlagen
INDIC	P3 NN NA ND NG
SUBJECT	hund ⟸ ehemaliges DOBJ
TREENODE	DECL1 "hans schlägt den hund" 'SCHLAGEN'
PRED	'SCHLAGEN'
ATTRS	DSUB DOBJ
DSUB	hans
PASSIVE	1 ⟸ interaktiv zugewiesen

→

VERB:

COPYOF	schlagen
INDIC	P3 NN NA ND NG
SUBJECT	hund
TREENODE	DECL1 "hans schlägt den hund" 'SCHLAGEN'
PRED	'WERDEN'
ATTRS	DSUB DOBJ
DSUB	hans
PASSIVE	1 ⟸ boolesches Attribut

VP:

COPYOF	schlagen
INDIC	PSTPRT NN NA ND NG
SUBJECT	hund
TREENODE	DECL1 "hans schlägt den hund" 'SCHLAGEN'
PRED	'SCHLAGEN'
ATTRS	DSUB DOBJ
DSUB	hans

Abb. 8: Vereinfachte Attribut-Wert-Records, erstellt durch die Generierungsregel aus Abb. 7

ruht nicht auf einem der gängigen Grammatikmodelle wie HPSG (Pollard & Sag 1987) u.ä., sondern auf Phrasenstrukturregeln, die auf der Basis von Records erweitert sind. Konzeption und Zielsetzung des Systems beeinflussen die Auffassung von der Generierung. Sie wird als Umwandlung von Records in Zeichenketten verstanden und läßt Aspekte wie das Planen einer Äußerung, die Festlegung des Inhalts, die Repräsentation von Sprecherintentionen etc. unbeachtet. Was generiert wird, wird wesentlich durch die Zuweisung entsprechender Attribute beeinflußt, die sowohl über die Generierungsregeln als auch interaktiv erfolgen kann und woran sich die zukünftige Arbeit knüpft. Die interaktive Methode der Attributzuweisung muß durch einen Formalismus ersetzt werden, der selbsttätig anhand von festgelegten Kriterien, z.B. Fokusbestimmungen, die Erzeugung einer bestimmten Variante aktiviert. Die Repräsentation solcher Kriterien ist natürlich erst dann sinnvoll, wenn das System in der Lage ist, nicht nur isolierte Sätze, sondern ganze Texte zu analysieren und zu generieren. Hierbei müssen dann auch die bereits definierten, aber noch viel zu wenig genutzten semantischen Attribute miteinbezogen werden.

Danksagung

An dieser Stelle möchte ich besonders Bettina Harriehausen- Mühlbauer danken. Ihre Anregungen und konstruktive Kritik haben mir bei der Ausarbeitung des Vortrages, auf dem dieses Manuskript basiert, sehr geholfen.

Literatur

[1] Appelt, Douglas E. (1985). Planning English Sentences. Cambridge: Cambridge University Press.

[2] Beurer, Andrea (1991). "Syntaxbasierte Satzgenerierung: Entwicklung eines Prototyps in PLNLP." Universität Trier. Magisterarbeit.

[3] Beurer , Andrea & Harriehausen Mühlbauer, B. (1991). "Choose Your Desired Structure or Generation in PLNLP." Proc. International Conference on Current Issues in Computational Linguistics. Penang, Malaysia: University of Sains Malaysia, Juni 1991, 391 -400.

[4] Chanod, Jean-Pierre, Harriehausen Mühlbauer, Montemagni, B (1990). "Post-processing Multilingual Argument Structures." IBM Scientific Center. Unpublished Paper.

[5] Heidorn, George E. (1972). "Natural Language Inputs to a Simulation Programming System." Monterey, CA: Naval Postgraduate School, Tech. Rep. NPS-55HD72101A.

[6] Hovy, Eduard H. (1988). "Generating Language with a Phrasal Lexicon." In: McDonald, David D. & Bole, Leonard eds. Natural Language Generation Systems. Berlin: Springer, 353 - 384.

[7] Jensen, Karen (1990). "Post-syntactic Computation of Arguments and Anaphora." Yorktown Heights, N.Y.: IBM Research Center. Unpublished Paper.

[8] McKeown, Kathleen R. (1985).] Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge: Cambridge University Press.

[9] Pollard, Carl & Sag, Ivan. (1987).] An information-based syntax and semantics. Vol. I: Fundamentals. Stanford: CSLI Lecture Notes 13.

LOG übersetzt worden sind. Sie können nur auf den Argumentpositionen von Prädikaten vorkommen.

Zwei Beispiele sollen die Behandlung von erweiterten Merkmalstrukturen illustrieren. Sie sind bewußt einfach gehalten, um dem Leser das Verständnis der Arbeitsweise von UBS nicht unnötig zu erschweren. Selbstverständlich kann UBS auch mit komplizierten Merkmalstrukturen umgehen; diese werden einfach in ihre Bestandteile zerlegt, und die Transformation wird Stück für Stück durchgeführt.

BEISPIEL 1: Der Ausdruck `a#b` steht für alle Ausdrücke, die mit `a` oder mit `b` unifizierbar sind. Da es sich in beiden Fällen um Konstanten handelt, sind dies gerade die beiden Ausdrücke `a` und `b` selbst. Betrachtet werden soll nun ein Programm, das aus nur einer einzigen Klausel mit dem einstelligen Prädikat `test1` besteht. Das Prädikat `test1` soll dann wahr sein, wenn sein einziges Argument gleich `a` oder `b` ist. Das Quellprogramm lautet also:

```
test1(a # b).
```

Um das gewünschte Verhalten des Programms zu bewirken, wird es in ein reguläres PROLOG-Programm transformiert. Der Ablauf der Transformation ist ungefähr der folgende:

1. Der Ausdruck `a#b` in UBS wird in einen regulären PROLOG-Ausdruck übersetzt, und zwar in diesem Fall in eine ungebundene Variable:

   ```
   X
   ```

2. Es wird Code zum Berechnen des Ausdrucks `a#b` erzeugt, und zwar:

   ```
   (X = a ; X = b)
   ```

   (Dabei ist ; der ODER-Operator von PROLOG.)

3. Der Code wird in die ursprüngliche Klausel eingeführt, so daß das transformierte Programm nun wie folgt aussieht:

   ```
   test1(X) :- (X = a ; X = b).
   ```

4. Tatsächlich ist das entstehende Programm noch etwas komplizierter. Den ursprünglichen Prädikaten werden nämlich drei zusätzliche Argumente angefügt, die zur Behandlung der Negation, von Mengen und von Funktionswerten gebraucht werden, so daß das Zielprogramm letztendlich die folgende Gestalt hat:

   ```
   test1(X, NEG, SET, FUN) :-
           (X = a ; X = b).
   ```

BEISPIEL 2: Nun soll ein Programm mit Negation betrachtet werden, das den Parameter `NEG` gebraucht. Es soll die Ausdrücke erkennen, die nicht mit der Konstanten `c` unifizierbar sind. Es kann in UBS ausgedrückt werden als

```
test2(^c).
```

und wird transformiert in

```
test2(X,NEG,SET,FUN) :- neg_(NEG,X,c).
```

Die Negation ist in UBS mit dem Prädikat `neg_/3` als konstruktive Negation (Walinsky 1987) implementiert, die auf einer dreiwertigen Logik basiert. Das Prädikat `neg_/3` überprüft, ob seine beiden letzten Argumente miteinander unifizierbar sind. Bei der Auswertung der Negation kann es jedoch vorkommen, daß es nicht entschieden werden kann, ob zwei Ausdrücke miteinander unifizierbar sind ("true"- oder "success"-Fall) oder nicht ("false"- oder "fail"-Fall), nämlich im Zusammenhang mit ungebundenen Variablen ("undef"-Fall). Im letzten Fall wird das Paar `(X,c)` in einer offenen Liste aufbewahrt, die an die Variable `NEG` gebunden ist, um später noch einmal untersucht zu werden.

Die drei Fälle sollen nun noch anhand von Beispielen erläutert werden. Gleichzeitig wird mit ihnen gezeigt, wie Programme in UBS aufgerufen werden können.

ANFRAGE:	ANTWORT:	FALL:
test2(c,NEG,_,_).	no	"true"
test2(d,NEG,_,_).	yes NEG = _	"false"
test2(X,NEG,_,_).	yes NEG = [(X,c)—_]	"undef"

(_ ist eine sogenannte anonyme Variable in PROLOG.)

Die Negation von komplexen Ausdrücken, z.B. Merkmalstrukturen, ist nicht allein mit dem Prädikat `neg_/3` lösbar. Um sie zu ermöglichen, werden komplexe Ausdrücke mit Negation umgeformt in äquivalente Ausdrücke, die nur noch negierte Konstanten enthalten.

4 Die Implementation

Bei der Implementation von UBS sind die wichtigsten Konstrukte von HPSG integriert worden. Dabei traten aber auch einige Probleme auf. Hier ein kurzer Überblick über Datenstrukturen und Operationen in UBS:

1. Datenstrukturen

▷ Die grundlegende Datenstruktur im Formalismus von HPSG ist die MERKMALSTRUKTUR. Jede Art von linguistischer Information wird mit ihrer Hilfe beschrieben. Merkmalstrukturen können in UBS (analog zu GULP) als Merkmal-Wert-Paare in beliebiger Reihenfolge formuliert werden.

UBS Eine Unifikationsbasierte Sprache zur Implementation von HPSG

FRIEDER STOLZENBURG; MARTIN VOLK Universität Koblenz-Landau Institut für Computerlinguistik Rheinau 3-4 *5400* Koblenz 0261-9119-469 E-Mail volk@brian.uni-koblenz.de

1 Einleitung

In der Computerlinguistik gewinnen unifikationsbasierte Ansätze immer größere Bedeutung für die Formulierung von Grammatiktheorien. Denn die Benutzung der Unifikation erlaubt es erstens, eine Vielzahl von Phänomenen der Syntax und zum Teil auch der Semantik natürlicher Sprachen elegant zu beschreiben; zweitens läßt sich die Unifikation mit mathematischer Exaktheit definieren, so daß die linguistischen Theorien auch als Computerprogramme implementiert werden können. Die in diesem Artikel vorgestellte Arbeit beschäftigt sich mit dem Formalismus von HPSG ("Head-Driven Phrase Structure Grammar", Pollard & Sag 1987) und mit dem Problem, ihn vollständig zu formalisieren und zu implementieren. Die zu diesem Zweck einsetzbare Programmiersprache UBS ("Unifikationsbasierte Sprache") ist im Rahmen einer Studienarbeit (Stolzenburg 1991) entwickelt und in ARITY /PROLOG 5.1 (Arity Corporation 1988) implementiert worden.

Der Formalismus von HPSG ist vielseitig. Er umfaßt mehr Datenstrukturen und Operationen als bisherige Grammatikformalismen. Zu nennen sind insbesondere Negation, Mengen und allgemeine Funktionen, die zusätzlich zu den grundlegenden Strukturen, der Unifikation und den Merkmalstrukturen hinzukommen.

2 Der Ansatz in GULP

Heutzutage gibt es bereits eine ganze Reihe von Werkzeugen, die es ermöglichen, Grammatiken, die in einem unifikationsbasierten Formalismus beschrieben sind, mit dem Computer zu bearbeiten und an Sätzen der natürlichen Sprache zu erproben. Zu diesem Zweck sind formale Sprachen entwickelt worden, die als Computerprogramm in einer bestimmten Programmiersprache implementiert sind.

Die Verwendung einer solchen formalen Sprache zur Beschreibung von Grammatiken bringt jedoch einige Nachteile mit sich: Ein Anwender, der eine Grammatik entwickeln will, muß zunächst die Syntax und andere Eigenschaften der Sprache lernen. Außerdem ist er in seinen Ausdrucksmöglichkeiten auf die Elemente der formalen Sprache beschränkt. Solche Sprachen besitzen meist nur gerade die Funktionen und Datenstrukturen, die unbedingt zur Behandlung einer Grammatik nötig sind, aber sie besitzen nicht die Mächtigkeit einer allgemeinen Programmiersprache.

Der Ansatz von Covington (1989) versucht, diese Nachteile zu beseitigen: Covington hat die Syntax von PROLOG so erweitert, daß Merkmalstrukturen leicht formuliert werden können. Alle PROLOG-Möglichkeiten bleiben erhalten. Der PROLOG-Interpreter braucht nicht verändert zu werden. Covington nennt sein System GULP ("Graph Unification Logic Programming").

Beim Einlesen von Klauseln in PROLOG aus einer Datei werden Merkmalstrukturen, für die eine besondere Schreibweise eingeführt wird, in ein internes Format, sogenannte Wertlisten (englisch "value lists"), überführt. Dadurch wird die Unifikation von Merkmalstrukturen auf die in PROLOG schon vorhandene Unifikation von Termen zurückgeführt, ohne daß zusätzlicher Code in die Klauseln eingefügt werden muß.

3 Die Sprache UBS

Bei der Entwicklung der Sprache UBS sind die gleichen Absichten wie in GULP verfolgt worden. Es handelt sich bei UBS um eine Erweiterung von GULP, die dem Anwender alle Ausdrucksmöglichkeiten des Formalismus von HPSG zur Verfügung stellen soll. Zur Behandlung der in HPSG verwendeten Strukturen muß zum Teil aber zusätzlicher Code in die Klauseln von PROLOG-Programmen, die UBS verwenden, eingefügt werden. Die Details dieses Vorgangs sollen im folgenden erläutert werden.

Wird ein Programm in UBS geladen, so wird es Term für Term gelesen, dann transformiert und der Wissensbasis hinzugefügt, nachdem die HPSG-spezifischen Konstrukte in reguläres PRO-

> Eine LISTE ist eine möglicherweise leere Folge von Beschreibungen meist ähnlicher Objekte, in HPSG z.B. die Valenzpartner eines Verbs in den Angaben zur Subkategorisierung des Verbs. In UBS können PROLOG-Listen als Werte von Merkmalen auftreten.

> In einer MENGE werden wie bei Listen mehrere Beschreibungen zusammengefaßt. Im Gegensatz zu Listen spielt bei Mengen die Reihenfolge der Elemente keine Rolle. Außerdem kann es sein, daß mehrere gleiche Beschreibungen dasselbe Objekt bezeichnen, also identisch sind. Die Formalisierung der Unifikation von Mengen erfolgt in UBS mit Hilfe von surjektiven Funktionen. Die Unifikation von Mengen stellt eine Erweiterung des Begriffs der Unifikation von Merkmalstrukturen dar. Pollard und Sag (1987) liefern keine formale Definition; Hinweise liefern Büttner (1986) und Kapur und Narendran (1986). Die Unifikation von Mengen ist in UBS noch nicht vollständig implementiert worden.

> Der Formalismus von HPSG läßt die Einführung von (Zeichen-) TYPEN zu. Sie sind hierarchisch angeordnet: Zeichen allgemein sind vom Typ "sign". Sie werden eingeteilt in lexikalische Zeichen und phrasale Zeichen. Ein typorientierter Ansatz zur Implementation von HPSG wird von Franz (1990) vertreten. Er betrachtet HPSG als eine Theorie, die aus Zeichentypen und damit verbundenen Bedingungen (englisch "constraints") besteht. In UBS ist ein komplementärer Ansatz verfolgt worden: Er ist mehr unifikationsbasiert. Typen sind nicht explizit verfügbar, denn PROLOG kennt auch keine Typen. UBS stellt jedoch MAKROS zur Verfügung, mit denen man die Notation von Merkmalstrukturen abkürzen kann. Mit Hilfe dieser MAKROS lassen sich auch Typen simulieren.

2. Operationen

0 Die grundlegende Operation im Formalismus von HPSG ist die UNIFIKATION. Die bekannte Definition (Shieber 1986) wird in UBS erweitert auf Listen und Mengen. Die Unifikation kann in UBS beliebig innerhalb von Merkmalstrukturen durchgeführt werden.

0 Zusätzlich eingeführt wird die DISJUNKTION von Werten. Sie wird verwendet, um auszudrücken, daß ein Attribut mehrere mögliche Werte haben kann, der Wert also nicht eindeutig festgelegt ist (vgl. Beispiel 1).

0 Auch die NEGATION von Werten ist möglich. Ein negierter Wert bedeutet, daß an dieser Stelle stehen darf, was nicht mit diesem Wert unifiziert werden kann. Die Interpretation der Negation ist problematisch: Sie kann nicht als "Negation by Failure" realisiert werden, denn das würde im Zusammenhang mit ungebundenen Variablen zu unerwünschten Ergebnissen führen (vgl. Beispiel 2). Die Negation ist deshalb in UBS als konstruktive Negation implementiert worden.

0 Ferner kann auch die IMPLIKATION ausgedrückt werden. Sie wird in UBS wie in der klassischen Logik auf Negation und Disjunktion zurückgeführt.

0 Schließlich können Merkmalwerte durch FUNKTIONEN berechnet werden. Z.B. können zwei Listen durch die zweistellige Funktion "append" aneinandergefügt werden. Das Ergebnis dieser Funktion ist also die Konkatenation der beiden Listen und kann als Wert einem Merkmal zugewiesen werden. Der Gebrauch von allgemeinen Funktionen erlaubt es, praktisch jede beliebige Manipulation an Merkmalwerten vornehmen zu können. In UBS ist die Möglichkeit geschaffen worden, beliebige PROLOG-Prädikate einzubinden. Sie berechnen die Funktionen, die auch allgemeine Relationen sein können, und werden verzögert ausgeführt.

5 Ausblick

UBS ist bereits im Rahmen einer Diplomarbeit an der Universität Koblenz erfolgreich eingesetzt worden, um ein Grammatikfragment des Deutschen in HPSG zu formalisieren. Dennoch soll UBS in nächster Zukunft erweitert werden: So ist daran gedacht, Typen explizit zur Verfügung zu stellen. Das würde die Effizienz bezüglich Zeit und Speicherplatzverbrauch steigern.

Wir haben versucht, die vielschichtigen Überlegungen bei Entwurf und Implementation einer formalen Sprache zur Behandlung von HPSG - einer linguistischen Theorie - zu skizzieren. Weitere Untersuchungen sind notwendig, um die praktische Verwertbarkeit von UBS auszutesten.

Literatur

[1] The Arity /Prolog Language Reference Manual. Concord, Massachusetts: Arity Corporation, 1988.

[2] Büttner, Wolfram: Unification in the Data Structure Sets. In: Goos, G.; Hartmanis, J. (Hrsg.): Proceedings of the 8th International Conference on Automated Deduction, Oxford, 1986, 489-495. (LNCS 230) Berlin; Heidelberg: Springer, 1986.

[3] Covington, Michael A.: GULP 2.0: An Ex tension of Prolog for Unification-Based Gram

mar. (Research Report AI-1989-01) Athens, Georgia: The University of Georgia. 1989.

[4] Franz, Alex: A Parser for HPSG. (CMULCL-90-3) Pittsburgh: Carnegie Mellon University, Laboratory for Computational Linguistics. Juli 1990.

[5] Kapur, Depak; Narendran, Paliath: NP - Completeness of the Set Unification and Matching Problems. In: Goos, G.; Hartmanis, J. (Hrsg.): Proceedings of the 8th International Conference on Automated Deduction, Oxford, 1986, 470-488. (LNCS 230) Berlin; Heidelberg: Springer, 1986.

[6] Pollard, Carl; Sag, Ivan A.: Information Based Syntax and Semantics. Volume 1: Fundamentals. (CSLI Lecture Notes 13) Leland Stanford Junior University: CSLI. 1987.

[7] Shieber, Stuart M.: An Introduction to Unification Based Approaches to grammar. (CSLI Lecture Notes 4) Leland Stanford Junior University: CSLI. 1986.

[8] Stolzenburg, Frieder: UBS - Eine unifikationsbasierte Sprache zur Implementation von HPSG. Studienarbeit. Koblenz: Universität Koblenz-Landau. 1991.

[9] Walinsky, Clifford: Constructive Negation in Logic Programs. Dissertation. Oregon Graduate Center. 1987.

Eine grammatikbasierte Integration von Hypertext und wissensbasierten Systemen

KLAUS PRÄTOR
Universität Düsseldorf

Für die folgenden Überlegungen gab es einen theoretischen und einen praktischen Anstoß. Der erste besteht in der These, daß Datenbanktheorie und Expertensysteme ein genuines Anwendungsfeld der Linguistik darstellen - und zwar nicht nur, wo es um natürlichsprachige Abfrage oder die Grammatiken der benutzten Programmier- oder Abfragesprachen geht. Der Grund ist ganz allgemein die Sprachförmigkeit der diesen Systemen immanenten Strukturen. (vgl. Prätor 1990) Obwohl diese These vor einem sprach philosophischen Hintergund nicht sonderlich überraschen kann, gibt es doch nur ein geringes Bewußtsein von ihr in den angesprochenen Anwendungsfeldern. Gleichwohl schlägt die Sprachförmigkeit auf die Begriffsbildung durch. Datensätze werden entweder als Objekte mit gewissen Attributen oder nach dem Muster der Prädikatenlogik als Sachverhalte in Bezug auf einen oder mehrere Gegenstände betrachtet.

Der praktische Anstoß resultiert aus der hohen Akzeptanz von Hypertextsystemen bei den Anwendern von wissensbasierten Systemen auf medizinischem Gebiet. Da deren Entwicklung mein gegenwärtiges Aufgabenfeld darstellt, kann mich das nicht gleichgültig lassen. In der Tat bilden Hypertextsysteme - wegen des geringeren Strukturierungsaufwands, der bequemen Handhabungsweise und der besseren Anschlußmöglichkeit an traditionelle Formen der Wissensdarstellung - häufig eine überlegenswerte Alternative zu wissensbasierten Systemen.

Die teilweise vielleicht etwas zu schematische Gegenüberstellung geht von der ursprünglichen Intention der Expertensysteme aus, einen Experten zu simulieren, was sich augenfällig im Stellen von vielen Fragen äußert, und kontrastiert sie mit dem Werkzeugcharakter der Hypertextsysteme, die die Führung ganz klar beim Benutzer belassen. Sie ordnet die Systeme, obwohl sie die jeweiligen Leistungen deutlich erweitern, den Grundtypen der Datenbanken mit hochstrukturierten Informationen und der Verarbeitung gering strukturierter Fließtexte zu. Daraus ergibt sich sowohl die Leistungsfahigkeit der Datenbanken für die Selektion und Umstrukturierung von Daten wie auch als Kehrseite ihr außerordentlich hoher Bedarf nach Standardisierung und Strukturierung der Informationen. Im Vergleich zu ihnen geben sich die Hypertexte mit einer flexibleren Strukturierung zufrieden. Sie ermöglichen so einen besseren Anschluß an die traditionellen Medien des wissenschaftlichen Informationsaustausches, besonders also die Druckwerke, und damit an die aktuelle fachliche Diskussion, häufig ein Schwachpunkt von Expertensystemen. In der abschließenden Charakterisierung der Wissensrepräsentation in Hypertextsystemen als menschenfreundlich ist die eigentliche Provokation, in diesem Zusammenhang überhaupt von Wissensrepräsentation zu sprechen. Nach dem Sprachgebrauch der Informatik liegt diese hier nicht vor. Man sollte aber nicht vergessen, daß in einem allgemeineren Verständnis Texte nicht nur einen Fall, sondern geradezu das Paradigma von Wissensrepräsentation darstellen, daß sie lediglich nicht den gegenwärtig realisierbaren Möglichkeiten automatischer Inferenz zugänglich sind. Man kann sogar umgekehrt behaupten, daß die Rede vom Übergang von Datenbanken zu Wissensbanken nur insofern Sinn macht, als darin die Annäherung der zunächst beliebigen Datenstrukturen an die semantischen Struktur von Sätzen, und damit von Textelementen, zum Ausdruck kommt.

Daß die Möglichkeiten der Hypertextsysteme im Hinblick auf die Behandlung von Wissen nicht so bescheiden sind, wie man auf den ersten Blick annehmen möchten, zeigt sich daran, daß man einfache Expertensysteme, nämlich solche mit determiniertem Suchbaum, durch Hypertextsysteme ersetzen kann. In der Abbildung 2 wird ein Ausschnitt aus einem kommerziell erwerbbaren System zur Leberdiagnostik in seiner Baumstruktur dargestellt, das in seiner Funktionalität vollständig durch ein Hypertextsystem ersetzt werden könnte.

Das Prinzip der Nachahmung ist einfach. An jedem Entscheidungsknoten ist das System auf Informationen angewiesen. Im Hypertextsystem werden dem Benutzer Fragen gestellt und er muß die der richtigen Antwort entsprechende Taste drücken und gelangt so entweder an die anschließende Frage oder, wenn der Entscheidungsbaum abgearbeitet ist, zur Antwort. Möglich ist das, wie bereits gesagt, nur bei einem relativ einfachen Typus von Expertensystemen. In komplizierteren Fällen ist die Ersetzung nicht mehr möglich.

Wünschenswert ist dann aber eine wechselseitige Ergänzung der Leistungen. Im Einzelfall können

Expertensystem	Charakter	**Hypertextsystem**
Subjektcharakter	Benutzerrolle	Werkzeugcharakter
eher passiv	charakteristische Datenstruktur	eher aktiv
Regeln	charakteristische Operation	Link
Inferenz	Sytemtypus	Bewegung in Text
Datenbank	Schematisches Operieren	Textverarbeitung
leistungsfähig	Strukturierungserfordernisse	wenig ausgeprägt
sehr hoch	Anschluß an traditionelle Formen der Wissensdarstellung	geringer
schwierig	Wissensrepräsentation	gut
maschinenfreundlich		menschenfreundlich

Abb. 1 Gegenüberstellung von Hypertext und Expertensystemen

dafür eine Vielzahl unterschiedlicher Gründe sprechen, von denen nur einige wichtige genannt werden sollen. Geht man von der Seite des Hypertextes aus, so kann Bedarf bestehen nach - einer punktuellen Unterstützung durch ein Expertensystem, z.B. zur Artbestimmung innerhalb eines biologischen Handbuchs. Vergleichbares ist für Reparaturanleitungen (Störfälle) oder medizinische Texte (Krankheitsbilder) vorstellbar. - zur Orientierung in komplexen Hypertextsystemen. Das lost in hyperspace" - Phänomen wird in fast jeder einschlägigen Veröffentlichung erwähnt. Auch für dieses Problem können expertensystemartige Unterstützungen eine Hilfe bieten.

Umgekehrt läßt sich ein Expertensystem ergänzen durch - die Möglichkeit der Erläuterung des Hintergrundes von Fragen des Systems oder allgemein von verwendeten Begriffen. So kann z.B. die Methodik eines durchzuführenden Tests erklärt werden. - die Verzweigung in Handbücher, Quellenangaben, Aufsätze als Bindeglied zu den geläufigen Formen der Fachkommunikation. - die Darstellung von im Expertensystem nicht oder schwer repräsentierbarem Wissen, z.B. der Ablauf von Stoffwechselvorgängen oder von anderen komplexen Abläufen. Durch grafische Darstellungen können topologische Zusammenhänge repräsentiert werden. - Möglichkeiten der Wissensakquisition, eine Ergänzung von ganz zentraler Bedeutung. Texte können in einer vorläufigen Strukturierung bereits der Benutzung zugänglich gemacht werden und durch weitere Erschließungs-schritte in die eigentlichen Expertensystemkompo-nenten übergeführt werden.

Diesem Wunsch nach Ergänzung steht die Tatsache entgegen, daß wissensbasierte und Hypertextsysteme ganz unterschiedliche Darstellungsweisen verkörpern. Beide Strukturtypen liegen sozusagen völlig windschief zueinander. Der Wunsch nach wechselseitiger Ergänzung beruht ja gerade darauf, daß man keinen einfachen Weg sieht, die Eigenheiten des einen Systems durch solche des anderen wiederzugeben. Man kann einwenden, daß einige Systeme existieren, die eine solche Ergänzung darstellen. So gibt es zum Beispiel eine Kopplung der Expertensystemshell Nexpert Object mit Hypercard und das System Knowledge Pro, das eine Integration beider Systemarten verspricht. Ohne die Leistungsfähigkeit dieser Systeme in Zweifel zu ziehen und ohne hier in Einzelheiten zu gehen, ist doch anzumerken, daß häufig entweder das Maß der Integration nur gering oder die Hypertextidee nur unzureichend verwirklicht ist, daß mithin die spezifischen Strukturen und Möglichkeiten textlicher Information zumindest teilweise ein-geschränkt werden. Insbesondere Hypercard und seine Nachahmer orientieren sich in ihrem Design mehr an der Datenbank- als an der Textmetaph0rik.

Relationen und wissensbasierte Systeme

Selbst wenn man zufriedenstellende Systeme zur Verfügung hat, kommt man, um diese vernünftig benutzen zu können, nicht umhin, sich selbst ein

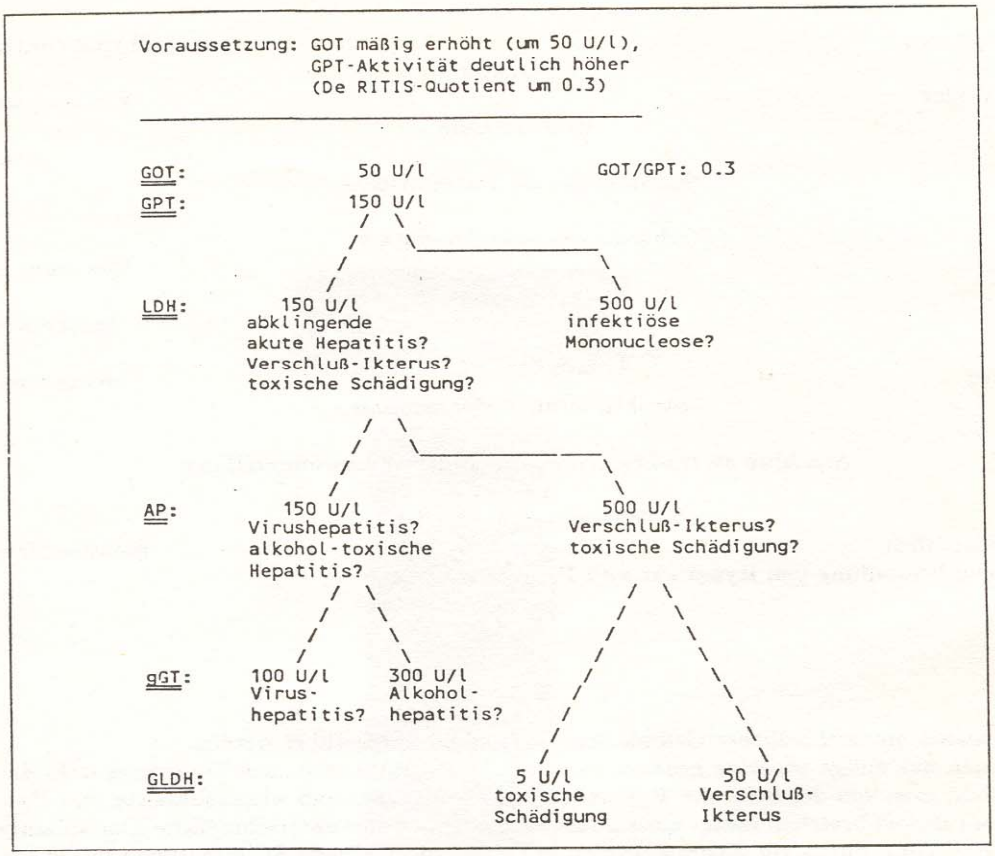


Abb. 2 Hypertext als Expertensystem

Bild von ihrem Zusammenspiel zu machen. Dazu muß erst ein gemeinsamer konzeptueller Nenner für ihre Integration gefunden werden. Als ein solcher wird hier ein relationaler Ansatz vorgeschlagen. Das Vorgehen besteht darin, zunächst zu fragen, wie man wissensbasierte Systeme und Hypertextsysteme in einem relationalen Modell darstellen kann und dann zu sehen, welche Möglichkeiten des Zusammenspiels beider Systemarten auf dieser Grundlage bestehen.

Ein relationaler Ansatz ist zur Zeit besonders im Bereich der Datenbanken vertraut, wo er derzeit den eindeutig dominierenden Typ darstellt. Sowohl die zu speichernden Objekte wie auch deren Beziehungen werden als Relationen abgelegt, die ersteren als Relationen ihrer Attribute. Die Darstellungsweise deckt sich völlig mit der in der Logikprogrammierung gebräuchlichen, deren Hauptvertreter gegenwärtig PROLOG darstellt. Die Prädikationen in der Prädikatenlogik stellen ein oder mehrstellige Relationen dar, die sich auch eins zu eins in relationale Datenbanken abbilden lassen.

Was die Logikprogrammierung und ganz generell auch die wissensbasierten Systeme von einfachen relationalen Datenbanken in technischer Hinsicht unterscheidet, ist die Verwendung von Regeln und der Einsatz von darauf beruhenden Schlußfolgerungen oder Inferenzen. Zweifellos stellen auch Regeln Relationen dar, nämlich Wenn-Dann-Beziehungen zwischen einem Antecedens und einem Succedens. Sie nehmen allerdings einen Sonderstatus ein, weil sie neben ihrer deklarativen Funktion auch einen prozeduralen oder operationalen Anteil haben. In ihrer wissensrepräsentierenden Funktion, d.h. in ihrem deklarativen Aspekt, lassen sich Regeln als Relationen darstellen. Das hat sogar den Vorteil, daß an die Stelle einer globalen Folgebeziehung inhaltlich qualifizierte Relationen treten können. Man kann dann beispielsweise unterscheiden zwischen Relationen, die Faustregeln bei der Lösung eines Problems darstellen, kausalen und semantischen Folgebeziehungen. Durch explizite Regeln wird dann nur noch bestimmt, wie der Schlußfolgerungsmechanismus das relational repräsentierte Wissen abarbeitet.

Relationen und Hypertext

Texte stellen eine völlig andere Datenstruktur dar. In der Regel werden sie in der Datenverarbeitung einfach als strings, als Zeichenketten wiedergegeben. Wählt man eine etwas höhere konzeptuelle

Pos	Wort		
1.1	Dies		
1.2	ist	100.1	Dies
1.3	der	100.2	ist
1.4	erste	100.3	eine
1.5	Satz	100.4	Erläuterung
1.6		100.5	zum
2.1	Dies	100.6	ersten
2.2	ist	100.7	Satz
2.3	der	100.8	
2.4	zweite		
2.5	Satz		
2.6			
3.1	Dies		
...			

Abb.3 Relationale Darstellung eines Textes

Ebene, so erscheinen sie als Listenstrukturen, als Listen von Wörtern etwa, oder auch in komplizierterer Form als Liste von Sätzen, die ihrerseits Wortlisten darstellen. Diese Struktur hat viele Vorteile. Sie eignet sich zum Beispiel gut zum Parsing. Für unserer Zwecke hat sie nur den Nachteil, von einer relationalen Darstellung, wie wir sie für die wissensbasierten Systeme erreicht haben, weit entfernt zu sein. Dabei ist es ein geringer Trost, daß Listen in KI-Sprachen wie LISP und PROLOG eigentlich als Relationen gehandhabt werden, nämlich als solche zwischen dem ersten Element der Liste und dem Rest. Auf dieser Basis wird die Gesamtliste rekursiv aufgebaut bzw. abgearbeitet. Dies ist aber eine sehr spezielle relationale Struktur mit der typischen Listeneigenschaft, nur element weise vom ersten Element her zugänglich zu sein. Demgegenüber hatten wir im Bereich der Datenbanken und wissensbasierten Systeme Gruppen von gleichförmigen, parallel angeordneten Relationen. Die Darstellung von Texten in dieser Form ist einfach und ausgefallen zugleich. In einer zweistelligen Relation werden die Wörter eines Textes und die jeweilige Position im Text aufgelistet, etwa wie das in Abbildung 3 dargestellt ist. Die Position könnte einfach durch fortlaufende Numerierung oder durch eine hierarchische Numerierung der Paragraphen, Sätze und Wörter erfolgen.

Diese Darstellungsweise hat hier nur den Status eines Denkmodells. Sie soll also nichts darüber besagen, wie eine hinreichend effektive Implementierung aussehen könnte. Versuche mit kleineren Texten in PROLOG, eine derartige Datei in gewohnter Gestalt am Bildschirm auszugeben, hatten allerdings recht gute Ergebnisse. Die Bezeichnung der Position kann schwierig werden, wenn Einfügungen und Streichungen sowie die Einbeziehung beliebig vieler und großer Dokumente möglich sein sollen. Hier existieren bereits Lösungen, wie sie z.B. von Nelson im Rahmen des Projektes XANADU ausgearbeitet wurden. (Nelson 1988) Mit diesen Methoden kann auch gleichermaßen auf einzelne Text

stellen (Wörter, evtl. auch Zeichen) zugegriffen werden, wie auch auf beliebige Textbereiche.

Ein angenehmer Nebeneffekt dieser Dateistruktur ergibt sich, wenn man den Index auf die zweite Spalte setzt, also statt nach den Positionen nach den Wörtern sortiert. Dann erhält man anstelle des fortlaufenden Textes den vollständigen Index des Dokuments, also die invertierte Datei.

Damit ist bisher nur die Umsetzung normaler Texte in die relationale Darstellung geklärt worden. Typische Hypertextbeziehungen sind entweder von assoziativer oder annotativer Art. Assoziative Beziehungen verknüpfen zwei Textstellen durch einen "link" in der Weise, daß von einer an die andere gesprungen werden kann. Annotative Verknüpfungen bringen beim "Anklicken" eines Textbereichs mit der Maus einen bisher unsichtbaren Textbereich dauerhaft oder zeitweise in den Blick. Beide Weisen können durch Relationen zwischen den Textadresssen, den Positionen der Wörter, repräsentiert werden, Eine assoziative Beziehung z.B. in prologähnlicher Schreibweise durch link (ausgangsadresse, zieladresse). Für eine annotative Beziehung ist die Basis gleichfalls eine zweistellige Relation zwischen Textstellen, nur muß in diesem Fall, nachdem der Sprung in den vorher unsichtbaren Textteil erfolgte, am Ende dieses Teils wieder in den Ursprungstext zurückgekehrt werden. Man kann die Verzweigung in den Hintergrundtext auch so auffassen, daß der Default- Wert für die nächste Textposition jeweils der fortlaufenden Numerierung entspricht, daß aber im Fall der Verzweigung statt des Defaultwertes ein anderer eingesetzt wird. (Wenn man hier bei Defaults an Frames denkt, ist man nicht auf der falschen Fährte)

Inhaltliche Strukturierung

Vergleichbar der relationalen Repräsentation von Regeln in wissensorientierten Systemen lassen sich auch die Relationen für die Hypertextlinks wie auch die verbundenen Textteile selbst inhaltlich qualifizieren. Ein Verweis auf das Literaturverzeichnis (Textrelation) kann als solcher gekennzeichnet werden und eine Fußnote oder eine Überschrift (Text teile) gleichfalls. Solche relativ formalen Textstrukturen können nahtlos in inhaltlichere übergehen, wenn etwa die Aspekte einer Literaturangabe (Autor, Titel,..) oder die festgelegten Teile einer bestimmten Vertragsform berücksichtigt werden. Selbstverständlich sind auch elaboriertere Qualifizierungen vorstellbar, aber es wird auch bei diesen Beispielen deutlich, daß durch die Aspektbildung Strukturen möglich werden, wie man sie von Datenbanken her gewohnt ist, daß der Übergang zu hochstrukturierten Datentypen mithin fließend ist.

Es wurde bereits angesprochen, daß sich bei dem vorgeschlagenen Modell der normalen Textdarstellung eine Anordnung in invertierter Form als Index und damit gleichzeitig als Wörterbuch des Dokuments gegenüberstellen läßt. Wie auf der Textseite eine inhaltliche Strukturierung durch zusätzliche Qualifizierungen denkbar ist, so können auf der Thesaurusseite Strukturierungen durch semantische Beziehungen der Wörter (Synonomie, Hyponomie,..) oder durch ihre Einordnung nach inhaltlichen oder funktionellen Gruppen erfolgen. Dies ist u. a. deshalb von Wichtigkeit, weil neben der Ergänzung durch Hypertexteigenschaften die Einbeziehung eines Thesaurus eines der wesentlichsten Desiderate wissensbasierter Systeme sein dürfte. Dies wird zwar nicht immer so gesehen, aber sobald mit unterschiedlichen Modulen und Anschlüssen an andersartige Informationsquellen gearbeitet wird, wird ein Thesaurus als Schnittstelle benötigt. Abgesehen davon ist in einem entwickelten Thesaurus ein großer Teil des über einen Welt ausschnitt vorhandenen Wissens inkorporiert.

Die für Dokument und Thesaurus unterschiedlichen Strukturen können auch übergreifend benutzt werden. So lassen sich beispielsweise Wörter durch ihre inhaltliche Einordnung im Thesaurus und zu gleich durch ihr Vorkommen in einer bestimmten inhaltlich oder funktional gekennzeichneten Textpassage charakterisieren, z.B. ein Symptom in einem Abschnitt zu einem bestimmten Schadstoff oder ein geografischer Name im Rahmen eines Literaturverweises. In jedem Fall gibt die Gesamtinformation hier mehr Information als die Summe ihrer Teile. Von Relevanz ist das in jedem Fall für die das Textretrieval, darüber hinaus aber auch der sprachlichen Analyse und der darüber möglichen Weiterverarbeitung von Wissensinhalten. Insgesamt eröffnet sich damit eine Perspektive, nicht nur wissensbasierte Systeme und Hypertext zu integrieren, sondern diese auch noch mit Thesaurus und Retrievalfunktionen verbinden zu können

Grammatiken für Textstrukturen

Die bisherigen Überlegungen zeigen, daß die Rede von der Unstrukturiertheit von Fließtexten nur aus dem Blickwinkel ihrer gegenwärtigen Behandlung in der automatischen Informationsverarbeitung richtig ist. Texte können über sehr ausgeprägte Strukturen verfügen, nur sind diese nicht gleichförmig genug, um sie mit den vergleichsweise groben Werk zeugen ihrer üblichen Computerhandhabung zu erfassen. Natürlich ist in vielen Bereichen ein Bewußtsein von diesen Strukturen vorhanden, beispielsweise im Bereich des Verlagswesens und Schriftsatzes. Dieser ist deshalb interessant, weil er eine Schnittstelle zwischen dem inhalt lichen und formalen Umgang mit Texten bildet. Hier haben sich Verfahren der Textauszeichnung entwickelt, Texte im Hinblick auf ihre grafische Gestaltung markieren. Dabei hat sich weitgehend die Praxis der inhaltlichen Auszeichnung durchgesetzt: Um Flexibilität bezüglich nachträglicher Änderung und mehrfacher Verwendung von Texten zu erreichen, werden nicht direkt grafische Attribute markiert, sondern funktionale oder inhaltliche. Dabei wird natürlich die konventionelle Textdarstellung verwendet und die Textcharakterisierungen werden als spezifische Zusatzzeichen in den Text eingebettet. Die Verfahren sind mittlerweile so elaboriert, daß man von Textauszeichnungssprachen sprechen kann, für deren Erzeugung und Prüfung auch formale Grammatiken existieren. Einen defacto Standard bildet heute die Standard Generalized Markup Language (SGML).

< *book* >
< *ti* > Organizational Burnout in Health Care Facilities
< *au* > Earl A. Simendinger < *deg* > Ph.D.
< *au* > Terence F. Moore
< *ehp* >< *no* > Chapter 1
< *cf* > Introduction to Organzational Burnout
..... Text ...< *fnr* > 1 < I > ... Text ...

< *h1* >Characteristics of Burned Out Organizations ... Text ...

< *h2* > Bickering
... Text... < *fn* >< *no* > 1 < *bb* >Health Service

Research
< *au* >Lloyd Connely< *au* > Dennis Pointer< *au* > Hirsh Ruschlin< *atl* >Viability and Hospital Failure: Methodology Considerations and Empirical Evidence< *7atl* >< *obi* >13 (Spring 1978): 27-36.< */fn*>
< */book* >

Abb. 5 Textauszeichnung mit SGML

Abbildung 5 zeigt markante Elemente des Dokumenttyps book. Die meisten Abkürzungen sind zu erraten:< ct > bezeichnet eine Kapitelüberschrift,< *fn* > eine Fußnote,< *fnr* > eine Referenz auf eine Fußnote, < *hl* > und< *h2* > eine Überschrift erster bzw. zweiter Ordnung, < *bb* > Bibliographie, < *atl* > Artikeltitel und < *obi*> weitere bibliographische Angaben. Das Ende einer Markierung wird durch< /... > gekennzeichnet oder ergibt sich eindeutig aus der folgenden Markierung.

Eben diese Textauszeichnungssprachen oder mark-up- languages sind gut geeignet, die im Zusammenhang eines relationalen Textmodells ange-

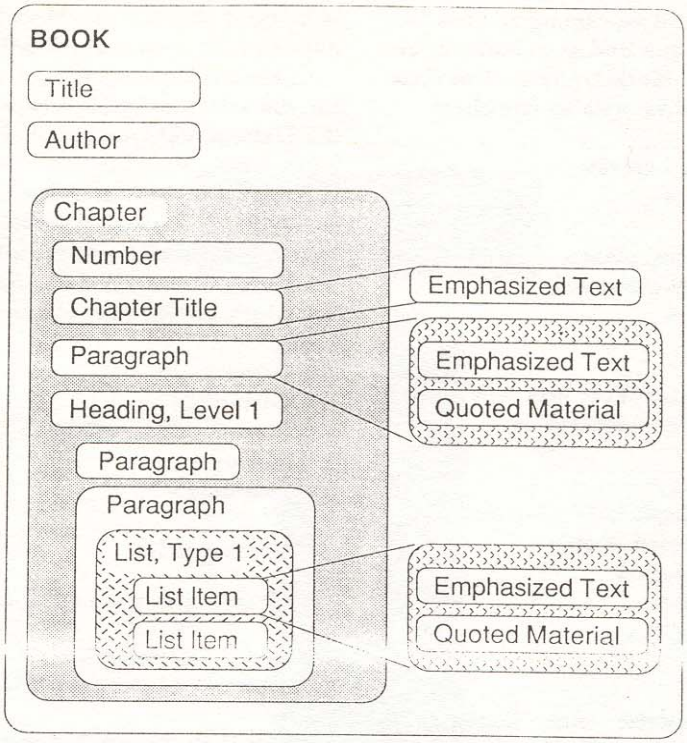


Abb. 4 Textstrukturen (nach Sperberg-McQueen1990)

sprochenen Qualifizierungen wiederzugeben. Zwischen beiden Modellen ist eine wechselseitige Abbildung möglich. Das heißt, daß auch diese Strukturierung flexibel hinsichtlich hoher und geringer Strukturierung ist. (Allerdings ist sie für hochstrukutrierte Daten nicht effizient.) Zu den üblichen Auszeichnungen gehören solche für hierarchisch geordnete Überschriften und die zugehörigen Textkörper sowie solche für Anmerkungen beziehungsweise Fußnoten. Beide können für die Umsetzung von annotativen Hypertextlinks benutzt werden Die jeweils untergeordnete Ebene kann relativ zur übergeordneten als zu verbergender Text betrachtet werden, der nach Belieben sichtbar gemacht und in sich natürlich wiederum verborgene Textteile enthalten kann. Anmerkungen würden in der Regel als temporär sichtbar werdende Textteile behandelt. Wo Auszeichnungsmöglichkeiten für Hypertexteigenschaften nicht standardmäßig vorhanden sind, lassen sie sich unschwierig hinzufügen. So schlägt die Text Encoding Initiative TEI (Sperberg-McQueen 1990) für assoziative Links, für Textsprünge also, eine Kodierung für Querverweise vor, die Einzeltextstellen und Textbereiche als Quelle und Ziel (auch in externen Dokumenten) handhabt, darüber hinaus Typisierung der Links (Qualifizierung) und die Verwaltung von Autor und Entstehungszeitpunkt der Links erlaubt. Textauszeichnungssprachen ebenso wie relationale Textstruktur erlauben die fast beliebige Ergänzung der Grundstrukturen durch jeweils für den Einzelfall zu definierende inhaltliche Kennzeichnungen.

Als ein weiterer Vorzug der mark-up languages erweist sich die Möglichkeit ihrer grammatikbasierten Erzeugung und Kontrolle. Abgesehen von der Nützlichkeit, die Korrektheit der Auszeichnung eines Textes zumindest formal kontrollieren zu können, eröffnen sich damit verbesserte Möglichkeiten der sprachlichen Analyse. An die Grammatik der Textstruktur kann sich eine in die in engeren Sinn sprachliche Struktur des Textes hineinreichende grammatische Analyse anschließen. Durch die Markierungen, besonders wenn sie inhaltlich angereichert sind, können zusätzliche Hilfen gegeben werden, die in vielen Fällen eine automatische Auswertung erst ermöglichen. Umgekehrt kann sprachliche Analyse auch die Auszeichnung der Texte erleichtern, indem sie, z.B. unterstützt von einem Thesaurus, nur in Zweifelsfällen eine explizite Markierung durch den Benutzer erforderlich macht.

Grundsätzlich hat eine Grammatik für Textauszeichnungssprachen zweierlei zu leisten. Zum einen hat sie die möglichen Ausprägungen eines Dokumenttyps zu kontrollieren und zum anderen die korrekte Anbringung der Markierungen. Das erste könnte im Fall eines Buches etwa so aussehen.

<book>→<title><author><text>
<text>→<chapter><text>
<text>→<chapter>
<chapter>→<chaptitle><chaptext>
<chaptext>→<paragraph><chaptext>
<chaptext>→<paragraph>

Eine inhaltlicher gefüllte Struktur ließe sich andeutungsweise folgendermaßen wiedergeben:

<Toxikologisches Handbuch>→
<Einleitung><Schadstoffkapitel> *[1]<Litverz>
<Schadstoffkapitel>→<Terminologie>
<allg.Stoffdaten><Toxikologie> .....
<Toxikologie>→<akute T.>
<chronische T.><spezielle T.>
...

Dabei wird üblicherweise eine kontextfreie Grammatik für die Erzeugung der Strukturen zugrundegelegt werden. Kontextfreie Sprachen erzeugen typischerweise lineare Zeichenketten. Eine interessante Alternative dazu bieten relationale Grammatiken (Heydthausen 1988, S. 73 ff.) Im Unterschied zu Chomsky- Grammatiken sind relationale Grammtiken nicht auf die Produktion linearer Zeichenketten, sondern auf die Generierung beliebiger Strukturen angelegt. Relationale Grammatiken erzeugen komplexe Gebilde durch die Beschreibung der Beziehungsstruktur ihrer Komponenten, die ihrerseits wieder komplex sein können. Interessant ist dieser Ansatz, weil er sich konzeptuell eng an den für die vorgetragenen Überlegungen zentralen Gedanken der relationalen Struktur anschließt, weil er semantischer orientiert ist als die Chomsky-Grammatiken und weil relationale Grammatiken auch die Erzeugung nichtlinearer Strukturen erlauben. Hypertexte sind aber per definitionem nichtlineare Datenstrukturen, auch wenn ihre einzelnen Sichtweisen jeweils linear darstellbar sind.

Grammatiken für hochstrukturierte Daten

Es klang bereits wiederholt an, daß die Möglichkeiten der Textmarkierung abgesehen von der Effizienz bis in Bereiche hochstrukturierter Informationen reichen, die üblicherweise mit relationalen Datenbanken, eventuell auch mit wissensbasierten Systemen gehandhabt werden. Da andererseits behauptet wurde, die Modelle der Textauszeichnungssprachen und der relationalen Textmodellierung seien ineinander abbildbar, liegt es nahe, auch für die relationalen Datenbanken eine Erzeugung der Datenstrukturen durch Grammatiken ins Auge zu fassen.

Dazu betrachtet man am besten eine hochstrukturierte Information in textlicher Darstellung, also etwa eine Tabelle mit physikalischen Daten oder ein Literaturverzeichnis. Diese würden mithilfe einer mark-up-Sprache etwas vergröbert so dargestellt:

<Litverzeichnis>
  <Liteintrag>
  <Autor><Titel><Biblio>
  </Liteintrag>
  <Liteintrag>
  .....
</Litverzeichnis>

Die Grammatik wäre etwa so geformt:

<Litverz>→<Liteintr> *
<Liteintr>→<Autor><Titel><Biblio>

Ersichtlich ist die Bedingung, daß so ein Übergang möglich ist, die Iteration von gleichartigen Elementen, also die Existenz einer listenförmigen Struktur. Es handelt sich also um einen Spezialfall einer textlichen Struktur. Ob sich eine listenförmige Struktur ergibt, hängt aber auch von der Größe und Art der gewählten Einheiten ab, von der Auflösung sozusagen. Wählt man als Grenzfall Wörter, Sätze oder Abschnitte so lassen sich leicht Listen und damit relational darstellbare Strukturen finden. Dies war der Trick bei der hier vorgeschlagenen relationalen Textdarstellung. Mit der Grammatik für das Literaturverzeichnis läßt sich auch eine Tabelle einer relationalen Datenbank darstellen, wobei die zweite Regel jeweils genau einen Datensatz erzeugen würde. In diesem Fall wird die interne Struktur einer Relation mithilfe der Grammatik abgeleitet. Es ist aber auch möglich, in ähnlicher Weise das Relationsgefüge einer Datenbank zu erzeugen. Ausgehend von der grammatischen Skizze einer Buchstruktur

<Buch>→<Titel><Kapitel> * <Litliste>
<Kapitel>→<Abschnitt> *
<Litliste>→<Liteintrag>
<Liteintrag>→<Autor><Titel><Biblio>

würde ein Gefüge von drei Tabellen (oder Listen von Datensätzen) entstehen, die aufeinander in der Weise bezogen sind, wie dies durch die Grammatik vorgegeben ist. (Abb. 6)

Hierarchien von textlichen Listen werden durch Hierarchien von Tabellen wiedergegeben. Dabei ist

[1] * ist eine Abkürzung für Wiederholbarkeit

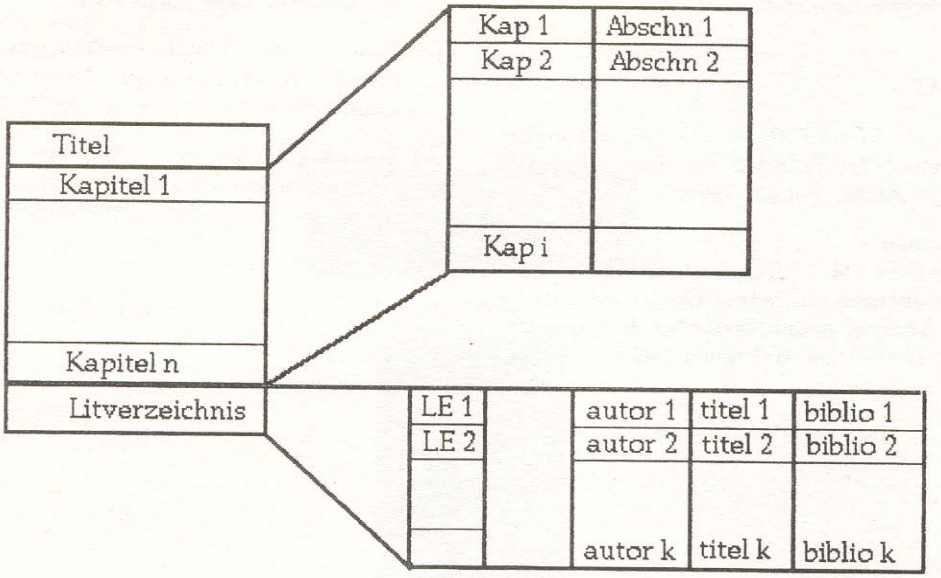


Abb. 6 Durch Grammatik erzeugte Dateistruktur

die Erzeugung der Tabellen mit Literatureinträgen (LE) nichtterminal. Die Datei wird nicht wirklich erzeugt. Sie ist nur ein Schritt auf dem Weg der grammatischen Produktion. Man sieht auch leicht, wie sich typische Hypertextstrukturen realisieren lassen: In sich verschachtelte Annotationen - Gliederungsstrukturen - bilden sich ab in eine Hierarchien von Relationen, die jeweils zwei Ebenen miteinander verknüpfen. Die so erzeugten Strukturen entsprechen teilweise denen, die in den Theorien der Normalisierung und Schlüsselbildung von Datenbanken oder auch in logisch-semantischen Datenbankmodellen behandelt werden. Neben dem konzeptionellen Brückenschlag ist dies ein zweites Argument, den in diesem Bereich wenig üblichen Gedanken der grammatischen Erzeugung weiterzuverfolgen. Ein weiterer Grund kommt hinzu: Datenstrukturen in dieser Perspektive zu betrachten, kann dazu beitragen, semantisch-syntaktische Aspekte stärker zu beachten und so den Übergang von einem datenorientierten zu einem wissensorientierten Ansatz zu befördern. Wissensbasierte Hypertextsysteme Die Übertragung des Gedankens der grammatischen Erzeugung auf hochstrukturierte Datensysteme findet ihre beste Entsprechung in der Frameidee, die eng mit semantischen Netzen und case grammars verwandt ist. Dieses Konzept unterscheidet sich nicht so sehr technisch von üblichen Datenobjekten - obwohl Differenzen in Form von Facetten und Triggern vorhanden sind -, sondern durch die Verknüpfung mit Konzepten der Wissensverarbeitung. Meines Erachtens ist es die semantisch-pragmatische Orientierung bei der Betrachtung der Funktionalität der einzelnen slots im Hinblick auf den Gesamtframe, die den wesentlichen Schritt dieses Konzepts hin zur Wissensverarbeitung ausmacht.

Betrachtet man ein Hypertextsystem in der referierten relationenartigen Darstellung aus dem Blickwinkel der Frameidee so erscheinen die textlichen Teile als besondere Slots. Diese können durch andere ergänzt werden, die den mithilfe der Textauszeichnungssprache angebrachten Zusatzqualifikationen entsprechen. Auf diese Typisierungen und auf die im Thesaurus niedergelegten Relationen stützen sich die Inferenzmöglichkeiten. Damit ist eine einheitliche konzeptuelle Darstellung der Wissenrepräsentation in Hypertext und wissensbasierten Systemen möglich. Mit geeigneten Werkzeugen läßt sich so ein integriertes System mit wissensbasierten und Hypertext-Elementen entwickeln. Die eigentlichen textlichen Teile sind nur für den menschlichen Benutzer interessant. Wegen ihrer geringen Strukturiertheit sind sie für die automatische Schlußfolgerung nicht zugänglich. Da sie aber in der vorgeschlagenen Strukturierung syntaktischer und lexikalischer Analyse gut zugänglich sind, besteht die Möglichkeit, ein derartiges System in Richtung auf höhere Strukturierung weiterzuentwickeln. Die zunehmende Aufbrechung textlicher Strukturen kann dabei als eine Form der Wissensakquisition betrachtet werden, die dem

Anschluß an die etablierten Formen der Wissensdarstellung entgegenkommt.

Literatur

[1] Carlson, D.A. und Ram, S., "HyperIntelligence: The Next Frontier", in: Communications of the ACM, Vol.33, 1990/3, S.311-321

[2] Heydthausen 1988
Heydthausen, M., "Diagnostische Sprache - Datenstrukturen und Algorithmen zur semantischen Analyse primärärztlicher Diagnosebezeichnungen," Diss. Hannover 1988

[3] Nelson, T. "Managing Immense Storage", in: BYTE Jan. 1988 S.225 - 242

[4] Prätor, K.M. "Die Sprachförmigkeit des Computers", in: Mitteilungen des Deutschen Germanistenverbandes, Jg. 37, 1990/3, S.15-19

[5] Sperberg-McQueen, C.M. und Burnard, L. (Hrsg.) "Guidelines For the Encoding and Interchange of Machine- Readable Texts", Chicago

Computergestützte Metalexikographie

Erfahrungen bei der Ermittlung der Wiederverwendbarkeit eines Wörterbuchs für maschinelle Sprachverarbeitung

MATTHIAS HEYN, OLIVER CHRIST UND ULRICH HEID Projekt *Polygloss*
Universität Stuttgart
maschinelle Sprachverarbeitung - Computerlinguistik, Azenbergstraße 12, D-W-7000
Stuttgart-1 e-mail: heid@informatik.uni-stuttgart.de

1 Einleitung

In den letzten Jahren wurde zunehmend über die Frage diskutiert, in welchem Umfang maschinenlesbar vorliegende Versionen einsprachiger Wörterbücher als Wissensquelle für sprachverarbeitende Systeme wiederverwendet werden können. Dabei wird davon ausgegangen, daß maschinenlesbar vorliegende Wörterbücher in ein für die weitere Verarbeitung geeignetes Format übersetzt *(reformatiert)* werden, und daß dann die für sprachverarbeitende Systeme notwendige Information extrahiert und in ein Wörterbuch für die betreffende Anwendung integriert wird. Dabei besteht theoretisch die Möglichkeit, entweder die Informationen der gesamten einzelnen Artikel des betreffenden Wörterbuchs für das sprachverabeitende System zugänglich zu machen (Transformation) oder selektiv auf Teilinformationen zuzugreifen (Extraktion). In beiden Fällen ist es notwendig, die lexikographische Beschreibung, die den Wörterbüchern zugrunde liegt, zu analysieren und zu *reinterpretieren,* d.h. die linguistische Beschreibungsintuition, die der Wörterbuchautor ausdrückt, zu rekonstruieren, damit die Angaben aus dem Papierwörterbuch für den theoretischen oder praktischen Ansatz nutzbar gemacht werden können, auf dem das jeweilige sprachverarbeitende System beruht.

Die Reformatierung und Reinterpretation von Wörterbuchartikeln sollte auf einer möglichst breiten Basis von Materialien des Ausgangswörterbuches aufsetzen. Einerseits muß die Struktur der Artikel des Ausgangswörterbuchs möglichst präzise erfaßt werden, damit eine halb- oder vollautomatische Reformatierung erfolgen kann. Andererseits erlaubt detailliertes Wissen über die in den Artikeln anzutreffenden Informationen und deren Präsentation erst die Abschätzung der qualitativen und quantitativen Nutzbarkeit des Wörterbuchinhalts für den angestrebten Zweck. Bevor eine Transformation oder eine Reihe von Extraktionsoperationen anhand eines gegebenen Wörterbuchs durchgeführt werden, sollte es also zunächst darum gehen, herauszufinden, ob die aus dem Wörterbuch extrahierbaren Informationen (für den angestrebten Einsatzzweck) adäquat und quantitativ relevant sind. Solche Untersuchungen fallen in den Bereich der Metalexikographie und sind in gewisser Weise der Wörterbuchkritik ähnlich: anhand des fertigen Wörterbuchs werden Struktur, Organisation und deskriptives Programm des Wörterbuchs beleuchtet. Der Unterschied zu üblichen Wörterbuchrezensionen liegt jedoch in der Repräsentationsform des Untersuchungsobjekts (maschinenlesbar), im Ziel der Untersuchung (der oben angesprochenen Überprüfung der Nutzbarkeit als Wissensquelle für NLP) und vor allem in den für die Untersuchung angewandten Methoden.

Ziel des vorliegenden Beitrags ist es, ein Beispiel für eine sehr detaillierte derartige "explorative" Untersuchung (anhand eines bestimmten Wörterbuchs) und - auf der Basis der dabei gesammelten Erfahrungen - Vorschläge für Arbeitsmethoden in diesem Bereich vorzustellen: wir haben das *Oxford Advanced Learner's Dictionary, 3rd, Electronic edition* mit computergestützten Methoden auf seine Verwendbarkeit als Wissensquelle für ein maschinelles Sprachgenerierungssystern untersucht. Wir konzentrieren uns hier auf die dabei angestellten methodischen Überlegungen. Wir geben zunächst einen Überblick über die Vorbereitung der metalexikographischen Untersuchung durch die Überführung des Wörterbuchs in ein in unserer Arbeitsumgebung leicht benutzbares Format. Dann gehen wir auf zwei Aspekte der metalexikographischen Untersuchung ein: zum einen auf die Struktur der Wörterbuchartikel und ihre Auswirkungen für die spätere Reinterpretation der Wörterbucheinträge, dann auf eine Typologie von Problemfällen für eine solche Reinterpretation. Der Nutzen dieser "Fehlertypologie" liegt vor allem darin, daß die Qualität der aus dem Wörterbuch ableitbaren Informationen abschätzbar wird.

Viele Wörterbuchrezensionen beruhen bisher auf der Untersuchung von Stichproben aus Wörterbuchtexten. In unserem Fall mußte jedoch eine nahezu exhaustive und sehr viel detailliertere Untersuchung des in Frage stehenden Wörterbuchs erfolgen: das gedachte Wiederverwendungsszena-

1 Wir danken Oxford University Press dafür, daß uns das Wörterbuch für Forschungszwecke zugänglich gemacht wurde.

rio betraf den Einsatz des von uns exemplarisch herangezogenen Wörterbuchs als Quelle von Information, die das Lexikon eines automatischen Textgenerierungssystems speisen sollte. Aufgabe der hier beschriebenen Arbeiten war es, eine konkrete Abschätzung des Aufwands zu liefern, der betrieben werden muß, um ein gegebenes maschinenlesbares Wörterbuch als Informationslieferant für die lexikalische Wissensquelle eines sprachverarbeitenden Systems aufzubereiten2.

Das *Oxford Advanced Learner's Dictionary* wurde dabei (seiner einfachen Verfügbarkeit wegen) als Beispielfall herangezogen. Wir diskutieren zwar im Folgenden die Detailergebnisse, die die Analyse dieses speziellen Wörterbuchs ergeben hat, jedoch liegt der Schwerpunkt der Aussage dieses Artikels auf den methodologischen Überlegungen und deren Interpretation vor dem Hintergrund der bisherigen metalexikographischen Diskussion, aber auch der Diskussion über die Wiederverwendung von machinenlesbaren Wörterbüchern für die Sprachverarbeitung3. Eines der "Nebenprodukte" der Untersuchung ist, daß auf ihrer Basis sehr detaillierte Regeln für die Umsetzung der elektronischen Edition dieses Wörterbuchs in ein für sprachverarbeitende Systeme direkt verarbeitbares Format formuliert werden können.

2 Vorbereitung der metalexikographischen Untersuchung

Das *Oxford Advanced Learner's Dictionary* 3 kann von Oxford University Press in der Form einer elektronischen Edition (OALD3e) erworben werden. Diese elektronische Edition ist mit einem SGML-ähnlichen Markup-System kodiert und wurde durch einen Parser aus dem Satzband des Wörterbuchs erzeugt. Dieser Parser benutzte den Text des Wörterbuches und die druckerspezifischen Anweisungen, um daraus die SGML-Version zu erzeugen. SGML (Standard Generalized Markup Language) ist eine deskriptive Auszeichnungssprache zur Beschreibung der Struktur von Texten oder Textbausteinen. Die Text Encoding Initiative (TEI) hat die Benutzung von SGML zur Beschreibung von Texten in Textcorpora und von gedruckten Wörterbüchern vorgeschlagen, und die Association of American Publishers (AAP) ist mit eigenen SGML-basierten Vorschlägen für die Auszeichnung des Texts gedruckter Wörterbücher gefolgt4. SGML-markierte Texte bestehen aus dem eigentlichen Text und einer *Document Type Definition* (DTD), die die Struktur des für den Text gültigen Markups festlegt und somit eine Art "Auszeichnungsgrammatik" darstellt. Im Falle des OALD3e fehlt eine solche Dokumenttypbeschreibung. SGML-Werkzeuge (wie z.B. SGML-Parser) standen in unserem Fall nicht zur Verfügung und sind ohnehin ohne zugehörige DTD für einen Text nicht anwendbar5. Um die Arbeit am OALD3e zu ermöglichen, haben wir deswegen den Text mit den üblichen UNIX-Werkzeugen in eine LISP-Notation überführt.

Diese Reformatierung ändert nichts an der Bedeutung der Repräsentation oder des Markups. Der große Vorteil der LISP-Repräsentation liegt in der einfachen Verarbeitungsmöglichkeit durch einen LISP-Interpreter, wodurch die Einheit der Repräsentations- und der Verarbeitungsmittel des Wörterbuchs durch LISP-Programme erreicht werden konnte. Auf dieser Repräsentationsebene konnten dann unter voller Berücksichtigung der durch das SGML-Markup erzeugten Artikelstruktur beliebig komplexe Such- und Filteranfragen in Form von LISP-Funktionen formuliert werden. Dieser Vorteil der Interpretierbarkeit der Artikelstruktur bei Suchanfragen mußte jedoch durch geringe Effizienz der Suchvorgänge erkauft werden. In den Fällen, wo die Eintragsstruktur für Suchanfragen keine große Rolle spielte, konnte das Wörterbuch sehr viel schneller mit gängigen Werkzeugen wie awk, `grep` und sed, sowie mit speziell für diesen Zweck erarbeiteten Werkzeugen auf PC-Basis bearbeitet werden.

Die auf LISP-, Unix-Tool- oder PC-Ebene entwickelten Suchmethoden wurden dann verwendet, um qualitative und quantitative Untersuchungen der Makro- und Mikrostruktur des Wörterbuches durchzuführen. Das Ergebnis einer solchen Suchanfrage ist in der Regel eine Liste von Beispielartikeln oder Teilen von Beispielartikeln, die ein bestimmtes untersuchtes -Phänomen illustrieren. Diese Listen konnten dann weiter gefiltert und/oder quantitativ ausgewertet werden.

Der wesentliche Vorteil der Computerunterstü-

2 Anders als z.B. im ACQUILEX-Projekt (vgl. [Cal et. a} 1990]) war hier nicht unbedingt a priori beabsichtigt, *alle* im Wörterbuch vorhandene Information in ein NLP-Wörterbuch zu integrieren (vollständige Transformation).

3 Vgl. die in der Zielstellung (Extraktion von Wissen aus einem maschinenlesbar vorliegenden Wörterbuch) ähnlichen Arbeiten von Boguraev und Mitarbeitem am LDOCE [Bog/Bri 1989] In Deutschland hat Lenders, [Len 1990] Arbeiten zum Duden Universalwörterbuch vorgelegt (Details sind in [Len 1991] dokumentiert worden); zu demselben Wörterbuch existieren auch weitere, detaillierte Untersuchungen: [Blä/Wer 1990].

4 Die Version 1.1 der *G1Lidelines* der TEI (vgl. [TEI 1991]) enthält einen sehr kurzen Abschnitt über Wörterbücher. Sehr viel detailliertere Vorschläge sind aber in Vorbereitung; einen kurzen Überblick gibt [Ide et. al. 92]. Es ist damit zu rechnen, daß sich SGML als Standardsyntax für die Markierung von linguistisch zu analysierendem Textmaterial durchsetzen wird. Die Benutzung von SGML für Wörterbuchtext wird auch im Abschlußbericht der Studie Eurotra-7 (vgl. [Hei/McN 1991]) empfohlen.

5 Versuche, eine DTD aus dem OALD3e zu konstruieren, sind aufgrund zahlreicher Hierarchisierungsfehler fehlgeschlagen (siehe auch Abschnitt 3.2).

tzung liegt bei diesem Vorgehen darin, daß bei jeder Suchanfrage das ganze Wörterbuch durchsucht werden kann. Dies bedeutet, daß *alle* Einträge des Wörterbuchs, die den Suchkriterien genügen, ermittelt werden können. Dies ermöglicht exakte quantitative Aussagen über die untersuchten Phänomene6.

3 Metalexikographische Untersuchungen

Bei unserer metalexikographischen Untersuchung des OALD3e stehen die Beschreibung des mikrostrukturellen Aufbaus der Wörterbuchartikel und der deskriptiven Qualität der Angaben im Vordergrund. Die Erforschung der mikrostrukturelIen Aufbauprinzipien von Wörterbuchartikeln führt zu Kenntnissen über die Geltungsbereiche der Angaben im Wörterbuch. Diese Kenntnisse ermöglichen die korrekte Zuordnung von Beschreibungen linguistischer Eigenschaften zu den beschriebenen linguistischen Objekten und sind deshalb Voraussetzung für die Reinterpretation der Daten. Strukturanalyse und deskriptive Untersuchung der Angaben sind unerläßlich für die Abschätzung des Aufwandes, der bei der Wiederverwendung der Daten betrieben werden muß.

Ausgangspunkt für unsere *mikrostrukturellen Analysen* sind die Methoden, die von \VIEGAND insbesondere in [Wie 1989a,b und 1991], vorgestellt wurden. Anhand von empirischen Untersuchungen zur Wörterbuchbenutzung hatte sich gezeigt, daß Wörterbuchbenutzer oft Schwierigkeiten beim Verständnis komplexer Wörterbuchartikel hatten. In aktuellen Benutzungssituationen konnten voneinander abhängige Textteile oft nicht einander zugeordnet werden. Um im Rahmen der metalexikographischen Forschung die textuellen Eigenschaften von Wörterbuchartikeln beschreiben zu können, die eine Ursache für solche Fehlbenutzungen sind, hat WIEGAND eine Beschreibungsmethode entwickelt, die den hierarchischen Aufbau von Wörterbuchartikeln aufdecken und in der Form von baumartigen partitiven Strukturgraphen sehr gen au darstellen kann. Bei diesen Analysen handelt es sich um die bisher detaillier testen Beschreibungen der hierarchischen Struktur von Wörterbuchartikeln und um eine umfangreiche Erfassung von in der deutschen Lexikographie vorkommenden Angabetypen.

In unserem Fall interessierte insbesondere der Aspekt der Geltungsbereiche von Angaben innerhalb einer komplexen Mikrostruktur , da die gegenseitigen Abhängigkeiten der Textelemente berücksichtigt werden müssen, damit die Artikel korrekt in die Zielrepräsentanten überführt oder aus ihnen jeweils zusammengehörige Teile vollständig extrahiert werden können. Dabei wurde von uns ein gegenüber WIEGAND modifizierter Analyseansatz entwickelt, der die Skopusrelationen - also die Abhängigkeiten von Artikelteilen - in den Vordergrund stellt.

Bei der *Erforschung der Angabetypen* ermöglicht der Computereinsatz neue Analysemethoden und damit Aussagen, wie sie der traditionellen Wörterbuchkritik nicht möglich sind (zum Beispiel quantitative Aussagen).

3.1 Die mikrostrukturelle Organisation der Wörterbuchartikel

Die Wörterbuchartikel des *Oxford Advanced Learner's Dictionary* sind komplexe hierarchisch organisierte Texte, die oft rekursive Eigenschaften aufweisen. Einem Hauptlemma kann eine bestimmte Definition zugeordnet sein; ihm kann außerdem in dieser Bedeutung ein Sublemma (z.B. ein Kompositum) zugeordnet sein, dem wieder eine eigene Bedeutungserklärung zugordnet ist usw.7

Bei dieser textuellen Organisation kann man prinzipiell zwei Arten von Angaben unterscheiden. Einmal Angaben, die die linguistischen Beschreibungen darstellen (z.B. eine Bedeutungsangabe) und zum anderen Angaben, die dem Zweck dienen, die Struktur eines Wörterbuchartikels, bzw. die Geltungsbereiche der Angaben anzuzeigen.

Angaben bestehen aus einem Angabetyp (z.B. Typ "Bedeutungsangabe") und einem Angabewert (z.B. dem Text der Bedeutungsangabe). Angaben, die linguistische Beschreibungen darstellen, lassen sich leicht in einer Attribut- Wert- Notation festhalten, wie sie in unifikationsbasiertierten Formalismen üblich ist.

Das Ziel unserer Analyse besteht darin, einige uns interessierende informationstragende Angaben in einer solchen Attribut- Wert-Notation zu isolieren. Um dies leisten zu können, müssen wir die strukturanzeigenden Angaben in der Weise (re-)interpretieren, daß die Geltungsbereiche der Angaben eine korrekte Zuordnung der jeweils relevanten Daten erlauben. Auf der Basis manueller

6 In den meisten metalexikographischen Untersuchungen, die ohne Computerunterstützung durchgeführt werden, ist es nicht möglich, eine exakte quantitative Abschätzung der beobachteten Phänomene zu liefern, da in der Regel manuell nur Stichproben aus einem Wörterbuch untersucht werden können. Hinzu kommt, daß bei der maschinellen Suche im gesamten Wörterbuch dem Linguisten auf einfache Weise umfangreiches Belegmaterial für einzelne Angabetypen zur Überprüfung bereitgestellt wird. Ein weiterer Vorteil liegt darin, daß die Suchanfragen inkrementell verfeinert werden können; außerdem können Anfragen und Suchergebnisse auf Sekundärspeichern des Rechnersystems abgelegt und für weitere Such-, Auswertungs- oder andere Verarbeitungsoperationen wiederverwendet werden.

7 Man redet bei dieser Zuordnung von Angaben oft auch davon, daß eine Angabe an eine andere "adressiert" ist (vgl. [Wie 1991, 442f.]) oder daß eine Angabe im "Skopus" einer anderen Angabe steht (vgl. [Wie 1991, 467f.]).

Reinterpretation stellen wir informelle Regeln auf, die der Ausgangspunkt für die Spezifikation eines entsprechenden Transformationsprozesses sind.

Die Analyse soll folgende Aufgabe erfüllen:

- ▷ komplexe Artikel sollen auf wenige wesentliche Struktureinheiten reduzierbar sein;
- ▷ der Geltungsbereich der Textteile soll explizit beschreibbar sein;
- ▷ für Artikel, deren mikrostrukturelle Organisation in der Realisation im Wörterbuch selbst nicht dem Wörterbuchkonzept entspricht („pathologische Mikrostrukturen"), soll sich der bei einer automatisierten Reinterpretation zu erwartende Informationsverlust beurteilen lassen;
- ▷ das Analyseergebnis soll Grundlage für die Reformatierung sein.

In einem ersten Schritt wird mittels der strukturanzeigenden Textelemente ein baumartiges Gerüst der Mikrostruktur des Wörterbuchartikels aufgebaut. In einem zweiten Schritt werden die in Attribut-Wert-Paar-Notation transformierbaren Textteile, die im Geltungsbereich eines Knotens im Baum der strukturanzeigenden Textteile liegen, in Attribut-Wert Paare aufgelöst und an diesen Knoten attributiert. Durch eine Top-Down-Analyse, die auf dem entstandenen Graphen operiert, werden alle Attribut-Wert-Paare des Artikeltextes aufgesammelt. Das Ergebnis dieses Prozesses nennen wir *Pfadinformationen* und die Analysemethode *Analyse in Pfadinformationen.*

Ein einfaches Beispiel einer solchen Analyse kann am folgenden Artikel zu dem Lemmazeichen *archer* vorgeführt werden[8]:

> **archer** /ˈɑːtʃə(r)/ *n* person who shoots with a bow and arrows. **arch-ery** /ˈɑːtʃərɪ/ *n* [U] (art of) shooting with a bow and arrows.

Dem Lemmazeichen *archer* sind eine Aussprachеangabe, eine Angabe zur Wortart und eine Bedeutungsangabe zugeordnet. Das von *archer* abgeleitete Wortbildungsprodukt *archery* ist in einem Untereintrag beschrieben, der seinerseits durch Angaben derselben drei Typen beschrieben ist. Zusätzlich findet sich bei dem Sublemma eine Angabe, die es der Klasse der unzählbaren (uncountable) Nomina zuweist.

Die entsprechende LISP-Kodierung des Artikels liegt wie folgt vor:

```
(ent :h "archer"
  (hwd "archer")
  (pr (ph  "\"A:tS@(r)"))
(hps :ps "n"
 (hsn
   (def "person who shoots
         with a bow and
         arrows")
   (cd
     (cp "arch|ery")
     (pr (ph  "\"A:tS@rI"))
     (cps  :ps "n"  :cu "U"
      (csn
      (def "(art of) shooting
      with a bow and
      arrows")))))))
```

Die LISP-Notation läßt sich leichter lesen, wenn sie in eine Baumnotation (vgl. Abbildung ??) umgeschrieben wird. Hier kann man leicht erkennen, daß der Artikelteil, der sich auf das Derivat bezieht (vgl. Subartikel in grauem Kästchen) nahezu die gleiche Artikelstruktur aufweist wie der Hauptartikel. Diesen rekursiven Artikel-Aufbau wird man sich sicher in einer Transformation zunutze machen, um (halbautomatisch) Subartikel aus den Artikeln herauszulösen.

Weiterhin kann man zwei verschiedene Arten von Knoten unterscheiden:

1. Knoten, deren Terminale linguistische Informationen tragen, wie z.B. `hwd, cp, ph, def`, bzw. Knoten, die mit linguistischen Informationen attributiert sind wie `hps` und `cps`;
2. Knoten, die die hierarchische Strukturierung des Artikels anzeigen wie `ent, cd, hsn` oder `csn`[9].

Die Knoten, deren Terminale linguistische Informationen tragen, können leicht in Attribut-Wert-Paare aufgelöst werden, die die strukturanzeigenden Knoten attributieren. Dabei sind umfangreiche Artikelanalysen notwendig, bevor Regeln festgelegt werden können, die die korrekte Zuweisung an die strukturanzeigenden Knoten zustande bringen. Ein einfaches Beispiel ist die Auflösung des `hwd`-Knotens in ein Attribut-Wert-Paar `[Form: Terminales Element]`, das dem Knoten `ent` zugewiesen wird. Als Attributnamen haben wir weitestgehend die Bezeichnungen übernommen, die im Projekt ACQUILEX eingeführt wurden (vgl. [**Cal et. al. 1990**]).

Eine Analyse des Beispielartikels führt zu dem in Abbildung ?? dargestellten Graphen.

Es zeigte sich bei unseren Analysen, daß funktionslose Knoten gelöscht werden können (z.B. `pr`) oder daß Knoten – aus Gründen, denen hier nicht weiter nachgegangen werden soll – "künstlich" eingeführt werden müssen, wie z.B. `SK` (für "semantischer Kommentar").

[8] Eine ausführliche Behandlung der Analysemethode mit ihren Handlungsschritten, Annahmen und zahlreichen Beispielen findet sich in [**Hey 1992**].

[9] Im Prinzip handelt es sich hier um die in der Metalexikographie getroffene Unterscheidung in strukturanzeigende und nicht-strukturanzeigende Textelemente.

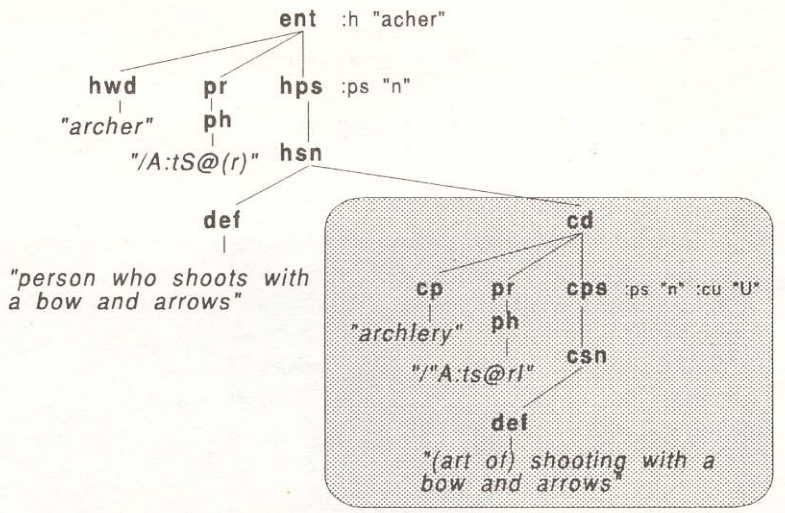


Abbildung 1: Hierarchie der Lisp-Kodierung des Artikels *archer*

In einem weiteren Arbeitsschritt können die zusammgengehörigen Attribut-Wert-Paare zusammengefaßt werden, wobei wiederum Regeln formuliert werden, die besagen, welche Informationen vererbt und welche überschrieben werden müssen. In dem vorliegenden einfachen Fall entstehen also zwei voneinander unabhängige Objekte.

Leider führt eine solche Analyse nicht bei allen Artikeln des OALD3e zu korrekten Ergebnissen. Die Ursache hierfür sind fehlerhafte Artikelstrukturen, deren Anteil an allen Artikeln des OALD3e wir mit circa 10% veranschlagen. Zu dieser Zahl kommen wir einmal aufgrund unserer umfangreichen manuellen Analysen und zum anderen aufgrund der Auswertung unseres ersten Versuches einer Transformation der Daten in eine objektorientierte Repräsentationsform.

Ein typisches Beispiel für solche Probleme, die wir *Hierarchisierungsfehler* nennen, ist der folgende Ausschnitt aus dem Artikel zu dem Lemmazeichen *air*:

```
(ent :h "air"
  ...
  (cd
    (cp "airway"...
      (def "route regularly
       followed by airliners")..)..)
  (id
    (ifg :gm "pl"
      (def "company operating a
            service of airliners"))))
```

Dem Hauptlemma (*air*) sind zwei Sublemmata zugeordnet. Ein Kompositum *airway*, und ein Pluraletantum, dessen Form implizit durch den Knoten `ifg` (= inflexional form group) eingeführt werden muß. Aufgrund der fehlerhaften Hierarchisierung wird aber bei einer automatischen Transformation unweigerlich die Form *airs* angesetzt werden. Die Bedeutungsbeschreibung zeigt, daß dies nicht gemeint sein kann. Eine richtige Hierarchisierung, die die Einbettung des zweiten Sublemmas in das erste Sublemma vornimmt und somit die korrekte Form *airways* erzwingt, müßte korrekterweise wie folgt aussehen:

```
(ent :h "air"
  ...
  (cd
    (cp "airway"...
      (def "route regularly
            followed by airliners")))
  (id
    (ifg :gm "pl"
      (def "company operating a
            service of airliners")...)))
```

Hierarchisierungsfehler können zahlreiche Ursachen haben. Oft fehlen den gedruckten Wörterbüchern strukturanzeigende Textelemente, die beim Parsen des Satzbandes benötigt würden, damit Einbettungen zustandekommen. Tiefere Einbettungen werden oft aus perzeptuellen Gründen nicht durch typographische Mittel ausgezeichnet, zumal der menschliche Benutzer implizite Zusammenhänge oft auch ohne typographische Unterstützung erfassen kann. Die Instruktionsbücher, die den Wörterbüchern zugrundeliegen sind oft in Details unpräzise. Handlungsanweisungen an den Lexikographen können fehlen oder mehrdeutig sein. Manchmal kommt es vor, daß sich ein Lexikograph nicht an die gesetzten Konventionen hält. Nicht zuletzt kann der Wörterbuchparser fehlerhaft konstruiert sein, was im vorliegenden Fall des OALD3e häufig die entscheidende Ursache für die Hierarchisierungsfehler war.

Besonders problematisch sind die Fälle, die sich – wie z.B. das obige Beispiel zum Artikel *air* – strukturell nicht von korrekten Artikeln unterschei-

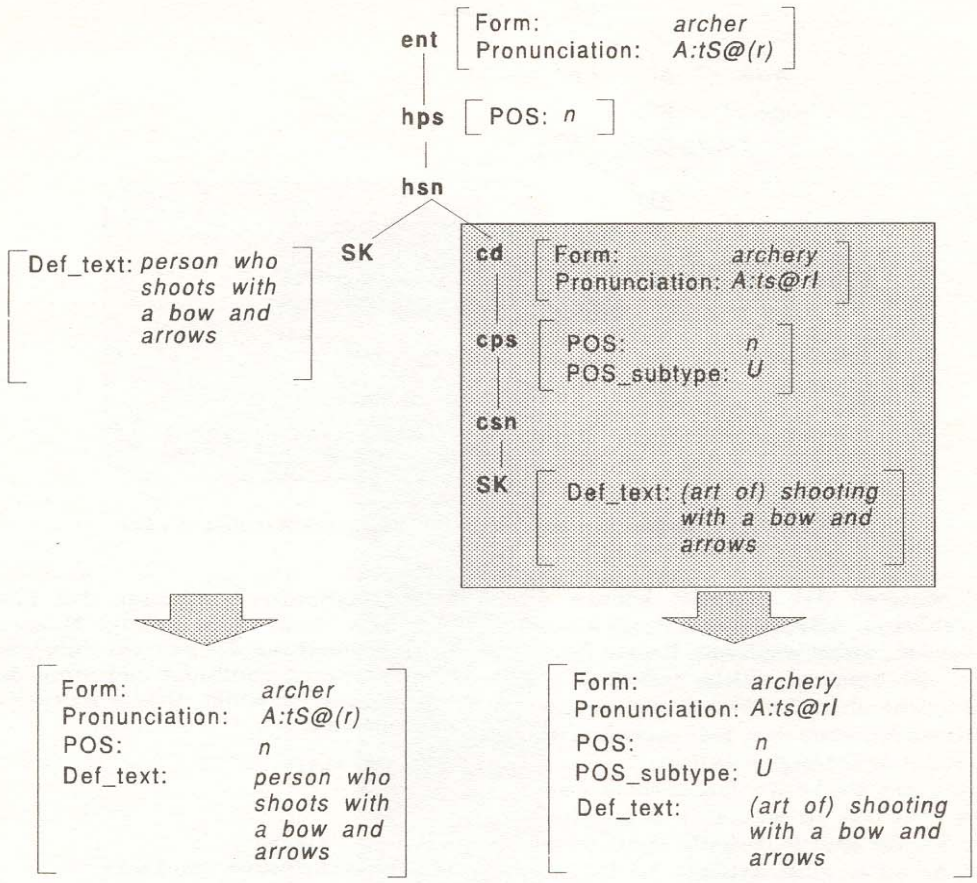


Abbildung 2: Pfadinformationen des Artikels *archer*

den. Daher sind solche "wohlgeformten", aber fehlerhaften Artikel nicht mit Hilfe des Computers erkennbar. Hieraus folgt aber unmittelbar, daß vollautomatische Extraktionen oder Transformationen nicht fehlerlos möglich sind.

3.2 Typen von inkonsistenter und fehlerhafter Information

Obwohl die Inkonsistenzen traditioneller Wörterbücher bekannt sind und oft kritisiert wurden, findet sich keine systematische Zusammenstellung von unterschiedlichen Typen dieser Inkonsistenzen. Im Rahmen einer computerunterstützten metalexikographischen Aufarbeitung des OALD3e erwies es sich als notwendig, unterschiedliche Problemfälle für die Reformatierung und Reinterpretation zu klassifizieren. Diese Klassifikation erlaubt es, die Problemtypen dahingehend zu beschreiben, ob sie maschinell erkannt werden können, und wie ggf. automatische Korrekturmaßnahmen aussehen könnten.

Wir unterscheiden folgende Typen:

▷ *Fehlen von Daten:* Häufig fehlen Angaben im Wörterbuch bzw. in der elektronischen Edition. Beispielsweise finden sich im OALD3e leere Knoten, die nur den Angabetyp, nicht aber den Angabewert spezifizieren.

```
(ent :h "awake" ...
  (hps :ps "vi ...
    (id) )))))
```

Im Artikel zum Lemmazeichen *awake* wird eine Mehrwortgruppe (`id`) eingeleitet, ohne daß ein entsprechender Textabschnitt folgt.

Zu jedem Verb wird eine Subkategorisierungsangabe in Form einer Liste von *verb patterns* erwartet. In 11% aller Fälle fehlt diese Angabe, wie der folgende Eintrag zeigt:

```
(ent :h "crucify" ...
  (hps :ps "vt"
    (hsn (def "put to death by
              nailing or binding
              to a cross")...
```

Das Fehlen von Daten ist mit dem Computer gut erkennbar, aber automatisch nicht korrigierbar. Je nach Umstand und An-

zahl der „Unterlassungssünden" des Wörterbuchs ist zu entscheiden, ob die Daten manuell rekonstruiert werden, ob eine vorläufige (Default-)Markierung die leeren Knoten ersetzen soll, oder ob auf die Information – zumindest für die vorliegende Quelle – ganz verzichtet werden soll.

▷ *Statistische Irrelevanz*: manche Angabewerte sind nur sehr selten im Wörterbuch vertreten. Vielleicht läßt sich argumentieren, daß dieser Sachverhalt deskriptiv linguistisch gerechtfertigt ist; das muß aber nicht gleichzeitig heißen, daß damit die Daten noch für eine Auswertung interessant sind. Meist liegen keine ideosynkratischen Sonderfälle vor, sondern Angabewerte, die bei der Wörterbucherstellung nicht systematisch für jede Bearbeitungseinheit überprüft wurden.

Die beiden prominentesten Vertreter dieses Problemtyps im OALD3e sind die Angaben zur Wortart und die Verteilung der *verb patterns*. Bei den 78 unterschiedlichen Angaben zur Wortart entfallen 98% auf die vier Hauptkategorien. Die verbleibenden 74 Klassen verteilen sich auf ganze 2%. 29 Angaben – wie z.B. `adverb_of_place_and_direction` – finden sich nur einmal im gesamten Wörterbuch. Es handelt sich folglich um unsystematische Ad-hoc-Beschreibungen, wohl ohne Absprache zwischen den am OALD arbeitenden Lexikographen. Ähnliches gilt für die 52 *verb patterns*, bei denen eine kleine Anzahl häufig vorkommender Vertreter einer großen Menge unterrepräsentierter Angabewerte gegenübersteht. Abbildung ?? gibt die Verhältnisse wieder.

Die statistische Relevanz von Angaben ist mit dem Computer erkennbar, aber es gibt keine Möglichkeit der automatischen Korrektur: Manuelle Korrekturen bedingen meist eine Reklassifizierung der Daten.

▷ *Polyfunktionalität*: wir sprechen von „polyfunktionalen Angaben", wenn *unterschiedliche Angabetypen unter gleichem Namen* vorgefunden werden.

Ein Beispiel ist die Angabe `:gm "pl"` im OALD3e, wobei `:gm` für *grammatical code* steht. Der Angabewert `"pl"` kann allerdings für zwei unterschiedliche Angabetypen stehen:

– *irreguläre Pluralbildung:*

```
(ent: h "bacillus"
  (hps :ps "n"
    (ifg :gm "pl"
      (if -"cilli"...
```

– *Pluraletantum:*

```
(ent :h "pliers"
  (hps :ps "n" :gm "pl" ...
```

Um die beiden Fälle auseinanderzuhalten, muß mehr Kontext hinzugezogen werden. So wird man die Einbettung des Angabewertes in die Angabe `ifg` (inflexional form group) beachten müssen.

Polyfunktionalität ist im Hinblick auf Wiederverwendung unerwünscht, da eventuell erst umfangreiche Inferenzen vorgenommen werden müssen, um die Daten korrekt zu kategorisieren. Solche Fälle sind dann aber mit dem Computer erkennbar und korrigierbar.

▷ *Gleichbedeutende Wertenamen*: dieser Fehler liegt dann vor, *wenn zwei unterschiedliche Wertenamen den gleichen Angabewert* benennen.

Der ausschließlich prädikative Gebrauch von Adjektiven wird im OALD3e dem ersten Anschein nach mit dem *grammatical code* `"pred"` ausgezeichnet:

```
(ent sick...
  ...(ifg :gm "pred"...
```

Eine Überprüfung der geläufigen prädikativen Adjektive wie *ill, fond, loath, afraid, ...* zeigt, daß diese Auszeichnung nicht konsequent vorliegt:

– *ill:* `(def "(usu pred)...`
– *fond:* `(un"pred only"...`
– *loath:* `(lab "pred only"...`
– *afraid:* `(hps :ps "pred_adj"...`

Beim OALD3e stößt man insbesondere bei den grammatischen Angaben sehr häufig auf diesen Problemtyp. Oft liegt dieser Problemtyp nicht so offensichtlich vor, wie beispielsweise bei den unterschiedlichen Angabewerten für die Angabe „3. Person Singular Präsens":

– `"3rd_person_sing_pres"`
– `"3rd_pers_sing"`
– `"3rd_pers,pres_t`
– `"3rd_p_sing_present_t"`
– `"3_rd_pers_pres_t"`

Daß verschiedene Wertenamen denselben Sachverhalt benennen, ist mit dem Computer nicht erkennbar. Erst nach einer manuellen und oft sehr aufwendigen Überprüfung können die fraglichen Wertenamen zusammengefaßt werden, so daß anschließend eine automatische Korrektur erfolgen kann.

▷ *implizite Angaben:* viele Angaben liegen im Wörterbuch nicht explizit vor. Beim Derivat *abomination* fehlt z.B. im OALD3e die Angabe zur Wortart. Der Lexikograph nimmt vom kundigen Wörterbuchbenutzer an, daß er mit dem Suffix des deverbalen Substantivs vertraut ist[10].

[10] Zumal das Sublemma als *uncountable* gekennzeichnet ist. Diese Angabe kommt nur bei Nomina vor.

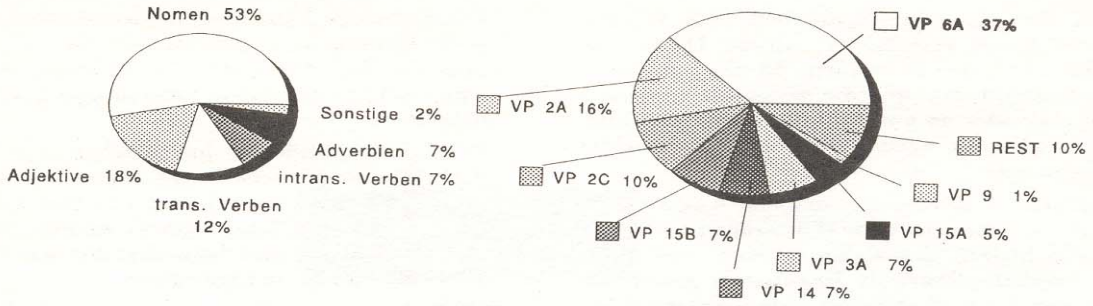


Abbildung 3: Verteilung der Angaben zur Wortart und zu *verb patterns* im OALD3e

Daß Angaben vom Wörterbuchbenutzer aufgrund anderer Angabetypen rekonstruiert, bzw. regelgeleitet ergänzt werden müssen, ist aus der Sicht der Wiederverwendung der Wörterbücher in NLP-Systemen unerwünscht. Solche Situationen sind oft nicht mit dem Computer erkennbar und können meist nur mit großem Aufwand automatisch korrigiert werden.

- *Deskriptive linguistische Mängel:* Probleme im Bereich der linguistischen Beschreibung sind natürlich mit automatischen Mitteln nicht erkennbar. Ihre Beurteilung wird traditionell von der Wörterbuchkritik geleistet. Der Vorteil computergestützter Methoden zur Aufdeckung solcher Mängel liegt vor allem darin, daß quer durch das Wörterbuch Einträge verglichen und aufbereitete Listen durchgesehen werden können. Als ein Beispiel aus dem OALD3e kann man an dieser Stelle die Subklassifikation der Pronomina anführen. Bei der Überprüfung der 13 Pronominalklassen zeigt sich, daß fehlerhafte Einordnungen vorliegen. In der Klasse der Personalpronomina finden wir *her, him, I, one, them, they,* aber nicht *you, we,....* Als Possesivpronomina sind nur *hers, mine* und *whose* klassifiziert; letzteres fehlt wiederum unter den Interrogativpronomina: diese Klasse besteht nur aus den drei Formen *who, which* und *whom.* Insgesamt kommt man zu dem Schluß, daß keine der 13 Klassen korrekt bearbeitet worden ist.

Tabelle ?? faßt die Ergebnisse unserer Klassifikation zusammen.

3.3 Das OALD3e auf dem Prüfstand

Bei der manuellen Analyse der Wörterbuchartikel des OALD3e in Pfadinformationen ergab sich, daß rund 10 Prozent aller Artikel pathologische Mikrostrukturen aufweisen. Da diese Fehler teilweise nicht maschinell erkennbar sind, wird auch auch eine rein maschinelle Reinterpretation nicht möglich sein. Der Aufwand, der bei einer halbautomatischen Reinterpretation notwendig wäre, wird von uns als (un-)verhältnismäßig hoch eingeschätzt.

Bei der Überprüfung der Angabetypen im Wörterbuch ergaben sich ähnliche Ergebnisse. Bearbeitet wurden von uns die folgenden Angabetypen:

- *phonologische Angaben:* Sie liegen mit auswertbaren Auszeichnungen vor und sind gut extrahierbar. Für rund 68% aller Sublemmata fehlt allerdings die Angabe zur Aussprache;
- *irreguläre Pluralbildung bei Substantiven:* Im Gegensatz zu vielen anderen Angaben sind diese Angaben problemlos auswertbar;
- *irreguläre Verbformenbildung:* Die meisten Angaben sind statistisch irrelevant, und daher sind nur wenige Formangaben auswertbar;
- *irreguläre Komparation:* Angaben sind nur für rund 10% der relevanten Fälle vorhanden;
- *attributiver vs. prädikativer Gebrauch von Adjektiven:* Da diese Angaben in hohem Maße unvollständig, polyfunktional und deskriptiv teilweise falsch sind, kann keine Auswertung erfolgen;
- *Wortartangaben:* von den 77 Klassen sind nur die vier Hauptkategorien auswertbar. Dies liegt an dem Vorkommen zahlreicher gleicher Wertenamen, vor allem dem hohen Anteil an statistisch irrelevanter Detailbeschreibungen und an deskriptiven Fehlern;
- *Subkategorisierungsinformationen:* Die Angabe von *verb patterns* fehlt bei 11% aller Verben. Da viele der *verb-pattern*-Typen sehr selten benutzt worden sind (statistisch irrelevant) und nach Stichproben festgestellt wurde, daß die Vergabe der *verb patterns* unsystematisch erfolgte, ist eine sinnvolle Auswertung sehr fraglich[11]. Nicht zuletzt wurden von Seiten der

[11] Es wurden in einigen Fällen sogar *verb patterns* vergeben, die gar nicht im Wörterbuchprogramm vorgesehen waren.

Problemtyp	Erkennbar mit dem Computer	Korrigierbar mit dem Computer
Fehlen von Daten	ja	nein
Statistische Irrelvanz	ja	nein
Polyfunktionalität	(ja) Inferenz	ja
Gleichbedeutende Wertenamen	nein	nach aufwendiger Aufbereitung
Implizite Daten	schwer	schwer
Deskriptive Fehler	nein	nein

Tabelle 1: Problemtypen beim Extraktionsvorgang

Wörterbuchkritik bereits zahlreiche Einwände hinsichtlich der deskriptiven Qualität dieses Angabetyps gemacht[12].

Desweiteren wurde von uns untersucht, ob an die reichhaltige Verweisklassifikation des OALD3e paradigmatische Relationen angeschlossen werden können. Hier zeigte sich, daß sich die Verweisklassifikation sehr stark an der (inkonsistenten) Typographie des Papierwörterbuches orientiert und somit keine weiteren deskriptiven Schlüsse erlaubt.

Die Ergebnisse der Untersuchung der elektronischen Edition der dritten Auflage des OALD zeigen eine Reihe von Problemen dieses Wörterbuchs, die unserer Meinung nach aber sinngemäß allen Wörterbüchern dieses Typs auftreten können. Ähnliche Ergebnisse sind für alle Wörterbücher der traditionellen Lexikographie zu erwarten. Dabei verstehen wir unter *traditioneller Lexikographie* eine Praxis, die ohne die konsistenzerzwingenden Mechanismen eines wie auch immer gearteten Computersystems für den Wörterbuchaufbau arbeitet.

Mittlerweile ist die vierte völlig überarbeitete Auflage des OALD als Papierversion erschienen. Dieses Wörterbuch wurde vollständig mit Unterstützung des Computers erstellt und ist inhaltlich, vor allem strukturell – wie wir durch zahlreiche manuelle Stichproben feststellen konnten – gegenüber dem Vorgänger wesentlich verbessert worden.

4 Zusammenfassung

Das *Oxford Advanced Learner's Dictionary 3* wurde in der in diesem Beitrag referierten Arbeit als Beispiel für ein Wörterbuch genommen, welches als Kandidat für die Extraktion von linguistischer Information für ein sprachverarbeitendes System in Frage kommt und auf seine Nützlichkeit für diesen Zweck untersucht werden soll. Da das Ziel der Arbeiten war, eine detaillierte Aussage zur Wiederverwendbarkeit einzelner Typen lexikalischer Beschreibungen aus diesem Wörterbuch machen zu können, wurde eine Methode entwickelt, mit der (die Struktur der) Wörterbuchartikel im Detail untersucht und Informationen aus ihnen „aufgesammelt" werden können. Die hier vorgestellte Analyse der Artikelstrukturen in Pfadinformationen ist eine für diesen Zweck geeignete Methode. Sie stellt direkt die Regularitäten heraus, deren spätere Implementierung als Transformationsregeln die kontrollierte Überführung komplexer Einträge aus der elektronischen Edition in eine andere Repräsentation, zum Beispiel in Objekte eines objektorientierten Systems oder in andere, partikularisierte Informationseinheiten, erlauben.

Das OALD hat sich in der von uns benutzen Fassung nicht als eine ideale Quelle für linguistische Beschreibungen für sprachverarbeitende Systeme erwiesen. Die Gründe dafür liegen unter anderem in dem relativ hohen Anteil problematischer Artikelstrukturen, beziehungsweise darin, daß in vielen Artikeln einzelne Textteile den verschiedenen Informationstypen nicht richtig zugeordnet werden können. Diese Situation geht ihrerseits auf zweierlei Gründe zurück, die völlig unterschiedlichen Bereichen zuzuordnen sind. Zum einen darf nicht vergessen werden, daß ein Wörterbuch wie das OALD für menschliche Benutzer konzipiert ist und mit grundsätzlich anderer Intention produziert wurde als derjenigen, linguistische Information für ein sprachverarbeitendes System extrahierbar zu machen. Die Arbeit in der traditionellen Lexikographie ist, soweit nicht interaktive Wörterbucherstellungs- und Dateneingabesysteme benutzt werden (wie dies bei der vierten Edition des OALD der Fall war), für Inkonsistenzen und „Freiheiten" der Lexikographen erheblich anfälliger, als dies aus der Sicht der Wiederverwendung für sprachverarbeitende Systeme wünschenswert ist. Der zweite Grund liegt darin, daß die von uns benutzte elektronische Edition von einem Parser produziert wurde, der an problematischen Stellen in den Wörterbuchartikeln zu inkonsistenten Ergebnissen gekommen ist.

Die Detailergebnisse, die im vorliegenden Aufsatz zusammengefaßt wurden, sind nicht nur als solche, sondern vor allem auch wegen der Methoden und Arbeitsmittel von Interesse, mit denen sie erzielt wurden. Die vorgestellte Arbeitsweise läßt sich auch auf andere Wörterbücher übertragen und erlaubt eine gezielte Aufwandsabschätzung dort,

[12] Vgl.: **[Akk 1989]**, **[Alt 1990]**, **[Hea 1982]**, **[Hea 1986]**, **[Her 1984]**, **[Her 1984]**, **[Lem/Wek 1986]**.

wo der Einsatz maschinenlesbarer Versionen von traditionellen Wörterbüchern als lexikalische Wissens quelle für ein sprachverarbeitendes System erwogen wird. Die Ergebnisse zeigen auch, daß jedem größerem Wiederverwendungsexperiment eine metalexikographische Untersuchung der potentiellen Quelle vorausgehen sollte. Entsprechend der Zielsetzung der geplanten Wiederverwendung kann auch selektiv überprüft werden, ob die Quelle die notwendige Information in der gewünschten Qualität bereitstellt. Elemente einer solchen Arbeitsmethode und Beispiele für die Anwendung der dazu notwendigen Werkzeuge wurden hier vorgestellt.

5 Literatur

[OALD3] Hornby, A.S. (Hrsg.): Oxford Advanced Learner's Dictionary, 3rd Edition. Oxford: Oxford University Press, 1974.

[OALD3e] Hornby, A.S. (Hrsg.): Oxford Advanced Learner's Dictionary, 3rd Edition, Electronic Version. Oxford: Oxford University Press, 1988.

[OALD4] Oxford Advanced Learner's Dictionary of Current English. Fourth Edition. Oxford: Oxford University Press 1990.

[Akk 1989] Akkerman, Eric. Independent analysis of the LDOCE grammar co ding system. In: [Bog/Bri 1989], SS. 65-83.

[Alt 1990] Altseimer, Susanne. Verbpatterns des OALD3e. Internes Papier. Projekt Polygloss. Universität Stuttgart: IMS-CL/IfI-AIS 1990.

[Blä/Wer 1990] Bläser, Brigitte j Wermke, Matthias: Projekt "Elektronische Wörterbücher j Lexika": Abschlußbericht der Definitionsphase. IWBS Report 145, Nov. 1990. Heidelberg: IBM Deutschland GmbH.

[Bog/Bri 1989] Boguraev, Branimir / Briscoe, Ted (Hrsg.): Computational Lexicography for Natural Language Processing. London, New York: Longman 1989.

[Cal et al. 1990] Calzolari, Nicoletta / Peters, Carol j Roventini, Adriana: Computational Model of the Dictionary Entry. Preliminary Report. Projekt ACQUILEX. Esprit Basic Research Action No. 3030. Pisa, April 1990. ILC-ACQ-I-90. 1990

[Chr 1990] Christ, Oliver. Die Nutzbarmachung eines maschinenlesbaren Standardlexikons durch Transformation in eine CLOS-basierte lexikalische Datenbasis. Studienarbeit Nr. 924 am Institut für Informatik der Universität Stuttgart, 1990.

[Hei/McN 1991] Heid, Ulrich / McNaught, John. Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of lexical and terminological resources in computerized applications. Final Report. Stuttgart / = Luxemburg: IMS-CL / Kommission der Europäischen Gemeinschaften, DGXIII B5, August 1991.

[Hea 1982] Heath, David. The Treatment of Grammar and Syntax in Monolingual English Dictionaries for Advanced Learners. In: Linguistik und Didaktik 49/50. 1982, SS. 95-107.

[Hea 1986] David Heath. Grammatische Angaben in Lernwörterbüchern des Englischen. In: Bergenholz, Henning und Mugdan, Joachim (Hrsg.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28-30.6.1984. (Lexicographica Series Maior 3). Tübingen, Niemeyer: 1985, SS. 332-345.

[Hey 1992] Heyn, Matthias. Wiederverwendung maschinenlesbarer Wörterbücher. Eine computergestützte metalexikographische Studie zur Wiederverwendung des Oxford Advanced Learner's Dictionary in NLP. Erscheint: Tübingen: Niemeyer 1992 (Lexicographica Series Maior).

[Her 1984] Herbst, Thomas. Bemerkungen zu den Patternsystemen des Advanced Learner's Dictionary und des Dictionary of Contemporary English. In: Götz, Dieter und Herbst Thomas (Hrsg.): Theoretische und praktische Probleme der Lexikographie. München: Hueber 1984, S. 139-145.

[Ide et. al. 1992] Ide, Nancy / Veronis, Jean / Warwick-Armstrong / Calzolari, Nicoletta: Principles of Encoding Machine readable Dictionaries. unveröff. Ms. 9. Seiten 1992 (eingereicht für EURALEX International Congress 1992, Tampere, Finnland 1992).

[Lem/Wek 1986] Lemmens, Marcel / Wekker, Herman. Grammar in English Learner's Dictionaries. Tübingen: Niemeyer 1986. [Lexicographica Series Maior 16].

[Len 1990] Lenders, Winfried: Semantische Relationen in Wörterbuch-Einträgen - Eine Computeranalyse des Duden-Universalwörterbuchs. In: Schaeder, Burkhard und Rieger, Burghard (Hrsg.): Lexikon und Lexikographie. Hildesheim, Zürich, New York: Olms, 1990, pp. 92-119.

[Len 1991] Lenders, Winfried: Überblick über den Forschungsstand auf dem Gebiet der lexikalischen Datenbanken und maschinenles-

baren Wörterbücher. In: IKP Arbeitsberichte, Abt. LDV Nr. 11, (Seitenangabe nicht möglich, da fehlerhafte Paginierung vorliegt).

[TEI 1991] Sperberg-McQueen, C. / Burnard, Lou (Hrsg.): Guidelines For the Encoding and Interchange of Machine-Readable Texts. Document Number TEI PI. Draft Version 1.1. Chicago, Oxford: TEL October 1990, 2nd Printing June 1991.

[Wie 1989a] Wiegand, Herbert Ernst: Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. Hrsg. von Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand und Ladislav Zgusta. Erster Teilband. Berlin / New York 1989 (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), SS. 462-501.

[Wie 1989b] Wiegand, Herbert Ernst. Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. Hrsg. von Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand und Ladislav Zgusta. Erster Teilband. Berlin/New York 1989 (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), SS. 409-461.

[Wie 1991] Wiegand, Herbert Ernst. Über die Strukturen der Artikeltexte im Frühneuhochdeutschen Wörterbuch. Zugleich ein Versuch zur Weiterentwicklung einer Theorie lexikographischer Texte. In: Ulrich Goebel und Oskar Reichmann (Hrsg.): Historical Lexicography of the German Language. Volume 2: Lewiston / Queenston / Lampeter: Edwin Mellen 1991. (Studies in German Language and Literature Vol. 6/Studies in Russian and German Nr. 3), SS. 341-672.

GTU - eine Grammatik Testumgebung mit Testsatzarchiv

MARTIN VOLK, HANNO RIDDER
Universität Koblenz-Landau Institut
für Computerlinguistik Rheinau 3-4
5400 Koblenz 0261-9119-469 E-
Mail volk@brian.uni-koblenz.de 14.
Januar 1992

Abstract: GTU ist eine Entwicklungs- und Testumgebung für Phrasenstrukturgrammatiken und ID /LP (Immediate Dominance/Linear Precedence) Grammatiken im Rahmen des Unifikationsparadigmas. GTU enthält eine morphologische Komponente, Lexikon und Testsatzarchiv für ausgewählte Probleme des Deutschen, sowie einen speziellen Editor zur Grammatikentwicklung und eine Ausgabe-Routine. In diesem Aufsatz diskutieren wir in erster Linie die Probleme beim Aufbau des Testsatzarchivs sowie die Stellung von GTU als linguistischer Lehrsoftware.

1 Einleitung

Für Linguisten ist der Computer zum unersetzlichen Werkzeug für Experimente mit den verschiedensten Theorien geworden. Besonders im Bereich der Syntaxanalyse natürlicher Sprache ist das Experimentieren mit Grammatikimplementationen in verschiedenen Formalismen zum integralen Bestandteil der Ausbildung geworden. Diese Experimente erfolgten traditionell mithilfe der Programmiersprachen LISP und Prolog. Dabei nimmt das Zusammenstellen des Lexikons und die Programmierung von übersichtlichen Ausgabestrukturen viel Zeit in Anspruch, die für die eigentliche Aufgabe - dem Erstellen einer Grammatik - nebenseitig sind. Darüberhinaus ist es oft mühsam, Testsätze zusammenzustellen, die ein Phänomen in der jeweiligen Sprache ausreichend beschreiben. GTU ist mit dem Ziel entstanden, diese Probleme zu lindern. Lexikon und Testsätze sind vorgegeben. Das Lexikon enthält z. Zt. rund 250 Stammformen, denen ein Vielfaches an Wortformen entspricht. Die Bestände des Lexikons sind auf die Testsätze abgestimmt. Der Benutzer kann also sofort mit der Formulierung der Syntaxregeln beginnen und muß diese dann lediglich mit Interface-Regeln an das Lexikon anbinden. Wie dies geschieht, wird im Folgenden erläutert.

2 Das Erstellen einer Grammatik mit GTU

Mit GTU kann man Grammatiken (Phrasenstrukturregeln im DCG Formalismus oder Regeln im ID /LP Formalismus) entwerfen und mit einfachen Interface-Regeln an das Lexikon anbinden. Zum Editieren der Grammatik steht ein Editor zur Verfügung, der spezielle Hilfefunktionen für die Grammatikentwicklung bereitstellt. Die Grammatikregeln können mit Merkmalstrukturen versehen werden. Die Syntax für die jeweiligen Formalismen ist flexibel anpaßbar an linguistische Konventionen.

Beispiel 1: Eine Grammatikregel für Präpositionalphrasen
PP[typ=P] -+ Prep[typ=P, kas=K] NP[kas=K].

PP, Prep und NP sind dabei Kategoriennamen, die Ausdrücke in den eckigen Klammern sind Merkmalstrukturen (Mengen von Merkmal-Wert-Paaren). Merkmalnamen und Bezeichnungen für Merkmalwerte können frei gewählt werden. Die Anzahl und Reihenfolge der Merkmale ist beliebig. Die Regeln können optionale Kategorien

und Terminalsymbole enthalten. Mit den Lexikon-Interface-Regeln legt der Benutzer fest, welche Information zu der jeweiligen Wortart aus dem Lexikon entnommen werden soll.

Beispiel 2: Eine Lexikon-Interface Regel für Präpositionen

```
If_in_lex (wortart=prep) then_in_gram
    Prep[ typ=#lemma, kas=#kasus].
```

Die Regel in Beispiel 2 besagt, daß für Präpositionen das Merkmal typ als Wert das jeweilige Lemma der morphologischen Analyse erhält, und daß das Merkmal kas den von der Präposition geforderten Kasus als Wert erhält. Die lexikalischen Regeln für die auftretenden Präpositionen werden dann während der morphologischen Analyse automatisch generiert.

3 Das Testsatzarchiv

Ist die Grammatik fertig editiert und geladen, muß überprüft werden, ob sie die erwarteten Strukturen erzeugt. Die Wichtigkeit der systematischen Überprüfung und Bewertung einer Grammatik kann leicht durch ein Zitat von Friedman (1989) motiviert werden: "Grammar writing is much more difficult than rule writing. The intricate interrelations of the individual rules of a grammar make grammar writing a complex and error-prone process, much like computer programming."

Zu diesem Zweck kann der Benutzer von GTU Testsätze manuell eingeben oder aus dem mitgelieferten Testsatzarchiv auswählen. Das Testsatzarchiv der GTU enthält rund 350 Testsätze in 15 Klassen. Der Aufbau des Testsatzarchivs geschah nach folgenden Kriterien:

I> Der Wortschatz der Testsätze kann beschränkt sein, da Methoden der Grammatikentwicklung eingeübt werden sollen und die Lexikonkomponente nur eine Servicefunktion hat.

I> Es gilt der Grundsatz *Vom Einfachen zum Komplexen,* da GTU als Lehr- und Lernsoftware konzipiert ist. Die Testsätze sollen aufeinander aufbauen, um eine inkrementelle Grammatikentwicklung zu fördern.

Die Testsätze der GTU wurden nach folgenden linguistischen Kriterien zusammengestellt: Zunächst gehen wir von einfachen Aussagesätzen aus, Nebensätze und andere Satztypen (Fragesätze) behandeln wir in fortgeschrittenen Lektionen. Innerhalb der Aussagesätze werden zunächst Kongruenzen (Subjekt-Prädikat, Artikel-Nomen) und Verbsubkategorisierungsrahmen variiert. Bei der Subkategorisierung sind anfangs nur obligatorische, später auch fakultative Mitspieler und freie Ergänzungen zu berücksichtigen. Im nächsten Schritt werden Verbformen und Nominalphrasen komplexer. Bei den Verbformen beginnen wir mit einfachen Vollverben, gehen dann zu zusammengesetzten Verbformen in Perfekt und Futur sowie Modalverb- und Kopulakonstruktionen. Nominalphrasen werden durch Adjektiv- und Genitivattribute erweitert. Danach kommen Besonderheiten des Deutschen wie abtrennbare Verbpräfixe, die verschiedenen Funktionen von *es* sowie Wortstellungsphänomene. Dann werden die bisher erarbeiteten Konstituenten durch Konjunktionen verbunden und zu komplexeren Konstruktionen zusammengefügt. Schließlich enthält das Archiv Satztypen, die andere Wortstellungen aufweisen als der Aussagesatz. Wir haben beispielhaft Fragesätze, daß-Sätze und Relativsätze ausgewählt.

Die Bereitstellung von vorbereiteten Testsätzen bietet eine Reihe von Vorteilen: 0 Die Testsätze können auf die Bestände des Lexikons abgestimmt werden.

0 Syntaktische Phänomene können systematisch aufgelistet werden.

0 Die Anzahl der Testläufe wird im Vergleich zu unspezifischen Eingabetexten minimiert.

0 Standardisierte Testsätze ermöglichen vergleichende Untersuchungen zwischen verschiedenen Formalismen bezüglich Effizienz und Kürze.

Bei der Einteilung der Testsätze in Problemklassen treten verschiedene Komplikationen auf: Die beschriebenen Phänomene sind mit zunehmender Komplexität schwieriger zu isolieren. So ist die Wortstellung ein Phänomen, das sich durch alle Testsätze zieht, aber von uns erst da aufgegriffen wird, wo es um variable Wortstellung geht (z.B. Mittelfeldanordnung der Nominalphrasen im Deutschen). Andere Phänomene, die wir ganz zu Anfang behandeln, treten im fortgeschrittenen Lektionen in anderer Form wieder auf (z.B. Kongruenz zwischen Artikel und Adjektiv in einer NP bzgl. der Adjektivdeklination). Man kann ein Phänomen nicht erschöpfend abdecken, ohne anderen Phänomenen vorwegzugreifen. Dennoch muß eine Einteilung erfolgen, damit ein schrittweises Vorgehen möglich ist.

Außer grammatisch korrekten Sätzen müssen auch ungrammatische Ketten aufgenommen werden, um eine Übergenerierung der Grammatik zu erkennen. Die Konstruktion der ungrammatischen Ketten kann spezifischer für das aktuelle Phänomen erfolgen als die Wahl eines grammatisch korrekten Satzes. Man wählt die Kette derart, daß ihre Ungrammatikalität lediglich von dem gewünschten Phänomen abhängt.

Beispiel 3: Eine ungrammatische Kette bezüglich Artikel-zu-Adjektiv Kongruenz

```
*Peter sieht das schönes Haus.
```

Darüber hinaus ist die Wahl der ungrammatischen Ketten abhängig von dem zu bearbeitenden Grammatikformalismus. Ist beispielsweise eine Grammatik im Rahmen des **ID** /LP Formalismus zu erstellen, so sind ungrammatische Ketten, die Wortstellungsregularitäten überprüfen, wichtiger als beim DCG Formalismus. Das ist darin begründet, daß beim ID/LP Formalismus die LP-Regeln explizit die Wortstellung fixieren, während die Wortstellung bei DCG implizit in den Regeln festgelegt ist.

Damit das Testsatzarchiv auch für große Sammlungen operational bleibt, müssen folgende Kriterien bedacht werden:

- Das Testsatzarchiv muß modular aufgebaut und leicht erweiterbar sein.
- Die Gründe für die Aufnahme eines Testsatzes in das Archiv müssen dem Benutzer transparent gemacht werden.
- Der Benutzer muß verschiedene Sichtweisen zur Verfügung haben, um sich auf unterschiedlichen Ebenen einen Überblick über das Archiv verschaffen zu können.
- Der Benutzer muß nach verschiedenen Gesichtspunkten Testsätze (auch über die gewählte Klasseneinteilung hinaus) auswählen und testen können.
- Protokollierung und Auswertung der Testergebnisse müssen dem Benutzer übersichtlich dargeboten werden. Das umfaßt die Präsentation der jeweils erfolgreich und fehlerhaft akzeptierten sowie abgelehnten Testsätze.

GTU erfüllt die meisten dieser Bedingungen. GTU unterstützt die Erweiterbarkeit des Testsatzarchivs um weitere Testsätze und Testsatzklassen. Zu diesem Zweck steht eine Musterdatei zur Verfügung, die das Eingabeformat für neue Testsätze erklärt. Die Datei wird automatisch geladen, wenn eine neue Testsatzklasse angelegt wird. Um zu dokumentieren, warum ein Testsatz aufgenommen wurde, wird jedem Testsatz eine Liste von Keywords mitgegeben, die auf ausführliche Begründungen verweisen, die an anderer Stelle abgelegt sind. Die Sammlung der Begründungen gibt einen Überblick über die behandelten Phänomene. Der Benutzer kann das GTU- Testsatzarchiv auf verschiedene Art einsehen. So kann er ein Verzeichnis aller Testsatzklassen, einzelne Testsatzklassen oder einzelne Testsätze mit ihrer jeweiligen Begründung einsehen. Übersichtliche Menüs erleichtern dabei die Auswahl. Mit diesen Optionen wählt der Benutzer auch Testsätze aus und übergibt sie dem System zur weiteren Verarbeitung. Protokollierung und Auswertung der Testergebnisse sind derzeit noch nicht implementiert.

4 Verarbeitung der Testsätze

Nach der Auswahl der Testsätze werden diese nacheinander abgearbeitet. Dazu wird zunächst eine morphologische Analyse sämtlicher Worte des Eingabesatzes durchgeführt und die entsprechenden lexikalischen Regeln werden erzeugt. Dann kann die eigentliche Syntaxanalyse beginnen. Grammatiken im DCG Formalismus werden dem Prolog Top-Down Left-To-Right Parser übergeben. Grammatiken im **ID** /LP Formalismus werden durch einen Bottom-Up Left-To-Right Chart-Parser abgearbeitet.

Bei erfolgreicher Analyse wird automatisch eine übersichtliche Baumstruktur erzeugt, ohne daß der Benutzer spezielle Parameter in seine Grammatik einbauen muß. Diese Ausgabestruktur enthält die Kategorienamen und Terminalsymbole. Danach wird in eingerückter Form die gesamte Information jeder Konstituente ausgegeben. Das heißt, die komplette Merkmalstruktur für jede Konstituente wird dem Benutzer dargeboten, damit er kontrollieren kann, ob die gewünschte Struktur erzeugt wurde. Ist ein Satz syntaktisch mehrdeutig, werden alle möglichen Strukturen erzeugt. Wird die Eingabe von der Grammatik nicht akzeptiert, erscheint eine entsprechende Meldung. Der Benutzer kann die Analyse einsehen, indem er die morphologische Analyse oder auch den Parsingvorgang im Trace-Modus verfolgt.

5 Der didaktische Ansatz von GTU

In den letzten Jahren sind in der Syntaxtheorie zwei Trends zu beobachten. Zum einen wird die Unifikation über Merkmalstrukturen zur wichtigsten Operation innerhalb der Theorien. Zum anderen haben neuere Formalismen (wie GPSG und HPSG) zunehmend deklarativen Charakter. Diesen Trends versucht GTU Rechnung zu tragen. Zwar kann mit GTU z. Zt. noch keine vollständige Implementierung von GPSG oder HPSG durchgeführt werden, aber die Studierenden lernen einerseits den Umgang mit Merkmalstrukturen und den Auswirkungen der Unifikation, und andererseits das deklarative Vorgehen mit DCG- und ID /LP-Grammatiken. Letztere stellen eine interessante Alternative für Sprachen mit variabler Wortstellung dar und sind somit zur Analyse der deutschen Sprache besonders geeignet. GTU ist als Hilfswerkzeug für Lernende konzipiert. Es wendet sich an StudentInnen der (Computer-)Linguistik und verwandter Fächer, die mit der Arbeit am PC in den Grundzügen vertraut sind und schon grundlegende Programmiererfahrung haben. Durch die automatische Visualisierung von Satzstrukturen wird die Motivation zur Arbeit in der Syntaxana-

lyse gestärkt. GTU erlaubt ein inkrementelles Vorgehen bei der Grammatikentwicklung, sodaß Zwischenergebnisse leicht überprüft werden können.

6 Abgrenzung zu existierenden Programmen

GTU unterscheidet sich von anderen Entwicklungsumgebungen durch seine Konzeption als Hilfswerkzeug für Lernende. Vergleichbare Programme wie z.B. GIATN (Brockmann, Fuchs 1990) für ATNs oder TAGDevEnv (Schifferer 1988) für Tree Adjoining Grammars sind dagegen als Werkzeuge für den Linguisten gedacht, der komplexe linguistische Systeme entwickeln will, die u. U. lexikalische, grammatische, semantische und pragmatische Wissensrepräsentation erfordern. GTU ist beschränkt auf die Entwicklung von Syntaxregeln und eignet sich besonders für die schnelle Einarbeitung in einen kleinen Problembereich.

Ein neuerer Ansatz von Erbach (1991) beschreibt ein System, das zum Experimentieren mit unterschiedlichen Parsingstrategien dient. Dabei kann der Benutzer Prioritäten für verschiedene Parsingaufgaben spezifizieren und damit den Parsingprozeß inkrementell optimieren. Wir glauben jedoch, daß ohne standardisierte Testsätze kein wirklicher Vergleich auf breiter Ebene möglich ist und sehen unseren Ansatz als komplementär zu dem von Erbach vorgestellten.

Einige Problembereiche sind in GTU durch die Testsatzklassen des Testsatzarchivs vorgegeben, wobei wir gleichzeitig die Verwaltung weiterer Testsatzklassen ermöglichen. Das war bei bisherigen, uns bekannten Systemen nicht der Fall. Nerbonne (1991) et al. beschreiben eine Arbeit, die in dieselbe Richtung geht. Sie sind dabei, eine
große Satzsammlung für das Deutsche in einer Datenbank zu organisieren. Auch in diesem System werden die Sätze zunächst nur nach syntaktischen Gesichtspunkten klassifiziert.

Schließlich sehen wir einen großen Vorteil unseres Systems in der Anbindung unterschiedlicher Formalismen an ein einheitliches Lexikon. Andere Systeme sind entweder monoformalistisch oder erfordern unterschiedliche Lexika für unterschiedliche Formalismen. Unser Vorgehen bietet nicht nur Vorteile für die Grammatikentwicklung, sondern spiegelt auch unsere Überzeugung wider, daß die Entwicklung von natürlich-sprachlichen Systemen nur durch systematische Nutzung der Lexikonresourcen vorangetrieben werden kann.

7 Ausblick

GTU wurde mehrfach erfolgreich in der Übung zu *Methoden der Syntaxanalyse* an der Universität Koblenz eingesetzt. Nach Aussage der betroffenen Studierenden ist das Programm leicht zu erlernen und stellt gegenüber einer Grammatikentwicklung in Prolog eine spürbare Erleichterung dar. Letzteres wird begründet mit der Verwendung der Lexikonkomponente sowie des automatisch erzeugten Strukturbaums bei der Ausgabe. Eine genaue Beschreibung der Funktionalität liefern Volk, Ridder (1991). Der Anschluß von weiteren Grammatikformalismen (LFG, GPSG) .ist in der Planung.

Literatur

[1] Brockmann, S.; Fuchs, U.: Eine ATN-Werkbank als erste Ausbaustufe für eine Graphic Interactive ATN Workstation. Diplomarbeit. Koblenz: Universität Koblenz-Landau. April 1990.

[2] Erbach, G.: An Environment for Experimentation with Parsing Strategies. (IWBS Report 167) Stuttgart: Wissenschaftliches Zentrum der IBM Deutschland. April 1991.

[3] Friedman, J.: Computational Testing of Linguistic Models in Syntax and Semantics. In: Batori, I. et al. (Eds.): Computational Linguistics. An international handbook on Computer Oriented Language Research and Applications. Berlin: Walter de Gruyter, 1989.

[4] Nerbonne, J. et al.: A diagnostic tool for German syntax. (Research Report RR-91-18) Saarbrücken: DFKI. Juli 1991.

[5] Schifferer, K.: TAGDevEnv. Eine Werkbank für TAGs. In: Batori, I. et al. (Hgg.): Computerlinguistik und ihre theoretischen Grundlagen. Berlin: Springer Verlag, 1988.

[6] Volk, M.; Ridder, H.: GTU (Grammatik Test Umgebung) Benutzerhandbuch. (Manuskript) Institut für Computerlinguistik. Universität Koblenz-Landau. 1991.

Tagungen

First International Quantitative Linguistics Conference (QUALICO 91)

In der Woche vom 23. bis 27. September 1991 fand an der Universität Trier die erste internationale *Quantitative Linguistics Conference* (QUALlCO 91) statt. Veranstalter dieser von der DFG finanziell sowie von zahlreichen nationalen und internationalen wissenschaftlichen Gesellschaften und Vereinigungen organisatorisch unterstützten Konferenz waren die GLDV, die im Rahmen der QUALlCO 91 auch ihre Jahrestagung abhielt (s. deren in diesem LDV-Forum, S.2-37, abgedruckte Beiträge), und die Herausgeber der Buchreihe *Quantitative Linguistics* (QL). Auf Initiative des derzeitigen Vorsitzenden der GLDV wurden beide Veranstaltungen als *joint conference* vom Fach Linguistische Datenverarbeitung/Computerlinguistik der Universität Trier ausgerichtet.

An der Konferenz nahmen über 100 Wissenschaftler aus 16 Ländern Europas, Asiens und Amerikas teil, wobei erfreulicherweise auch die osteuropäische quantitative Linguistik - wenngleich, gemessen an Umfang und Bedeutung allein der in den Staaten der ehemaligen Sowjetunion betriebenen quantitativ-linguistischen Forschung, stark unterrepräsentiert - mehrfach vertreten war. In acht Sektionen ("Dialectometry" , "Models and Explanation", "Phonemics and Phonetics", "Process Dynamics and Semiotics", "Quantification and Measurement", "Reports, Projects and Results", "Statistical Studies", "Textual Structures and Processing") wurde auf dieser Konferenz das gesamte Spektrum quantitativ-linguistischer Forschungsrichtungen und -ansätze präsentiert und diskutiert, wobei insbesondere die eingeladenen Vorträge die derzeitigen Schwerpunkte der weltweiten Forschungsaktivitäten hervortreten ließen:

- Prof. Dr. Gabriel Altmann (Universität Bochum), seit mehr als zwei Jahrzehnten die zentrale Figur (nicht nur) der deutschen quantitativen Linguistik, der in seinem Vortrag "Science and Linguistics" eine wissenschaftstheoretische Standortbestimmung der quantitativen Linguistik und eine Analyse ihrer Relation zu anderen sprachwissenschaftlichen Disziplinen unternahm;
- Dr. Kenneth W. Church (AT&T Bell Labs, Murray Hill, USA), der Möglichkeiten und Fruchtbarkeit der Anwendung quantitativer Methoden in der Lexikographie anhand sehr großer Corpora demonstrierte: "Using Statistics in Lexicographic Analysis";
- Prof. Dr. Hans Goebl (Universität Salzburg, Österreich), mit einem Vortrag zu Möglichkeiten und Methoden der rechnerunterstützten Dialektometrie: "Computational Dialectometry";
- Prof. Dr. John S. Nicolis (University of Patras, Griechenland), zur Modellierung dynamischer Zeichenprozesse und ihrer chaostheoretischen Erklärung: "Chaotic Dynamics of the Linguistic Processes: At the Syntactical, Semantic and Pragmatic Levels";
- Prof. Dr. Mildred G. Shaw und Brian R. Gaines (University of Calgary, Alberta, Kanada), die in ihrem Vortrag" A Methodology for Analyzing Terminological and Conceptual Differences in Language Use across Communities", die Resultate ihrer langjährigen Arbeiten zur Entwicklung einer kognitionspsychologisch fundierten Methodologie der quantitativ-empirischen, semantischen Analyse sprachlicher Vermittlung und deren Implementierung vorstellten.

Die 31 Vorträge der acht QUALICO-Sektionen können in folgender Weise thematisch gruppiert und charakterisiert werden:

i. allgemeine Fragen quantitativ-linguistischer Methodologie: R. Grotjahn diskutierte exemplarisch die wesentlichen methodologischen Probleme der Modellierung der Verteilung sprachlicher Einheiten am Beispiel der Wortlänge. Der Vortrag von J. Krilik behandelte Methoden der Skalierung und Klassifikation von Texten.

ii. quantitative Beiträge 'unter Gesetzesniveau' (zu Metrisierung, Klassifikation, Tests, Methoden der Datenerhebung und -repräsentation, deskriptiv-statistischen Befunden etc.) zu einzelnen sprachwissenschaftlichen Teilbereichen und zu 'verwandten' Bereichen: Dialektometrie: multi dimensionale Skalierung als dialektometrische Methode (S. Embleton); Morphosyntax: Metrisierung dekodierungsrelevanter Aspekte von Kongruenz, Rektion etc. zu sprachtheoretischen wie sprachtypologischen Zwecken (P. Schmidt); Verbvalenz/Satzbaupläne: sprachstatistische

Resultate eines polnischen Lexikonprojekts (M. Swidzinski); Verblexikon/-semantik: Merkmalsanalyse und quantitative Klassifikation englischer und deutscher Verben auf der Basis von Merkmalen mehrerer Sprachebenen (G. Sil'nickij); Analyse von Bedeutungswörterbüchern: Charakterisierung von Güte und Repräsentativität auf der Basis 'externer' (Seitenmengen pro Initial, Lemmatamenge pro Initial, etc.) quantitativer Indikatoren (S. Schierholz/ E. Windisch); Sprachgütebeurteilung/ Audiometrie: quantitative Parameter der phonetischen/phonologischen Ausbalanciertheit von Testmaterialien (W. Sendlmeier); konnektionistische Modelle der Sprachproduktion: Modellierung von Selbstkorrekturen (" Reparaturen") von Sprechern in Äußerungen (U. Schade); Mensch-Maschine-Kommunikation: empirische Untersuchungen zum menschlichen Sprachverhalten in Mensch-Maschine-Dialogen (C. Womser-Hacker); quantitative lexikalische Kollokationsanalyse als Hilfsmittel historisch-soziologischer Forschung (M. Olsen); Klassifikation literarischer Prosatexte mit clusteranalytischen Methoden (N. Bolz); Entzifferung: quantitativer Annäherungsversuch an den berühmt-berüchtigten Diskos von Phaistos (D. Rumpel).

iii. quantitative Studien 'auf explanatorischem Niveau', d.h. solche Beiträge, die bereits bekannte probabilistische Sprachgesetze involvieren oder gesetzesartige probabilistische Hypothesen formulieren: Im Beitrag von K. Ejiri und A. Smith wurde ein 'genetisch' auf ZIPFS Gesetz (Frequenzrang-Frequenz-Verteilung von Einheiten) zurückgehendes Maß zur Charakterisierung des inhaltlichen Reichtums von Texten formuliert und validiert. A. Fenk und G. Fenk-Oczlon präsentierten eine sprachtypologische Studie zum MENZERATHschen Gesetz ("Je komplexer eine Einheit, desto kürzer ihre Konstituenten.") als Explanans der Interrelation von Kernsatzlänge, Wortlänge und Silbenkomplexität und einen kognitivistischen Motivationsversuch des MENzERATHschen Gesetzes als Explanandum aus informationstheoretischen Ökonomieprinzipien. L. Hrebiicek skizzierte seinen Ansatz zur Beschreibung von Texten als Strukturen von Koreferenzketten ("Aggregationen") als Textkonstituenten und seine darauf basierenden Untersuchungen (Aggregationen und MENZERATHsches Gesetz, systemtheoretische Betrachtungen zur Textstruktur und -dynamik). A. Polikarpov stellte ein quantitatives Modell des Bedeutungswandels von Wörtern als 'Alterungsprozeß' vor, bei dem deren Polysemiepotential (Fähigkeit zur Annahme neuer Bedeutungen) mit steigender Polysemie abnimmt.

iv. statistische Modelle (hier: MARKOV-Modelle im weiteren Sinne) in der automatischen Sprachverarbeitung: Algorithmen zur Wortartenklassifikation (E. Dermatas/ G. Kokkinakis, R. Kneser/ H. Ney), Effizienzsteigerung von Spracherkennungssystemen durch textabhängige Dynamisierung der Information zu Übergangswahrscheinlichkeiten von Einheiten (U. Essen/ H. Ney), *Hidden Markov Models* der Phonem-Graphem-Zuordnung (P. Rentzepopoulos/ A. Tsopanoglou/ G. Kokkinakis), quantitative Parameter zur Bewertung probabilistischer Sprachmodelle (M. Refice / M. Savino). v. systemtheoretisch (inspiriert)e Ansätze:

a) Der von B. Rieger seit Anfang der 80er Jahre entwickelte prozedural-rekonstruktive Ansatz, der Sprecher-Hörer-Prozesse der Bedeutungskonstitution und des Verstehens in selbstorganisierenden Systemen modelliert und deren kognitive Leistungen als strukturbildende und -verändernde Resultate dynamischer Informations- und Wissensverarbeitung erklärt, war repräsentiert durch Riegers programmatische Darlegung des Forschungsansatzes einer *dynamischen Semiotik* und durch die Präsentation eines kategorientheoretischen Modells der Entstehung und Entwicklung lexikalischer Bedeutung und seiner Implementierung (B. Rieger/C. Thiopoulos).

b) Die Mitte der 80er Jahre von R. Köhler und G. Altmann begründete *synergetische Linguistik*, die einen quantitativ-linguistischen Forschungsansatz zur Modellierung natürlicher Sprachen als selbstorganisierende dynamische Systeme bietet mit der Aussicht, zu einer erklärenden Theorie ihrer Strukturen und Entwicklungen zu gelangen, war vertreten durch Köhlers programmatischen Abriß des synergetischen Ansatzes in der Linguistik, durch eine Variante eines synergetischen Modells der Lexik von R. Hammerl und durch Sambor/Hammerl (s.u).

vi. Überblicksberichte (über größere Projekte oder nationale quantitativ-linguistische Forschung): F. Dupuis, D. Gosselin und B. Habert zu einem Korpus des Mittelfranzösichen, F. Qian zum durch typologische Eigenarten des Chinesischen motivierten Modell der *C[hinese] P[hrase] S[tructure]* G[rammar], J. Reitsma zum Projekt eines Korpus des Friesischen, J. Sambor und R. Hammerl zu Arbeiten zum Polnischen im Rahmen des Bochumer Projekts "Sprachliche Synergetik", P. Saukkonen zur quantitativen Linguistik in Finnland.

Die Organisation und Durchführung einer internationalen und interdisziplinären Konferenz zur quantitativen Linguistik, welche 1990-91 die GLDV und die Herausgeber der QL-Reihe dankenswerterweise übernommen hatten, stellte zweifellos seit längerer Zeit ein Desiderat dar. Denn über eine Reihe von Jahren ließ sich in verschiedenen theoretischen wie angewandten sprach- und kognitionswissenschaftlichen Disziplinen eine zunehmende Tendenz zur Entwicklung von im weitesten Sinne *quantitativ* zu nennenden Ansätzen beobachten. Hierfür scheinen sehr unterschiedliche Motivationen (Effizienzgesichtspunkte, prinzi-

pielle qualitative Schwellen bei der Modellierung von Sprachverhalten, psychologischer oder sogar biologischer Realismus, Einsicht in die essentielle- makro- wie mikroskopische - Plastizität, Variabilität, Unschärfe, Adaptivität und Dynamik natürlicher Sprachen) bestimmend zu sein, womit diese Neuansätze erkennbar teils komplementäre, teils alternative Entwicklungen gegenüber konventionellen Positionen traditioneller Orientierung in den verschiedenen Disziplinen vertreten und verfolgen. Speziell scheint das in verschiedenen konkreten Manifestationen auftretende und sich verbreitende systemtheoretische Paradigma zum ersten Mal die Möglichkeit der Konstruktion einer Sprachtheorie zu bieten, die zugleich erklärende Kraft beanspruchen und diesen Anspruch gemäß den Standards der exakten Wissenschaften auch einzulösen vermag, wobei sie das bestenfalls partielle Erklärungsmuster des CHOMSKYschen Innatismus komplementieren und einbetten könnte.

QUALICO 91 war gekennzeichnet durch ihre Interdisziplinarität und die repräsentative Vielfalt der vertretenen Aspekte der quantitativ-linguistischen Forschung - darunter verschiedene vielversprechende quantitative Ansätze paradigma*tischen*, forschungsleitenden Zuschnitts-, durch fruchtbaren und regen wissenschaftlichen Austausch, eine perfekte Organisation und positive Konferenzatmosphäre (wobei hier der attraktive Rahmen der Konferenz in ihrem reichhaltigen *Social* Program mit Empfang durch die Stadt Trier, Stadtbesichtigungen, Weinprobe, Dampferfahrt nach Saarburg und abendlichem Diner auf der Burg ausdrücklich zu erwähnen und einzuschließen ist). Der beschriebenen, sich verstärkenden quantitativ-linguistischen Tendenz ist mit QUALICO ein internationales und interdisziplinäres wissenschaftliches Forum entstanden, von dem zu hoffen ist, daß es sich institutionalisieren wird und Folgekonferenzen dieser Art erleichtert.

PETER SCHMIDT, *Universität Konstanz*

Tagung:

Information und Klassifikation

Die 16. Jahrestagung der Gesellschaft für Klassifikation fand vom 1.-3. April 1992 an der Universität Dortmund statt. Unter dem Motto 'Information und Klassifikation. Konzepte, Methoden und Anwendungen' fanden eine Reihe von Disziplinen zueinander, die ansonsten kaum miteinander in Berührung kommen: Mathematik, Statistik, Informatik, Medizin, Biologie, Bibliothekswissenschaft, Informationswissenschaft, Psychologie, Sprachwissenschaft, Computerlinguistik, Wirtschaftswissenschaft, Sozialwissenschaft, Musikwissenschaft, Archäologie. Der organisatorische Rahmen umfaßte Plenar- und Übersichtsvorträge, parallele Sektionsvorträge, Workshops, Tutorials und Softwaredemonstrationen-ein kompaktes Programm für nur drei Tage. Die Verfasser sehen sich deshalb außerstande, darüber in aller Ausführlichkeit zu berichten. Der Tagungsbericht gibt deshalb nur die subjektive Auswahl der Verfasser wieder. Wir haben uns auf Beiträge aus den Bereichen Informatik, Informationswissenschaft, Musikwissenschaft, Sprachwissenschaft und Computerlinguistik konzentriert.

Der insgesamt doch recht heterogene Charakter der Tagung spiegelte sich u.a. in der Zusammensetzung der einzelnen Sektionen wieder. So vereinte z.B. das Tutorium "Grundlagen und Nutzungsmöglichkeiten von Computergrammatiken und semantischen Repräsentationsformalismen" einen Bericht über die in den letzten Jahrzehnten gesammelten Erfahrungen in der thesaurusbasierten Auswertung von medizinischen Befunden (W. Giere), einen Überblick über neuere Grammatikformalismen (S. Naumann) und die Vorstellung eines sprachverarbeitenden Systems (SMART) zur semantischen Analyse medizinischer Nominalgruppen und -komposita (J. Ingenerf).

Im Mittelpunkt der beiden Workshops zur medizinischen Linguistik standen Beiträge, in denen über Erfahrungen im Einsatz von medizinlinguistischen Systemen (wie z.B. MUMPS und SNOMED) berichtet wurde. Über zwei in diesem Bereich angesiedelte computerlinguistische Projekte, deren Ziel die praktische Evaluierung semantischer und texttheoretischer Ansätze bildet, berichteten H. Kranzdorf "Automatische Generierung von Sprachtherapieberichten" und M. Schulz "Ein natürlichsprachliches Interface für eine komplexrelationale Datenbank".

Jürgen Kristophson versuchte in 'Ein neuer Beitrag zur Sprach bund diskussion' die Definition von 'Sprachbund: auf eine quantitative Grundlage zu stellen. Über die Textfrequenz bestimmter Merkmale (z.B. postdeterminierend, prädeterminierend) lassen sich Indizien gewinnen, die auf einen möglichen Sprachbund hinweisen.

Stefan J. Schierholz 'Zur Klassifikation von Substantiven nach ihren Determinatoren' lieferte eine Beschreibung der Kookkurrent von Substantiven und Artikeln im Text. Diese Beschreibung steht in Verbindung mit der Entwicklung eines maschinellen Grammatik-Checkers. Als Ergänzung zu einer Klassifikation auf struktureller Grundlage schlug Schierholz eine Angabe von Vorkommenswahrscheinlichkeiten vor.

Heinz J. Weber gab in 'Generierung themenbasierter Links in einem Hypertext-System für Pressenachrichten' einen Überblick über das experimentelle System t-X-t. Auf der Grundlage einer durch Textparsing erstellten Topik-Hierarchie für jeden Nachrichtentext werden inhaltlich ähnliche Topiks ermittelt und miteinander vernetzt.

Eine in Teilen vergleichbare Zielsetzung stellte der Übersichtsvortrag von Gerard Salton 'Automatic Text Linking and Text Grouping Methods' vor. Teiltexte (z.B. Kapitel, Paragraph, Fußnote) aus einer großen Datenbasis (25.000 Artikel einer Enzyklopädie) sollen aufgrund von statistisch ermittelten inhaltlichen Ähnlichkeiten in Affinitätsklassen zusammengefaßt und durch Links miteinander verbunden werden. In einer Retrieval- Umgebung bieten diese Verbindungen Zugänge zu spezifischen Informationen.

In der Sektion *Informationssysteme* 5 stellten D. Fensel und J. Klein drei Algorithmen *Relax, H*-Relax und *I-Relax zur* Generierung allgemeiner Regeln aus einer Menge positiver und negativer Beispiele vor. Ausgehend von den positiven Beispielen wird eine maximal-allgemeine Beschreibung der Zielklasse erzeugt. Im Gegensatz zu anderen Verfahren wie AQ, CABRO oder ID3 verwendet Relax Generalisierung als Suchstrategie. Darüber hinaus wurde noch aufgezeigt, wie Verfahren des maschinellen Lernens mit statistischen Verfahren kombiniert werden können.

In der Sektion *Informationssysteme* 6 - *Wissen-sakquisition* berichtete A. Ultsch über die Integration von selbst-organisierenden neuronalen Netzen (insb. KOHONEN-feature-map) und regelbasierten Expertensystemen. Der Einsatz neuronaler Netze beschränkt sich bei dem vorgestellten Ansatz auf die Extraktion gewisser Regularitäten in den Beispieldaten. Diese Regularitäten werden dann von einem Regelextraktor in PROLOG-Regeln überführt, die dann zusammen mit dem Netzwerk in das Expertensystem als Wissenbasen integriert werden. J. Schrepp beschrieb ein Verfahren, das aufgrund einer Menge natürlichsprachlicher Texte eine kontextfreie Syntax für diese Textmenge erzeugt.

In der Sektion *Neural Networks for EDA and Classification* beschrieb A. Ultsch die Anwendungsmöglichkeiten selbst-organisierender neuronaler Kohonen-Netzwerke im Bereich der explorativen Datenanalyse. Eine fundamentale Eigenschaft der KOHONEN-Netze ist die strukturerhaltende Abbildung des n-dimensionalen Datenraumes auf den 2-dimensionalen Raum der Verarbeitungselemente (units) des Netzes. Um die extrahierte Struktur sichtbar zu machen, wurde die U-Matrix-Methode vorgestellt, die es erlaubt, die Eingabedaten zu klassifizieren.

In der leider nur sehr spärlich besetzten Abteilung *Musikwissenschaft* diskutierte M. G. Boroda die Zweckmäßigkeit verschiedener Gleichheitskriterien für F-Motive. F-Motive sind elementare musikalische Einheiten, die im Gegensatz zum herkömmlichen intuitiven Motivbegriff eindeutig definiert sind. Das ZIPFsche Gesetz liefert ein Indiz dafür, daß Motive nur dann als gleich angesehen werden können, wenn sie bis auf Transposition übereinstimmen. Aus anderen Gleichheitsrelationen, wie z.B. Gleichheit bei rhythmischer Identität resultiert eine Häufigkeitsverteilung, die nicht mit dem ZIPFschen Gesetz vereinbar ist.

Ulrich Franzke demonstrierte recht eindrucksvoll, wie aus einfachen Melodien verschiedene Merkmale abstrahiert werden können, die es erlauben, neue Melodien maschinell zu synthetisieren, ohne daß ihre Künstlichkeit wahrnehmbar ist. Dies ist ein wichtiger Beitrag zur formalen Beschreibung von Eigenschaften, die Melodien als spezielle Klasse von Tonfolgen auszeichnen.

E. LEOPOLD, S. NAUMANN, J. SCHREPP, H.-J. WEBER, *Universität Trier*

COMPUTERLINGUISTIK ANDERSWO

EINHARD KÖHLER, UNIVERSITÄT TRIER
FB II, LINGUISTISCHE DATENVERARBEITUNG

Im September 1991 unternahm der Autor dieses Beitrags eine Vortrags- und Informationsreise nach Leningrad (jetzt St. Petersburg), Smolensk und Moskau 1, in deren Verlauf er auch Gelegenheit hatte, eine Reihe von Forschungseinrichtungen zu besuchen. Obwohl seit vielen Jahren enge Kontakte und ein ständiger wissenschaftlicher Austausch zwischen deutschen und osteuropäischen Wissenschaftlern - wenigstens im Bereich der quantitativen Linguistik und besonders zwischen den Herausgebern der Reihe QUANTITATIVE LINGUISTICS und ihrem Gegenstück KVANTITATIVNAJA LINGVISTIKA I AVTOMATIÖESKIJ ANALIZ TEKSTA (Tartu)- bestanden hatte, ergab der Besuch Informationen über eine überraschende Vielfalt von Aktivitäten und eine Vielzahl von Instituten, Gruppen und Zentren, deren Arbeit bei uns nicht oder nur wenig bekannt sind.

Daraus resultierte die Idee, im LDVForum eine Rubrik einzurichten, in der aufgrund von Informationsreisen über die Situation von LDV und LDV-bezogenen Forschungen berichtet wird.

Die Forschungsgruppe Statistika reči in St. Petersburg

Die Gruppe statistika reči ("Sprachstatistik") ist eine der bedeutendsten linguistischen Forschungseinrichtungen auf dem Gebiet der ehemaligen Sowjetunion. Sie wurde im Jahre 1956 von R. G. Piotrowski und. L. Novak gegründet. Heute gehören mehr als 200 Wissenschaftler zu ihr, von denen viele auch in Hochschulen und Forschungseinrichtungen anderer Länder (USA, Israel, Großbritannien, Frankreich u. a.) tätig sind. Die Gruppe besteht aus einer Reihe regionaler Untergruppen, die unter der Leitung von Prof. Piotrowski von St. Petersburg aus koordiniert werden, darunter die Gruppe St. Petersburg (Leitung Prof. Piotrowski, Prof. P. Alekseev, Dr. L. Belaeva), die Weißrussische Gruppe (Leitung Prof. A. Zubov, Dr. I. Sovpel), die Moldauische Gruppe (Leitung Prof. W. Czyzakowski, Doz. V. Goncarenko), die Kasachische Gruppe (Leitung Prof. K. Bektaev) und einige kleinere Gruppen (Wolga, Irkutsk, Usbekistan, Georgien).

Im Rahmen dieser Einrichtung finden Forschungen vor allem auf dem Gebiet der quantitativen Linguistik unter verschiedenen Anwendungsgesichtspunkten statt; unter den aktuellen Arbeitsschwerpunkten sind besonders die Projekte zur maschinellen Übersetzung hervorzuheben. Es handelt sich um pragmatisch orientierte Ansätze mit ausgeprägtem Lexikonanteil und einer statistischprobabilistischen Komponente. Die z. Zt. behandelten Sprachpaare sind Russisch ↔ Englisch, Russisch ↔ Französisch und Russisch ↔ Deutsch, wobei die Entwicklung des letzten noch in den Anfängen steckt. In diesem Zusammenhang beklagt R. Piotrowski den Mangel an ausgebildeten Fachkräften mit guten Deutsch-Kenntnissen; auch aus diesem Grund besteht auf russischer Seite ein Interesse an Kooperation. Entsprechende Kontakte und Vorbereitungen zur Aufnahme einer offiziellen Zusammenarbeit mit der Universität St. Petersburg hat das Fach Linguistische Datenverarbeitung der Universität Trier bereits eingeleitet.

Zur Zeit konzentrieren sich die theoretischen und praktischen Entwicklungsarbeiten auf den "polyfunktionalen linguistischen Automaten", der das Resultat mehrjähriger interdisziplinärer KI-orientierter Forschung unter Beteiligung von Mathematikern, Linguisten, Psychologen und anderen darstellt.2

Ein weiterer Schwerpunkt der Gruppe besteht weiterhin in den Untersuchungen zur Sprachstatistik und quantitativer Textanalyse sowie das Studium von pathologischer Sprache mit Anwendungen in der Computerlinguistik und anderen Fächern wie Medizin, Recht, Technik, Pädagogik und Wirtschaftswissenschaften.

1 Auch an dieser Stelle sei der Deutschen Forschungsgemeinschaft für die gewährte Unterstützung gedankt.

2 Ein Bericht über diesen Aspekt der St. Petersburger Arbeiten und eine Darstellung des Konzepts des linguistischen Automaten von W. Czyzakowski und R. G. Piotrowski ist in Vorbereitung und wird in einer der nächsten Ausgaben von GLOTTOMETRIKA (in der Reihe QUANTITATIVE LINGUISTCS) erscheinen.

COMPUTERLINGUISTIK ANDERSWO

REINHARD KÖHLER, UNIVERSITÄT TRIER
FB II, LINGUISTISCHE DATENVERARBEITUNG

Im September 1991 unternahm der Autor dieses Beitrags eine Vortrags- und Informationsreise nach Leningrad (jetzt St. Petersburg), Smolensk und Moskaul, in deren Verlauf er auch Gelegenheit hatte, eine Reihe von Forschungseinrichtungen zu besuchen. Obwohl seit vielen Jahren enge Kontakte und ein ständiger wissenschaftlicher Austausch zwischen deutschen und osteuropäischen Wissenschaftlern - wenigstens im Bereich der quantitativen Linguistik und besonders zwischen den Herausgebern der Reihe QUANTITATIVE LINGUISTICS und ihrem Gegenstück KVANTITATIVNAJA LINGVISTIKA I AVTOMATIÖESKIJ ANALIZ TEKSTA (Tartu)- bestanden hatte, ergab der Besuch Informationen über eine überraschende Vielfalt von Aktivitäten und eine Vielzahl von Instituten, Gruppen und Zentren, deren Arbeit bei uns nicht oder nur wenig bekannt sind.

Daraus resultierte die Idee, im LDVForum eine Rubrik einzurichten, in der aufgrund von Informationsreisen über die Situation von LDV und LDV-bezogenen Forschungen berichtet wird.

Die Forschungsgruppe Statistika reci in St. Petersburg

Die Gruppe statistika reCi ("Sprachstatistik") ist eine der bedeutendsten linguistischen Forschungs-einrichtungen auf dem Gebiet der ehemaligen Sowjetunion. Sie wurde im Jahre 1956 von R. G. Piotrowski und. L. Novak gegründet. Heute gehören mehr als 200 Wissenschaftler zu ihr, von denen viele auch in Hochschulen und Forschungseinrichtungen anderer Länder (USA, Israel, Großbritannien, Frankreich u.a.) tätig sind. Die Gruppe besteht aus einer Reihe regionaler Untergruppen, die unter der Leitung von Prof. Piotrowski von St. Petersburg aus koordiniert werden, darunter die Gruppe St. Petersburg (Leitung Prof. Piotrowski, Prof. P. Alekseev, Dr. L. Belaeva), die Weißrussische Gruppe (Leitung Prof. A. Zubov, Dr. I. Sovpel), die Moldauische Gruppe (Leitung Prof. W. Czyzakowski, Doz. V. Goncarenko), die Kasachische Gruppe (Leitung Prof. K. Bektaev) und einige kleinere Gruppen (Wolga, Irkutsk, Usbekistan, Georgien).

Im Rahmen dieser Einrichtung finden Forschungen vor allem auf dem Gebiet der quantitativen Linguistik unter verschiedenen Anwendungsgesichtspunkten statt; unter den aktuellen Arbeitsschwerpunkten sin'd besonders die Projekte zur maschinellen Übersetzung hervorzuheben. Es handelt sich um pragmatisch orientierte Ansätze mit ausgeprägtem Lexikonanteil und einer statistischprobabilistischen Komponente. Die z.Zt. behandelten Sprachpaare sind Russisch +-+ Englisch, Russisch +-+ Französisch und Russisch +-+ Deutsch, wobei die Entwicklung des letzten noch in den Anfängen steckt. In diesem Zusammenhang beklagt R. Piotrowski den Mangel an ausgebildeten Fachkräften mit guten Deutsch-Kenntnissen; auch aus diesem Grund besteht auf russischer Seite ein Interesse an Kooperation. Entsprechende Kontakte und Vorbereitungen zur Aufnahme einer offiziellen Zusammenarbeit mit der Universität St. Petersburg hat das Fach Linguistische Datenverarbeitung der Universität Trier bereits eingeleitet.

Zur Zeit konzentrieren sich die theoretischen und praktischen Entwicklungsarbeiten auf den "polyfunktionalen linguistischen Automaten" , der das Resultat mehrjähriger interdisziplinärer KIorientierter Forschung unter Beteiligung von Mathematikern, Linguisten, Psychologen und anderen darstellt.2

Ein weiterer Schwerpunkt der Gruppe besteht weiterhin in den Untersuchungen zur Sprachstatistik und quantitativer Textanalyse sowie das Studium von pathologischer Sprache mit Anwendungen in der Computerlinguistik und anderen Fächern wie Medizin, Recht, Technik, Pädagogik und Wirtschaftswissenschaften.

L Auch an dieser Stelle sei der Deutschen Forschungsgemeinschaft für die gewährte Unterstützung gedankt.

2 Ein Bericht über diesen Aspekt der St. Petersburger Arbeiten und eine Darstellung des Konzepts des linguistischen Automaten von W. Czyzakowski und R. G. Piotrowski ist in Vorbereitung und wird in einer der nächsten Ausgaben von GLOTTOMETRIKA (in der Reihe QUANTITATIVE LINGUlSTCS) erscheinen.

Bericht über die Sitzung des Arbeitskreises Lexikographie in der GLDV am 27.9.91 Universität Trier

DR. W. HELLMANN
INSTITUT FÜR DEUTSCHE SPRACHE, MANNHEIM

Die Sitzung des AK Lexikographie fand in der Universität Trier während der Jahrestagung der GLDV am Freitag, 27.9.91, von 14 bis 16.30 Uhr statt.

Anwesend: Sieben Mitglieder des Arbeitkreises (davon ein neues) von insgesamt ca. 25 in Trier noch anwesenden GLDV-Mitgliedern.

Teilnehmerliste in der Anlage.

Tagesordnung

1. Kurzvorstellung

der anwesenden AK-Mitglieder mit derzeitigen Arbeitsschwerpunkten.

2. Situation im AK

Eine Sitzung des AK hat seit der letzten Jahrestagung in Gießen nicht mehr stattgefunden. Der bisherige Leiter des AK, Gerd Frackenpohl, ist nach Wissen einiger GLDV-Mitgliedern beruflich stark belastet. Es scheint erforderlich, dem AK einen neuen Impuls zu geben, wobei möglicherweise die Arbeit des AK thematisch stärker zu bündeln bzw. zu profilieren ist. Im Anschluß an die auf der QUALICO/GLDV-Tagung sichtbar gewordenen Interessenslagen wird als neuer Themenschwerpunkt die *Entwicklung und Standarisierung von lexikographischen Werkzeugen* sowohl zur Erstellung von neuen Wörterbüchern wie auch zur Evaluierung vorhandener Wörterbücher (makro- und mikrostrukturell) vorgeschlagen. Eine Fachtagung/ein Fachgespräch des AK zum Thema Werkzeugentwicklung wird für wünschenswert gehalten. Vertreter von institutionell-wissenschaftlichen und kommerziellen Anwendern (Verlagen) und gegebenenfalls Softwarehäusern sollten einbezogen werden.

3. Organisatorisches

Der AK fordert Manfred W. Hellmann, IDS Mannheim, auf, als (vorläufiger?) Sprecher des AK entsprechende Initiativen zu entwickeln. Dieser erklärt sich dazu bereit, wenn folgende Fragen geklärt sind:

1. Logistisch-organisatorische Unterstützung durch das IDS (auch im Hinblick auf die gewünschte Fachtagung/Fachgespräch zum Thema Werkzeugentwicklung)
2. Unterstützung durch den Vorstand der GLDV
3. Einigung mit Gerd Frackenpohl.

Die Teilnehmer Ulrich Heid, Universität Stuttgart, und Andrea Beurer, Universität Trier, erklärten sich bereit, den Sprecher des AK beratend und gegebenalls organisatorisch zu unterstützen.

Teilnehmer an der Sitzung des Arbeitskeises Lexikographie vom 27.9.91

1. Beurer, Andrea, Universität Trier, Brühlstr. 34, 5500 Trier, Tel.0651/27963.
   Interessenschwerpunkt (IS): Programmgesteuerte syntaxbasierte Satzgenerierung, morpho-syntaktische Fehlererkennung und Fehlerkorrekturen.
2. Seewald, Uta, Universität Hannover, Mommsenstr. 8, 3000 Hannover 1, Tel. 0511/808401

IS: Morphosemantische Analyse des (französischen) Wortschatzes; automatische Segmentierung mit Inhaltsbeschreibung von Derivaten aus ihren Strukturelementen.

3. Wenzel, Friedrich (Prof. Dr.), Universität Hannover, Postfach 6009, 3000 Hannover 1, Tel. 0511/762-3401
IS: Automatische Segmentierung von Fachwörtern, automatische Generierung und Analyse von Derivaten im Bereich der (russischen naturwissenschaftlichen) Fachlexik .
Praktisches Interesse: Erstellung kleiner Fachwörterbücher für den Lehrbetrieb.

4. Heid, Ulrich, Universität Stuttgard (IMS-CL), Kepplerstr. 17, 7000 Stuttgard 1
IS: Entwicklung von Analyse- und Evaluierungsverfahren zur Prüfung der Konsistenz von Wortartikeln (mikro und makro). Speziell: Entwicklung eines Programmsystems für die Wortartikelerstellung mit Parallel-Editor (Eingabestruktur /Grafikmodus).

5. Weber, Heinz J., Universität Trier (FB **II,** LDV-Computerlinguistik), Postfach 3825, 5500 Trier, Tel. 0651/74628, d.: 201-253.

6. Hitzenberger, Universität Regensburg
IS: Entwicklung einer lexikographischen Datenbank für Lehrzwecke, einschließlich Werkzeugen für Konsistenzprüfungen.

7. Hellmann, Manfred W., IDS Mannheim, Postfach 101621, 6800 Mannheim 1, Tel. 0621/4401-278
IS: Korpusgestützte Lexikographie; Entwicklung von Werkzeugen zur Bearbeitung großer Belegmengen, zur Erstellung von Wortartikeln und für Konsistenzprüfungen (makro- und mikrostruktuell) .

Veranstaltungen

Veranstaltungskalender

1.4.–3.4.92 in Trient Tagung über Angewandte Sprachverarbeitung.
Auskunft: Lyn Bates, BBN Systems & Tech Corporation, 10 Mouton Street, Cambridge MA 02238, U.S.A.

1.4.–4.4.92 in Paris
9th International Humor Conference
Auskunft: Françoise Barioud, 39 rue Ste Croix de la Bretonnerie, B 23,
F–75004 Paris.

03.04.–05.04.1992, Buffalo USA Cognition and Representation
Info: Center for Cognitive Science,
651 Baldy Hall, State University of New York at Buffalo, Buffalo,
New York 14260
Email: dcp@cs.buffalo.edu

05.04.–09.04.1992, Oxford GB
ALLC-ACH'92 Joint Conference
This event covers literary, linguistic and humanities computing.
Info: Centre for Humanities Computing, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, England. voice: +44-8865-273 200
Fax: ...273 275
ALLCACH@VAX.OX.AC.UK

06.04.–09.04.1992, Makuhari, Japan
Communication Tokyo'92
Veranstalter: Communications Industry Association of Japan

13.04.–14.04.1992, München 6. Workshop **Planen und Konfigurieren**
Veranstalter: Forschungszentrum für Wissensbasierte Systeme, GI-FG 1.5.3 (Planen und Konfigurieren)
Info: Tilo Messer, Forschungszentrum für Wissensbasierte Systeme, Orleansstr. 34, W-8000 München 80

20.04.–23.04.1992, Washington, D. C., USA
Georgtown University Roundtable on Linguistics

21.04.–24.04.1992, Dessau, Deutschland
Umweltklasse im Bauhaus Dessau
Themen: Management, Vollzug und Gestaltung von Umweltschutz
Auskunft: Joachim Borner, Bauhaus Dessau, Gropius–Allee 38, Postfach 160, D–O4500 Dessau

21.04.–24.04.1992, Univ Vienna, AT Eur. Meeting on Cybernetics and Systems Research This is the 11th meeting from a series starting in 1972. There will be a number of symposia. Submit 7 single spaced, max 50 lines A4 by Oct 15. Papers and info: Austrian Society for Cybernetic Studies, Schottengasse 3, A-1010 Wien 1, Austria.
voice +43 1 5353 2810,
fax: +43 1 630 652
sec%ai-vie.uucp@relay.eu.net

24.04.–23.04.1992, Washington, D. C. ,USA
International Linguistic Association Annual Meeting, Georgetown University
Information: Prof. Ruth Brend, 3363 Burbank Drive, Ann Arbor, MI 48105, USA

30.04.–03.05.1992, Reading, USA 6th Round Table on Law and Semiotics
Auskunft: Roberta Kevelson, Center for Semiotic Research in Law, Goverment and Economics, The Pennsylvania State Universita at Berks, Reading, PA 19610–6009, USA

11.05–15.05.1992, Vancouver, Canada
AI'92 Canadian AI Conf
This is the 9th bienial conference on AI sponsored by the Canadian Society for Computational Studies of Intelligence. Papers and info: Janice Glasgow,Dept. of Computing and Information Science, Queens University, Kings-

ton, Ontario, K7L 3N6, Canada.
voice +16135456058
fax:...545 6513
janice@qucis.queensu.ca

19.05–22.05.1992, Groningen, Netherlands
10th Workshop of European Society for the Study of Cognitive Systems **ESSCS**
Information: ESSCS (Dr. G. J. Dalenoort), Dept. of Psychology, University of Gronningen
P. O. Box 72, 9700 AB Gronningen, Netherlands
Voice Netherlands +31–50–536472 (or ..6448)
Fax: +31–50–636304
E–mail <DAL at RUG86.RUG.NL> (or: DAL at RUG.NL)

18.05.–22.05.1992, Zürich, Schweiz 6th Europepan Knowledge Acquisition Workshop **EKAW92**
Veranstalter: GI–FA1.5(Experteensysteme)
Information: Dr. Thomas Wetter, IBM Deutschland GmbH, Wissenschaftliches Zentrum, Wilckenstr. 1a, W–6900 Heidelberg

27.5.–30.5.92 in Glottertal Arbeitstagung über biosemiotische Modelle in der Medizin. Thema: Subjektivität und Intersubjektivität von Krankheitszeichen.
Auskunft: Jörg M. Herrmann, Klinik für Rehabilitation, Glotterbad, W–7804 Glottertal.

01.06.–05.06.1992, Tokyo, Japan
International Conference on "Fifth Generation Computer Systems" (FGCS'92)
Information: FGCS'92 Secretariat
Institute for New Generation Computer Technology (ICOT)
Mita Kokusai Bldg. 21F
4–28 Mita 1 -chome, Minato-ku
Tokyo, 108, Japan
Tel: +813/3456–3195
Fax: +813/3456–1618

01.06.–05.06.1992, Avignon, FR
Avignon'92 12th Int Conf
AI-Expert Systems-Natural Language; TAMA'92
This 12th conference adopts a new formula. It will comprise one scientific conference devoted to tools, techniques and applications used for Knowledge-Based Systems, one conference on Natural Language Processing and its Applications.
Avignion'92 will also host TAMA'92, the second TERMNet symposion on terminological applications on micro-computers.
Submission Nov 29.
Papers and info: Jean Claude Rault, EC-2, 269 Rue de la Garenne, 92024 Nanterre Cedex,France.
voice +33 1 4780 7000
fax: +33 1 4780 6629

5.6.–6.6.92 in Avignon, France 12th International Avignon Conference **Terminology in Advandced Microcomputer Applications** *TAMA 92*
Info: Jean–Claude Rault
EC2
269, rue de la Garenne
Phone: (33.1)47 80 70 00
Telex: 612 469
Fax : (33.1)47 80 66 29

5.6.–7.6.92 in Stendal Kolloquium über die Beziehung zwischen Text und Bild im Werk Stendals.
Auskunft: Michael Nerlich, Institut für französische Literaturwissenschaft, Technische Universität Berlin, Sekr. TEL 3, Straße des 17. Juni 135, W–1000 Berlin 12.

10.06.–12.06.1992, Paris, Frankreich
European Symposium **Information Technology in tomorrow's Europe – Opportunities and Dangers**
Veranstalter: Centr de coordination pour la recherche et l'Enseignement en Informatique et Société
Information: Secrétariat du colloque CREIS 92, Boite 165 – Tour 55–65, Bureau 309, UniversitéParis VI, 4, place Jussieu, F–75252 Paris Cedex 05

10.06.–12.06.1992, Montreal, Canada
ITS'92 2nd Int Conf on Intelligent Tutoring Systems
The conf will focus on a board spectrum of research concerned with how AI and other advandced technologies can be applied to education and training.
Info: Prof. Claude Frasson,
Dept d'I.R.O., Universite de Montreal, C.P.6128,Succ."A",Montreal (Quebec), Canada, H3C 3J7
Tel.: +1 514 343 7019,
Fax: ... 343 5834,
Email: frasson@iro.umontreal.ca

10.6.–14.6.92 in Göttingen Tagung über Medien des internationalen Literatur Transfers. Thema: Übersetzungs-Anthologien.
Auskunft: Arnim Paul Frank, Sonderforschungsbereich literarische Übersetzung, Georg–August–Universität, Humboldtallee 17, W–3400 Göttingen.

17.06.–20.06.1992, Nova Scotia, Canada
ICCAL'92 4th Int Conf on Computers and Learning
ICCAL is devoted to theory and practice of computers and learning.
Submission by Sept 15 not accept. Dec 15, crc Febr 15.
Info: Dr. Ivan Tomek, Jodrey School of Comp.Sc., Acadia Uni, Wolfville, Nova Scotia, BOP1XO, Canada.
ICCAL@AcadiaU.ca
voice: +1 902 542 2201
fax: ...542 7224

22.06.–23.06.1992, Stockholm, SE
Translation and the European Communities In particular, papers are welcomed on the following issues: A pluralingual community and its impact on the national languages; Terminological support and language control; The translation market after 1992;
Submit by Jan 15, not accept. March 1, crc May 15.
Info: Eurofat AB, Skeppsbron 26, S-11130 Stockholm, Schweden.
fax: + 46 8 796 9639
voice: +46 8 789 6683
coling@com.qz.se

01.07.–03.07.1992, Tokyo, Japan
2nd Int conf'l. on Artificial Reality and Tele–Existence
Veranstalter: Japan TechnologieTechnologie Transfer Association

01.07.–03.07.1992, Aberdeen, Scotland 9th Int Conf Maschine Learning
subm: early January 1992
Info: ML-92, Dept. of Computing Science, King's College University of Aberdeen, Aberdeen, AB92UB Scotland,
Tel: +44 224 272296
fax: +44 224 487048
Janet: ml92@uk.ac.abdn.cs

13.07–17.07.1992, Imatra, Finnland
Jahrestagung der Finnischen Semiotik-Gesellschaft
Themen: Das Konzept der Grenze; die Semiotik der Aufführenden Künste
Auskunft: Eero Tarasti, Musikologisches Institut, Universität Helsinki, Vironkatu 1, SF–00170 Helsinki

07.07.–09.07.1992, Tokyo, Japan Voice Systems Worldwide'92 Japan
Zuständig: Nihon Keizai Shimbun

17.07.–22.07.1992, Iizuka, Fukuoka
Int'l Conf. on Fuzzy Logic and Neural Networks **IIZUKA'92** Zuständig: Fuzzy Logic Systems Institution (**FLSI**)

23.07.–28.07.1992, Nantes, France
COLING -92 14th Int Conf on Computational Linguistics
The conf will last 5 full days (not counting sunday).
Pre-COLING tutorials: 22-22 July(2-1/2 days).
All topics in Computational Linguistics are acceptable.
Submission: Nov 1. Send 6 A4 or 8.5x11 inch copies of the full paper to the Chair: Prof. A.Zampolli, Universita di Pisa, ILC, via della Faggiola 32, I-56100 Pisa,Italy;
voice: +3950 560481
fax: +39 50 589055. Not acc March 1, crc May 1.
Info: GETA-IMAG, COLING-92, BP 53X, F-38041 Grenoble France
Exhib-info: EC2, G.d'Aumale, 269 Rue de la Garenne, Nanterre Cedex, France.
voice: +33 1 47 80 7000,
fax: ...80 6629

31.7.–2.8.92 in Canterbury Conference on Literary Semantics.
Auskunft: Trevor Eaton, Honeywood Cottage, 35 Seaton Avenue, Hythe, Kent, CT21 5HH, England.

03.08.–07.08.1992, Vienna AT
ECAI'92 The European Conference on Artificial Intelligence
ECAI's are held biannually and organized by the European Coordinating Committee on AI. This conference covers all the aspects of AI research and brings together basic research and applied research.
Paper submission: 5 hardcopies, long (5000 w. or 10 single-sp p.) or short (2000 w./5 p.), by Jan 17th to the Progr. Chair., not accept April 1, crc May 15.
Prog. Chair: Bernd Neumann,
FB Informatik Universität Hamburg,
Bodenstädtstr.16
W-2000 Hamburg 50, Germany
neumann@rz.informatik.uni-hamburg.dbp.de
voice: +49 40 4123 6130
fax: +...4123 6530
Conf off.: ADV c/o ECAI'92,
Trattnerhof 2,
A-1010 Vienna, Austria.
voice: +43 1 533 0913 74
fax: ...0913 77

09.08.–14.08.1992, San Diego, CA US
AAAI-92
Info: AAAI, 445 Burgess Drive, Melno Park Cal 94025, US.
fax: +1 415 321 4457

Email: ncai@aaai.org

17.08.–19.08.1992, Jurmala, Latvia
International Symposium on Terminology Science and Terminology Planning
Info: Dr. G.Budin, Infoterm,
Heinestr. 38, 1021 Vienna, Austria

26.08.–28.08.1992, Madras, India Second ISKO Conference on "Cognitive Paradigms in Knowledge Organization"
Information: Dr. Sushila Kumar
President, Madras Library Association
No. 5, Sivaganga Road
Madras 600034, India

26.8.–29.8.92 in Uppsala International Conference on Discourse and the Professions.
Auskunft: Britt–Louise Gunnarsson,
Uppsala University, P.O.Box 1834, S–751 48 Uppsala.

31.08.–04.09.1992, Bonn German Workshop on **Artificial Intelligance GWAI–92**
Veranstalter: GI, GMD
Information: Christine Harms, GMD, Postfach 1316, W–5205 Sankt Augustin

8.8.–15.8.93 in Leipzig 10th World Congress of Applied Linguistics. Thema:
Language in a Multicultural Society.
Auskunft: Johann F. Matter, Vakgroep TTW–VU, 10A–28, P.O.Box 7161, NL–1007 MC Amsterdam.

04.09.–07.09.1992, Brighton GB
ICANN'92 Int Conf on
Artificial Neural Networks
The ICANN series covers theory, implementation and applications of Artificial Neural Networks.
Info: Prof. I.Aleksander, Dept. of Electrical Ingeneering, Imperial College, Exihibition Road London SW7 2 BT, U.K.
voice: +44 715895111 ext 5100
fax: +44 718238125
Laleksander@vaxa.ccimperial.ac.uk

14.09.–16.09.1992, München Workshop **Hypertext und Hypermedia 1992: Konzepte und Anwendungen auf dem Weg in die Praxis**
Veranstalter: GI–FG 4.9.2 (Multimediale Elektronische Dokumente)
Information: Dr. Ralf Cordes, Telenorma GmbH, Abt. EVO 3, Mainzer Landstr. 128–146, W-6000 Frankfurt 1

07.10.–09.10.1992, Nürnberg Tagung Verarbeitung natürlicher Sprache **KONVENS-92**
Termine: Beiträge erbeten bis 15.04.1992
Veranstalter: GI-FG 1.3.1 (Natürlichsprachliche Systeme), GI-FG 1.3.2 (Gesprochene Sprache)
Information: Christine Harms, c/o GMD, Postfach 13 16, W-5205 Sankt Augustin 1
Tel: (02241) 14-2473 Fax: (02241) 14-2618
email: christine.harms@kmx.gmd.dpe.de

17.02.–19.02.1993, Hamburg, Deutschland
2. Deutsche Tagung über Expertensysteme **XPS-93**
Veranstalter: GI-FA 1.5 (Expertensysteme)
Information: Prof. Dr. Wolff von Gudenberg, Universität Würzburg, Institut f"r Informatik, Am Hubland, W-8700 Würzburg

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

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

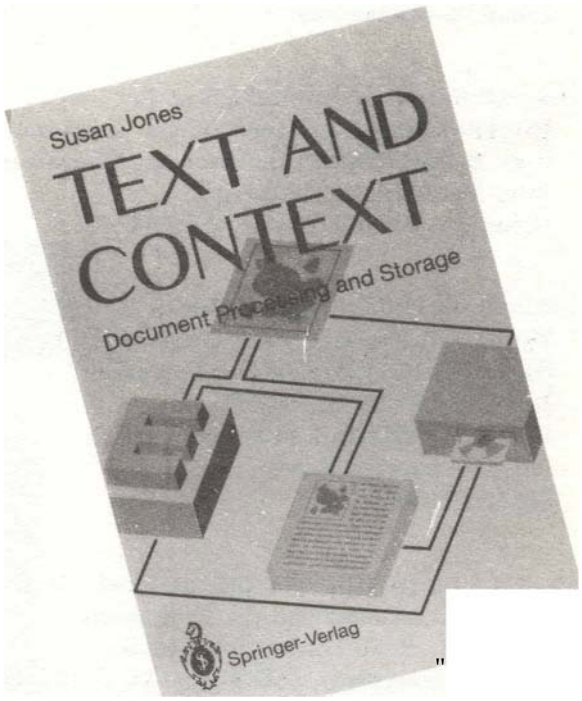
4.10.–7.10.93 in Frankfurt a.M.
7. Internationaler Kongreß der Deutschen Gesellschaft für Semiotik.
Auskunft: Brigitte Schlieben-Lange,
Mittelweg 1B, W-6368 Bad Vilbel.

Sprachdatenverarbeitung für Übersetzer und Dolmetscher

Symposium zum Abschluß des Saarbrücker Modellversuchs 28./29. September 1992

Der Saarbrücker Modellversuch "Studienkomponente Sprachdatenverarbeitung in der Übersetzer- und Dolmetscherausbildung" läuft zum 31. März 1993 aus. Aus diesem Grunde findet zum Abschluß des Modellversuchs am 28./29. September 1992 ein Symposium statt. Hierbei sollen die Ergebnisse und Erfahrungen des Modellversuchs vorgestellt und die Themenschwerpunkte "Rechnergestütze Terminologiearbeit", "Maschinengestütze und Maschinelle Übersetzung" sowie "Weitere Einsatzbereiche der Sprachdatenverarbeitung für Übersetzer und Dolmetscher" in Vorträgen und Diskussionen behandelt werden. Vor dem Hintergrund dieser Themenschwerpunkte sollen im Rahmen einer abschließenden Podiumsdikussion mit Vertretern aus Praxis und Lehre Perspektiven für die Aus- und Weiterbildung sowie Entwicklungsmöglichkeiten des Berufsbildes diskutiert werden.

Veranstalter:	Fachrichtung 8.6 der Universität des Saarlandes
Termin:	28./29: September 1992
Tagunsgebühr:	DM 50,00
Informationen bei:	DM 30,00 für Studierende Karl-Heinz Freigang, Felix Mayer Fachrichtung 8.6, Universität des Saarlandes Im Stadtwald, D-6600 Saarbrücken 11
Tel:	0681/302-2929 oder 0681/302-3204
Fax:	0681/302-4440
E-mail:	Sl16wwb@rz-uni-sb.de



1991. XIII, 298 pp: 66 figs. Softcover DM 54,ISBN 3-540-19604-8

Text and Context: Document Processing and Storage describes information processing techniques, inc1uding those which do not appear in conventional textbooks on database systems. It focuses on the input, storage, retrieval and presentation of primarily textual information, together with auxiliary material about graphic and video data. There are chapters on text analysis as a basis for lexicography, full-text databases and information retrieval, the use of optical storage for both ASCII text and scanned document images, hypertext and multi-media systems, abstract document definition, and document formatting and imaging.

The material is presented informally, with an emphasis on current text-processing applications and software. There are, among others, case studies from Reuters, British Airways, Sony, and S1. Bartholomew's Hospital in collaboration with McMaster University (Canada). Relevant industry standards are discussed inc1uding ISO 9660 for CD-ROM file storage, CCITT Group 4 data compression, the Standard Generalised Markup Language and Office Document Architecture, and the Postscript language.

- Heidelberger Platz 3. W-1000 Berlin 33. F.R. Germany
- 175 Fifth Ave., New York, NY 10010, USA
- 8 Alexandra Rd., London SWI9 7JZ, England
- 26, rue des Carmes, F-75OO5 Paris, France
- 37-3, Hongo3-cbome, Bunkyo-ku, Tokyo 113, Japan
- Room 701, Mirror Tower, 61 Mody Road, Tsimshatsui, Kowloon, Hong Kong
- Avinguda Diagonal, 468-40 C, E-08oo6 Barcelona, Spain D Wesselényi u. 28, H-1075 Budapest, Hungary

Mitteilungen aus der GLDV

1. Im Rahmen der diesjährigen ersten gemeinsamen *Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)* in Nürnberg (07.-09.10.) wird auch die Mitgliederversammlung 1992 der GLDV abgehalten und zwar am

Mittwoch: 08.10.1992, 16.30 Uhr;

die satzungsgemäße Einladung dazu wird rechtzeitig erfolgen.

2. Nach dem Beschluß der Mitgliederversammlung-1991 wird die Jahrestagung-1993 in Kiel stattfinden. Auf der letzten Sitzung des Vorstands wurde daraufhin nach Vorschlag des Beirats der thematische Schwerpunkt für die Jahrestagung diskutiert und beraten. Danach wird die *Kieler GLDV-Jahrestagung-1993* dem Thema gelten:

Sprachtechnologie.
`Methoden, Werkzeuge, Perspektiven`

Als Termin ist die 9. Kalenderwoche ins Auge gefaßt, wobei bisher der Zeitraum

Montag, 01.03. - Mittwoch, 03.03.1993

angesetzt wurde, um eine terminliche Überlappung mit der DGfS-Jahrestagung in Jena in der gleichen Woche (nach derzeitigem Planungsstand: 03.05.03.1993) möglichst gering zu halten.
Organisation und Planung liegt bei *Horst* P. *Pütz* (Kiel), der auch zusammen mit Johann *Haller* (Saarbrücken) den Call-for-Papers vorbereitet, welcher Ende Juni 1992 versandt werden wird.
Das Programm-Kommitee setzt sich aus den Mitgliedern des Beirats sowie einzelnen Fachreferenten zusammen. Vorgesehen ist, daß die *Sektionen* der Jahrestagung von einzelnen Beiratsmitgliedern und Fachleuten vorbereitet, organisiert und z. T. auch geleiten werden. Es soll damit sichergerstellt werden, daß der Tagungsband (bei *Olms*, Hildesheim) zur Tagung erscheinen kann. Hier die wichtigsten Daten (des vorläufigen Zeitrahmens):

0 Deadline: 15.10.92 Kurzversionen der einzureichenden Beiträge (min.5 bis max. 10 Seiten)

0 Benachrichtigung: 30.11.1992 über Annahme

0 Ablieferung: 31.12.1992 der druckfertigen Manuskripte Die endgültigen Termine werden im CfP mitgeteilt.

3. Mit dem Wunsch nach schnellerem Meinungsaustausch, GLDVinterner Diskussionen wichtiger Themen, Standpunkte, Meinungen, aber auch nach aktuellerer Information über kurzfristig aufkommende Nachrichten, auslaufende Fristen, fällige Daten, Erinnerungen, etc. im Rahmen des `GLDV-Newsletter` *(gldv-nl)*, verbindet der Vorstand die Bitte. an alle Mitglieder der G LDV, Email Anschriften soweit vorhanden elektronisch mitzuteilen an:

ldvforum%utrurt@unido.uucp.de oder
unido!utrurt!ldvforum.

Denn Dietmar Rösner plant, alle GLDV-Mitglieder in ein *gldv-nl-Abonnement* aufzunehmen und hofft, daß die inzwischen ja beträchtlich vergrößerte Zahl der Mitglieder mit Email-Anschluß auch in zunehmend größeren Aktivitäten innerhalb dieses von ihm betreuten GLDV-Angebots sich umsetzen wird. Und darüber hinaus würde es auch der Vorstand begrüßen, wenn er (schon aus Gründen der Arbeitsvereinfachung) auch diejenigen der (meist langjährigen) GLDV-Mitglieder elektronisch erreichen könnte, deren inzwischen vorhandene Email-Anschriften ihm bisher unbekannt geblieben sind.

4. Richtigstellung
In der letzten Ausgabe des GLDV-Info-Faltblatts (Dezember 1991) stellt sich die *GLDV* u. a. mit Ihren Publikationen vor. Versehentlich ist dort von einem Studienführer für die "deutschsprachigen Universitäten" die Rede. Diesen Anspruch kann und will die GLDV nicht erheben, vielmehr hat der Arbeitskreis 'Ausbildungs- und Berufsperspektiven' eine bescheidenere Aufgabe gelöst und den "Studienführer LDV/CL für die *deutschen* Universitäten" vorgelegt (alle weitergehenden Ambitionen lagen dem Arbeitskreis fern).