Editorial

LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-FORUM 5(1987/88)4
Forum der Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)

Herausgeber
Gesellschaft für Linguistische Datenverarbeitung (GLDV)

Redaktion
Prof. Dr. G. Knorz, FH Darmstadt, Fachbereich IuD, Schöfferstraße 8, D-6100 Darmstadt
Tel.: (06151) 317881)
Netz: xid2knor@ddathd21.bitnet
unter Beteiligung der Projektgruppe "Systementwicklung für elektronisches Publizieren" am Fachbereich IuD der FH Darmstadt.
Die Formatierung der Beiträge für den elektronische Satz lag überwiegend in den Händen von Frau Antje Kress und Herrn Claus Ritzler, die Datenbank–Erfassung für den Veranstaltungskalender übernahm Frau Birgit Anderson.

Wissenschaftlicher Beirat des LDV-Forum:
I. S. Bátori, R. Drewek, Ch. Habel, P. Hellwig, G. Knorz, J. Krause, R. Kuhlen, H.D. Lutz, D. Rösner

Wissenschaftliche Zusammenarbeit
Prof. Dr. J. Raben

Erscheinungsweise:
halbjährlich, bis Bd. 5, Heft 1 einschließlich: zum 30. Juni und 30. Dezember – ab Bd. 5, Heft 2/3: zum 31. März und 30. September

Bezugsbedingungen:
Der Bezug des LDV-Forum ist für Mitglieder der GLDV im Jahresbeitrag eingeschlossen. Jahresabonnements

Editorial

Leser bekommen in der Regel recht schnell einen *Blick* für das typische Erscheinungsbild einer Zeitschrift, worauf die Verantwortlichen auch tatsächlich nicht geringen Wert legen. Wer will schon gerne namenlos bleiben und als unterschiedslos zur Konkurrenz beurteilt werden? Darüber hinaus ist das verläßliche und stabile Gestaltungsprinzip einer Zeitschrift auch ein Dienst am Leser, der sich auf diese Weise ohne langes Suchen und Rätseln in den ihm bekannten Informationsstrukturen zurechtfindet. In diesem Sinne hat sich auch das *LDV-Forum* bisher um eine angemessene äußere Erscheinung bemüht: die Gestaltung des Titelblattes, das schlanke Format, die Aufmachung der fachlichen und redaktionellen Beiträge. Bereits auf den ersten Seiten werden Sie allerdings diesmal eine deutliche Änderung des bisherigen Layouts feststellen können und TEX–Kenner werden unschwer den Grund erraten: Die Redaktion, unterstützt durch die Projektgruppe *Systementwicklung für elektronisches Publizieren* am Fachbeieich IuD der FH Drmstadt, hat versucht, auf die Satzerstellung mit LATEX umzustellen. Wie (fast) nicht anders zu erwarten, war die Umstellung längst noch nicht abschließend vorbereitet, als die Entscheidung für die neue Vorgehensweise getroffen werden mußte. Die Folge ist, daß manches im Layout weniger den Vorstellungen der Redaktion, sondern vielmehr den einfachsten Möglichkeiten von LATEX entspricht. Ein bis auf weiteres stabiles Layout ist also erst ab der nächsten Ausgabe zu erwarten. Die Vorteile der Umstellung sollten in einer besseren Lesbarkeit und in einer Vereinfachung der redaktionellen Arbeitsschritte liegen. Verbesserungsvorschläge seitens der LeserInnen sind ausdrücklich erwünscht! Autoren fordern möglichst frühzeitig die neuen Autorenrichtlinien an!

Beim Blick auf das Inhaltsverzeichnis wird Ihnen eine weitere Besonderheit auffallen: Viele Rubriken fehlen diesmal vollständig. Der Grund: Neben ca. 40 Seiten Fachbeiträgen waren aus aktuellem (und zum Teil sehr kurzfristig angemeldetem) Anlaß etwa genauso viele Seiten als Ergebnis von Aktivitäten der GLDV und seiner Mitglieder vorzusehen! Auch eine Art von (erfreulicher) Schwerpunktbildung, wenn auch mit gewissen redaktionellen Problemen!

Neben dem ausführlichen Bericht von U. Heid über das Arbeitstreffen des Lexikographie–Arbeitskreises und dem anregenden Diskussionsbeitrag von R. Drewek ist der letzte Beitrag dieses Heftes besonders hervorzuheben. Er ist so formatiert, daß er sich leicht aus dem Heft herauslösen läßt und selbständig behandelt werden kann. Unter dem Titel **Zur Konturierung des Faches Computerlinguistik** finden Sie dort das Ergebnis der Beratungen der *Arbeitstreffen "CL–Studiengänge"*. Die Kapitel 1 bis 3 stellen eine Überarbeitung dessen dar, was im *LDV-Forum* Bd. 5, Nr. 1., S. 76-78 vorab veröffentlicht worden war. Die Kapitel 4 bis 7 sind dem-

gegenüber neu. Zu diesen Kapiteln finden Sie außerdem unter der Rubrik *Arbeitskreise* "Hintergründe und Erläuterungen" von Magdalene Lutz-Hensel.

Die *Arbeitstreffen "CL–Studiengänge"* stellen das Ergebnis ihrer Arbeit hiermit zur Diskussion. Auf der Mitgliederversammlung der GLDV im Februar/März 1988 besteht die Möglichkeit zur Aussprache und Verabschiedung.

Die nächste Ausgabe des *LDV-Forum* soll den diesmal zu kurz gekommenen Aspekten Gerechtigkeit widerfahren lassen: Die für diesmal zurückgestellten Tagungsberichte sind nur aufgeschoben, für die Nachrichtenrubrik liegt mehr als genug Material bereit, und ich erhoffe mir eine umfangreiche Rubrik mit Buchbesprechungen. Folgende Rezensionsexemplare können bei mir angefordert werden:

- Aus der Reihe *Communication in Artificial Intelligence* (Pinter Publishers, London)
  - M. Zock, G. Sabah (Hr.): Advantages in Natural Language generation. An interdisciplinary Perspective. Band 1 und 2. Juli 1988
  - E. Steiner, P. Schmidt, C. Zelinsky–Wibbelt (Hr.): From Syntax to Semantics. Insights from Machine Translation
- Aus der Reihe *Sprache und Information* (Niemeyer, Tübingen)
  - R. Kuhlen: Informationslinguistik. Theoretische, experimentelle, currikulare und prognostische Aspekte einer informationswissenschaftlichen Teildisziplin. Bd. 15, 1986
  - W. Wilss, K.-D. Schmitz (Hr.): Maschinelle Übersetzungs–Methoden und Werkzeuge. Bd. 16, 1987
  - H.-D. Luckhardt: Der Transfer in der maschinellen Sprachübersetzung. Bd. 18, 1987
- A. Kučera, A. Rey, H.E. Wiegand, L. Zgusta (Hr.): Lexicographica. Internationales Jahrbuch für Lexikographie. Tübingen: Niemeyer, 1987

Selbstverständlich können auch andere einschlägige Bücher zur Rezension vorgeschlagen werden. Von besonderem Interesse sollte beim nächsten Mal der Stand der Beantragung des DFG–Schwerpunktprogramms *Information aus sprachlich repräsentiertem Wissen* sein. Mit dem ausgesprochen interessanten Fachbeitrag von B. Endres–Niggemeyer über *Informationsorientiertes Schreiben* profitiert bereits die vorliegende Ausgabe von den vorbereitenden Arbeiten. Eine Fortsetzung derartiger positiver Seiteneffekte kann ich bereits heute versprechen.

G.K.

P.S.: Wenn jemand auf die Drohung hin *"Geld oder Leben"* sein Bargeld herausrückt, wird man ihm deswegen kaum eine besondere Großzügigkeit in Gelddingen nachsagen dürfen. Die analoge Drohung der letzten Tage heißt: *"Abschluß der Arbeiten jetzt – oder massive Verschiebung des Erscheinungstermins"*. Sagen Sie mir bitte keinen fehlenden Sinn für Sorgfalt nach!

Studienführer LDV

Lutz-Hensel, M.: Studienführer Linguistische Datenverarbeitung (LDV) für die wissenschaftlichen Hochschulen der Bundesrepublik Deutschland, 1985.
zu beziehen über: Prof. Dr. J. Krause, Universität Regensburg, Linguistische Informationswissenschaft, Postfach 397, D-8400 Regensburg

Arbeitskreise:

Ausbildung und Berufsperspektiven; LDV und Nachbarn; Maschinelle Lexikographie und Lexikologie; Maschinelle Übersetzung; Spracherkennung, Sprachgenerierung und phonetische Datenbanken; Textanalyse

Titelgestaltung:

Fred Zimmermann, *FH Darmstadt*

Rubriken:

Die mit Namen gekennzeichneten Beiträge geben ausschließlich die Meinung der Autoren wieder. Einreichungen sind an die Redaktion zu richten.

Fachbeiträge:

Fachbeiträge, die zur Veröffentlichung im LDV-Forum eingereicht sind, werden von mindestens einem Mitglied des wissenschaftlichen Beirats oder (im Ausnahmefall) von einem/einer beauftragten exteren Wissenschaftler/in begutachtet. Über das Ergebnis wird der Autor unverzüglich informiert. Manuskripte sind grundsätzlich bei der Redaktion einzureichen. Durch die Anmel-Übermittlung von Beiträgen angefordert und beachtet werden. Ein Fachbeitrag hat im Regelfall eine Länge von ca. 6 bis 8 Seiten. Für Beitragsreihen gelten besondere Randbedingungen (vgl. LDV-Forum 4(1986)2: S. 25).

Redaktionsschluß:

Redaktionsschluß für das März-Heft: 15. Jan., für das September-Heft: 15. Juli

Herstellung:

Verlagsdruckerei Hoppenstedt, Havelstraße 9, D-6100 Darmstadt

Auflage:

550 Exemplare

Anzeigen:

Media-Information kann bei der Redaktion angefordert werden.

Bankverbindung:

Sparkasse Darmstadt, BLZ 508 501 50, Kto.-Nr. 554 090

Gesellschaft für Linguistische Datenverarbeitung (GLDV)

Anschrift:

Prof. Dr. Brigitte Endres-Niggemeyer, FH Hannover, Fachbereich BID, Hannomagstr. 8, D-3000 Hannover 91
Tel. (0511) 444344

Mitgliedsbeiträge:

Für Studierende: DM 10,00; für natürliche Personen: DM 50,00; für wissenschaftliche Institute: DM 100,00; für gewerbliche Unternehmen, Behörden und andere juristische Personen: DM 250,00

Vorstand:

B. Endres-Niggemeyer *(1. Vorsitzende)*, K. G. Schweisthal *(2. Vorsitzender)*, B. Schaeder *(Schatzmeister)*, Ch. Schneider *(Schriftführerin)*

können zum Preis von DM 30,00 (incl. Versand), Einzelexemplare zum Preis von DM 15,00 (zuzügl. Versandkosten) bei der Redaktion bestellt werden.
Back-up-Exemplare bis einschließlich Ausgabe 2/85 sind zum Preis von DM 10,00, ab Ausgabe 2/86 zum regulären Preis von DM 15,00 (zuzügl. Porto) zu bestellen (Die Ausgabe 1/86 ist vergriffen).

Fachbeitrag

Strategien zur Entwicklung effizienter Analyseverfahren für die Massentextverarbeitung

Gerda Ruge
ZT ZTI INF 323
Siemens AG München

Kurzfassung

Linguistische Verfahren können heute schon in der Massentextverarbeitung eingesetzt werden, wenn sie effektiv und effizient sind. So lassen sich z.B. Repräsentationen von Texten durch Dependenzstrukturen, die einige Probleme der Massentextverarbeitung lösen, mit Hilfe einer lokalen syntaktischen Analyse gewinnen. Ein solches hier vorgestelltes Verfahren ist so schnell, daß große kommerzielle Textdatenbanken in vertretbaren Zeiten verarbeitet werden können. Am Beispiel von Konjunktionen wird gezeigt, wie linguistische Verfahren empirisch durch die Minimierung von Fehlertoken so entwickelt werden können, daß sie den Anforderungen von Effektivität und Effizienz genügen.

1 Einleitung

Damit in Literaturdatenbanken Texte gefunden werden können, muß es computergerechte Informationen über diese Texte geben. Sie können manuell oder maschinell erstellt werden. Manuell erstellte Textprofile haben den Vorteil einer guten Beschreibungsqualität, sind aber teuer und subjektiv. Die heute von Datenbankanbietern eingesetzten automatischen Verfahren können den Inhalt von Texten nur bedingt darstellen, weil sie mit Texten genauso umgehen wie mit Zahlen.

Schwierigkeiten bereitet die Analyse von Texten in Textdatenbanken einerseits, weil große Mengen verarbeitet werden müssen. Die Analyse sollte ohne Dialog auskommen, so daß Rückfragen nicht möglich sind. Andererseits gibt es keine Einschränkungen, denen die Texte unterliegen, weder bzgl. ihres Bereichs, noch bzgl. der darin vorkommenden sprachlichen Phänomene.

Die differenzierte Repräsentation des Inhalts von Freitexten, das sind Texte, die weder formalen noch inhaltlichen Beschränkungen unterliegen, ist ein noch lange nicht erreichtes Ziel. Trotzdem können linguistische Verarbeitungsmethoden heute schon nützlich für die Verarbeitung von Massentexten sein.

2 Suchverfahren im Information Retrieval

Avancierte Verfahren im Information Retrieval (IR) sind weitgehend statistisch orientiert. Statt einer Syntaxanalyse benutzen sie Wortabstände, d.h. die Anzahl von Wörtern, die zwischen zwei Wörtern in einem Text stehen; statt einer morphologischen Analyse wird die Trunkierungen verwendet, d.h. es werden Wortmengen mit gleichen Anfängen gebildet. Einen Überblick über die vorwiegend statistischen Verfahren im IR gibt [Salton 87]. Im folgenden werden die grundlegenden Techniken der Freitextsuche kurz dargestellt.

Gewöhnlich werden Dokumente durch Schlagworte charakterisiert. Bei Freitextsystemen sind alle Worte, die im Text vorkommen, außer **der, ein, und** usw., Schlagworte. Eine Suchfrage, die relevante Dokumente beschreiben soll, besteht aus Schlagworten, die mit logischen Operatoren verbunden werden können. Um Literatur mit dem Thema "Textanalyse" zu finden, könnte man die folgende Anfrage stellen:

LDV-Forum Bd. 5, Jg. 1987/88, Nr. 4, September 1988: S. 3 – 11

(1) *(ANALYSE /\ TEXT)*
V *TEXTANALYSE*

Wird eine solche Frage gestellt, bildet ein IR-System die Menge aller Dokumente, in denen Analyse gemeinsam mit Text vorkommt oder in denen Textanalyse vorkommt. Etwa Dokumente mit folgenden Ausschnitten:

(2) ... die Analyse von einem Text...

(3) ... Schwierigkeiten mit der Textanalyse sind zu erwarten...
(4) ... Nachdem der Text ausgegeben wurde, ist die Analyse der Fehler beendet...

Der Text (4) ist nicht relevant, wird aber als relevanter Text angegeben. Andererseits gibt es viele Dokumente, die relevant sind und nicht gefunden werden:

(5) **...** die Analyse von Texten ...

(6) ... Analysen eines Textes ...

Um auch Texte, die (5) und (6) enthalten, zu erreichen, kann man alle Erweiterungen von Worten mit einbeziehen, indem man sie trunkiert. Die Anfrage

(7) *(AN ALY SE? /\ TEXT?)*
V *TEXTANALYSE?*

liefert dann (5) und (6), aber auch:

(8) ... die Analyse von Textilien ...

Hier ist ein grundsätzliches Problem im IR sichtbar, die Gegenläufigkeit von Recall und Precision. Wenn eine Anfrage so modifiziert wird, daß mehr Dokumente damit beschrieben werden, wenn man also den Recall erhöht, dann erhält man auch mehr irrelevante Dokumente. Wenn man die Menge der Dokumente stärker spezifiziert, also die Precison erhöht, dann verliert man viele Dokumente.

Die Einbeziehung der Worterweiterungen in die Suche ist ein Mittel, den Recall zu erhöhen. Ein Precision erhöhendes Instrument ist die Möglichkeit, Dokumente nur dann auszuwählen, wenn nur eine beschränkte Zahl von Worten zwischen den Schlagworten steht.

(9) *(ANALYSE? adj (2) TEXT?)*
V *TEXTANALYSE?*

Damit wäre zwar (4) aussortiert, aber nicht (10), außerdem ist (11) ein relevantes Dokument, das nicht mehr gefunden wird.

(10) ... dieser Text beschreibt die Analyse von Stickoxiden **. . .**

(11) ... die Analyse von uneingeschränkten, beliebig langen Texten.. .

Ein weiteres Problem entsteht durch Synonyme, unterschiedliche Schreibweisen und Abkürzungen [Doszkocs 86]. Texte, in denen Synonyme der Suchworte benutzt werden, werden nicht gefunden.

3 Linguistische Hilfsmittel im IR

3.1 Möglichkeiten linguistischer Hilfsmittel

Die Fehler der oben beschriebenen Verfahren entstehen dadurch, daß maschinell leicht ermittelbare Näherungsrelationen für Relationen wie "hat den gleichen Wortstamm wie" oder "steht in syntagmatischer Beziehung zu" benutzt werden.

Noch ist kein Verfahren in der Lage, Freitexte aus großen Datenbanken so zu analysieren, daß die Bedeutung der Texte dargestellt werden kann. Ausgefeiltere Analyseverfahren behandeln immer nur Texte aus einem kleinen Bereich. Selbst die Syntaxanalyse für Frei
texte birgt große Schwierigkeiten, weil keine Rückfragen möglich sind und der Parser nie in einen Fehlerzustand laufen darf. Alles, sei es noch so unkonventionell oder sogar fehlerhaft, etwa bei nicht- muttersprachlichen Autoren, muß akzeptiert werden. Konventionelle formale Beschreibungen der Sprache geben an, was akzeptabel ist. Eine solche Philosophie ist bei Freitexten aufgrund ihrer oben genannten Eigenschaften nicht angemessen. "Thinking Machines" plante, einen ATN Parser im IR zu benutzen. Er wurde zwar entwickelt, aber nicht eingesetzt [Waltz 87].

Vielleicht ist es nicht nötig, für die Zwecke des IR eine tiefe Analyse durchzuführen [Doszkocs 86b]. Gewöhnlich rechnet man

in einer Literaturrecherche damit, daß die Hälfte der relevanten Texte nicht gefunden wird, während die Hälfte der angebotenen Texte nicht relevant ist. Nach einer U ntersuchung von [Smeaton 86] verbessert schon die Information, ob zwei Worte im gleichen Satz vorkommen, die Precision. Das gilt in noch stärkerem Maß für die Information, ob sie in der gleichen Phrase stehen. In der Untersuchung von Smeaton wurden die Phrasengrenzen von Hand festgelegt. Dieses zu automatisieren, ist bereits ein Problem, wenn es auf Freitexte angewendet werden soll. Besonders im Englischen wird man sich mit syntaktischen Phänomenen auseinandersetzen müssen, weil Verben, die zwischen den Phrasen stehen, oft Homographen sind.

Seit langem überwiegt die Ansicht, daß Phrasen ein Dokument besser charakterisieren können als einzelne Schlagworte [Seelbach 75]. Wenn [Salton 86] bisher nur wenig Erfolg damit feststellen kann, so liegt das vielleicht daran, daß die getesteten Verfahren vorwiegend ohne linguistische Mittel arbeiten.

Werden Phrasen zur Charakterisierung von Textinhalten gewählt, so stellt sich die Frage, wie eine Normierung, die Paraphrasen auf einen gemeinsamen Repräsentanten ab bildet, erreicht werden kann.

(12) ... content of water ...

(13) ... content in water ...

(14) ... water content. ..

(15) ... studies in text analysis...

(16) ... studies on text analysis. ..

Sollen z.B. Präpositionen in den Repräsentanten der Paraphrasen aufgenommen werden? Damit würden (12) und (13) als nicht bedeutungsgleich dargestellt, aber auch (15) und (16). Wie kann man (14) so umformen, daß es zu (12) paßt, oder sollte (12) an (14) angepaßt werden?

[Berrut 86] läßt die Präpositionen in einer Phrase, beachtet aber nicht, daß Paraphrasen unterschiedliche Wortreihenfolgen haben können. [Andreewsky 77] versucht außer Worten, die nebeneinander stehen, noch zusätzlich Wortpaare zu finden. Auch er behält die Präpositionen bei, unterscheidet aber Präpositionalphrasen von den zugehörigen Komposita.

Die bei Siemens entwickelten syntaktischen Verfahren für das **IR** stellen alle Spezifikationen zwischen Worten in einer Phrase dar [Schwarz 86b], die Phrasen werden in Dependenzstrukturen transformiert. Dabei gehen Unterscheidungen durch verschiedene Präpositionen verloren, doch werden alle Phrasen als gleich betrachtet, die den gleichen Head mit gleichen Spezifikationen haben. Hier werden (12), (13) und (14) nicht unterschieden, gen au wie (15) und (16). Phrasen, die sich nur durch eine Präposition unterscheiden, zu identifizieren, führt nur dann zu einem Fehler, wenn ein und dasselbe Wort ein anderes in verschiedenen Rollen spezifizieren kann. Bei der Evaluierung der Recherche mit Dependenzstrukturen in dem System COPSY wurden kaum derartige Fehler festgestellt [Schwarz 86a].

3.2 Anforderungen an linguistische Hilfsmittel

Probleme entstehen bei der Analyse von Freitexten nicht nur dadurch, daß Texte uneingeschränkten Inhalts verarbeitet werden müssen (s.o.), auch Effizenzfragen schränken bei den großen Textmengen die möglichen Verfahren ein. Für die Zwecke des **IR** ist es nicht unbedingt nötig, eine tiefgehende Analyse, die jedes Phänomen behandelt, vorzunehmen.

Linguisten sind es gewohnt, zu jeder Theorie Gegenbeispiele zu finden, auch wenn die dabei entstehenden Sätze zuweilen etwas ungewöhnlich sind. Die so gezeigten Phänomene kommen oft sehr selten vor; die Analyse der Probleme dient im allgemeinen zur Bildung einer Sprachtheorie.

Empirische Untersuchungen zeigen, daß einige wenige Konstruktionen schon den größten Teil der Token im Text bilden. [Schüler 86] stellt beispielsweise bei einer Untersuchung von Nominalgruppen in unterschiedlichen Texten fest, daß fast die Hälfte der Token aus einzelnen Nomen oder Nomen mit einem Determinator bestehen. Es zeigt sich auch bei [Schwarz 88], daß mit einigen wenigen Regeln ein guter Grundstock für ein Analysesystem entsteht. Je komplexer und schwieriger beschreibbar Konstrukte sind, um so seltener kommen sie vor. Diese empirische

ermittelte Eigenschaft von Freitexten führt zusammen mit der Notwendigkeit von Effizienz zu den folgenden Anforderungen, die wir an Regeln, die in ein System übernommen werden sollen, stellen:

- Die Regeln sind korrekt, d.h. sie wurden empirisch evaluiert.
- Die Regeln sind effektiv, d.h. die betroffene Konstruktion muß so häufig vorkommen, das sie einen signifikanten Einfluß auf die Fehlerquote hat.
- Die Regeln bilden ein effizientes System - auch breit wirkende Regeln können nicht benutzt werden, wenn dadurch die Rechenzeit übermäßig groß wird.

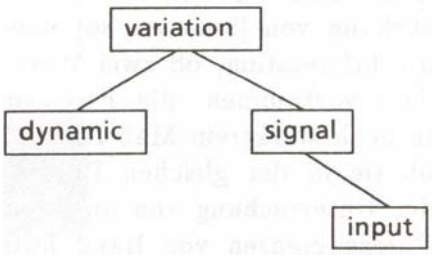
4 Generierung von Dependenzstrukturen aus Freitexten

4.1 Dependenzstrukturen

Um englische Texte für die Literaturrecherche zu repräsentieren, wurden bei Siemens im System COPSY die Dependenzstrukturen aller im Abstract des jeweiligen Textes enthaltenen Phrasen gewählt. Damit können beispielsweise Texte, in denen die einzelnen Worte einer natürlichsprachlichen Suchanfrage vorkommen, nach ihrer Relevanz bzgl. dieser Suchfrage geordnet werden, indem die Dependenzstrukturen der Suchfrage und der Texte verglichen werden.

COPSY benutzt keine vollständige Syntaxanalyse für Sätze, sondern Kontextmuster. Zuerst werden die Phrasen isoliert. Als Phrasengrenzen werden im wesentlichen Interpunktionszeichen und Verben betrachtet, allerdings nur unter bestimmten Kontextbedingungen. Die Dependenzstruktur der Phrase entsteht durch das Eintragen von Kanten zwischen Worten, die direkt in einer Spezifikationsrelation zueinander stehen. Einen Einblick in die Problematik der Isolation der Phrasen, der Lemmatisierung und Kategorisierung der Worte und der Bestimmung der Spezifikatonsrelationen gibt [Schwarz 88]. In Beispiel (17) sind einige der häufigsten Konstruktionen, mit denen Spezifikationsrelationen induziert werden, dargestellt: durch Adjektive,

dynamic variations of the input siQnal



Beispiel einer Dependenzstruktur

Abbildung 1: (17) dynamic variation of the input signal

Komposition und Präpositionen. Unter Spezifikation ist zu verstehen, daß ein Term durch einen anderen modifiziert wird. Geschieht das mit syntaktischen Mitteln, so nennen wir das einen Link zwischen den Worten. Ob ein Link in einer Phrase enthalten ist, kann getestet werden, indem die Wortpaare, die sich aus der Phrase bilden lassen, betrachtet werden. In den Wortpaaren steht der Head, das spezifizierte Wort, an zweiter Stelle, in den Dependenzstrukturen ist er das obere Wort. In Beispiel (17) sind das: dynamic variation, input signal und signal variation, aber nicht input variation oder variation signal.

4.2 Die Leistungsfähigkeit von empirisch entwickelten syntaktischen Regeln

Die Regeln, die zu Kontextmustern die darin enthaltenen Links angeben, wurden empirisch entwickelt. Hypothesen über solche Regeln wurden anhand von Patentabstracts des amerikanischen Patentamts entwickelt. Diese Texte haben den Vorteil, daß sie von verschiedenen Themen handeln und von unterschiedlichen, teilweise nicht muttersprachlichen Sprechern verfaßt wurden.

Falls eine Hypothese nicht verworfen wird, jedoch nicht ausreicht, führt die Analyse ihrer Fehlerfälle wieder zu neuen Hypothesen. Durch sukzessive Verbesserung des Systems an den Stellen der größten Fehler entstand schließlich ein Version, die 85% aller Links er

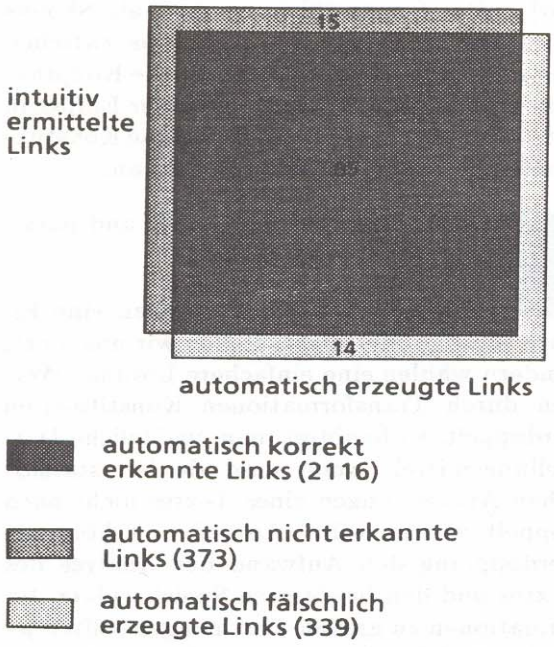


Abbildung 2: Korrelation von intuitiv gesetzten und automatisch erzeugten Links

kennt und dabei nur 14% Fehler macht. Zur Evaluierung des Systems wurden die intuitiv von Menschen gesetzten Links M mit den von der letzten Version der Dependenzstrukturgenerierung erzeugten, der Menge S, verglichen. Die Mengen M und S wurden für 50 Abstracts, die nicht in die Entwicklung der Regeln einbezogen waren, ermittelt. Die Menge M entspricht mit ihren ca. 2500 Links dem Umfang, der im PADOK-Test [Krause 86] gewählt wurde.

Da in M Links enthalten sind, die nicht in S sind, also fehlende Links, und umgekehrt in S Links enthalten sind, die es in M nicht gibt, also falsche Links, kann weder M noch S als Basis für ein Gütemaß, das aus nur einem Wert besteht, genommen werden. Statt dessen wird die Güte des Verfahrens durch die Korrelation zwischen M und S angegeben, d.h. dem Maß der Überlappung von M und S.

$$k(M,S) = \frac{|M \cap S|}{|M \cup S|}$$

Der Wert einer Korrelation liegt zwischen 1 und 0 und ist dimensionslos. Eine Korrelation von 1 bedeutet äquivalenz, ist also der anzustrebende Wert, während eine Korrelation von 0 entsteht, wenn sich die Mengen nicht überlappen. Die intuitiv und automatisch erzeugten Links korrelieren in unserem Fall mit dem Wert 0,75.

Die Effizienz des Systems ist hoch. Die Transformation leistet etwa 5 DIN A 4 Seiten Text pro CPUsec auf einem Siemens BS2000 Großrechner. Damit werden 20 MB Text in ca. einer Stunde Realzeit verarbeitet, so daß die Verarbeitung von ganzen Textdatenbanken möglich wird. Diese Leistung konnte nur dadurch erreicht werden, daß der Aufwand für die Analyse gegen ihren Effekt abgewogen wurde.

5 Verhältnis von Aufwand und Wirkung am Beispiel des Konjunktionsskopus

5.1 Darstellung von Konjunktionen in Dependenzstrukturen

Den größten Anteil an der Fehlerrate bei dem System COPSY hatte die Behandlung von Konjunktionen, die vorläufig als Phrasengrenzen behandelt wurden [Schwarz 86b]. Unter der Voraussetzung, daß die Disambiguierung von konjugierten Phrasen nur mit semantischen Mitteln möglich ist - diese Hypothese drängt sich auf, wenn man sich ein paar Beispiele überlegt -, wäre der Aufwand dafür unvertretbar. Es würde ein ausführliches Lexikon der ganzen Sprache mit allen ihren Fachsprachen benötigt. Daß aber heuristische Regeln für die Textanalyse empirisch entwickelt werden können, weil ein großer Teil von natürlichen Texten eine einfache Struktur hat, soll hier beispielhaft für die Behandlung von Konjunktionen gezeigt werden.

Werden Konjunktionen, damit sind hier nur **and** und **or** gemeint, behandelt, so müssen die Links, die über eine Konjunktion reichen, mit in die Dependenzstrukturen aufgenommen werden.

Die Disambiguierung der konjungierten Phrasen reduziert sich auf das Bestimmen des Bereichs, in dem die Phrasenteile liegen, die durch die Konjunktion zusammengefaßt werden. Diesen Bereich nennen wir den Skopus der Konjunktion. In den Beispielen (18) und (19) in Abb. 3. und Abb. 4. sind das **[start and stop]** bzw. **[actuated and control-**

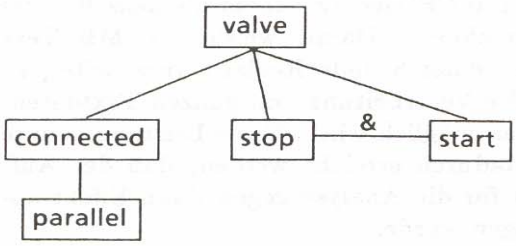


Abbildung 3: Beispiel für Dependenzstruktur mit Konjunktionen (18) *parallel connected start and stop valves* Links, die über die Konjunktion greifen: start valve

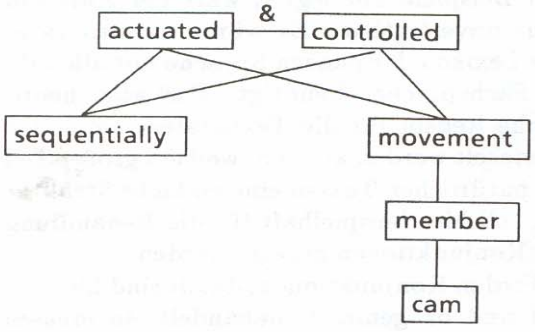


Abbildung 4: Beispiel für Dependenzstruktur mit Konjunktionen (19) *sequentially actuated and controlled by the movement of cam members* Links, die über die Konjunktion greifen: sequentially controlled, movement actuated

led]. [Das-Gupta 87] betrachtet als Skopus eines **and** grundsätzlich alles, was zwischen den zwei Präpositionen liegt, die die Konjunktion umschließen. Er geht von einer Ellipse in (18) aus und wendet dann auf solche Konstuktionen die Transformation in (20) an.

(20) parallel connected start valve and parallel connected stop valve

Die Frage, ob bei Konjunktionen eine Ellipse vorliegt oder nicht, stellen wir uns nicht, sondern wählen eine einfachere Lösung. Werden durch Transformationen Konstituenten verdoppelt, so benötigt man zusätzliche Darstellungsmittel, wenn man sie bei statistischen Auswertungen eines Textes nicht auch doppelt werten will. Außerdem erhöht die Verdopplung den Aufwand der Analyse des Textes und den benötigten Speicherplatz, Informationen zu großen Textmengen sollten jedoch möglichst komprimiert dargestellt werden.

Wir legen den Skopus so fest, daß keine Transformationen mehr nötig sind. Der Skopus der Konjunktion ist dann der Bereich, der in einer Paraphrase wie (20) nicht verdoppelt würde. Die Dependenzstruktur kann bei bekanntem Skopus leicht erzeugt werden, indem je eine Teilstruktur für die linke und rechte Seite innerhalb des Skopus erzeugt wird und die Heads der beiden konjugierten Teilstrukturen gemeinsam so in die Struktur eingefügt werden, wie ein einzelnes Wort der gleichen Kategorie.

Nicht zu jeder Konjunktion gibt es Links, die über sie hinausreichen, weil nicht alle Konjunktionen innerhalb einer Nominalphrase liegen, z.B. Satzkonjunktionen oder Konjunktionen, die Verbalphrasen verbinden. Da das System nicht mit einer vollständigen Syntaxanalyse arbeitet, sondern mit lokalen Kontextregeln, können Links wie in (21) zwischen **positioned** und **inlet** nicht erkannt werden, wenn die Worte weit auseinander liegen.

(21) A unique diffusion plate containing a large number of small diameter diffusion holes is positioned under the ice cream or soft drink mixture and above the inlet for the compressed gas, ...

In solchen Fällen wird eine Konjunktion dann wieder als Phrasenende behandelt. Es

zeigt sich, daß trotzdem gute Ergebnisse für die Konjunktionsbehandlung möglich sind, weil der größte Teil der Token von Konjunktionen in viel einfacheren Konstruktionen vorkommt.

Da es für die Generierung von Dependenzstrukturen genügt, zur Analyse von Konjunktionen festzustellen, ob sie Phrasenteile verbinden, und wenn, wie weit ihr Skopus reicht, wird im folgenden nur noch der Konjunktionsskopus betrachtet.

5.2 Entwicklung der Fehlerquote bei Steigerung des Aufwandes

Bei der Entwicklung der Regeln für die Bestimmung des Konjunktionsskopus untersuchten vier linguistisch vorgebildete Entwickler jeweils ca. 1000 Token von Konjunktionen aus Patentabstracts bzgl. unterschiedlicher Regelhypothesen in diversen Arbeitsgängen. Die während der Untersuchungen entstanden Statistiken bauen nicht systematisch aufeinander auf, da Kriterien immer wieder geändert wurden. Die folgenden Angaben über Fehlerraten oder Häufigkeiten beziehen sich, falls nicht anders angegeben, auf 230 aufeinander folgende Token einer systematischen Nachuntersuchung. Dabei wurde nicht die absolute Anzahl der Links ermittelt, sondern die Quote der korrekt behandelten Token.

Zur Generierung der Dependenzstrukturen müssen die Nominalphrasen aus den Texten isoliert werden. 30% der Konjunktionen verbinden Verbalphrasen oder Sätze. Die vorläufige Lösung, Konjunktionen als Phrasengrenzen zu betrachten, ergibt also schon fast ein Drittel für unsere Zwecke korrekt behandelte Fälle. Wenn die Konjunktion als solche erkannt ist, erfordert diese vorläufige Lösung keinen weiteren Aufwand.

Das einfachste Muster von Konjunktionen sind Fälle, in denen zwei Worte gleicher Kategorie links und rechts neben der Konjunktion stehen. Es macht 20% aller Token aus. Die zwei Worte entsprechen, wenn sie keine Nomen sind, fast immer dem Skopus.

(22) which is [rotating or rolling] about its axis

(23) the apparatus is [arranged and constructed]

Wird das als Regel in die Konjunktionsbehandlung einbezogen, so werden bereits 45% der Fälle richtig behandelt. Es sind nicht 50%, weil die Regel für Komposita, die durch Konjunktionen verbunden sind, nicht immer gilt und auch einige der zuerst getrennten Fälle jetzt fälschlich verknüpft werden.

27,5% der Token betreffen Konjunktionen von Nominalphrasen, bei denen die Kategorien der Worte rechts und links von der Konjunktion nicht gleich sind, und die restlichen 22,5% betreffen komplexere Fälle, wie z.B. between ... and ... oder die Konjunktion von Präpositionalphrasen.

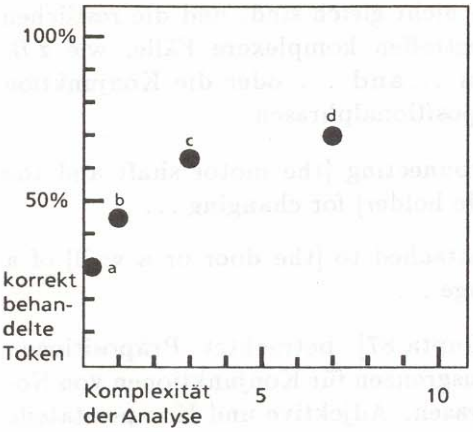
(24) ... connecting [the motor shaft and the blade holder] for changing ...

(25) ... attached to [the door or a wall] of a garage. . .

[Das-Gupta 87] betrachtet Präpositionen als Skopusgrenzen für Konjunktionen von N 0-minalphrasen. Adjektive und Kompositateile innerhalb dieses Skopus spezifizieren in der Regel nach rechts, während die durch Präpositionen angebundenen Worte außerhalb dieses Skopus nach links spezifizieren. Die Skopusgrenzen an die Stellen, wo sich die Spezifikationsrichtung umdreht, zu legen, ist deshalb eine Verallgemeinerung von Das-Guptas Regel. Damit werden auch Fälle wie (24) richtig behandelt. Benutzt man die verallgemeinerte Regel zusammen mit der Kategoriengleichheit, so werden 61 % der Token richtig behandelt.

Für die letzte implementierte Version wurden spezielle Muster, wie between **...** and **...** und konjungierte Komposita hinzugenommen. Außerdem konnten einige Fälle von Konjunktionen zwischen Nominalphrasen, die jedoch Ende und Anfang von zwei Sätzen sind, als solche identifiziert werden. Mit dieser letzten implementierten Methode können 70% der Token korrekt behandelt werden.

In Abb. 5. ist angegeben, wie sich die Quote der korrekt behandelten Fälle auf den Aufwand auswirkt. Je nach Implementierung ist der Aufwand verschieden, wenn er in Vergleichen gemessen wird. Hier wurde als Grundlage für die Aufwandsabschätzung die Anzahl der festzustellenden Fakten betrachtet. Das Feststellen, ob zwei Worte gleiche Kategorien haben, ob Numeruskongruenz



(a) Konjunktionen werden grundsätzlich als Phrasengrenzen behandelt

(b) Ausnahme zu (a) sind Fälle mit Worten gleicher Kategorie neben der Konjunktion

(c) zusätzlich zu (b) werden alle Fälle aufgenommen, in denen Nominalphrasen verbunden werden

(d) neben (c) werden Sonderfälle wie between..and und ein größerer Kontext betrachtet

Abbildung 5: Abhängigkeit der Korrektheit der Analyse des Konjunktionsskopus von seiner Komplexität

vorliegt, oder ob es sich um ein bestimmtes Wort handelt, zählen als eine Aufwandseinheit. Angegeben ist das gewichtete Mittel des Aufwands für die Bestimmung des Skopus eines Konjunktionstokens. Es wird davon ausgegangen, daß die Konjunktion be reits als solche erkannt ist und alle Informationen über den Text außer dem Konjunktionsskopus, wie z.B. Wortkategorien, vorliegen. Das leisten die anderen Komponenten unseres Verfahrens ähnlich effizient.

Eine Extrapolation wie z.B. für Abb. 5. hat keine Beweisfunktion. Sie unterstützt jedoch die Vermutung, daß die Möglichkeiten, lokaler syntaktischer Methoden zur Analyse von Konjunktionen nicht viel mehr als 70% der Fälle abdecken können. Das heißt aber auch, daß 70% aller Konjuntionen ohne die Hilfe eines Lexikons mit semantischen Informationen korrekt behandelt werden können.

Anmerkung Die beschriebenen Ergebnisse entstanden im Rahmen des Projekts TINA (**Tin**haltsanalyse), an dem A. Baumer, G. Ruge, C. Schwarz unter Überstrapazierung von vielen Werkstudenten arbeiten.

Literaturverzeichnis

[Andreewsky 77] Andreewsky A., Fluhr C., Debili F.: Computational Learning of Semantic Lexical Relations for the Generation and Automatical Analysis of Content. Information Processing 77, 667-672

[Berrut 86] Berrut C., Palmer P.: Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing. In: ACM Conference on Research and Development in Information Retrieval, 1986, 123-130

[Das-Gupta 87] Das-Gupta P.: Boolean Interpretation of Conjunctions for Document Retrieval. J.ASIS 38/4, Juli 1987, 245-254

[Doszkocs 86] Doszkocs T. E.: IR, NLP, AI and UFOS: or IR-Relevance, Natural Language Problems, Artiful Intelligence and User-Friendly Online Systems. In: ACM Conference on Research and Development in Information Retrieval, 1986, 49-57

[Doszkocs 86b] Doszkocs T. E.: Natural Language Processing in Information Retrieval. J.ASIS 37/4, Juli 1986, 191-196

[Krause 86] Krause, J.: PADOK, Test und Vergleich von Texterschließungssystemen für das Deutsche Patent- und Fachinformationswesen, Endbericht. Universität Regensburg, Linguistische Informationsverarbeitung, 1986, 86-87

[Salton 86] Salton G., McGill, M.J.: Information Retrieval. McGraw–Hill, Hamburg 1987

[Salton 87] Salton G.: On the use of Term Association in Automatic Information Retrieval. In: COLING 86, 380-386

[Seelbach 75] Seelbach D., Computer-Linguistik und Dokumentation: Key-Phrases in Dokumentationsprozessen. Verlag Dokumentation Saur KG, München, 1975

[Schüler 86] Schüler F., Statistsiche Textuntersuchungen als Hilfsmittel bei der Entwicklung linguistischer Regeln. In [Schwarz 86]

[Schwarz 86] (Hr.) Schwarz C., Thurmair G., Informationslinguistische Texterschließung. Georg Olms Verlag, Hildesheim 1986

[Schwarz 86a] : Schwarz C., COPSY: Ein Verfahren zur vollautomatischen syntaktischen Indexierung und Recherche. In [Schwarz 86]

[Schwarz 86b] : Schwarz C., Textrecherche mit Mehrwortbegriffen. In: Sprachverarbeitung in Information und Dokumentation. Informatik Fachberichte 114, Berlin Heidelberg New York Tokyo: Springer 1986, 216 – 225

[Schwarz 88] : Schwarz C., Computerlinguistische Probleme bei der Bearbeitung großer Textmengen. LDV-Forum, 5/2, März 1988, 10 – 16

[Smeaton 86] : Smeaton A.F., Incorporation Syntactic Information into a Document Retrieval Strategy: An Investigation. In: ACM Conference on Research and Development in Information Retrieval, 1986, 103 – 113

[Waltz 87] : Waltz D. L., Applications of the Connection Machine. Computer, Januar 1987, 85 – 97

Computergestützte Übersetzung in der Anwendung: Das Projekt MARIS und das Saarbrücker Translationssystem STS

Heinz-Dirk Luckhardt Institut für Angewandte Informationsforschung (IAI) Martin-Luther-Str.14 D-6600 Saarbrücken

Kurzfassung

Die Autoren stellen in ihrem Beitrag den im Projekt MARIS an der Fachrichtung Informationswissenschaft der Universität des Saarlandes entwickelten Service rur computergestützte Übersetzung (STS) vor. Es werden in dem vom BMFT geförderten Projekt maschinelle und intellektuelle Übersetzung in einer gemeinsamen Systemumgebung (Übersetzerarbeitsplatz) miteinander verknüpft. MARIS setzt die entwickelten Verfahren und Systeme bei der Übersetzung (Deutsch-Englisch) von Titeln, Deskriptoren und Abstracts aus deutschen Datenbanken praktisch ein. Bisher wurden ca. 2 Mio. Wörter übersetzt.

The paper presents the Saarbrücken Computer-Aided Translation Service (STS) being developed in the MARIS project at the Information Science Department of the University of Saarbrücken. Intellectual and machine translation (German-English) are combined in a joint system surrounding (translator's workstation). MARIS applies the methods and (sub)systems developed to titles, abstracts, and descriptors form German databases. To date around 2 million words have been translated. The MARIS project is funded by the Federal Ministry of Science and Technology.

1 Die STS-Verfahren zur computergestützten Uber¨setzung

Die im folgenden beschriebenen Forschungs- und Entwicklungsarbeiten werden im Rahmen des Projekts MARIS (Multilinguale Anwendung von Referenz- Informationssystemen, Laufzeit: 5/85 - 12/88) an der Fachrichtung Informationswissenschaft der Universität des Saarlandes und am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung an der Universität des Saarlandes (IAI) durchgeführt. MARIS wird vom Bundesministerium für Forschung und Technologie (Förderkennzeichen 1013209 2) gefördert. Ziel des Projekts ist die exemplarische technische und organisatorische Entwicklung eines Systems für computergestützte Übersetzungen im Bereich Fachinformation, zu dem als Rahmen das Saarbrücken Translationssystem als Service (STS) eingerichtet wurde (vgl. [Zimmermann & Kroupa & Luckhardt 1987]).

Von Projektbeginn an wurden deutschsprachige Titel aus Datenbanken ins Englische übersetzt und wieder in die Datenbanken integriert. Das Zusammenwirken von intellektueller und maschineller Übersetzung sowie die Terminologiearbeit sollen im folgenden beschrieben werden.

1.1 Ausgangslage

Ausgangsmaterial für die computergestützte Übersetzung Deutsch-Englisch sind deutschsprachige Titel und Deskriptoren in Datenbanken (seit Anfang 1988 auch Abstracts). Diese Datenbestände liegen maschinenlesbar vor, zum Teil existieren bereits Übersetzungen ins Englische bzw. aus dem Englischen ins Deutsche (und in andere Sprachen). Zu den Datenbanken gibt es in der Regel kontrollierte Vokabulare (Thesauri), mit deren Hilfe der Inhalt zusätzlich intellektuell erschlossen wird und die bislang nur teilweise mehrsprachig verfügbar sind.

LDV-Forum Bd. 5. Jg. 1987/88. Nr. 4. September 1988: S. 12 - 20

Bei fachspezifischen Übersetzungen stellt insbesondere die Terminologiearbeit den Übersetzer vor große Probleme. Demgegenüber sollen die Kosten für die Übersetzung den erheblichen Aufwand bei der Erschließung der Dokumente nicht ungebührlich erhöhen.

Inzwischen sind technische Hilfsmittel entwickelt worden, die eine rationellere Durchführung der Übersetzungsaufgaben ermöglichen, deren Einsatz sich jedoch erst ab einem gewissen Mindestvolumen an Übersetzungen lohnt. Diese Hilfsmittel unterstützen die fremdsprachige Textgenerierung (Übersetzung, Postedition) und führen im terminologischen Bereich zu größerer Konsistenz.

Im Terminologiebereich können zudem bei "zentraler" Sammlung und Nutzung für einen Übersetzungsservice weitere Vorteile entstehen: während der bisherigen Projektlaufzeit
von MARIS zeigte sich bereits an den Fachgebieten "Bauwesen" und "Technische Regeln", aber auch für "Bauwesen" und "Umweltschutz", daß sich erhebliche Berührungspunkte ergeben (z.B.: Baurecht, Verwaltungsvorschriften, Technische Regeln, Bauausführung, Normen etc.). Insgesamt steht zu erwarten, daß durch das kumulierte Sammeln der technischen Terminologien eine erhebliche Unterstützung des Übersetzungsprozesses möglich wird.

1.2 Computergestützte intellektuelle Übersetzung (CAT-H)

Es werden im folgenden zwei Varianten der computergestützten Übersetzung (computeraided translation; CAT), d.h. der Unterstützung der Arbeit des Übersetzers durch computergestützte Verfahren v.a. zu Textverarbeitung und zu Terminologieverwendung, unterschieden:

(a) CAT-H: Eine vorwiegend intellektuell ge fertigte (Human-) Übersetzung;

(b) CAT-C: Eine vom Menschen durch Präedition, Interaktion und/oder Postedition gesteuerte bzw. korrigierte Übersetzung durch den Computer.

Da von Projektbeginn an Übersetzungen für deutsche Fachinformationszentren anzufertigen waren, wurde zunächst ein herkömmliches (Human-) Übersetzungsverfahren mit Computerunterstützung entwickelt und eingesetzt, gleichzeitig aber die Integration eines maschinellen Übersetzungsverfahrens in einen Übersetzungsservice konzipiert und implementiert:

(a) Das als Kernsystem für das maschinelle Übersetzungsverfahren verwendete maschinelle Übersetzungssystem SUSY wurde von der universitätseigenen Siemensanlage (Betriebssystem BS2000) auf die projekt eigene Nixdorfanlage Targon /35 (Betriebssystem UNIX) migriert.

(b) Da es sich bei SUSY um ein forschungsorientiertes System handelte, mußte eine produktionsorientierte Systemumgebung geschaffen werden.

(c) Da keines der zu bearbeitenden Fachgebiete durch das SUSY- Übersetzungswörterbuch abgedeckt war (Stand zu Projekt beginn: Testwörterbuch mit 7.000 Einträgen), mußten Verfahren zur (teil-) automatischen Lexikonerweiterung entwickelt werden (Stand des Übersetzungswörterbuchs, d.h. des Terminologiepools (4/88): 200.000 Einträge).

Die Organisation der intellektuellen Übersetzung in einem computergestützten Service - wie sie im Rahmen des Projekts MARIS für maschinenlesbare Daten entwickelt wurde läßt sich wie folgt beschreiben:

Datenübergabe

Die Auftraggeber liefern die zu übersetzenden Dokumente per Magnetband oder Diskette an. Ein Dokument ist in der Regel ein Datenbankeintrag mit einem Identifikationscode, einem Titel, einem Abstract, einem Fachgebietsschlüssel und evtl. intellektuell vergebenen Deskriptoren.

Datenumsetzung

Die Daten werden auf das STS-Format umgesetzt. Die zu bearbeitenden Dokumente werden dabei in eine Form gebracht, die eine angemessene Weiterverarbeitung erlaubt. Für den Vorgang nicht benötigte Teile der Dokumente werden ausgeblendet: Es wird ein Bereich für das Eintragen der Übersetzungen bereitgestellt. Jede Lieferung (zwischen 500 und

5000 Dokumenten) wird in einzelne "Pakete" zu 50 oder 200 Dokumenten aufgegliedert. Jedes Paket wird ausgedruckt.

Paketverteilung und -übersetzung

Jeder Übersetzer erhält eine Anzahl von Paketen zur Übersetzung. Entsprechend den implementierten Verfahren kann diese auf zweierlei Arten geschehen:

(a) Die Übersetzungen werden vom Übersetzer auf Papier vorgeschrieben und in einem zweiten Schritt im Dialog mit der Siemensanlage mit dem zeilenorientierten Editor erfaßt. (Dieses Verfahren wird nicht mehr eingesetzt).

(b) Nachdem das Projekt inzwischen über eine ausreichende Anzahl an PCs verfügt, werden die Pakete auf Disketten gelegt und mit einem bequemen Textprozessor bearbeitet.

Terminologieunterstützung

Die Übersetzer erhalten zu den Grundformen der Originaltexte alle Übersetzungsäquivalente aus dem Terminologiepool. Diese werden durch automatische Lemmatisierung der Texte und automatische Wörterbuchsuche nach dem ALT-Verfahren (automatic lemmatisation and translation) ermittelt. Die Übersetzer werden dabei auf bevorzugte Varianten (z.B. in-house-Terminologie der Auftraggeber) hingewiesen.

Datenrücksendung

N ach einer abschließenden Kontrolle werden die übersetzten Pakete wieder zu einer großen Datei zusammengefügt und an den Auftraggeber zurückgeschickt.

1.3 Computergestützte (maschinelle) Übersetzung (CAT-C)

Mit der Migration von -SUSY auf die Nixdorf-Anlage wurde 1986 die Voraussetzung für die Einrichtung des computergestützten Saarbrücker Translationsservice STS geschaffen. Weitere Voraussetzungen bzgl. der Präedition, der Auswahl von Übersetzungsäquivalenten und der Postedition waren zu erfüllen:

1.3.1 Präedition

Bevor die Daten maschinell übersetzt werden konnten, mußten drei Probleme gelöst werden: Beseitigung von Rechtschreibfehlern, Auflösung der Mehrdeutigkeit Abkürzungspunkt/Satzendepunkt, Selektion fremdsprachiger Titel.

Die sprachlich-formale Qualität der angelieferten Daten läßt z. T. zu wünschen übrig. Für den menschlichen Übersetzer bedeuten Rechtschreibfehler in der Regel keine unüberwindlichen Hindernisse. Die MÜist jedoch auf vollkommene Korrektheit des Eingabetextes angewiesen. Alle zu bearbeitenden Titel werden demzufolge mit der automatischen Rechtschreibhilfe PRIMUS überprüft.

Bezüglich der *Punkte,* die in den u. U. aus mehreren Teilen bestehenden Titeln vorkommen, ergibt sich das Problem der Entscheidung "Satzendepunkt oder Abkürzungspunkt?". Derzeit in der Erprobung befindliche automatische Verfahren werden zur gegebenen Zeit in die Systemumgebung integriert. Bis dahin führen die Übersetzer die Vereindeutigung in einem halbautomatischen Verfahrensschritt durch.

Bei einigen Anwendern kommen in den angelieferten Daten außer den deutschen auch fremdsprachige Titel vor (außer in Französisch u.a. auch in Italienisch, Spanisch, Finnisch etc.), die von der Übersetzung ausgeschlossen werden müssen. Auch dies geschieht z. Zt. intellektuell, für später sind Sprachidentifikationsverfahren vorgesehen.

1.3.2 Auswahl von Übersetzungsäquivalenten

Mit dem Anwachsen des Terminologiepools stellt sich mehr und mehr das Problem der Auswahl zwischen mehreren verschiedenen Übersetzungsäquivalenten (vgl. [Luckhardt 1987]), die im Laufe des Projekts teils aus konkreten Übersetzungen, teils aus maschinenlesbar vorliegenden Sammlungen gewonnen wurden (vgl. 2.). Bei der Titelübersetzung werden von Übersetzern Äquivalente in Abhängigkeit vom Fachgebiet, von Anwendervorschriften und/oder vom Kontext vergeben, wobei ggf. alle drei Gesichtspunkte ineinandergreifen. Die Möglichkeit des Abwägens der verschiedenen Kriterien hat

der Übersetzer dem Computer voraus, da diese intellektuelle Leistung kaum formalisiert ist. Insbesondere ergeben sich die folgenden Probleme:

(a) Fachgebiet Die Titel eines Auftraggebers können nicht ohne weiteres einem fest umrissenen Fachgebiet zugeordnet werden. Ein Beispiel dafür stellen die Daten des Umweltbundesamtes dar, die den Fachgebieten Umweltschutz, Chemie, Biologie, Land- und Forstwirtschaft, Recht, Wirtschaft, Raumordnung, Bauwesen etc. zuzuordnen sind, wobei u. U. *in einem Titel* mehrere von ihnen vertreten sind.

In der Regel sind die Dokumente (Titel/ Abstracts) der Datenbankanbieter klassifiziert, d.h. in Fach- oder Themengebiete eingeordnet. Dies bedeutet *an sich* eine wertvolle Unterstützung bei der Disambiguierung. Ein Problem stellen jedoch die unterschiedlichen Fachgebietsklassifikationen der verschiedenen Anwender dar. Jeder Anwender stellt eine eigene Klassifikation nach den für ihn wesentlichen Kriterien auf und markiert die Klassen an den Stellen besonders detailliert, an den.en er es für sinnvoll erachtet. Dies steht den Anforderungen einer *allgemeinen Klassifikation* entgegen, wie sie für die Ordnung von Fachgebieten innerhalb eines maschinellen Übersetzungswörterbuchs für verschiedene Anwender/Fachgebiete (eines Terminologiepools) sinnvoll erscheint. Dazu das folgende Beispiel:

Das Umweltbundesamt hat die folgende Grobklassifikation gewählt (die "Umweltgebiete" d.h. die "Säulen" der Klassifikation):

LU LUFT

WA WASSER

BO BODEN

NL NATUR UND LANDSCHAFT

LF LAND- UND FORSTWIRTSCHAFT / NAHRUNGSMITTEL

AB ABFALL

LE LÄRM/ERSCHÜTTERUNGEN

CH UMWELTCHEMIKALIEN/SCHADSTOFFE

SR STRAHLUNG

EN ENERGIE UND ROHSTOFFE

UW UMWELTÖKONOMIE

UA ALLGEMEINE UND UBERGREIFENDE UMWELTFRAGEN

Diese Gebiete sind in zwei Hierarchiestufen weiter klassifiziert (d.h. facettiert), z.B.: "Boden" in:

BO 50 technische Bodenschutzmaßnahmen

BO 60 planerische Bodenschutzmaßnahmen

BO 70 Theorie, Grundlagen und allg. Fragen

BO 71 Bodenkunde und Geologie

BO 72 Bodenbiologie

Das heißt: Die Fachgebiete selbst (wie Biologie, Geologie etc.) tauchen auf einer niedrigen Hierarchieebene auf, und zwar an verschiedenen Stellen der Klassifikation. Das Umweltbundesamt hat einen anderen Querschnitt durch Wissenschaft und Technik angesetzt, als es z.B. das Informationszentrum RAUM und BAU tut. Dermaßen verschiedenartige Klassifizierungen sind nur schwer miteinander bzw. mit einer allgemeinen Klassifizierung zu kompatibilisieren.

(b) Anwender Die Nutzer von MÜ-Systemen (oder auch allgemeiner: die Auftraggeber von Übersetzungen) legen in der Regel großen Wert darauf, daß die in ihrem Hause übliche Terminologie (Inhouse-Terminologie) verwendet wird. So hat das Auswahlkriterium "Benutzerpriorität" Vorrang vor anderen, muß aber natürlich mit dem Kriterium "Fachgebiet" in Einklang gebracht werden. Der Auswahlalgorithmus muß also z.B. die folgenden Pfade verfolgen:

- stimmen Benutzercode des Textes und eines vorliegenden Lexikoneintrags überein?
- wenn ja: stimmen auch die Fachgebietscodes überein?
- wenn nein: stimmen wenigstens die Fachgebietscodes überein?
- etc. (vgL auch [Luckhardt 1987]).

c) Kontext Titel werden in STS losgelöst von den dazugehörigen Texten übersetzt. "Kontext" bedeutet im vorliegenden Falle also in der Regel "knappe, meist nur nominale Strukturen, allenfalls eine einfache Verbform (fin. Verb, Infinitiv, Partizip)". Doch auch hier hat der Übersetzer dank seines Weltwissens, seines Assoziationsvermögens, seiner Phantasie etc. dem Computer - der z.Zt. allein auf linguistisch-formaler Ebene entscheidet schwer formalisierbare intellektuelle Leistungen voraus. Ein Beispiel:

In einem zweisprachigen Glossar ist" Altstadt" mit "old parts of a town" übersetzt. Dieses Äquivalent wird in ein deutschenglisches Computerlexikon übernommen.
Damit würde aber der Computer die Phrase "Altstadt von Saarbrücken" mit "old parts of a town of Saarbrücken" übersetzen. Ein anderes Beispiel:

"ENTSORGUNG" =} "WASTE DISPOSAL"

"ENTSORGUNG VON ABFALL" =} "WASTE DISPOSAL OF WASTE" ?

Solche Probleme lassen sich zwar auch im Rahmen des vorliegenden Systems lösen, jedoch nur mit einem erheblichen Aufwand an Regelformulierung und Lexikonkodierung.

1.3.3 Postedition

Die maschinell übersetzten Titel werden von Übersetzern an PCs postediert. In einer ersten Implementation des Posteditionsverfahrens wurden die MÜ- Ergebnisse auf Disketten gelegt und vom Posteditor mit dem Textprozessor WordStar2000 direkt bearbeitet. Dabei entstanden zwei Probleme:

- Nicht übersetzte Wörter mußten während des Posteditionsvorgangs recherchiert werden, so daß der PC unnötig lange blockiert wurde;
- Übersetzungsvorschläge des Computers mußten darauf überprüft werden, ob die Inhouse-Terminologie des Auftraggebers berücksichtigt war.

In einer zweiten Implementationsstufe wurden diese Probleme inzwischen wie folgt gelöst:

- Nicht übersetzte Wörter bzw. Übersetzungsvorschläge des Computers für Komposita und abgeleitete Wörter werden gesondert ausgegeben und können vom Posteditor in einer Recherchesitzung vor der Posteditionsphase übersetzt bzw. überprüft werden;
- bei der automatischen Auswahl von Übersetzungsäquivalenten wird jetzt der Benutzercode ausgewertet, so daß der Posteditor sicher sein kann, daß im maschinellen Output die Termini des Anwenders berücksichtigt sind.

2 Der STS-Terminologiepool

Der STS- Terminologiepool umfaßt derzeit (April 1988) ca. 200.000 deutsch- englische Übersetzungsäquivalente, die gewonnen wurden durch:

- Einspielen fremder bzw. anwendereigener Terminologiesammlungen;
- Extraktion von Äquivalenten aus fertigen Übersetzungen nach dem halbautomatischen STS-CTX- Verfahren;
- intellektuelle Extraktion aus vorliegen den Übersetzungen;
- titelunabhängige Übersetzung von z.B. Schlagwortverzeichnissen bzw. Thesauri.

2.1 Terminologiegewinnung

2.1.1 Fremde bzw. anwendereigene Terminologiesammlungen

Wenn beim Anwender fremdsprachige Terminologie vorliegt, hat diese Vorrang bei den anwenderbezogenen Übersetzungen. So wurden im Falle des Informationszentrums RAUM und BAU (IRB) und des Deutschen Informationszentrums für technische Regeln (DITR) haus eigene deutsch- englische Terminologiesammlungen in das entsprechende STS-Computerlexikon überspielt. Diese Sammlungen sind allerdings nicht ohne Anpassungen zu übernehmen. Einige Probleme sollen kurz angedeutet werden:

Derartige Einträge sind nicht immer als Übersetzungsäquivalente zu verstehen und zu

benutzen. Dazu einige Beispiele aus den Quellen:

MESSWESEN > MEASUREMENT, TESTING, AND INSTRUMENTS

UMRECHNUNG > CONVERSION (UNITS OF MEASUREMENT)

ORTUNGSGERÄT > DETECTORS

KIPPEN > BUCKLING

Zum einen handelt es sich oft um Begriffe, die (z.B. durch Klammerausdrücke) genauer spezifiziert sind, weil sie sonst als *Deskriptoren* (und das ist ja ihr eigentlicher Zweck) nicht aussagekräftig genug sind. Zum anderen werden die englischen Äquivalente entsprechend dem anglo-amerikanischen Usus nach Möglichkeit im Plural wiedergegeben. Dies ist jedoch für ein MU-Lexikon problematisch, weil hier *Grundformen* bzw. Stämme erwartet werden. Es kann aber erst über ein Deflektionsverfahren entschieden werden, ob ein auslautendes *-s* eine Pluralen dung ist oder zum Wortstamm gehört. Aus einem Verb der Quellsprache ("kippen") ist ein Verb der Zielsprache abzuleiten ("buckle") und nicht ein Gerundium ("buckling"). (Daß die Substantivierung - hier: das "Kippen" - gemeint ist, ist eine andere Frage, die wiederum aus der unterschiedlichen Problemstellung abgeleitet ist).

2.1.2 Extraktion aus vorliegenden intell. Übersetzungen

Die Daten der Anwender werden automatisch auf das Eingabeformat für ein automatisches Texterschließungs- und Indexierungssystem (CTX, vgl. [Kroupa 1982]) gebracht, mit dem die Grundformen aus den Eingabedaten ermittelt werden. Diese können dann mit dem Terminologiepool verglichen werden. Das Vergleichsprogramm speichert diejenigen Grundformen in einer besonderen Datei, für die es kein zielsprachliches Äquivalent findet. Wenn die intellektuellen Übersetzungen fertig sind, ordnen Übersetzer/Terminologen den dem Pool noch "unbekannten" Begriffen zielsprachliche Äquivalente aus den entsprechenden Titelübersetzungen zu.

2.1.3 Titelunabhängige Terminiübersetzung

Seit Mai 1987 wird im laufenden Projekt MARIS das deutsche Stich- und Schlagwortverzeichnis des Deutschen Patentamts zur internationalen Patentklassifikation IPC (ca. 130.000 Begriffe) ins Englische übersetzt. Nach Fertigstellung dieser Arbeit wird ein umfassender fachsprachlicher Wortschatz zur Verfügung stehen, der weitgehend die Grundbegriffe der gesamten Technik abdeckt.

Das Fehlen eines umfassenderen Kontextes schafft hier neue Bedingungen: Allerdings gibt die Beschreibung der IPC-Klasse, die jedem Stichwort zugeordnet ist, dem Übersetzer/Terminologen einen wichtigen Hinweis auf die vorliegende Lesart eines (ggf. mehrdeutigen) Begriffs und somit auf das auszuwählende zielsprachliche Äquivalent. So kann der Begriff "Schnecke" im Bereich *Backwarenherstellung* mit "worm" und im Bereich *Gartenbau* mit "snail" eindeutig übersetzt werden.

Ein Problem stellen echte Synonyme dar, die z.B. ins Spiel kommen, wenn in der in gedruckter Form mit herangezogenen englischen Version der IPC Begriffe an verschiedenen Stellen verschieden benannt werden, z.B. "baler" oder "baling preß" für "Ballenpresse" . Hier kann keine eindeutige Entscheidung zugunsten des einen oder des anderen Aquivalents getroffen werden.

Ein weiteres Problem wurde schon oben angedeutet: Die Einträge in den Thesauri/Stichwortverzeichnissen der MARIS-Anwender stellen *Deskriptoren* dar, wie sie für das Information Retrieval vergeben werden. So kommen, z.B., im Thesaurus der DOMA-Datenbank (Fachinformationszentrum Technik: Maschinenbau) die folgenden Einträge vor:

Abfall (Fertigung)
Abfall (Kerntechnik)
Abfall (Land wirtschaft)
Abfall (Müll)
Abfall (Papier)
Abfall hochaktiv
Abfall mittel aktiv
Abfall schwach aktiv

Die Differenzierungen sind in diesen Fällen für den Aufbau des Terminologiepools irrele-

vant, da es sich in allen Fällen um "refuse" ("waste") handelt. Klammerausdrücke in der Quelle können allerdings auch der terminologierelevanten Bedeutungsdifferenzierung dienen, wie in den folgenden Beispielen:

Anhänger	⇒	supporter
Anhänger (Fahrzeug)	⇒	trailer
Anker (Schiff)	⇒	anchor
Anker (el. Maschine)	⇒	armature

2.2 Fachgebietsklassifikation

Die Problematik von Fachgebietsklassifizierungen ist bereits zur Sprache gekommen. Es erschien wenig sinnvoll, die Disambiguierung in STS unmittelbar auf den Anwenderklassifikationen aufzubauen. Dies würde v.a. in der Entwicklungsphase zu einer Reihe von Problemen führen, u.a. in bezug auf das Copyright. Für STS ist daher eine "neutrale", möglichst allgemein gehaltene Klassifikation gewählt worden. Sie erlaubt eine Grobauswahl zwischen Übersetzungsäquivalenten. Auf eine genauere Klassifizierung wurde für STS aus den oben (vgl. 1.2.2) genannten und aus einigen weiteren Gründen verzichtet. Am Beispiel der Internationalen Patentklassifikation IPC soll das erläutert werden.

Die IPC ist in Sektionen, Untersektionen, Klassen, Unterklassen, Haupt- und Untergruppen unterteilt, z.B.:

Sektion **A**: täglicher Lebensbedarf
Untersektion "Nahrungsmittel und Tabak"
Klasse **A21**: Backen; eßbare Teigwaren
Unterklasse **A21D**: Behandeln von Mehl oder Teig, Backverfahren; Bäckereierzeugnisse; deren Haltbarmachung
Hauptgruppe **A21D 15/00**: Konservieren von Bäckereierzeugnissen
Untergruppe **A21D 15/02**: Konservierung von Bäckereierzeugnissen durch Kühlen

Es ist offensichtlich, daß nicht die gesamte Klassifikation für die maschinelle Übersetzung übernommen werden könnte, wenn man bedenkt, daß die IPC 8 Bände mit bis zu 300 Seiten Umfang pro Band umfaßt. Die meisten Begriffe des Lexikons müßten u.U. Hunderten von Untergruppen zugeordnet werden. Selbst die Hauptgruppen und Unterklassen geben eine zu feine Klassifizierung ab, da z.B. bestimmte Maschinenteile intellektuell verschiedenen Unterklassen zugewiesen werden müßten. Die Kodierung neuer Lexikoneinträge würde unverhältnismäßig viel Zeit kosten, und der Nutzen wäre überdies zweifelhaft, da es keine Texte gibt (außer den Texten im Patentwesen), die entsprechend markiert sind, so daß ein Vergleich der Markierung des Textes und der Lexikoneinträge zur eindeutigen Zuweisung von zielsprachlichen Äquivalenten führen könnte.

Oberhalb der Unterklassen sind in der IPC-Hierarchie die "Klassen" angesiedelt. Selbst hier gibt es noch unerwünschte Überschneidungen, z.B. zwischen A22 (u.a. Fisch- und Fleischverarbeitung) und A23 (u.a. Fisch- und Fleischkonservierung); A21 (u.a. Backöfen), A47 (u.a. häusliche Backausrüstung), F23 (Feuerungen) und F24 (u.a. häusliche Heiz- oder Kochherde, die ganz oder teilweise als Backöfen ausgebildet sind).

Wohlverstanden: dies ist auch hier keine Kritik an der IPC, die "vor allem als ein wirksames Recherchenwerkzeug für das Wiederauffinden von Patentdokumenten durch die Patentämter und sonstigen Anwender, zur Feststellung der Neuheit und zur Beurteilung der Erfindungshöhe ... von Patentanmeldungen" und einigen weiteren Zwecken (vgl. [IPC (1979)], Band 9, S.7) dient. Für die allgemeine Klassifizierung von Lexikoneinträgen und Texten zum Zwecke der maschinellen Übersetzung scheint jedoch vorerst nur die zweitoberste Hierarchieebene geeignet, nämlich die der "Untersektionen", die in einigen Fällen in Klassen aufgeteilt werden kann, z.B. die Untersektion "Transportieren" in die "Klassen" *Eisenbahnen, Gleislose Landfahrzeuge, Schiffe, Luftfahrzeuge, Transportverfahren und Sattlerei/Polsterei.* Dies schließt übrigens nicht aus, daß es in Einzelfällen sinnvoll sein kann, zur Disambiguierung eines spezifischen Wortes auch eine feinere Klassifikation mit heranzuziehen, soweit dies in einem solchen Falle zur Verbesserung der Übersetzungsergebnisse führt.

Auf jeden Fall bietet *jede Art von Klassifikation* "Angriffsflächen" für den genauen Betrachter. Es kommt jedoch immer auf die Verwendung der Klassifikation an. Ein menschlicher Anwender (z.B. ein Prüfer des Patentamts), wird dank seiner Intelligenz, Assozia-

tionsfähigkeit, vor allem aber seines Weltwissens mit einer Klassifikation eher zurechtkommen, als die Maschine, die auf genau umrissene Klassen angewiesen ist, solange es nicht möglich ist, *Klassenübergänge* zu formalisieren.

Die STS-Klassifikation ist sachgebietsorientiert und hierarchisch aufgebaut. Nicht alle Fachgebiete, sondern nur solche, für die heute schon Daten vorliegen, sind vertreten (vgl. [Luckhardt 1987a]).

2.3 Nutzung des Terminologiepools DEENWO

Unter Terminologiepool wird in MARIS die für die maschinelle Übersetzung zugreifbare Terminologie für alle Fachgebiete/Anwender verstanden. Daneben existieren zwei dbase-Datenbanken (Sozialwissenschaften bzw. Raum und Bau) für die intellektuelle Übersetzung, die von Zeit zu Zeit mit dem Terminologiepool kompatibilisiert werden.

Auf den Terminologiepool greifen verschiedene Verfahren zu:

a. die automatische Übersetzung von Deskriptoren aus Datenbanken

b. die Wortübersetzung in der Teilkomponente "Transfer" der maschinellen Übersetzung von Titeln und Abstracts

c. die automatische Indexierung CTX und anschließende Übersetzung der erzeugten Deskriptoren

U.U. reicht die Angabe des Auftraggebers zur Auswahl der korrekten zielsprachlichen Äquivalente aus. Man versieht die zu übersetzenden Deskriptoren bzw. Dokumente mit dem Code des Auftraggebers, und der Suchalgorithmus kann die mit dem gleichen Code versehenen zielsprachlichen Äquivalente zuordnen.

Die Fachgebietsmarkierung kommt dann ins Spiel, wenn entweder kein Übersetzungsäquivalent mit dem passenden Anwendercode vorhanden ist oder der Pool für einen Anwender mehrere Äquivalente anbietet. Es ist offensichtlich, daß auch Pooleinträge mit Anwendercode für Texte anderer Anwender nutzbar sein müssen. Wenn der Begriff "Umweltverschmutzung" für das Umweltbundesamt mit "environmental pollution" übersetzt wird, soll diese Übersetzung auch für andere Anwender verfügbar sein, ohne daß sie dupliziert und/oder mit anderen Anwendercodes versehen wird. Wenn mehrere Äquivalente zur Verfügung stehen, wird das gewählt, dessen Fachgebietscode mit dem des Dokuments/Textes übereinstimmt.

2.4 Umfang der STS-Systemlexika

Die STS-Systemlexika haben derzeit (April 1988) den folgenden Inhalt:

	Einträge
dt. morpho-synt. Analyselexikon:	143.546
dt. Kompositalexikon:	158.900
dt. semantisches Lexikon:	75.853
dt./engl. Transferlexikon:	200.000
engl. Syntheselexikon:	2.896

Dazu kommen die folgenden PC-Datenbanken für Übersetzer:

1. dBaseIII (IZ,IRB,DITR)

   17.000 Einträge

2. GOLEM (IRB und DITR)

   13.000 Einträge

3 Ausblick

Der Erfolg maschineller Übersetzungssysteme wird m.E. in Zukunft von drei Faktoren abhängen (vgl. auch [Zimmermann 1987]):

- der Qualität des MÜ-Kernsystems
- der Qualität der Systemumgebung
- der Qualität der Terminologieunterstützung

In allen drei Bereichen besteht bzgl. Forschung und Entwicklung ein erheblicher Nachholbedarf, der erwarten läßt, daß die legendäre "FAHQT" (fully automatic high quality translation) erst im nächsten Jahrhundert verfügbar sein wird. Bis dahin können die verfügbaren Systeme eingesetzt werden,

für die ins besondere in den beiden zuletzt genannten Bereichen noch sehr viel getan werden kann. Die Qualität der MÜ-Kernsysteme - d.h. ihre computerlinguistische Basis und ihre daraus resultierenden Fähigkeiten der Verarbeitung und Disambiguierung natürlicher Sprachen - kann durchaus noch verbessert werden. Hier sind jedoch besonders die in der Entwicklung befindlichen Systeme wie EUROTRA und andere computerlinguistische Forschungsprojekte gefordert. Verbesserungen in den Bereichen *Terminologieunterstützung* und *Systemumgebung* werden die bestehenden Systeme allgemeiner verwendbar und v.a. auch weiteren Anwendern verfügbar machen. Darüberhinaus werden die Ergebnisse dieser Arbeiten weitestgehend auch rur Systeme der nächsten Generation nutzbar sein.

Literaturverzeichnis

[IPC (1979)] IPC (1979): Internationale Patent klassifikation. München: Carl-Heymanns . Verlag

[Kroupa 1982] Kroupa, E.: Strategien der Dokumentrepräsentation bei CTX. Ein Verfahren zur computergestützten Texterschließung und Textwiedergewinnung. In: I. Batori, S. Krause, H.-D. Lutz (Hrsg.). Linguistische Datenverarbeitung. Sprache und Information Band 4, Tübingen, 155-161, 1982

[Luckhardt 1987] Luckhar4t, H.-D.: Probleme bei der Auswahl von Übersetzungsäquivalenten. In: [Zimmermann&Kroupa&Luckhardt, 1987]

[Luckhardt 1987a] Luckhardt, H.-D.: Terminologieerfassung und -nutzung im computergestützten Saarbrücker Translationssystem STS. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarlandes, 1987

[Luckhardt 1987b] Luckhardt, H.-D.: Der Transfer in der maschinellen Sprachübersetzung. Tübingen: Niemeyer, 1987

[Luckhardt 1988] Luckhardt, H.-D.: Generation of Sentences from a Syntactic Deep Structure with a Semantic Component. In: McDonald/ Bole (Hrsg.). Natural Language Generation Systems. Reihe Symbolic Computation. New York: Springer, 1988

[Peters 1987] Peters, J .P.: Die Mono-/Multilingualität von Datenbanken. In: [Zimmermann&Kroupa&Luckhardt, 1987], 10-22

[Zimmermann 1987] Zimmermann H.H.: Linguistic- Technical Aspects of Machine Translation. In: Proceedings of the AGARD TIP Meeting on 'Barriers to Information Transfer and Approaches toward their Reduction', Washington D.C., 23. - 24. September 1987, 1987

[Zimmermann & Kroupa & Luckhardt 1987] Zimmermann, H. H., E. Kroupa, H.-D. Luckhardt: Das Saarbrücker Translationssystem STS - Eine Konzeption zur computergestützten Übersetzung. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarland es, 1987

Informationsorientiertes Schreiben oder die Produktion von textuell dargestelltem Wissen
Eine Übersicht über den Stand der Kenntnis in den Fachgebieten Wissenschaftssoziologie, Technisches Schreiben, Schreibforschung und Textverarbeitung.

Brigitte Endres-Niggemeyer
Fachhochschule Hannover
Fachbereich BID
Hanomagstr.8 D-3000
Hannover 91

Kurzfassung

Das Schreiben informationsorientierter Texte ist eine besonders wichtige Form der Wissensproduktion. Dazu wird zwecks Weiterverwendung in Computerlinguistik und Informationswissenschaft der Kenntnisstand aus den Fachgebieten Wissenschaftssoziologie, Technisches Schreiben, Schreibforschung und Textverarbeitung zusammengetragen.

1 Einleitung

Wissen muß fixiert werden, wenn es zwecks späterer Verwendung gespeichert werden soll. Die Verschriftlichung erfüllt diese Anforderung in eminenter Weise. Darum fällt die Produktion von Wissen praktisch vielfach mit der Aufgabe zusammen, informationsorientierte Texte zu schreiben. ([Rauch 1982], bes. 30-31), schließt den intellektuellen Prozeß der Gedankenfindung und der Formulierung noch aus der "Informationsgenerierung" aus. Zu dieser Einschränkung besteht jedoch dann kein Anlaß, wenn der gesamte Prozeß der Wissensproduktion von der Setzung des Themas an von derselben Person - etwa einem Sachbearbeiter oder einer Wissenschaftlerin - erledigt wird. Darum wird das informationsorientierte Schreiben hier mit der Textproduktion einschließlich der Erzeugung der Textbedeutung gleichgesetzt. Es ist dann eine besonders wichtige Form der Wissensproduktion.

Die Aufgabe, Wissen zwecks Speicherung und Weitergabe aufzuschreiben, gehört in einer modernen Dienstleistungs- und Informationsgesellschaft zu den Grundanforderungen in allen Schreibtischberufen. Eine ausreichende Schreibkompetenz muß heute von einem hohen Prozentsatz der Bevölkerung erworben werden (vgl. [Faigley et al. 1985],78). Vor dem Hintergrund langfristig steigender Anforderungen an die Schreib- und Kommunikationskompetenz ist eine aktuelle technische Entwicklung zu sehen: Durch den Übergang zum Elektronischen Schreiben und Publizieren (Überblick bei [Hierppe 1986]) ändern sich die Rahmenbedingungen bei der Produktion und Rezeption von Texten. Damit eröffnen sich besonders dann neue Gestaltungschancen für die Wissensproduktion und das Schreiben informationsorientierter Texte, wenn sich zum technischen Fortschritt die erforderliche wissenschaftliche und konzeptionelle Klärung der neu zu gestaltenden Prozesse gesellt. Durch konvergierende methodische und technische Fortschritte kann eine neue Qualität des Schreibens entstehen ([Collins & Gentner 1980], vgL auch [Krause 1988]). Wer dazu beitragen will, wird sich zunächst nach dem Stand unserer Kenntnis über das Schreiben fragen, soweit es mit der Wissensproduktion einhergeht.

Aus der Leitfrage

"Wie wird das Wissen in Texte gepackt?"

lassen sich viele interessante Teilfragen ableiten (vgl. auch [Reither 1985], 623):

LDV-Forum Bd. 5, Jg. 1987/88. Nr. 4. September 1988: S. 21 - 37

. Welche Rolle spielt das Wissen bei der Produktion informierender Texte? Welche das Kommunikationsumfeld ? Welche das Zielpublikum?

. Welche Konventionen steuern das Schreiben?

. Woher kommt das Faktenwissen? Und das Sprachwissen?

. Unter welchen Bedingungen wird geschrieben?

. Für wen wird geschrieben?

. Welche Hilfsmittel werden benutzt?

. Welche Sorten informierender Texte gibt es?

. Welche Eigenschaften kennzeichnen sie?

. Welche Eigenschaften bestimmen ihre Informationsqualität?

. Was bedeutet es für Menschen, informative Texte zu schreiben?

. Wie schreiben die Leute informierende Texte?

. Welche Fähigkeiten setzen sie dabei ein?

. Wie/wann wurden diese Fähigkeiten erworben/erlernt/ vermittelt?

. Wie verhält sich das Schreiben zu anderen kognitiven Fähigkeiten wie Lesen, Rechnen, Verstehen usw.?

. Gibt es Methoden und Strategien des Schreibens?

. Läßt sich der Schreibprozeß in Teilaufgaben, -funktionen oder -prozesse zerlegen?

. Wenn ja, in welche?

. Wo liegen die Schwierigkeiten beim Schreiben?

. Gehen hochkompetente Schreiber(innen) anders vor als weniger kompetente?

. Verläuft ein Schreibprozeß je nach Person, Aufgabe usw. verschieden?

Da die Situation in der Textproduktion durch neue Gestaltungsmöglichkeiten auf sich aufmerksam macht, ist man schnell bei einer zweiten Leitfrage:

> "Was ist zu tun, damit Wissen besser und leichter in Texte gepackt werden kann?"

Auch zu dieser Frage lassen sich Subfragen stellen:

. Fehlen bei der Textproduktion Werkzeuge?

. Haben die vorhandenen Werkzeuge Mängel?

. Brauchen die Personen mehr Wissen oder zusätzliche Fähigkeiten?

. Wie werden neue Hilfsmittel in den Schreibprozess integriert?

. Kommt eine echte Optimierung zustande?

. Welche Systemlösungen gibt es schon?

. Welche Systemlösungen werden gebraucht?

. Welche Systemlösungen sind machbar?

. Inwieweit läßt sich das Schreiben automatisieren?

Es ist wichtig, nicht im eigenen fachlichen Horizont zu verharren, wenn man das vorhandene Wissen zu den genannten Fragen sichtet und auf seine Perspektiven untersucht. Die Forschung in Linguistik, Computerlinguistik, sprachorientierter KI und Informationswissenschaft kann von Beiträgen aus anderen Wissenschaftsgebieten nur profitieren. Das Anliegen dieses Artikels ist es, zur Erforschung des informationsorientierten Schreibens Kenntnisse aus folgenden Fachgebieten zugänglich zu machen:

. Wissenschaftssoziologie

. Technisches Schreiben (Technical Writing)

. Schreibforschung (Writing Research Composition Research)

. Textverarbeitung (Word Processing, Text Processing)

2 Wissenschaftssoziologie

Im wissenschaftlichen Alltag beginnt der forschende Wissenserwerb und die schriftliche Fixierung des Wissens mit der Setzung des Forschungsthemas:

> " . . . writing and research are part of the same overall process which begins with attention and problem selection"

([Bazerman 1983],167).

Der Inhalt einer wissenschaftlichen Publikation entsteht in einem besonders langwierigen Prozeß des Pre-Writing (s. unten). Die Produktion von informationsorientierten Texten ist in vielen Fällen gleichzusetzen mit der wissenschaftlichen Arbeit, deren Produkt eine wissenschaftliche Publikation ist. Aus dieser Perspektive ist die Wissensproduktion ein Forschungsthema in der Wissenschaftssoziol0gie (Überblick bei [Bazerman 1983]) .

([Latour & Woolgar 1979], 151-183) schildern ausführlich, wie in dem von ihnen beobachteten biochemischen Forschungslabor der Inhalt eines späteren Aufsatzes nach und nach konstituiert wird. ([Knorr-Cetina 1984], 175239) beschreibt die Fabrikation von Erkenntnis am Beispiel eines Aufsatzes zur Proteinsynthese, den sie durch sämtliche Stadien seiner Entstehung verfolgt. Sie findet einen Vorgang der Wissenstransformation und -selektion vor, der von den unterschiedlichsten Faktoren beeinflußt wird. [Woolgar 1981] geht den umgekehrten Weg. Er zeigt am Beispiel einer Festrede zur Verleihung des Nobelpreises auf, wie eine wissenschaftliche Entdeckung im Text erscheint.

Die wissenschaftssoziologische Auseinandersetzung mit der Beziehung von Forschung und Textproduktion ist erst in jüngster Zeit in Gang gekommen, obwohl der Vorgang für die wissenschaftliche Praxis zentral ist ([Bazerman 1983], 168).

3 Technisches Schreiben (Technical Writing)

Das Wissensgebiet "Technisches Schreiben" (historischer Überblick bei [Connors 1982]) verdankt seinen Aufschwung den steigenden Anforderungen, die an Dokumentationen, Handbücher und Gebrauchsanleitungen technischer Systeme gestellt werden. Das Gebiet dringt langsam von einer vorwiegend praktisch orientierten zu einer wissenschaftlichen Behandlung seines Gegenstandsbereiches vor (Überblick über den Stand der wissenschaftlichen Diskussion bei [Anderson et a1. 1983]). Der Gegenstandsbereich weitet sich aus. Er umfaßt heute die wissenschaftliche, technische und professionelle Kommunikation.

Die professionelle Schreibkompetenz wird technischen Autoren und Kommunikatoren gezielt vermittelt. Die vorliegenden Lehrprogramme (z.B. [Sanders 1987], [Univ. of Michigan 1982], Coll. of Engineering, Überblick bei (Winkler & Mizuno 1985]) vermitteln Kenntnisse über Literaturtypen (etwa EDV- Handbücher), über die einzelnen Arbeitsgänge beim Schreiben (z.B. das Editieren von Texten), über das in der Praxis häufige arbeitsteilige Produzieren von Dokumenten, über Stilistik und Rhetorik sowie den Einsatz von Textsystemen.

Geschrieben wird in einer professionellen Routinesituation, in der eine systematische Optimierung des Schreibprozesses und die Entwicklung einer hohen Schreibkompetenz lohnt. Das Wissen über ein effizientes Vorgehen beim Schreiben wird insbesondere in den zahlreichen Handbüchern des Faches
((Mills & Walter 1970], (Michaelson 1982], [Weiss 1985], (Houghton-Alico 1985] und viele andere) festgehalten. Die Handbücher sind gut informiert: Die enge Beziehung zwischen den mentalen Prozessen des wissenschaftlichen Denkens und des wissenschaftlichen Schreibens sowie die zentrale Rolle der Bedeutungsorganisation für die Informationsqualität des Textes hebt schon (Peterson 1961],iii hervor:

> "This book is based on a postulate. The postulate is that you cannot write good expository prose without knowing and applying the general logic of science. ... The relationship of scientific logic to scientific writing is considered in terms of organization. Good organization,
> not only of the whole re port but of all the parts, is the key to good writing. It may be heresy to say so, but

> bad grammar, misspellings, and incorrect punctuation may tarnish but cannot destroy the value of a well organized scientific article, treatise, talk, or progress report."

Es gibt eine vielfältige Literatur zu einzelnen Problemen des technischen Schreibens. Darunter sind auch praktisch orientierte Artikel über den Einsatz von Textsystemen beim Schreiben (z. B. [Raskin 1986]).

4 Schreibforschung (Writing Research, Composition Research)

Die Schreibforschung befindet sich noch im Pionierzustand ([Matsuhashi 1982], 270). ([Faigleyet al. 1985], 3-89) nennen in ihrem ausführlichen State-of-the-Art-Bericht 1963 als terminus post quem der Disziplin. Seitdem wurden beachtliche Erfolge bei der Erforschung des Schreibens von Texten erzielt, unter anderem durch die geschickte Adaptierung von Erkenntnissen anderer Wissenschaftsgebiete wie Rhetorik, Sprachpädagogik, Kognitionswissenschaft, Psycho- und Textlinguistik. Einen guten Eindruck von dem in der Schreibforschung erreichten Kenntnisstand vermittelt auch [Molitor 1984]. Über die empirischen und experimentellen Methoden zur Erforschung von Textverarbeitungsprozessen berichten ([Rickheit & Strohner 1985], 28-41). Die Verfahren der Schreibforschung sind Thema eines Sammelbandes von [Mosenthai et al. 1983]. Zur Methodik von Fallstudien kann man [Molitor 1985] und [Molitor 1987] heranziehen.

([Collins & Gentner 1980], 51) charakterisieren die forschungspraktische Situation der Schreibforschung als vielversprechend:

> "A major breakthrough in the teaching of writing has been made possible by the convergence of two recent developments in seience and technology. Cognitive science, which brings together the disciplines of cognitive psychology, artificial intelligence, and linguistics, has begun to provide us with the theoretical means for constructing formal process theories of human cognition. Thus we now have many of the tools needed for constructing a process theory of writing. At the same time, technology is being developed to manipulate and edit text. Soon this technology will be relatively inexpensive and commonplace. It should be possible to merge these two developments into a computer-based "Writing Land" for teaching and assisting people in writing."

Einen Überblick über kognitive Ansätze in der Schreibforschung vermitteln die Sammelbände von [Gregg & Steinberg 1980] und [Nystrand 1982] und andere mehr.

[Bracewell 1980] stellt das Schreiben als komplexen kognitiven Vorgang anderen Denktätigkeiten wie dem Lösen von Rechenaufgaben, dem Lesen oder Zuhören gegenüber. Im Vergleich zum Lösen mathematischer Probleme ist das Schreiben weniger von der vorgegebenen Aufgabe bestimmt. Die schreibende Person muß die Problemsituation stärker selbst erfinden und auch selbst festlegen, wann sie mit der Lösung zufrieden ist. Schreiben, Sprechen, Zuhören und Lesen stützen sich auf gemeinsame Grundfähigkeiten. Diese Grundfähigkeiten lassen sich transferieren. Bei Kindern läßt sich beobachten, wie die Schreibfähigkeit zunächst als Fähigkeit zum Aufschreiben gesprochener Rede entsteht und sich nach und nach vom Vorbild der gesprochenen Kommunikation löst. Für [Kucer 1985] stellt das Lesen und das Schreiben einen intentionalen Prozess dar, in dem vom Autor oder Leser eine Textwelt konstruiert wird. Beide Vorgänge verlaufen weitgehend parallel und benutzen vielfach verwendbare kognitive Grundfunktionen.

[Molitor 1984] betont ebenfalls die enge Interaktion von Lese- und Schreibprozessen bei der Texterzeugung.

Im Gegensatz zum Sprechen muß die Kulturtechnik Schreiben in einem langen und mühsamen Prozeß erlernt werden. Menschen bilden individuell verschiedene Grade der Schreibkompetenz aus. Je mehr es ihnen gelingt, die Basisfertigkeiten des Schreibens (z.B. Buchstaben schreiben, die Regeln

der Orthographie und Zeichensetzung einhalten) zu automatisieren, desto mehr können sie ihre bewußte Aufmerksamkeit auf die "höheren" Denktätigkeiten konzentrieren (Bildung des Textinhaltes, kohärente Textorganisation, Ausrichtung auf den Leser usw.).

Das Schreiben eines Textes unterscheidet sich von der Produktion eines gesprochenen Diskurses unter anderem dadurch, daß mehr bewußte Gestaltung möglich ist und erwartet wird. Ein großer Teil der kognitiven Leistung beim Schreiben liegt in der Durchgestaltung des Textes. Im Vergleich zur normalen mündlichen Kommunikation stellt das Schreiben anspruchsvollerer Texte erhebliche Zusatzanforderungen an die Denktätigkeit.

([Scardamaglia et al. 1982], 203-205) postulieren anhand ihrer Analyse der Schwierigkeiten, die Kinder beim Schreiben haben, eine Vielzahl von Textrepräsentationen, die in ihrer Genauigkeit verschieden sein können. Sie werden normalerweise konstruiert oder rekonstruiert, sooft sie bei der intellektuellen Verarbeitung benötigt werden. Dies ist mit mentaler Anstrengung verbunden.

. Personen mit hoher Schreibkompetenz setzen sich beim Schreiben besonders ausgeprägt und bewußt Ziele. Sie wechseln flexibel zwischen der Behandlung von lokalen Subzielen und Globalzielen hin und her. Die Strategien bei der Textplanung sind flexibel. Insbesondere kommen auch opportunistische Strategien [Hayes-Roth & Hayes-Roth 1979] vor. Hochkompetente Erwachsene entwickeln auch detaillierte Vorstellungen von ihren Lesern, wenn sie einen Text gestalten [Berenkotter 1981]. Die Anzahl der leserbezogenen Statements in den Schreibprotokollen war am größten, wenn es die Funktion des Textes war, zu überzeugen, geringer bei informierenden und erzählenden Texten. Auch [Flower & Hayes 1980a] kommen zu dem Ergebnis, daß besonders kompetente Schreiber(innen) sich bei der Planung ihres
Textes stark auf die rhetorische Situation und das Zielpublikum beziehen, während weniger kompetente Personen stärker auf das Textthema fixiert sind. [Faigley & Miller 1982] und [Odell & Goswami 1982] zeigen, wie bewußt Schreiber in nicht-akademischen Arbeitsumgebungen die gesamte rhetorische Situation (das Bild der eigenen Person, das Thema, die Kommunikationsumgebung und den Leser) in ihre Textplanung einbeziehen.

4.1 Modelle des Schreibprozesses

Die bekannten Modelle des Schreibprozesses wurden meist anhand empirischer Daten entwickelt.

Populär und einflußreich geworden ist das erste einfache Modell des Schreibens von Rohman Wlecke [Rohman 1965]. Es sieht drei nacheinander ablaufende Phasen des Schreibens vor: "Pre-Writing", "Writing" und "ReWriting". Beim Pre-Writing wird der Textinhalt geplant. Die Ideen werden oft notiert und zu einer Gliederung verarbeitet. Im eigentlichen Schreibprozeß wird der geplante Inhalt als zusammenhängender Text niedergeschrieben. Anschließend folgt mit dem "ReWriting" die Überarbeitung des Rohtextes, bis die endgültige Fassung hergestellt ist.

Ein differenzierteres Modell schlug [Emig 1971] vor. Sie wertete Verbalprotokolle von 8 Schülern der zwölften Klasse aus. Nach ihren Beobachtungen beeinflußt das Arbeitsumfeld - und besonders der Lehrer - den Arbeitsablauf. Als Subprozesse des Schreibens treten Prewriting, Planen, Anfangen, Textorganisation, Reformulieren und Aufhören auf.

[Collins & Gentner 1980] entwickeln einen Beschreibungsrahmen für den Schreibvorgang. Sie strukturieren das Schreiben in Subprozesse, die konkreten Teilaufgaben entsprechen. Global unterscheiden sie die Ideenproduktion und die Textproduktion. Zum Erarbeiten von Ideen nennen sie Verfahren, die aus der Literatur bekannt sind: etwas mit ähnlichen Dingen vergleichen, seine Wirkung in Gedanken simulieren usw. Nach Collins Gentner geht die Textproduktion von einer detaillierten Gliederung aus, die nach und nach bis zur Satzebene in Text umgewandelt wird. Benutzt werden dabei Operatoren, die auf Text-, Paragraphen- und Satzebene arbeiten. Implizite Globalziele (z.B. das Ziel, den Text verständlich zu machen) steuern die Textgestaltung. Die "Devices" der Textgestaltung können strukturbezogen, stilistisch und inhaltsbezogen arbeiten.

Das Modell des Schreibens von Hayes & Flower (prägnante Darstellung in [Hayes & Flower 1980b]) wurde wie das von Emig mithilfe von Verbalprotokollen des Schreibens entwickelt. Im Hayes & Flower- Modell (Abb. 1) bildet "alles, was außerhalb der Haut der schreibenden Person" liegt, das "task environment": das begonnene Schriftstück gehört ebenso dazu wie die Schreibaufgabe mit ihrem Thema, der Zielgruppe und der Motivation, die die schreibende Person aus der aktuellen Situation bezieht. Das Langzeitgedächtnis liefert das Wissen über das Thema und die Zielgruppe und eventuell fertige Textpläne.

Der Schreibvorgang umfaßt die Komponenten Planen, sprachliche Darstellung und Revidieren. Das Planen hat die Unterkomponenten Generieren, Organisieren und Zielbestimmung. Beim Revidieren werden Lesen und Editieren als Subprozesse aufgeführt. Ein einfacher, als Produktionssystem dargestellter Monitor steuert den Einsatz der Subkomponenten. Dadurch wird der Ablauf anpassungsfähig. Der Monitor realisiert vier Grundtypen des Vorgehens beim Schreiben. Im einzelnen können die Subprozesse in vielen sinnvollen Abfolgen durchlaufen werden. Jeder Subprozeß ist intern durch ein Ablaufdiagramm strukturiert.

Das Modell von Hayes & Flower beschreibt die Produktion eines einfachen Sachtextes. Es beruht auf der Vorstellung, daß die einzelnen Subprozesse des Schreibens in variabler Reihenfolge, aber immer nacheinander ablaufen. Seine interne Organisation des Schreibprozesses ist ausreichend realitätsnah für einen empirischen Test an einem besonders geeigneten Schreibprotokoll, den [Hayes & Flower 1980a] erfolgreich durchführen.

([Beaugrande 1984], 87-146) schlägt eine Organisation des Schreibens vor, bei der die Subprozesse simultan ablaufen und miteinander interagieren können. Er gibt die Einfachheit eines relativ gut strukturierten modularen Systems, wie Hayes Flower es vertreten, zugunsten einer empirisch reicheren, aber weniger bestimmten Modellierung auf. Ein parallel und interaktiv arbeitendes Modell kann berücksichtigen, daß unterschiedliche Vorgänge beim Schreiben parallel ablaufen (etwa Formulieren und motorisches Niederschreiben) und daß die Vorgänge einander beeinflussen. Die Phasen des Schreibprozes-

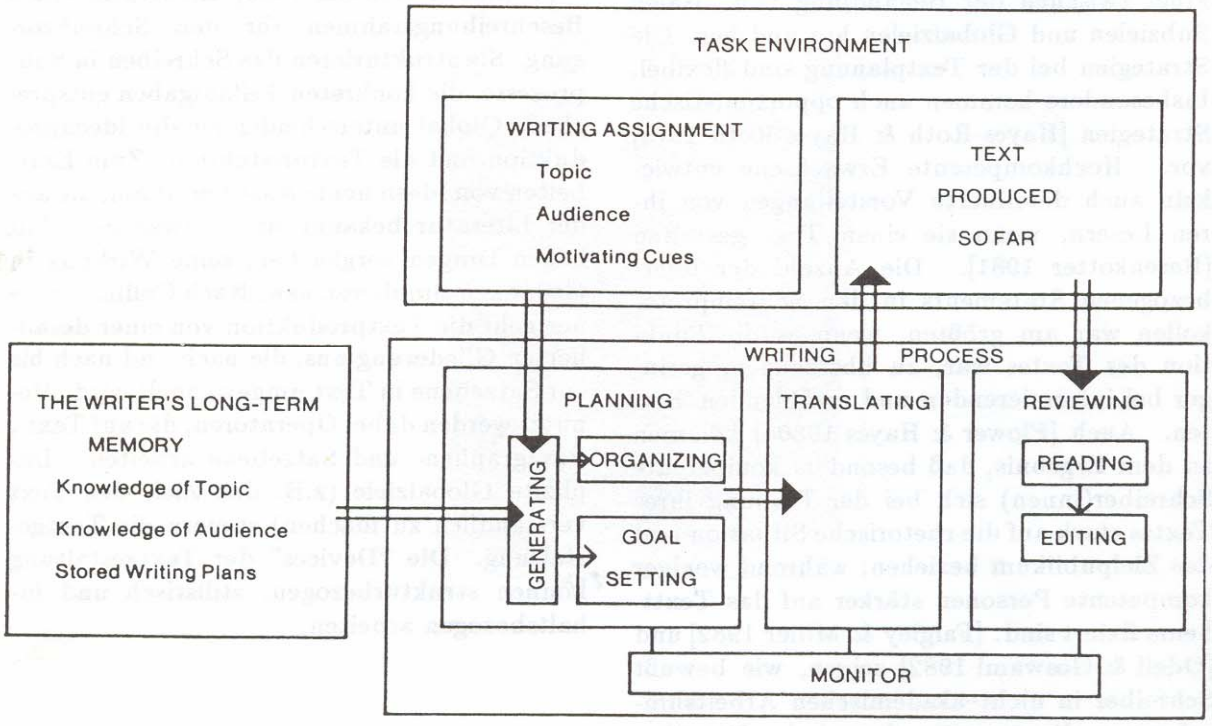


Abb.1: Modell des Schreibprozesses nach Hayes & Flower (Überblick)

ses sind durch ihre Funktion bestimmt. Man kann sie sich als Gruppen von Experten vorstellen. Sie treten in Aktion, wenn die Daten es nahelegen und setzen sich wieder zur Ruhe, wenn ihre Aufgabe beendet ist. Da beim Schreiben die Begrenztheit menschlicher Ressourcen eine wesentliche Rolle spielt, wird jeweils eine Phase (oder allenfalls wenige Phasen) in der Verarbeitung dominant sein. Sie erhält die notwendigen Ressourcen auch um den Preis, daß andere Phasen nur verzögert, fehlerhaft oder seicht arbeiten können.

Die Phasen werden mit einer globalen Zielrichtung, aber im einzelnen flexibel und interaktiv durchlaufen:

In der Zielsetzungsphase wird geplant, welche Diskursziele mit dem Text verfolgt werden sollen, wie sie organisiert werden, welche Diskurshandlungen sich daraus ergeben. Die Rolle des Textes wird festgelegt, der Texttyp wird bestimmt. Die schreibende Person macht sich klar, wie gut sie auf den Text vorbereitet ist, und plant dessen Informationsdichte. Der Stil des Textes ist damit weitgehend entwickelt.

In der Phase der Ideenfindung wird der begriffliche Gehalt konstituiert, der als Zentrum des Text-Welt-Modells dient, der gesamten Konfiguration von Wissen, die für die Behandlung des Textes benötigt wird. In dieser Phase wird im Gedächtnis gesucht. Die gefundenen Konzepte werden semantisch angereichert und integriert. Die Textplanung wird als ein Spezialfall des Problemlösens aufgefaßt: Es wird eine Konfiguration von Inhalten hergestellt, die für ein bestimmtes Publikum interessant ist.

Die Ausdrucksphase faßt alle Operationen zusammen, die die gefundene begriffliche Konfiguration mit der darunter liegenden Bedeutung in sprachliche Ausdrücke (Wörter, idiomatische Ausdrücke) überführt.

Die Linearisierungsphase sorgt dafür, daß die Ausdrücke in die beim Sprechen und Schreiben erforderliche Reihenfolge gebracht werden. Dabei werden syntaktische Formate (Satz, Phrase) und handlungsorientierte Formate (Sprechtakt, Äußerung, Aussage usw.) benutzt.

4.2 Teilprozesse des Schreibens

Die Modelle des Schreibens stimmen in vielen Grobannahmen über die Stadien des Schreibprozesses überein. Ihr Grundraster dient oft als Orientierungsrahmen, um Forschungsfragen zu stellen und zu beantworten.

4.2.1 Prewriting: Ideengenerierung und Textplanung

Beim "orthodoxen" Prewriting wird der Textinhalt aufgebaut und meist mithilfe von Stichpunkten und oder einer Gliederung festgehalten. Dann beginnt das eigentliche Schreiben. In den meisten Fällen durchziehen Prewriting-Aktivitäten wie Ideengenerierung und Textplanung den gesamten Schreibprozess. Häufig wird der bereits geschriebene Text noch einmal durchgelesen, um neue Punkte anzuknüpfen.

[Matsuhashi 1982] und [Matsuhashi & Quinn 1984] beobachten die Planungsprozesse während der Produktion. Sie werten zwei Aufsätze des Studenten John und die Videoaufnahmen seiner Schreibarbeit aus. Dabei konzentrieren sie sich auf die Analyse der Pausen. Sie nehmen an, daß Pausen Planungs- und Entscheidungsprozesse verraten. In [Matsuhashi 1982] werden besonders die Augen- und Handbewegungen von John herangezogen, um die Pausen zu interpretieren. [Matsuhashi & Quinn 1984] erhalten unterschiedliche Pausenlängen vor Propositionen des Typs Prädikat, Modifikation und Konnektiv. Sie schließen daraus auf die Aufwendigkeit der lokalen Planungsprozesse. Insbesondere ist viel Denkarbeit erforderlich, wenn die Idee, die ausgedrückt werden soll, noch weit von der Codierbarkeit entfernt ist.

[Caccamise 1981] konnte bei Schreibexperimenten beobachten, daß Studenten unter eingeschränkteren Bedingungen weniger Ideen erzeugen, die auch weniger kohärent sind. Ideen wurden nicht nur zielorientiert, sondern auch nach einem trial-and-error-Verfahren generiert.

Verfahren, die die Ideenerzeugung unterstützen, sind bekannt und wirksam ([Faigley 1985], 38-3); ([Collins & Gentner 1980], 5458); ([Beaugrande 1984], 113-114). Beispielsweise wird die tagmemische Entdeckungsprozedur für Ideen von [Young et al. 1970], die

auch [Schwartz 1984] in ihrem System zur Schreibunterstützung verwendet, von [Young & Koen 1973], [Ode1l1974] und [Burns 1979] getestet.

[Hillocks 1982] entwirft ein Modell der systematischen Befragung des Gegenstandes. Er zeigt am Beispiel des Bienenforschers Kar! von Frisch, daß die Fragestrategien seines Modells auch in der naturwissenschaftlichen Forschung benutzt werden. Schüler, die aktiv mit seinen Strategien arbeiten, erzielen bessere Ergebnisse.

[Flower & Hayes 1984] verfolgen die Entstehung von Ideen. Um ihren empirischen Beobachtungen gerecht zu werden, vertreten sie die "multiple representation hypothesis": Von der ersten bildhaften Darstellung bis zu einer propositionalen Form, die sprachlich ausgedrückt werden kann, durchlaufen Ideen unterschiedliche Darstellungsformen. Vage Gedanken und Vorstellungen werden Schritt für Schritt zu konkreten Äußerungen ausgearbeitet.

Den Einfluß unterschiedlicher Informationsquellen (Schulbildung, Massenmedien, Gespräche, eigene Erfahrung) auf die Textqualitätuntersuchen [Graesser et a1. 1984]. Die Autoren vermissen in der Literatur vergleichbare Ansätze.

Insgesamt ist über die Erzeugung des Textinhaltes in der Schreibforschung weniger gearbeitet worden als über andere Komponenten des Schreibprozesses ([Faigley et a1. 1985], 84, 86).

4.2.2 Writing: die Textproduktion

Nach [Clark & Clark 1977] wird die eigentliche Produktion beim Schreiben oft als die Übersetzung von Ideen aus einer Propositionaldarstellung in wohlgeformte Sätze definiert ([Faigley et al. 1985], 45). Dieses Transkriptionsmodell der Textproduktion wird von [Nystrand 1982] weiterentwickelt: Die Produktion stellt sich ihm als eine Interaktion zwischen den sprachlichen Ausdrucksmöglichkeiten und dem auszudrückenden Gedanken dar. Die Bedingungen der Diskursproduktion (Medium, Sprache, Textsorte, Situation, Zielgruppe) wirken auf den Prozeß der sprachlichen Realisierung ein. In diesem Vorgang wird vielfach der Kern des Schreibprozesses gesehen. Er ist empirisch schwer zu beobachten.

([Faigley et al. 1985], 44, 51-52) vermissen brauchbare psycholinguistische Theorien, die erklären, wie aus einem bestimmten Kommunikationsziel in einer bestimmten Situation eine bestimmte Äußerung entsteht. Die Forschungslage zur Produktion ist im Vergleich zur umfangreichen Literatur über das Sprachverstehen defizitär.

4.2.3 Revision: die Überarbeitung des Textes

Hohe Schreibkompetenz wird vielfach mit der aktiven Überarbeitung der eigenen Texte in Verbindung gebracht (vg1. [Huft' 1983], 800, 816): "Writing is essentially rewriting" ([Bartlett 1982], 346).

In der Revisionsphase wird ein schon existierendes Textstück verbessert. Die Überarbeitung findet besonders nach der Produktion des Textes statt. Prinzipiell werden jedoch alle Teile und Vorprodukte des zukünftigen Textes ("Prätexte" - [Witte 1985]) auch revidiert. [Bartlett 1982] findet als Teilprozesse der Revision vor allem die Entdeckung von Fehlern, die Problemidentifikation und die Strategien zur Fehlerkorrektur. Die Fehler können sehr verschieden sein: Tippfehler, falsche Referenzen, Aussagen mit nicht intendierter Bedeutung, ungünstige Textorganisation, falsche Argumentation, Fehleinschätzung der Adressaten usw.

[Huft' 1983] schlägt drei Phasen der Textüberarbeitung vor, die mehrfach durchlaufen werden können: "zero drafting", " problem solving drafting" und "final drafting". Die Null version des Textes ist die erste zusammenhängende Ausführung des Themas nach der Planungsphase. Beim "problem solving drafting" werden konzeptionelle und textorganisatorische Fehler korrigiert. Erst beim "final drafting" wird ein auch äuß erlichhandwerklich akzeptabler Text angestrebt. Fragenkataloge steuern die Arbeit in den einzelnen Phasen.

[Flower et a1. 1983] entwickeln am Beispiel von Verwaltungsvorschriften ein Szenario-Prinzip für die Revision von Gebrauchstexten. Die Autoren beobachten, wie Leser die Verwaltungvorschriften nach ihrer Handlungsrelevanz in der konkreten Situation befragen, wenn sie sie zu verstehen versu-

chen. Die Leser müssen beim Verstehen den Text durchgreifend umorganisieren. Damit holen sie de facto einen Schritt der Textgestaltung nach, der von den Verfassern der Verwaltungsvorschriften versäumt wurde. [Flower et al. 1983] ziehen die methodischen Konsequenzen: Handlungsanleitende Texte sind nach dem Szenario-Prinzip um die Handlungen einer Person in einer bestimmten Situation zu organisieren.

4.3 Spezielle Aspekte des Schreibens

4.3.1 Schreiben im sozialen Umfeld

Das Umfeld und besonders das Zielpublikum bestimmen den Schreibprozeß mit. Zur Schreibkompetenz gehört es, sich im Kommunikationsumfeld adäquat verhalten zu können. Neben Wissenschaftlergemeinden (vgl. [Reither 1985], 625) sind Organisationen und Unternehmen Kontexte, die die Textproduktion besonders nachdrücklich beeinflussen. Die Mittel der Beeinflussung reichen von organisationsspezifischen Schreibvorschriften bis zu impliziten Wertvorstellungen, die die schreibende Person über deren soziale Wahrnehmung ([Piche & Roen 1987]) erreichen. [LaRoche & Pearson 1985] und [Harrison 1987] integrieren neuere rhetorische Vorstellungen über das Publikum und Theorien über die Kommunikation in Unternehmen mit dem Ergebnis, daß Unternehmen als rhetorische Kontexte aufgefaßt werden können.

4.3.2 Evaluierung von Schreibprozessen

[Faigley et al. 1985] beschäftigen sich ausführlich mit der Bewertung von Schreibprozessen. Sie beziehen auch das Wissen ein, das während des Schreibprozesses aktiviert wird. Als Beispiel für die Evaluierung der Leistungen von Schülern, die beim Prewriting von einem wissens- basierten System anstelle eines menschlichen Tutors angeleitet wurden, kann man [Gillis 1987] heranziehen. Der wissensbasierte Tutor erreichte geringfügig bessere Ergebnisse. [Bridwell et al. 1985] zeigen am Beispiel von fünf Student(inn)en auf, wie unterschiedlich Personen reagieren, wenn sie zum Schreiben und besonders Revidieren mit einem Textsystem übergehen.

4.3.3 Das Erlernen des Schreibens

Einen guten Überblick über pädagogische Möglichkeiten, die Entfaltung von Schreibfähigkeiten zu fordern, gibt [Warshauer & Friedman 1985].

Von [Bereiter 1980] erfährt man, wie Kinder unter den Anforderungen des Schulsystems ihr Fähigkeit zum schriftlichen Ausdruck entwickeln. Sie haben beim Schreiben viele unterschiedliche Teilaufgaben oft gleichzeitig zu lösen. Da bei Kindern anders als bei Erwachsenen viele Teilfunktionen noch nicht automatisiert sind und bewußte Aufmerksamkeit beanspruchen, müssen Kinder beim Schreiben einfacher vorgehen und erreichen erst nach und nach die Leistungen von Erwachsenen. [Gundlach 1982] befaßt sich ausführlich mit den ersten Schritten von Kindern auf dem Weg zur schriftsprachlichen Ausdrucksfähigkeit. [Scinto 1982] konzentriert sich in seiner ausführlichen empirischen Untersuchung auf die Frage, wie Kinder die Fähigkeit erwerben, kohärente Texte zu produzieren.

Nach ([Daiute 1985], bes. 138) führen auch Erwachsene beim Schreiben einen inneren Dialog, der aus dem echten Dialog ihrer Kinderzeit entstanden ist (vgl. [Vygotsky 1934]). Dieser innere Dialog dient der Selbstüberwachung. Wird das Frage-und-Antwort-Spiel von zwei Personen übernommen, indem z.B. ein Lehrer mit gezielten Fragen die Schreib arbeit eines Schülers anregt und steuert, dann wird mehr geschrieben.

4.3.4 Computerunterstütztes Schreiben

In amerikanischen Schulen und Hochschulen findet computerunterstütztes Schreiben in großem Stil statt: Nach [Frase et al. 1985] benutzen jährlich 4.000 Studenten an der University of Colorado Writer's Workbench. Zur Orientierung in den vielfältigen Ansätzen und Projekten kann man [Wresch 1983] und [Wresch 1984] heranziehen.

Wo die methodische Produktion von Texten das erklärte Unterrichts ziel ist, fällt es auf, wenn "normale" Textsysteme die mecha-

nischen und handwerklichen Aufgaben beim Schreiben unterstützen, nicht aber die schwierigeren Denkarbeiten bei der inhaltlichen Erarbeitung des Textes. pädagogisch motivierte Programme stoßen in diese Lücke. Ein Beispiel ist das System von [Schwartz 1984]. Das Programm sorgt durch seine gezielten Fragen dafür, daß sich die schreibende Person die wichtigsten Faktoren bewußt macht, die bei der Textorganisation zu berücksichtigen sind: Thema, These, Zielgruppe und Argumentationsziel. Dann führt das System durch die Ideen, die im Text unterzubringen sind. Es fragt nach Ähnlichkeiten, Gegensätzen usw. und schlägt schließlich vor, wie man den Text aufbauen könnte. Die Anliegen der Schreibforschung und -pädagogik stehen auch hinter kommerzialisierten Systemen wie dem HBJ Writer, der auf dem WANDAH-System [Von Blum & Cohen 1984] beruht, und dem Writer's Helper, der auf William Wresch zurückgeht.

5 Textverarbeitung (Word Processing, Text Processing)

Eine umfassende Bibliographie der sehr aktiven Forschung und Entwicklung zur Textverarbeitung bietet [Morgan 1987]. [Krause 1988] analysiert und bewertet die Forschungssituation.

Die Textverarbeitung ist schnell eine typische PC-Anwendung geworden. Darum wendet sich kommerziell erhältliche Textverarbeitungssoftware heute meist an Benutzer(innen) von Mikrocomputern. Auch die Systementwicklung orientiert sich vorwiegend am PC-Markt. Seit 1985 werden Systeme zum Desktop Publishing angeboten. Damit kann der Autor den Text auch selbst setzen. Der gesetzte Text wird für neue Anwendungsfälle erreichbar. Systeme zum Desktop Publishing haben die Entwicklung auf dem Markt für Textverarbeitungssysteme weiter beschleunigt.

5.1 Kommerzielle Textverarbeitungssysteme und Tools für Autoren

Einen Überblick über kommerzielle Textverarbeitungssysteme und ihre Funktionen findet man bei [Gralla 1987]. Rechtschreibhilfen gehören oft schon zur Normalausstattung (s. Tabelle bei [Mullins & Faust 1987],66-69). Standalone-Spelling Checker werden ebenfalls angeboten, z.B. Turbo Lightning und Strike, besprochen von [Ramsey 1986].

Writer's Workbench ([McDonald 1982], Erfahrungsbericht bei [Frase 1985]) wird als klassisches Großrechnersystem zur Unterstützung des Schreibens in großem Umfang benutzt und hat viele PC-Softwareentwicklungen beeinflußt. Das Programmpaket ist in erster Linie stilistisch orientiert: Es liefert statistische Kennwerte für die Lesbarkeit des Textes. Zu unnötig wortreichen Phrasen bietet es kürzere Alternativen an. Zusätzlich erleichtert es mit einem einfachen Verfahren die Textorganisation: Es bietet die ersten und letzten Sätze aus Abschnitten zu einer zusammenhängenden Lektüre an. Wenn - wie oft empfohlen wird - am Absatzbeginn ein Themasatz steht und am Absatzende ein zusammenfassender Satz, dann kann so die Textkohärenz auf einer Makroebene geprüft werden.

Das bei IBM entwickelte Großrechnersystem EPISTLE CRITIQUE ([Heidorn 1982], vgl. auch [Krause 1988]) setzt den Akzent ebenfalls auf die Stilkritik. Seine Diagnosen und Verbesserungsvorschläge beruhen auf der syntaktischen Analyse des vorliegenden Textes. Die stilistischen Werturteile sollen jeweils der Anwendungsumgebung des Systems angepaßt werden [Miller 1986].

Die PC-Systeme Rightwriter (getestet von [Pfaffenberger 1986], [Gralla 1987] und [Buckingham 1987]), PC-Style und Grammatik (s. [Raskin 1986], [Gralla 1987]) bieten eine Stilkritik in der Nachfolge von Writer's Workbench.

Der Word Finder (getestet von [Morgan 1986]) ist ein Thesaurus, der vor allem bei der Suche nach Synonymen hilft.

Index-Aid (getestet von [Grosch 1986]) erleichtert es, zu einem Buch das Register zu machen.

"Idea tools" sind ebenfalls auf dem Markt. Das PC Magazine vom 25.3.86, 199-220, vergleicht PC-Outline, Da Vinci, Fact Cruncher, MaxThink, Ready!, ThinkTank und Thor, [Brown 1987] bespricht Manuscript. Die Systeme erlauben es, Ideen zu notieren, sie zu

organisieren, eine formal korrekte Gliederung aufzuschreiben, die Gliederung auf verschiedenen Spezifikationsebenen zu lesen und sie umzustellen.

Mit ForComment, Red Pencil und CompareRite (besprochen von [Rich 1987]) können Autorenteams besser elektronisch publizieren.

ForComment vergleicht nicht nur verschiedene Versionen eines Schriftstücks. Es verfolgt auch die Stadien in der Entwicklung des Manuskriptes.

Die bisher genannten kommerziellen Systeme sind Werkzeuge in der Hand des Autors. Sie unterstützen Teilfunktionen des Schreibens. pädagogisch motivierte Programme wie der aus dem WANDAH-System hervorgegangene HBJ Writer (rezensiert von [NydahI1986]) und Writer's Helper (besprochen von [Augustine 1987]) setzen den Schwerpunkt anders. Sie leiten zu einem methodischen Vorgehen beim Schreiben an.

Die Rezensenten von Systemen, die Autoren bei Teilaufgaben des Schreibens entlasten sollen, müssen öfters elementare Funktionsdefizite beanstanden: Die Programme lassen beispielsweise wichtige Funktionen vermissen. Sie kennen gängige Termini nicht. Darum verdächtigen sie sie als orthographisch falsch oder beanstanden sie als schwer verständlich. Ad-hoc-Lösungen reichen als Grundlage für verläßliche Systemleistungen längst nicht immer aus. So parst kein kommerzielles System die Sätze des Benutzers vollständig, bevor es deren Grammatikalität beurteilt.

5.2 Hypertextsysteme

Die ersten PC-Hypertextsysteme sind auf dem Markt (so das GUIDESystem [Harriman 1987]; Erfahrungsbericht
über die Erstellung von EDV-Handbüchern mit GUIDE bei [Brown 1987]). Damit gewinnen Hypertextsysteme praktische Relevanz und neue Aktualität (vgL [Smith 1988]).

Die Geschichte des Hypertextes (Überblick bei [Rostek & Fischer 1985]) begann 1945. ([N eis on 1967], 195f.) prägte den heute üblichen Begriff "Hypertext":

> ". . . "hypertext" is the generic term for any text which cannot be printed (or printed conveniently) on a conventional page... "Nonlinear text" might be a fair approximation. Hypertext may differ from ordinary texts in its sequencing (it may branch into trees and networks), its organisation (it may contain moving or manipulable illustrations, moving or flashing typography), and so on. ... Hypertexts will require no special training to
> use, and rather little special training to create . . . "

Hypertextsysteme sind besonders an der Brown University und bei Xerox entwickelt worden (Überblick bei [Yankelovich et al. 1985] und bei [Rostek & Fischer 1985]).

[Yankelovich et al 1985], 18 betonen die Vorzüge von Hypertexten, sie sehen aber auch noch Funktionsmängel:

> "In short, electronic document systems using today's hardware and software offer substantial advantages over paper books in providing aids for connectivity, audiovisualization, dynamics, customizability, interactivity, and rapid information retrieval, but also have a number of drawbacks in providing spatial orientation, historical tracing, joint editing, visual clarity, portability, and cost."

5.3 Autorenarbeitsplätze

Daß viele disparate Tools zur Unterstützung der Manuskripterstellung nicht das Optimum sind, wird in der neuesten Literatur erkannt (vgL auch [Krause 1988]). ([Pfaffenberger 1987], 10) fordert aus einer ganz praktischen Perspektive "the professional writer's workstation":

> "Here's my idea of the exemplary writing tool: the green box on your screen is not merely aspace on which to write; it's also a gateway to a world of writing accessories, all of which are available at a keystroke; an outlining program (which helps you maintain high standards of document organiszation), a dictionary (not just a list of words, mind you, but areal dictionary, with definitions), a text-oriented data base manager, a thesaurus, a context sensitive style guide (now there's a programming challenge for you!), and

> style analysis software (along the lines of the famed Writec's Workbench). What I'm describing is the ideal integrated word processing package . . ."

[Huntley 1986] kommt aus seiner Sicht eines Autors und Dozenten für wissenschaftliches Schreiben zu einem ähnlichen Ergebnis. Sein Autorensystem soll die Grundaktiväten beim Schreiben unterstützen: Planen, Entwerfen, Revidieren, Evaluieren und zur Veröffentlichung Aufbereiten.

6 Informationsorientiertes Schreiben und Wissensproduktion: Forschungssituation und Forschungsbedarf

Wer Wissen mit der Absicht zu informieren durch komplexe Systeme schleust, muß sich mit dem Schreiben von Texten befassen: Dabei findet die Ur-Fixierung von Wissen statt. Wenn sie nicht voll gelingt, belastet dieses Manko die gesamte weitere Verarbeitung und den angestrebten Informationserfolg.

Über informationsorientiertes Schreiben ist jedoch noch nicht allzuviel bekannt. Was an empirischem Wissen über das Schreiben verfügbar. ist, bezieht sich vor allem auf Schüler. Die Untersuchungen sind überwiegend pädagogisch motiviert. Wenn es um die Produktion von Wissen geht, schreiben jedoch Erwachsene unter professionellen Bedingungen. Ihre Schreibperformanz ist noch zu wenig erforscht. Eine Hilfe ist das praktisch orientierte Methodenwissen des technischen Schreibens.

Die Informationsqualität von Texten hängt eng mit deren Inhalt zusammen. Wie die Textbedeutung zustande kommt, ist noch zu wenig erforscht. Alternativen zur Darstellung oder Vertextung von Wissen werden in neuester Zeit entwickelt [Wahlster 1987].

Für den Vorgang der eigentlichen Textproduktion fehlen empirisch basierte Theorien. Hier müssen linguistische Theorien und Modelle in eine praktisch verwend bare Form gebracht werden, damit ihre Plausibilität beurteilt werden kann.

Die vorliegenden Modelle des Schreibens haben den Vorzug, kognitive Prozeßmodelle zu sein. Sie sind empirisch begründet, jedoch bei weitem noch nicht so entwickelt und validiert, daß man sie als praktisch tauglich ansehen könnte. Für das Schreiben informationsorientierter Texte reichen sie zudem nicht aus:

1. Sie berücksichtigen nicht das Vorgehen eines professionellen Schreibers, der bei der Arbeit eine höhere Kompetenz und Hilfsmittel einsetzt.

2. Sie modellieren die Konstituierung der Textbedeutung zu schwach. Die Planung eines informationsorientierten Textes muß eine Wissenserwerbskomponente vorsehen, die einen umfangreichen Forschungs- oder Ermittlungsvorgang und den Einsatz von professionellen Techniken und Hilfsmitteln zuläßt. Die für informierende Textarten typische starke Intertextualität muß wiedergegeben werden.

3. Sie beschreiben die in der Praxis häufige Produktion von Texten in Teamarbeit nicht.

4. Sie gehen nicht auf die unterschiedlichen Medien ein, für die ein Text geschrieben wird.

Die kommerziellen Systeme zur Unterstützung des Schreibens bewegen sich in der Welt des Desktop Publishing. Ihre Modellierung von Texten reicht so weit, wie es zur Gestaltung des Druckbildes nötig ist. Eine Texttheorie, die tiefer in die Textbedeutung eindringt, fehlt. Darum können die "höheren" Denkleistungen beim Schreiben nur sehr äußerlich unterstützt werden.

Typischerweise sind die kommerziellen Systeme Tools, die einzelne Arbeitsgänge beim Schreiben unterstützen. Funktionsmängel sind häufig auf zu vordergründige Systemlösungen zurückzuführen. Integrierte Autorenarbeitsplätze stehen noch aus. Erste Systeme ermöglichen die Erstellung von einfachen Hypertexten. Wissensbasierte Systeme sind noch nicht auf dem Markt.

An der lebhaften Forschung und Entwicklung in der Textverarbeitung, den Installa-

tionszahlen von Textsystemen und der funktionalen Vielfältigkeit des Softwareangebotes kann man ablesen, wie groß das Interesse an wissenschaftlich gut fundierten und funktionstüchtigen Systemen zur Unterstützung des professionellen Schreibens sein muß. Wieviel aus der Sicht der Systementwicklung noch zu tun ist, bis auch die sprachliche und inhaltliche Gestaltung von Texten wirksam unterstützt werden kann, kann man sich von ([Miller 1986], 204) sagen lassen:

> ". . . Specific content, organization, coherence, relevance to topic, information value, clarity. None of these can be assessed by computer software at the present time. And one pales at the thought of how much science and art have yet to be invented to make computer critiquing of this kind possible."

Literaturverzeichnis

[Anderson et al. 1983] Anderson, P. V.; Brockman, R. J.; Miller, C. R. (Hrsg.): New Essays in Technical and Scientific Communication. Farmingdale NY, Baywood, (1983)

[Augustine 1987] Augustine, D.: Writers Helper . Computers & Humanities 21 (1987) 119-121

[Bartlett 1982] Bartlett, E. J.: (1982): Learning to revise: Some component processes. 324364 In: [Nystrand 1982]

[Bazerman 1983] Bazerman, Ch.: Scientific Writing as a Social Act: A Review of the Literature of the Sociology of Science. 156-184 In: Anderson et al. 1983]

[Bazerman 1981] Bazerman, Ch.: What Written Knowledge Does: Three Examples of Academic Discourse. Philosophy of the Soc. Sciences 11 (1981) 361-388

[Beaugrande 1984] Beaugrande, R. de: Text Production. Toward a Science of Composition. Norwood NJ, Ablex, (1984)

[Bereiter 1980] Bereiter, C.: Development in Writing. 73-96 In: Gregg, 1. W.; Steinberg, E. R. (Hrsg.), (1980)

[Berenkotter 1981] Berenkotter C.: Understanding a Writers A wareness of Audience. College Composition an Communication 32 (1981) 388-399

[Bishop 1975] Bishop, R. L.: The JOURNALISM programs: Help for the weary writer. Creative Comput. 1 (1975) 28-30

[Bracewell1980] Bracewell, R. J. : Writing as Cognitive Activity. Visible Language 14 (1980) 4, 400-422

[Bridwell et al. 1985] Bridwell, L.; Sirc, G.; Brooke, R.: Revising and Computing: Case Studies of Student Writers. 172-194 In: [Warshauer Friedman 1985]

[Brown 1987] Brown, E.: Building Manuscripts Block by Block. PC world, (May 1987), 214223

[Brown 1985] Brown, P. J.: A Simple Mechanism for Authorship of Dynamic Documents. 3442 In: [van Vliet 1986]

[Buckingham 1987] Buckingham, W. J.: Rightwriter (Software Review). Computers and the Humanities 21 (1987) 69-70

[Bums 1979] Bums, H.: Stimulating rhetorical invention through computer-assisted instruction. Diss. Univ. of Texas, Austin, (1979)

[Caccamise 1981] Caccamise, D. J.: Cognitive processes in writing: Idea generation and integration. Diss. Univ. of Colorado, (1981)

[Carney 1~87] Camey, T.: Desktop Publishing and the Writer: An Outline of the Future. Res.. on Word Process. Newsl. 5 (1987) 1, 110

[Chapman 1986] Chapman, V.: Ready!: A scratchpad for outIines and notes. Electronic Library 4 (1986) 6, 324-325

[Cherry 1982] Cherry, L.: Writing Tools. IEEE Trans. Comm. COM-30 (1982) 1, 100-105

[Clark & Clark 1977] Clark, H. ; Clark, E.: Psychology and language: An introduction to psycholinguistics. N ew Y ork, Harcourt Brace Jovanovitch, (1977)

[Collins & Gentner 1980] Collins, A.; Gentner, D.: A Framework for a Cognitive Theory of Writing. 51-72 In: Gregg, L. W.; Steinberg, E. W. (Hrsg.) (1980)

[Connors 1982] Connors, R. J. : The Rise of Technical Writing Instruction. In: America. J. Technical Writing and Communication 12 (1982) 4, 329-351

[Daiute 1985] Daiute, C.: Do Writers talk to Themselves? In: [Warshauer Friedman 1985]

[Durharn et al. 1983] Durharn, 1.; Lamb, D. A.; Saxe, J. B.: Spelling Correction in User In terfaces. Comm. ACM 26 (1983) 764-773

[Emig 1971] Emig, J.: The Composing Process of 12th Graders. Urbana IL, NCTE Res. Rep. 13, (1971)

[Faigley et al. 1985] Faigley, L.; Cherry, R. D.; Jolliffe, D. A.; Skinner, A. M.: Assessing Writers' Knowledge and Processes of Composing. Norwood NJ, Ablex, (1985)

[Faigley & Miller 1982] Faigley, L.; Miller, T.: What we learn from writing on the job. Coll. Engl. 44 (1982) 557-569

[Flower & Hayes 1980] Flower, L. S.; Hayes, J. R.: The Dynamics of Composing: Making Plans and Juggling Constraints. 31-50 In: [Gregg & Steinberg 1980]

[Flower & Hayes 1980a] Flower, L. S.; Hayes, J. R.: The cognition of discovery: Defining a rhetorical problem. Coll. Comp. and Comm. 31 (1980) 21-32

[Flower & Hayes 1984] Flower, L. S.; Hayes, J. R.: Images, Plans, and Prose: The Representation of Meaning in Writing. Written Communication 1 (1984) 1, 120-160

[Flower et al. 1983] Flower, L. S.; Hayes, J. R.; Swarts, H.: Revising Functional Documents: The Scenario Principle. 41-58 In: [Anderson et al. 1983]

[Frase et al. 1985] Frase, 1. T. ; Kiefer, K. E.; Smith, C. R.; Fox, M. L.: Theory and Practice in Computer-Aided Composition. 195210 In: [Warshauer Friedman 1985]

[Gillis 1987] Gillis, P. D.: Using Computer Technology to Teach and Evaluate Prewriting. Comp. and the Humanities 21 (1987) 1, 319

[Graesser et al. 1984] Graesser A. C.: The Input Of Different Information Sources On Idea Generation. Written Communication 1 (1984) 341-364

[Gralla 1987] Gralla, P.: Focus on Word Processing. PC week 4 (1987) 21, May 26, 121-131

[Gregg & Steinberg 1980] Gregg, L. W.; Stein berg, E. W. (Hrsg.): Cognitive Processes in Writing. Hillsdale NJ, Erlbaum, (1980)

[Grosch 1986] Grosch, A. N.: Index-Aid: Computer-assisted back-of-the-book indexing. Electronic Library 4 (1986) 5, 278-280

[Gundlach 1982] Gundlach, R. A. : Children as Writers: The Beginnings of Learning to Write. 129-148 In: [Nystrand 1982]

[Harriman 1987] Harriman, C.: Hypertext Comes to the Small Screen. Seybold Outlook on Prof. Comput. 5 (1987) 7, 14-18

[Harrison 1987] Harrison, T. M. : Frameworks for the Study of Writing in Organizational Contexts. Written Comm. 4 (1987) 1, 3-23

[Harty 1985] Harty, K. J.: Strategies for Business and Technical Writing. San Diego CA, Harcourt, (1985)

[Hayes & Flower 1980a] Hayes, J. R.; Flower, 1. S.: Identifying the Organization of Writing Processes. 3-30 In: [Gregg & Steinberg 1980]

[Hayes & Flower 1980b] Hayes, J. R.; Flower, L. S.: Writing as Problem Solving. Visible Language 14 (1980) 4, 388-399

[Hayes-Roth & Hayes-Roth 1979] Hayes-Roth, B; Hayes-Roth, F.: A cognitive model of planning. Cognitive Science 3 (1979) 275-310

[Heidorn et al. 1982] Heidorn, G. E.; Jensen, K.; Miller, L. A.; Byrd, R. J.; Chodorov, M. S.: The EPISTLE Text-Critiquing System. IBM Syst. J. 21 (1982) 305-326

[Hierppe 1986] Hierppe, R.: Electronic Publishing: Writing Machines and Machine Writing. ARIST 21 (1986) 123-166

[Hillocks 1982] Hillocks, G.: Inquiry and the Composing Process: Theory and Research. Coll. Engl. 44 (1982) 7, 659-673

[Houghton-Alico 1985] Houghton-Alico, D. (1985): Creating computer software user guides. New York, MacGraw Hill, (1985)

[Huff 1983] Huff, R. K.: Teaching Revision: A Model of the Drafting Process. Coll. Engl. 45 (1983) 8, 800-816

[Huntley 1986] Huntley, J. F.: Beyond Word Processing: Computer Software for Writing Effective Prose. EDUCOM Bull. (Spring 1986), 18-22

[Kincaid et al. 1981] Kincaid, J. P.; Aagard, J. A.; O'Hara, J. W.; Cottrell, L. K.: Computer Readabilty Editing Systems. IEEE Trans. Prof. Comm. PC 24 (1981) 38-41

[Knorr 1979] Knorr, K. D.: Tinkering Toward Success: Prelude to a Theory of Scientific Practice. Theory and Society 8 (1979) 347376

[Knorr 1981] Knorr, K.; Krohn, R.; Whitley, R. (Hrsg.): The Social Process of Scientific Investigation. Dordrecht, Reidel, (1981)

[Knorr-Cetina 1984] Knorr-Cetina, K. D.: Die Fabrikation von Erkenntnis: Zur Anthr~ pologie der Naturwissenschaften. Frankfurt, Suhrkamp, (1984)

[Krause 1988] Krause, J.: Sprachanalytische Komponenten in der Textverarbeitung. In: [Röhrich 1988]

[Kucer 1985] Kucer, S. L.: The Making of Meaning. Reading and Writing as Parallel Pr~ cesses. Written Comm. 2 (1985) 3,317-336

[LaRoche & Pearson 1985] LaRoche, M. G.; Pearson, S. S.: Rhetoric and Rational Enterprises. Written Comm. 2 (1985) 3, 246-268

[Latour & Woolgar 1979] Latour, B.; Woolgar, S.: Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills CA, Sage, (1979)

[Matsuhashi 1982] Matsuhashi, A. : Explorations in the real-time production of written discourse. 269-290 In: [Nystrand 1982]

[Matsuhashi & Quinn 1984] Matsuhashi, A.; Quinn, K.: Cognitive Questions from Discourse Analysis. Written Comm. 1 (1984) 3, 307-339

[McDonald et al. 1982] McDonald, N.; Frase, L.; Gingrich, P.; Kennan, S.: The Writer's Workbench: Computer Aids for Text Analysis. IEEE Trans. Comm. (1982), 105-110

[Michaelson 1982] Michaelson, H. B.: How to Write and Publish Engineering Papers and Reports. Chicago, ISI Press, (1982)

[Miller 1984] Miller, J. J. J. : PRO TEXT I: Proceedings of the First International Conference on Text Processing Systems. Dublin, Boole Press, (1984)

[Miller 1985] Miller, J. J. J.: PROTEXT II: Proceedings of the Second International Conference on Text Processing Systems. Dublin, Boole Press, (1985)

[Miller 1986] Miller, L. A.: Computers for Composition: A Stage Model Approach to Helping. Visible Language 20 (1986) 2, 188-218

[Mills & Walter 1970] Mills, G. H.; Walter, J. A.: Technical Writing. New York, Holt, Rinehart and Winston, (1970)

[Molitor 1987] Molitor: Weiterentwicklung eines Textproduktionsmodells durch Fallstudien. Unterrichtswissenschaft (1987), 4, 396-409

[Molitor 1985] Molitor: Personen- und aufgabenspezifische Schreibstrategien. Fünf Fallstudien. Unterrichtswissenschaft (1985), 4, 334345.

[Molitor 1984] Molitor: Kognitive Prozesse beim Schreiben. Univ. Tübingen, DIFF - Forschungsbericht 31, (Okt. 1984)

[Morgan 1987] Morgan, B. A.: From Word Processing to Desktop Publishing and CD ROM: A Five-Year Bibliographic Perspective on the Impact of Computers on Writing and Research. Res. in Word Process. Newsl. 5 (1987) 5, 2-40

[Morgan 1986] Morgan, L.: Word Finder: an automated thesaurus. Online Review 10 (1986) 2, 99-101

[Mosenthal et al. 1983] Mosenthal, P.; Tamor, L.; Walmsley, S. A. (Hrsg.) : Research on Writing. New York, Longman, (1983)

[Mullins & Faust 1987] Mullins, C. J.; Faust, M.: Buyers' Guide: Multilingual Word Processing Programs. PC Week 4 (1987) 33, August 18,61-63,66-69,72,75-77

[Nelson 1967] Nelson, T. H.: Getting it out of your system. 191-209 In: [Schecter 1967]

[NydahI1986] Nydahl, J.: Writing-Instruction Software with HBJ Writer. Res. in Word Process. Newsl. 4 (1986) 4, 12-15

[Nystrand 1982] Nystrand, M. (Hrsg.) : What Writers Know. New York, Acad. Press, (1982)

[Odell1974] Odell, L.: Measuring the effect of instruction in pre- writing. Res. in the Teaching of English 8 (1974) 228-240

[Odell & Goswami 1982] Odell, L.; Goswami, D.: Writing in a non-academic setting. Res. in the Teaching of English 16 (1982) 201-223

[Peterson 1961] Peterson, M. S. : Scientific Thinking and Scientific Writing. New York, Reinhold Publ. Corp., (1961)

[Pfaffenberger 1987] Pfaffenberger, B.: The Professional Writer's Workstation: Integrated Word Processing Software: Has ii Arrived? Res. in Word Process. Newsl. 5 (1987) 2, 10-11

[Pfaffenberger 1986] Pfaffenberger, B.: The Scholar's Software Library: Rightwriter version 2.0., Rightwords version 2.0. Res. in Word Process. Newsl. 4 (1986) 4, 27-31

[Piche & Roen 1987] Piche, G. L.; Roen, D.: Social Cognition and Writing. Written Comm. 4 (1987) 1, 68-89

[Ramsey 1986] Ramsey, R.: Turbo Lightning and Strike (Software Review). BYTE, (November 1986), 289-290

[Raskin 1986] Raskin, R.: The Quest for Style. IEEE Trans. Prof. Comm. PC 29 (1986) 3, 10-18

[Rauch 1982] Rauch, W. D.: Büroinformations-systeme. Wien, Böhlau, (1982)

[Reither 1985] Reither, J. A.: Writing and Knowing: Toward Redefining the Writing Process. Coll. Engl. 47 (1985) 6, 620-628

[Rich 1987] Rich, J.: Author, Author, Author! PC World, (August 1987),250-256

[Rickheit & Strohner 1985] Rickheit, G.; Strohner, H.: Psycholinguistik der Textverarbeitung. Studium Linguistik 17/18 (1985) 1-78

[Rohman 1965] Rohman, D. G.: Pre-Writing: The Stage of Discovery in the Writing Process. Coll. Comp. & Comm. 16 (1965) 106112

[Röhrich 1988] Röhrich, J. (Hrsg.): Handbuch der Textverarbeitung (im Druck), (1988)

[Rostek & Fischer 1985] Rostek, L.; Fischer, D.: An author's workstation - visions, views and activities. 82-95 In: [Miller 1985]

[Sanders 1987] Sanders, S. P. : Subjective Objectivity ... What I'm Teaching Now. IEEE Trans. Prof. Comm. PC 30 (1987) 2, 91-95

[Scardamaglia et al. 1982] Scardamaglia, M.; Bereiter, C.; Goelman, H.: The Role of Production Factors in Writing Ability. 173-210 In: [Nystrand 1982]

[Schauer Samuels 1982] Schauer Samuels, M.: Scientific Logic: A Reader-Oriented Approach to Technical Writing. J. Techn. Writing & Comm. 12 (1982) 4, 307-328

[Scheceter 1967] Schecter, G. (Hrsg.): Information Retrieval: A critical view. London, Academic Press

[Schwartz 1985] Schwartz, H. J.: Interactive Writing: Composing with a Word Processor. New York: Holt, Rinehart and Winston, (1985)

[Schwartz 1984] Schwartz, H. J. : Teaching Writing with Computer Aid. Coll. English 46 (1984) 3, 239-247

[Scinto 1982] Scinto, L. F. M. : The Acquisition of Functional Composition Strategies for Text. Hamburg, Buske, (1982)

[Shanklin 1982] Shanklin, N.: Relating Reading and Writing: Developing a Transactional Theory of the Writing Process. Bloomington, Ind. Univ. Press, (1982)

[Smith 1988] Smith, K. E.: Hypertext - Linking to the Future. Online, (March 1988), 32-40

[Stevenson 1981] Stevenson, D. W. (Hrsg.) : Courses, Components and Exercises in Technical Communication. Urbana IL, NCTE (1981) 87

[Univ. of Michigan 1982] Univ. of Michigan, College of Engineering(1982): Written Communication for Engineers, Scientists and Technical Writers. Engineering Summer Conference July 19-23, (1982)

[van Vliet 1986] Vliet, J. C. van (Hrsg.): Text processing and document manipulation: Proceedings ot the International Conf. Cambridge Univ. Press, (1986)

[Von Blum & Cohen 1984] Von Blum, R.; Cohen, M. E.: WANDAH: Writing Aid AND Author's Helper. 154-173 In: [Wresch 1984]

[Vygotsky 1934] Vygotsky, L.: Denken und Sprechen. Berlin, (1964)

[Wahlster 1987] Wahlster, W.: Wissensbasierte Informationspräsentation. Projektskizze. Univ. Saarbrücken, (Sept. 1987)

[Warshauer Friedman 1985] Warshauer Friedman, S.(Hrsg.) : The Acquisition of Written Language. Response and Revision. Norwood NJ, Ablex, 1985

[Weiss 1985] Weiss, E. H.: How to write a usable user manual. Philadelphia, ISI Press, (1985)

[Winkler & Mizuno 1985] Winkler, V. M.; Mizuno, J. 1.: Advanced Courses in Technical Writing: Review of the Literature (19771984). The Techn. Writing Teacher 12 (1985) 1, 33-49

[Witte 1985] Witte, S. P.: Revising, Composing Theory, and Research Design. 250-284 In: [Warshauer Friedman 1985]

[Woolgar 1981] Woolgar, S.: Logic and Sequence in a Scientific Text. 239-268 In: [Knorr 1981]

[Words 1986] Words: The Business of Outliners. PC Magazine (1986), March 25, 199-220

[Wresch 1983] Wresch, W.: Computers and Composition Instruction: An Update. Coll. Engl. 445 (1983) 8, 794-799

[Wresch 1984] Wresch, W. (Hrsg.) : The Computer in Composition Instruction: A Writer's Tool. Urbana Il, NCTE, (1984)

[Yankelovich et al. 1985] Yankelovich, N.; Meyrowitz, N.; van Dam, A.: Reading and Writing the Electronic Book. IEEE Computer (Oct. 1985), 15-29

[Young et al. 1970] Young, R.; Becker, A.; Pike, K.: Rhetoric, discovery, and change. New York, Harcourt Brace Jovanovitch, (1970)

[Young & Koen 1973] Young, R.; Koen, F.: The tagmemic discovery procedure: An evaluation of its uses in the teaching of rhetoric. Ann Arbor, MI, Univ. of Michigan (ERIC No. ED 084 517), (1973)

Phonetische Beiträge zur maschinellen Spracherkennung

Parameter der Sprachanalyse im Zeitbereich ein Schritt in Richtung Transkriptionsmaschine?

Patrick Schweisthal Walter Kopetzky Arbeitskreis Spracherkennung, Sprachgenerierung und phonetische Datenbanken der Gesellschaft für linguistische Datenverarbeitung Institut für Phonetik und sprachliche Kommunikation Universität München, Schellingstr. D-8000 München 40

In den beiden letzten Ausgaben des LDVForum wurde die Extremwertanalyse im Zeitbereich als Verfahren der automatischen Sprachanalyse vorgestellt; anhand ausgewählter Intervallogramme und Sonagramme derselben Äußerungen erfolgte eine Gegenüberstellung von Zeitbereichs- und Frequenzbereichsanalyse. Der vorliegende Beitrag behandelt, ein von den Autoren entwickeltes Zeitanalyseverfahren. Anhand des Zeitsignals wurden Parameter herausgearbeitet, die eine Zuordnung von akustischem Produkt und Wahrnehmungskategorien ermöglichen sollen. Besonders zu beachten ist der Stimmhaftigkeits- und Tonhöhenalgorithmus, der in leicht abgewandelter Form auch zur Beschreibung der Realformanten herangezogen wird.

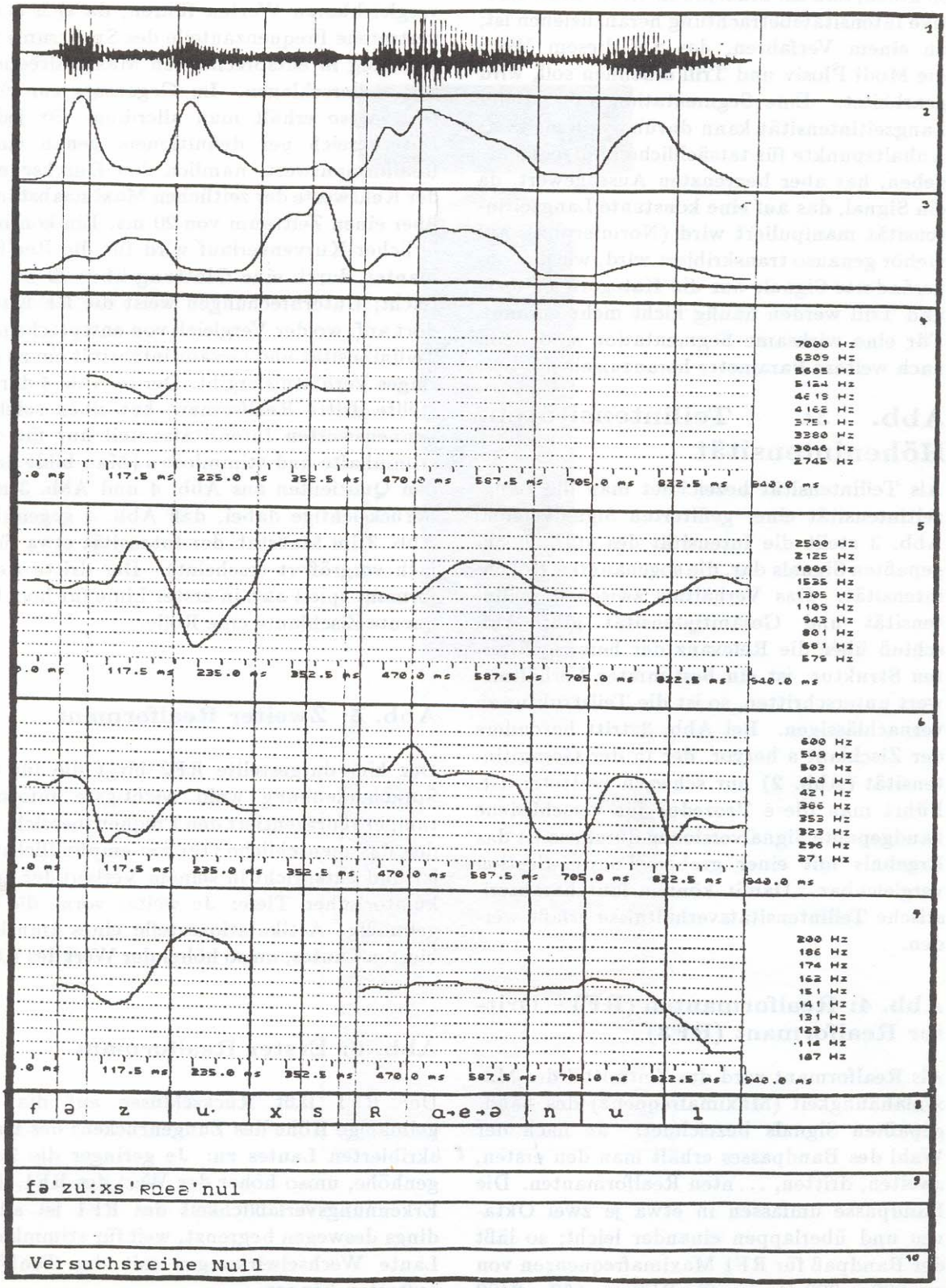
Die Parameter sind in zwei Gruppen zu unterteilen, nämlich Intensitätsbetrachtungen (Abb. 2 und 3) und Extremwertdichtebetrachtungen (Abb. 4 bis 7). Im folgenden werden die Parameter erklärt, wobei auf ihre mögliche Bedeutung für die Spracherkennung hingewiesen wird.

Abb. 1: Zeitsignal, Oszillogramm

Das Zeitsignal stellt die Druckveränderung in Abhängigkeit von der Zeit dar, die wir als Schall wahrnehmen. Die Form des Zeitsignals kann für dieselbe Äußerung - abhängig von Aufnahmedecoder und Mikrofoneigenschaften - erheblich variieren; wegen dieser hohen Varianz hat das Zeitsignal den Ruch, für die Sprachanalyse weitgehend ungeeignet zu sein, zumal entsprechende Verfahren, die sich auf den gesamten Kurvenverlauf beziehen, den Eindruck der hohen Varianz noch bekräftigen und einen unpraktikablen Rechenaufwand mit sich bringen. Demgegenüber sind geübte Transkribenten, die sich regelmäßig mit Computersegmentation an Sprachsignal beschäftigen, mit Einschränkungen in der Lage, das Oszillogramm ohne akustische Wiedergabe zu "lesen". Das brachte die Autoren auf die Idee, sich den Signalpartien zuzuwenden, an denen sich der Segmentator orientiert, nämlich den Extremwerten. Die' vorliegende Darstellung ist allerdings zeitlich so gestaucht, daß Einzelheiten nicht erkennbar sind. Der maschinellen Sprachanalyse wird das Zeitsignal durch die Auf teilung in Unterstrukturen (Realformantbereiche), Intensitäts- und Extremwertbetrachten zugänglich.

Abb. 2: Langzeitintensität, Gesamtintensität

Die Langzeitintensität stellt ein Zeitmittel der im Oszillogramm angezeigten Druckveränderungen dar. Das hier verwendete Verfahren richtet das Oszillogramm gleich und mittelt die so erhaltene Kurve zweimal über einen Zeitraum von 20 ms. Zu beachten ist, daß die absolute Intensität für die Spracherkennung nur insofern Bedeutung hat, als bestimmte Grenzwerte nach oben (Lärm) und nach unten (Stille) nicht überschritten werden können, ohne die Wahrnehmung zu beeinträchtigen; die absolute Intensität hängt schließlich von


1
2
3
4
7000 Hz
6309 Hz
5685 Hz
5124 Hz
4619 Hz
4162 Hz
3751 Hz
3380 Hz
3046 Hz
2745 Hz
5
2500 Hz
2125 Hz
1806 Hz
1535 Hz
1305 Hz
1109 Hz
943 Hz
801 Hz
681 Hz
579 Hz
6
600 Hz
549 Hz
503 Hz
460 Hz
421 Hz
386 Hz
353 Hz
323 Hz
296 Hz
271 Hz
7
200 Hz
186 Hz
174 Hz
162 Hz
151 Hz
141 Hz
131 Hz
123 Hz
114 Hz
107 Hz
0.0 ms
117.5 ms
235.0 ms
352.5 ms
470.0 ms
587.5 ms
705.0 ms
822.5 ms
940.0 ms
8
f ə z u· x s ʁ a→e→ə n u l
9
fə'zu:xs'ʁaeə'nul
10
Versuchsreihe Null

den Aufnahmebedingungen ab; auch ist die Intensitätswahrnehmung hörer- und frequenzabhängig. Aus diesen Gründen ist ersichtlich, daß als Maßstab in erster Linie relative Intensitätsbetrachtung heranzuziehen ist; an einem Verfahren, das auf diesem Wege die Modi Plosiv und Trill erkennen soll, wird gearbeitet. Eine Segmentation anhand der Langzeitintensität kann darüber hinaus zwar Anhaltspunkte für tatsächliche Lautsegmente geben, hat aber begrenzten Aussagewert, da ein Signal, das auf eine konstante Langzeitintensität manipuliert wird (Normierung), auf Gehör genauso transkribiert wird, wie das unveränderte Signal; nur die Kategorien Plosiv und Trill werden häufig nicht mehr erkannt. Für eine wirksame Segmentation sind demnach weitere Parameter heranzuziehen.

Abb. 3: Teilintensitäten, Höhenintensität

Als Teilintensität bezeichnet man die Langzeitintensität einer gefilterten Signalversion. Abb. 3 stellt die Intensität des stark hochgepaßten Signals dar, die sogenannten Höhenintensität. Das Verhältnis zwischen Teilintensität und Gesamtintensität gibt Aufschluß über die Relevanz der herausgefilterten Struktur; ist ein bestimmter Verhältniswert unterschritten, so ist die Teilstruktur zu vernachlässigen. Bei Abb. 3 tritt besonders der Zischlaut s hervor, der in der Gesamtintensität (Abb. 2) nur schwach vertreten ist. Führt man diese Prozedur für verschiedene bandgepaßte Signalversionen durch, so ist das Ergebnis mit einer groben Fourier-Section vergleichbar. Damit können lautcharakteristische Teilintensitätsverhältnisse erfaßt werden.

Abb. 4: Realformanten (RF) - Dritter Realformant (RF3)

Als Realformant wird das Zeitmittel der Maximahäufigkeit (Maximafrequenz) des bandgepaßten Signals bezeichnet. Je nach der Wahl des Bandpasses erhält man den ersten, zweiten, dritten, .. . nten Realformanten. Die Bandpässe umfassen in etwa je zwei Oktaven und überlappen einander leicht; so läßt der Bandpaß für RFl Maximafrequenzen von ca. 200 - 800 Hz zu, für RF2 ca. 600 - 2400 Hz, für RF3 ca. 2000 - 8000 Hz. Zum verwendeten Filter siehe LDV-Forum Juni 87. Im Vergleich mit den Fourierformanten ist festzustellen, daß die Realformantverläufe zu vergleichbaren Werten führen, da sich stark vertretene Frequenzanteile des Spektrums regelmäßig in entsprechenden Maximafrequenzen niederschlagen. Im Gegensatz zur Fourieranalyse erhält man allerdings für jeden Filterbereich per definitionem genau einen Realformantwert, nämlich den Durchschnitt der Kehrwerte der zeitlichen Maximaabstände über einen Zeitraum von 20 ms. Ein kontinuierlicher Kurvenverlauf wird für die Realformanten durch eine Glättung über 20 ms erreicht; Unterbrechungen weist die RF-Kurve dort auf, wo der Vergleich von entsprechender Teilintensität und Gesamtintensität ein zu geringes Verhältnis ergibt. Der in Abb. 4 dargestellte dritte Realformant hat einen solchen nennenswerten Intensitätsanteil nur für das stimmhafte und stimmlose s (Man bilde dazu den Quotienten aus Abb. 4 und Abb. 3 und berücksichtige dabei, daß Abb. 4 gegenüber Abb. 3 im Maßstab der Intensität etwa fünffach vergrößert erscheint). Der dritte Realformant spielt also in erster Linie für hochfrequente Zischlaute eine Rolle.

Abb. 5: Zweiter Realformant

Der hier dargestellte RF2 birgt die für die Spracherkennung wohl wichtigste Information; er repräsentiert den Frequenzbereich, für den das menschliche Ohr am empfindlichsten ist und entspricht in seinem Verlauf der artikulatorischen Tiefe: Je weiter vorne die regelmäßige Artikulationsstelle eines transkribierten Lautes, umso höher der Wert des RF2.

Abb. 6: Erster Realformant

Der RFl läßt Rückschlüsse auf die regelmäßige Höhe des Zungenrückens des transkribierten Lautes zu: Je geringer die Zungenhöhe, umso höher der Wert des RFl. Die Erkennungsverläßlichkeit des RFl ist allerdings deswegen begrenzt, weil für stimmhafte Laute Wechselwirkungen mit der Tonhöhe auftreten können.

Abb. 7: Die Tonhöhenkurve

Die Tonhöhenkurve gibt das Zeitmittel der Kehrwerte der zeitlichen Dauern quasiperiodischer Einheiten der Signalkurve wieder; meßbar werden die Periodendauern durch einseitige Gleichrichtung des Oszillogramms und eine Kombination spezieller Digitalfilter gemessen werden die Maximaabstände der so gewonnen Kurve. Anschließend entscheidet ein Auswahlalgorithmus über die Periodizität: N ur die Abstände werden berücksichtigt, die den Periodizitätsanforderungen gerecht werden; aus ihren Kehrwerten wird eine Kurve erstellt, die zweimal über 20 ms geglättet wird. Das Ergebnis entspricht einer Periodensegmentation von Hand. Das Tonhöheverfahren arbeitet für einen Bereich von bis zu zwei Oktaven sehr zuverlässig und liefert damit verbindliche Angaben über Stimmhaftigkeit und Betonung.

Abb. 8: Segmentelle Transkription

Hier wurden Segmenten nach Gehör des Segmentators Lautqualitäten zugeordnet; Ziel der Arbeitsgruppe ist es, eine solche Transkription vom Rechner erstellen zu lassen.

Abb. 9: Phonetische Transkription

Abb. 9 zeigt die fertige phonetische Transkription mit Akzenten als adäquate Verschriftung der Äußerung.

Abb. 10: Orthographische Umschriftung

Das Ziel der Spracherkennung ist, fließend gesprochene Sprache automatisch in ihre orthographische Norm zu überführen. Das kann unserer Ansicht nach aber nur über eine praktikable, automatische, phonetische Transkription gelingen.

Computerlinguistik und ihre theoretischen Grundlagen.

Bericht über die Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung

1 Überblick

Vom Mittwoch, dem 9. bis Freitag, dem 11. März fand das Symposium zum Thema *"Computerlinguistik und ihre theoretischen Grundlagen"* in Saarbrücken in Räumen der Universität des Saarlandes statt. Veranstalter waren die Gesellschaft für linguistische Datenverarbeitung, DGfS, Sektion 'Computerlinguistik' und GI, 'Fachgruppe natürlichsprachige Systeme'. Das Symposium wurde von ca. 140 Teilnehmern besucht.

Es gab 12 Referate und eine Podiumsdiskussion. Für die einzelnen Referate standen ca. 40 Minuten, für die jeweils anschließenden Diskussionen ca. 10 Minuten zu Verfügung. Der Zeitplan wurde genau eingehalten. Die meisten Vorträge wurden in deutscher Sprache gehalten. Den Teilnehmern standen in vielen Fällen Abstracts und Kopien der Folien der Referenten zur Verfügung. Besonders stark vertreten waren EUROTRA ([d], [h]), STUF (Stuttgart Type Unifkation Formalism)([a], [f]) und TAG (Tree Adjoining Grammar) ([h], [i]).

Am Rande des Symposiums fanden die jährliche GLDV-Mitgliederversammlung, GLDVArbeitskreise und ein DGD- Workshop statt. Bei den das Symposium begleitenden Vorführungen computerunterstützter Systeme zur Sprachdatenverarbeitung waren naturgemäß die in Saarbrücken beheimateten Projekte gut repräsentiert.

Die Qualität der Beiträge zum Symposium war recht unterschiedlich, die Auswahlkriterien für die Vorträge waren nicht ohne weiteres einsichtig. Im allgemeinen wurde verständlich vorgetragen, es kam aber auch vor, daß Text schnell abgespult wurde (z. B. [a] oder entlegener Terminologie und Abkürzungen nicht eingeführt wurden. Das Publikum reagierte insgesamt zurückhaltend auf die Referate, auch mit kritikwürdigen Referaten wurde rücksichtsvoll verfahren.

Wer Kontakte pflegen oder herstellen wollte, profitierte davon, daß das Program gerade genügend Zeit dafür übrig ließ. Die Organisation des Symposiums klappte unauffällig und reibungslos.

2 Thema

Das Thema des Symposiums "Computerlinguistik und ihre theoretischen Grundlagen" wurde dank seiner unscharfen Formulierung von allen Beiträgen berührt. Eine engere Auslegung des Themas hätte dazu führen können, daß speziell das Verhältnis von der Computerlinguistik zu ihren theoretischen Grundlagen untersucht worden wäre. In diesem Falle wäre eine Bestimmung sowohl dessen, was unter 'Computerlinguistik', als auch dessen, was unter 'theoretischen Grundlagen' verstanden werden sollte, zweckmäßig gewesen; dann hätten Wirkungen und Rückwirkungen untersucht werden können. Dieser Ansatz wurde jedoch nicht systematisch verfolgt.

Tatsächlich gab das Thema den weiten Rahmen für Referate, in denen aus Projekten, unterschiedlicher Fundierung, in denen (mit Ausnahme von [e] und [kD ein praktisches Projektziel im Vordergrund stand berichtet wurde. Die theoretischen Grundlagen standen dabei oft im Hintergrund, sie wurden nicht als gemeinsamer Ansatzpunkt behandelt. Um so interessanter waren die Referate, in denen grundsätzliche Fragestellungen problematisiert wurden.

3 Referenten, Diskutanten, Themen

Von allgemeinem, vielleicht auch von grundsätzliche Interesse war der erste Vortrag ([a]), aus einem umfangreichen Projekt; demgegenüber zeigte sich, daß auch ein kleineres Projekt ([g]) gut fundierbar ist.

Das Verhältnis der theoretischen Grundlagen zur Computerlinguistik wurden fast nur in einer Podiumsdiskusion am Donnerstag Nachmittag ([j]) abgehandelt.

Die Beiträge in der Reihenfolge des Programms:

[a] H. Uszkoreit: *Syntaktische und semantische Kategorien im linguistischen Typenverband*

Die hier vorgestellten im Zusammenhang mit dem Projekt STUF (Stuttgart Type Unification

Formalism) entstandenen Überlegungen zur Typenformalisierung und zum Strukturabgleich sind von grundlegendem Interesse. Sie gaben einen Überblick zur Datenstrukturierung bildeteten ein Raster, das zum Vergleich mit anderen Projekten hätte geeignet sein können.

(Leider lagen zu diesem Vortrag keine schriftlichen Unterlagen vor.)

[b] M. Aulich, G. Drexler, G. Rickheit, H. Strohner: *Input Wort. Ansätze der Simulation wortweiser Texterarbeitung*

Das thematisierte Teilproblem, die Imitation des Kuzzeitgedächtnisses bei der Köhärenzbildung setzt bei der Aufnahme schriftlicher Informationen (Leseforschung) an und führt erst dann zu lexikalischen, semantischen und syntaktischen Prozeduren. Die Referenten stellten zwar dar, daß Fragen offen blieben, nutzten die Chance aber nicht, differenzierte Fragen an das Auditorium zu formulieren.

[c] J. - Grabowski-Gellert: *Einige psychologische Überlegungen zu Aufforderungsinteraktionen zwischen Mensch und Computer*

Die hier vorgeführten Überlegungen orientieren sich an der Kommunikation von Mensch zu Mensch. Dazu wurden Bedingungen für und Komponenten "on Aufforderungssituationen, sowohl verbaler, als auch nonverbaler Art, ermittelt und zu formalisieren versucht.

[d] P. Schmidt: *Lösungsansätze für die Behandlung von "unbounded dependencies" in der deutschen Komponente von EUROTRA*

Der CAT-Formalismus, der verschiedene Regeln für Struktur und Attribute zur Verfügung stellt, dient zur Konzentration auf lokaler Ebene. Eine sinnvolle Ergänzung für diesen Vortrag dürfte das LDV-Forum Juni 1987 mit seinem Themenschwerpunkt *"maschinelle Übersetzung"* enthalten.

[e] T. Janssen: *A Mathematical View on Compositionality in EUROTRA*

Ausgehend von der Überlegung, daß eine direkte Übersetzung von Quell- zur Zielsprache ohne Stratifikation einfacher ist, wurde, linguistische Überlegungen außer acht lassend, das Konzept von EUROTRA abgelehnt.

[f] C. Beierle, U. Pletat, H. Uszkoreit: *An Algebraic Characterization of STUF*

Im Unterschied zum ersten STUF betreffenden Referat ([a] wurden hier der Zweck (Notation und Bearbeitung) beispielhaft dargestellt. Dabei wurde die zentrale Bedeutung algebraischer Spezifikation für diesen Ansatz demonstriert.

[g] A. Erben: *Ein Konzept zur Komposition der Semantik aus bedeutungstragenden Teilen einer Äußerung und zur Behandlung alternativer Interpretationen*

Die Beschränkung auf einen überschaubaren Diskursbereich und überschaubare Operationen soll die Erschließung und Verarbeitung von Mehrdeutigkeiten ermöglichen; es werden Disjunktionsmengen gebildet und mit fortschreitendem Parsing reduziert.

[h] K. Harbusch: *Effiziente Analyse natürlicher Sprache mit TAGs*

TAGs (Tree Adjoining Grammar) dienen der formale Beschreibung sprachlichen Wissens; besonderer Wert wird auf die Optimierung der Laufzeit, die Vertständlichkeit der Regeln und ihre Formulierbarkeit gelegt. Die Transparenz der Darstellung des Formalismus in Baumstrukturen wurde im Rahmen des Referates demonstriert.

[i] K. Schifferer: *TAGevEnv, eine Werkbank für TAGs*

Diese Werkbank ist Ergebnis sprachorientierter KI-Forschung. Sie soll einen graphischer Strukturreditor, Testhilfen für Konsistenzsicherung, Aufbau und Pflege von Lexika, sowie Parser für syntaktische Analyse liefern und graphisch unterstützt darstellen.

[j] Podiumsdiskussion: M. Bierwisch, Ch. Babel, B. Uszkoreit, W. Wahlter, B. Rieger (Moderation): Theoretische Grundlagen

Die Strukturierung des Gebiets 'Computerlinguistik' und ihrer Quellen wurde vor Beginn der Diskussion in Kurzreferaten durch die Diskutanten versucht. In der Podiumsdiskussion schloß die eingespielte Verständigung der Teilnehmer das Auditorium kaum mit ein.

[k] R. Moore: *Unification-based Semantic Interpretation*

Exemplarisch wurden Unifikationsoperationen dargestellt und dabei der Applikationsstil reflektiert. Besonderes Augenmerk fiel auf long distance dependencies und wh-movement.

[I] A. Rotkegel: *ModelIierungen in der Maschinellen Übersetzung*

Auf grundsätzliche Weise wurde hier maschinelle Übersetzung in Komponenten und Verfahren strukturiert und nicht nur als Anwendungsfall für sprachwissenschaftliches Wissen sondern auch als deren Quelle thematisiert. Danach wurden Beispiele für die formale Bearbeitung von Sätzen gegeben.

[m] A. Scheller: *PARTIKO. Kontextsensitive, wissensbasierte Schreibfehleranalyse und -korrektur*

Die Mensch-Maschine-Kommunikation soll durch dieses fehlertolerante oder -korrigierende System erleichtert werden. Der Ansatz ist pragmatisch orientiert und stützt sich auf eigene nicht immer in linguistischer Tradition stehende Überlegungen.

4 Ergebnisse

Das Symposium hat gezeigt, daß sich unter dem Thema" Computerlinguistik und ihre theoretischen Grundlagen" unterschiedlichste Ansätze versammeln lassen. Harte und allgemeinverbindliche Kriterien über Art und Umfang formaler und linguistischer Fundierung gibt es nicht.

Zu lebhaften Auseinandersetzungen kam es kaum, Gemeinsamkeiten und Widersprüche wurden selten formuliert. Insgesamt schien es, waren Interessenlage und Informationsstand der Teilnehmer oft nicht homogen genug für Auseinandersetzungen über Details und nicht heterogen genug für pointierte Auseinandersetzugen über konzeptuelle Fragen. Stattdessen bot das Symposium einen Fundus, der von den Teilnehmern individuell zu nutzen war.

Zu begrüßen ist die Publikation der Referate, die in Buchform erfolgen soll.

Michael Schmidtke Karl-Peters-Str. 13 4300 Essen 11

Linguistic Approach to Artificial Intelligence

13. International L. A. U. D. - Symposium in Duisburg 23.-26.03.1988

"Linguistic Approaches to Artificial Intelligence" wurden auf dem 13. L. A. U. D. - Symposium an der Universität - Gesamthochschule - Duisburg vorgestellt. In der altmodisch gemütlichen Atmosphäre eines ehemaligen Kurhotels und unter der Gesamtleitung von Rene Dirven (Duisburg) hörten und diskutierten rund 80 Teilnehmer aus 13 Ländern 28 Vorträge zu fast allen Themenbereichen im Grenzgebiet von Sprachwissenschaft und Künstlicher-Intelligenz-Forschung. Das Spektrum reichte von klassisch-linguistischen über kognitionswissenschaftliche und psychologische Ansätze bis vor allem zum KI-Paradigma im engeren Sinne, von einführenden Überblicken bis zu sehr speziellen Arbeitsberichten aus laufenden Projekten, von der Diskussion theoretischer Grundlagen bis zur Vorstellung implementierter Systeme.

Petr Sgall (Prag) behandelte logische Probleme der Satzbedeutung. *Edit Doron* (Jerusalem) bestimmte den Unterschied von gerundivem Nominal und Handlungsnominal aus der Sicht der Situationssemantik. Am Beispiel der Rolle lexikalischer Dekomposition für die Darstellung von Wortbedeutungen unterbreitete *James Pustejovsky* (Waltham, USA) Vorschläge zur Integration theoretisch-semantischer und praktisch sprachverarbeitender Interessen. *Yorick Wilks* (Las Cruces, USA) stellte laufende Forschungsarbeiten zur kollektiven Semantik in natürlicher Sprachverarbeitung vor und zeigte in einem zweiten Vortrag, wie Wörterbücher als Quellen für preference semantics genutzt werden können. *Jörg Siekmann* (Kaiserslautern) führt in Geschichte, Begriff und aktuelle Leistungsfähigkeit der Unifikationstheorie ein. *Hans Haugeneder* (München) gab einen systematischen Abriß der Probleme, die beim Bau von Parsern für natürliche Sprache entstehen und verglich die Leistungsfähigkeit verschiedener grammatiktheoretischer Ansätze unter diesem Gesichtspunkt. *Harada Yasunari* (Tokio) sprach über Probleme bei der deklarativen Beschreibung der Syntax natürlicher Sprachen in Unifikationsgrammatiken. *Hans Uszkoreit* (Stuttgart) nutzt die Spannung zwischen deklarativtaxonomischer Wissensdarstellung und prozeßorientierten Performanzbeschränkungen für eine vereinfachte Beschreibung letzterer. *M asaru Tomita* (Pittsburgh) zeigte den praktischen Nutzen der Trennung sprachbezogenen syntaktischen und bereichsbezogenen semantischen Wissens bei der Entwicklung anspruchsvoller sprachverstehender Systeme etwa im Bereich der maschinellen Übersetzung, während *Jürgen Kunze* (Berlin/DDR) sein Modell semantischer Repräsentation als ein Bindeglied zwischen Parsing und Wissensrepräsentation vorstellte. *Harold Somers* (Manchester) wog Vorzüge und Nachteile verschiedener Flexibilisierungsmöglichkeiten von frames zur Wissensdarstellung gegeneinander ab. *AI/red Kobsa* (Saarbrücken) stellte SB-ONE vor, ein Werkzeug zur Darstellung semantischen und bereichsbezogenen Wissens sowie zur Benutzermodellierung im Rahmen natürlich-sprachlichen Zugangs zu Expertensystemen.

Ulrich Hedtstück (Stuttgart) erläuterte Form und Leistungsfähigkeit der STUF -Graphen, die im LILOG-Projekt zum Verstehen natürlichsprachlicher Texte entwickelt werden. Am Beispiel der Diagnose und Reparatur bestimmter Maschinenschäden zeigte *Ralph Grishman* (New York), wie das zum automatischen Textverstehen notwendige Bereichswissen organisiert

und verarbeitet werden muß. *Annely Rothkegel* (Saarbrücken) diskutierte textlinguistische Grundlagen des TEXPERT-Modells, das bei der computergestützten Textanalyse und Textproduktion bereichsspezifisches mit textstrukturbezogenem Wissen verknüpft. Mit Fernblick auf automatische Texterzeugung, maschinelle Übersetzung und maschinelles Lernen untersuchte *Greg Myers* (Bradford) Kohäsionsmuster in englischen Wissenschaftstexten, um die Beziehung zwischen lexikalischer Kohäsion und dem Wissen des Lesers genauer beschreiben zu können. Mit zahlreichen Argumenten und Beispielen plädierte *Walther von Hahn* (Hamburg) für gründlichere Erforschung diskurserzeugender Prozesse und nichtlinguistischer Grundlagen von Textkohärenz.

Dietmar Rösner (Stuttgart) demonstrierte, auf welche Weise und zu welchen Zwecken im SEMSYN-Projekt Texte aus semantischen Repräsentationen erzeugt werden können. Umgekehrt stellte *Eva Hajicova* (Prag) ein System zur automatischen Konstruktion semantischer Netzwerke aus natürlichsprachlichen Texten vor. *Ulrich Reimer* (Konstanz) gab einen Überblick über das wissensbasierte Volltext- Informationssystem TOPIC, das deutschsprachige Fachtexte teilweise zusammenfassen und die erzeugten Abstracts graphisch darstellen kann. *Tamar Feuerstein* (Tel Aviv/Jerusalem) führte einen Algorithmus zur automatischen Erkennung sinntragender Wörter in Texten vor', der auch für anspruchsvolle Lückentests im Fremdsprachenunterricht verwendet werden kann.

Mit linguistischen und wahrnehmungspsychologischen Mitteln entwarf *Roben Hoffman* (New York) Grundzüge einer bisher fehlenden geordneten Terminologie zur Beschreibung topographischer Zustände, die langfristig die KI-unterstützte Auswertung von Luftbildaufnahmen erleichtern soll. *Gerhard Strobe* (Bochum) berichtete über noch nicht abgeschlossene psycholinguistische Experimente, die zeigen sollen, wie menschliche Sprecher sich semantische und pragmatische Information beim Parsen von Sätzen zunutze machen, und *Janusz Bien* (Warschau) diskutierte informationstheoretisch begründete Modelle des Gedächtnisses unter besonderer Berücksichtigung seiner Hauptleistung, nämlich des Vergessens.

Sergei Nirenburg (Pittsburgh) beschrieb das intelligente und interaktive Wissenserwerbssystem ONTOS, das den Aufbau umfangreicher Bereichsmodelle, Wörterbücher und sonstiger Wissensbanken unterstützen soll. Anhand des ALP-Materials und vor dem Hintergrund von Fergusons Sprechregister-Begriff untersuchte *Jürgen Krause* (Regensburg) die Unterschiede zwischen Mensch-Mensch- und Mensch-Maschine-Kommunikation.

Erwin Stegentritt (Saarbrücken) führte einige Probleme bei der Übersetzung des einfachen natürlich-sprachlichen Zugangs zu Datenbanken im kommerziellen PC-Programm "Q&A" vor.

Entsprechend der schnell und teilweise wuchernd expandierenden Forschung im Umkreis des KI-Paradigmas kamen zahlreiche netzartige Querverbindungen zwischen mehr oder minder umfangreichen Einzelaspekten der Fragestellungen und Lösungsvorschläge der meisten Vorträge in den Blick, aber nur wenig wirklich gemeinsame Entwicklungslinien. Insbesondere scheint es noch immer erst wenige Brücken zwischen den klassisch-sprachwissenschaftlichen Interessen und den im engeren Sinne KI-bezogenen Herausforderungen zu geben.

Unter dem Titel des Symposiums erscheinen die überarbeiteten (englischsprachigen) Vortragstexte im Herbst 1988 als Band 2 der neuen Reihe "Duisburger Arbeiten zur Sprach- und Kommunikationsforschung" im Verlag Peter Lang.

Andreas Kunz und Ulrich Schmitz

tekom * tekom * tekom

Bericht über die Frühjahrstagung der Gesellschaft für technische Kommunikation e.V. Stuttgart, 21.-23. April 1988

Die Gesellschaft für technische Kommunikation (tekom) ist ein Fachverband für Verfasser und Hersteller von technischer Literatur (Handbücher, Bedienungsanleitungen, Wartungsunterlagen etc.). tekom-Tagungen erheben den Anspruch, besondere Tagungen zu sein. Sie richten sich an Praktiker, die den Kontakt mit Kollegen und Geschäftspartnern suchen und sich weiterbilden wollen. tekom-Tagungen sind teilnehmerorientiert. Sie kombinieren Workshops, Diskussionen, Firmenbesichtigungen, eine Ausstellung, ein Speaker's Corner mit einer Informationsbörse, die Vorträge umfaßt. Die Vorträge werden mehrfach wiederholt, um allen eine individuelle Planung und das Zuhören in gesprächsfahigen Gruppen zu ermöglichen.

Die Themen der durchweg praktisch orientierten Beiträge bei der tekom-Frühjahrstagung reichten vom neuen EG-Recht bis zur Kalkulation im Desktop-Publishing, von der Wissensvermittlung in Text und Bild bis zu einer Übersicht über das aktuelle Ausbildungsangebot für technische Autoren.

Die Berichterstatterinnen besuchten die Tagung mit der Absicht, sich über den methodischen Stand des technischen Schreibens und über die

Notwendigkeit und Chancen eines Ausbildungsgangs für technische Autoren zu informieren. Darum war der Überblick über das Ausbildungsangebot von vorrangigem Interesse. Im Gegensatz z.B. zu den USA existiert bisher in der Bundesrepublik keine reguläre staatliche Ausbildung für technische Autoren. An verschiedenen Hochschulen gibt es jedoch einzelne Lehrveranstaltungen. In Karlsruhe, Landau, Aachen, Berlin, Hannover, Paderborn und Köln bemüht man sich um die Einrichtung von Studiengängen. Im Moment liegt die Aus- und Fortbildung technischer Autoren in der Hand von interessierten Privatfirmen (u.a. Siemens) und Verbänden (z.B. die Deutsche Angestelltengewerkschaft und die tekom).

Im gleichen Zusammenhang war die Diskussion über das Berufsbild des technischen Autors oder Redakteurs aufschlußr0ich, das der Bildungsausschuß der tekom entwickelt hat. Das Berufsbild geht von den Tätigkeiten aus, die technische Autoren praktisch erledigen: Sie schreiben Bedienungsanleitungen für Eieruhren ebenso wie Wartungspläne für Passagierflugzeuge. Insgesamt erfüllen sie laut Berufsbild (S.l) folgende Aufgaben:

1. Selbständiges und verantwortliches Erstellen von Informationsmitteln zu Kenntnis und Umgang mit einem technischen Produkt.

2. Zusammenstellen, Systematisieren, Veredeln und Bereitstellen von Informationsmaterialien zu Kenntnis und Umgang.

3. Verwaltung, Distribution und Aktualisierung der Informationsmittel und Informationsmaterialien.

Es spricht also nichts dagegen, den technischen Autor als speziellen Informationsberuf aufzufassen. In der Diskussion über das Berufsbild mit dessen Autoren *Herbert Herzke* und *Dietrich Juhl* wurde deutlich, wie stark praktisch arbeitenden technischen Autoren bewußt ist, daß sie mehr methodisches Wissen brauchen könnten als sie haben. Das vorliegende Berufsbild nimmt die sprachlichen Fertigkeiten eines technischen Autors als naturgegeben an. Wissenschaftliche Grundkenntnisse braucht er/sie in den Bereichen Didaktik, Psychologie, Kommunikation und Verständlichkeit (S.13). *H. Herzke* und *D. Juhl* betonten in der Diskussion ihr Interesse, den fachlichen und wissenschaftlichen Bezugsrahmen ihres Berufsbildes zu erweitern.

Einen Ansatz, mithilfe eines Redakteurhandbuchs die methodische Situation zu verbessern, schilderten *W. Hoffmann* und *W. Schlummer* von Siemens in München. Das firmeninterne Redakteurhandbuch trägt das Verfahrenswissen und das Selbstverständnis einer größeren Gruppe von technischen Autoren zusammen. Kennzeichnend für die pioniermäßige Arbeitssituation ist auch hier der geringe Bezug zu fachlichen Quellen außerhalb der eigenen Gruppe.

Technische Autoren arbeiten mit modernen Medien, Desktop-Publishing ist für sie ein etabliertes Diskussionsthema, Textsysteme werden selbstverständlich benutzt. Hypertextsysteme werden als interessante Realisierungsform von Online-Dokumentationen diskutiert. Die Praktiker in einer Diskussionsrunde, die den Bezug des technischen Schreibens zur Forschung diskutierte, konnten sich als zusätzliche Systemhilfen beim Schreiben vor allem terminologische Datenbanken vorstellen. Eine als wissensbasiert es System realisierte Autorenwerkbank stieß bei ihnen noch auf gemäßigte Skepsis, wenn auch die prinzipielle Offenheit für weitergehende Systemlösungen betont wurde, die den Autor unterstützen und nicht gängeln.

Ein Schwerpunkt des Tagungsprogramms lag bei der Nutzung von Bildinformationen. Es wurde über "Bedienungsanleitungen ohne Worte" berichtet *(G. Sedlaczek),* die Kunden von Daimler-Benz, die die Landessprache nicht lesen können, zumindest einfache Handlungsanleitungen vermitteln.

W. Zieten stellte dar, wie Bilder zu gestalten sind, damit sie schnell und reibungslos Informationen vermitteln. Insbesondere sollen die Bilder den Lesefluß möglichst wenig unterbrechen. Sie sollen keine unnötigen Informationen und nicht mehr als fünf Informationseinheiten enthalten, damit sie die Aufmerksamkeit beim Lesen nicht zu sehr beanspruchen. Eine systematische Beschreibung der Interaktion von Text und Bild bei der Wissensvermittlung trugen *H. Mandl* und *S.-P. Ballstedt* aus Tübingen bei. Im Mittelpunkt dieses Vortrags stand die Erörterung der beiden Fragestellungen: 1. Was können die Lernmedien Bild und Text füreinander leisten? 2. Wie müssen Bild-Text-Kombinationen gestaltet sein, damit sie aufeinander bezogen verarbeitet werden? Bild und Text können in Bezug auf das andere Medium verschiedene Funktionen innehaben: Das Bild veranschaulicht, organisiert oder abstrahiert Textinhalte und kann motivierend auf die Lese- und Lernbereitschaft einwirken. Der Text interpretiert oder abstrahiert die Bildinformationen und steuert die Abfolge und Intensität der Bildverarbeitung. Bild-Text- Kombinationen, die integrativ verarbeitet werden, folgen in ihrer Gestaltung dem Prinzip der Redundanz oder dem der Komplementarität. Die redundante Darstellung präsentiert die Informationseinheiten in beiden Medien mit dem Text als Leitmedium und dem Bild als zweiter, zusätzlicher Informationsquelle. Da die Bildinformationen leicht als überflüssig empfunden und übersehen werden, muß der Leser durch explizite Hinweise zu einer anspruchsvollen Bildauswertung angeregt werden, um die integrative

Bild-Text-Auswertung zu gewährleisten. Bei der "komplementären" Bild- Text-Darstellung sind die Lerninhalte derart auf die beiden Medien verteilt, daß sie jeweils "Leerstellen" enthalten, die das andere Medium ausfüllt. Zum Gesamtverständnis müssen beide Informationsquellen herangezogen und ausgewertet werden. Dem Aspekt der Komplementarität kommt bei der Wissensvermittlung mit Text und Bild besondere Bedeutung zu.

Instruktiv war die Teilnahme an einem METAPLAN- Workshop unter Leitung von *Freddy Puschkai* (IBM Herrenberg). METAPLAN ist ein Verfahren zur Ideenfindung und -organisation in Gruppen. Es verschriftlicht die Diskussion: Vorschläge werden aufgeschrieben und an die Tafel geheftet. Der Moderator sorgt dafür, daß' sie zu Clustern gruppiert werden. Die Beiträge werden in wortloser Abstimmung bewertet: Die Teilnehmer(innen) kleben Punkte an die Cluster von Karten an der Tafel. Das Diskussionsergebnis an der Tafel kann in einem nächsten Schritt praktisch umgesetzt werden. Besonders günstig ist METAPLAN, wenn es darum geht, in stark hierarchisch geprägten Gruppen von allen eine Meinungsäußerung zu bekommen. Wer keinen mündlichen Diskussionsbeitrag riskieren würde, wird doch seine Idee auf eine Karte schreiben. Die Abstimmung durch tätiges Handeln (einen Punkt ankleben gehen) vermeidet wieder den Diskurs. Da alle Gruppenmitglieder an der Meinungsbildung beteiligt sind, werden sie sich leichter mit dem Ergebnis identifizieren können. In der Unterstützung der Konsensbildung liegt ein weiterer Vorteil des Verfahrens. Wer sich in der Diskussion zuhause fühlt, wird an der Langsamkeit und der Diskursunterdrückung einer METAPLAN-Sitzung allerdings schier verzweifeln.

Brigitte Endres-Niggemeyer und Astrid Meyer, Fachhochschule Hannover

Bericht über den 11th German Workshop on Artifical Intelligence

Schloß Eringerfeld, Geseke, 28. Sept. - 2. Okt. 1987

1 Einleitung/ Zielsetzung dieses Berichts

Die GWAI 87 (11th German Workshop on Artificial Intelligence) fand vom 28. 9. bis 2. 10. 87 in Schloß Eringerfeld bei Geseke statt. Tagungsleiterin war Katharina Morik, Organisationsleiter Dimitris Karagiannis (beide TU Berlin). Sowohl fachlich als auch organisatorisch war die Konferenz, was sich bereits im Vorfeld zeigte, sehr gut vorbereitet. Entsprechend fruchtbar verlief die Tagung.

Die GWAI 87 zeichnete sich durch einen Schwerpunkt im Bereich "Natürlichsprachliche Systeme" aus. Es existierten nicht nur spezielle Sektionen zu den Themen "Generierung in natürlichsprachlichen Systemen" und" Repräsentationssysteme fUr Grammatik und Lexikon", auch im allgemeinen Programm war das Themengebiet "Natürlichsprache Systeme" mit einem eingeladenen und zahlreichen weiteren Vorträgen vertreten. Unverkennbar ist die Computer-Linguistik im Bereich der AI-Forschung als Teilbereich "Natürlichsprachliche Systeme" ein Gebiet mit wachsender Bedeutung.

Dieser Bericht faßt die Ergebnisse derjenigen Vorträge auf der GWAI 87 zusammen, die relevant für Aktivitäten im Bereich der wissensbasierten Sprachverarbeitung (insbesondere Generierung) sind. Abschließend werden eine kurze Charakterisierung der Tendenzen und eine Gesamtbeurteilung gegeben. Sowohl die Schwerpunktsetzung beim Referieren der Vorträge als auch die Beurteilung erfolgen auf der Grundlage unserer eigenen Arbeit, die derzeit im Rahmen des BMFT-Projekts WISBER erfolgt und zur Erstellung des (produktreifen) Textgenerierungssystems NUGGET geführt hat, das zur Generierung beliebiger Texte auf der Grundlage einer semantischpragmatischen Repräsentation (codiert in einer propositionalen Sprache) dient.

2 Vorträge und Workshop-Beiträge der GWAI 87

2.1 Wissensrepräsentation

Kai *von* Luck (Ko-Autor Bernhard Nebel, TU Berlin CIS/KIT) stellte einige Charakteristika des Wissensrepräsentationssystems BACK dar. Die

Repräsentation begrifflichen Wissens geschieht auf der Grundlage eines KL-ONE-Derivats. Diese TBOX wird durch die Repräsentation assertionalen Wissens ergänzt. Dabei kommt es auf eine sinnvolle Kopplungsstrategie an. Wenn nämlich Assertionen zur Wissensbasis hinzugefügt werden, soll dies Schlußfolgerungen auf der Grundlage der T-Box-Definitionen auslösen. Aus eingegebenen Assertionen werden über FORWARD- und BACKWARD-PROPAGATION Schlußfolgerungen gezogen. Dabei wird die Konzepthierarchie aufsteigend und absteigend abgearbeitet. Diese Hierarchie drückt die Spezialisierungsbeziehungen zwischen einzelnen Konzepten aus und gibt an, wie über VALUE- und NUMBER-Restriktion der verschiedenen Rollen eine Unterscheidung der Konzepte möglich ist. Insbesondere kann der Rückgriff auf die T-BOX dazu dienen, der Wissensbank hinzuzufügende Assertionen vorher auf Konsistenz zu überprüfen.

Repräsentationsprobleme bezüglich des Zeitaspekts waren das Thema des Vortrags von Klaus *Groth* (Danet GmbH). Dabei wurden Überlegungen zur mathematischen Formalisierung zeitlicher Aspekte vorgetragen. In der Forschung wird teilweise von Situationen in Vorgänger-/Nachfolger-Relationen, von Ereignissen oder von Zeitintervallen gesprochen. Unterscheidungen werden häufig zwischen Ereignissen, die in einem atomaren Teilintervall stattfinden, und Prozessen, die sich über mehrere davon erstrecken, vorgenommen.

Mit der Problematik der Prädikatenlogik höherer Stufe und den zur Lösung der Probleme vorgeschlagenen Meta-Prädikaten setzte sich Stefan *Wrobel* (TU Berlin KIT/LERNER) auseinander, wobei die Wissensrepräsentation des BLIP-Wissensakquisitionssystems zur Sprache kam. Meist wird ein Wissensrepräsentationsformalismus streng auf die Prädikatenlogik 1. Stufe beschränkt, also darauf, Eigenschaften von Basis-Objekten auszudrücken. Werden diese Eigenschaften wiederum als Objekte behandelt, über die etwas ausgesagt wird, spricht man von Prädikatenlogik höherer Stufe. Damit mit solchen Konzepten gearbeitet werden kann, sind spezielle "high-order" - Inferenzen notwendig. Metaprädikate beziehen sich *nun* auf Prädikate, wenn für diese bestimmte Regeln gelten. Beispielsweise kann transitive/p) dadurch definiert werden, daß aus p(x,y) AND p(y,z) folgt, daß p(x,z). Mit den Metaprädikaten kann ein Metafakt assertiert werden, z.B. transitive (kleiner). Grundidee ist, Prädikate höherer Ordnung durch Regeln auf der daruterliegenden Stufe zu definieren. Dies wird als Meta-Wissen verstanden (nicht zu verwechseln mit konsultativem Wissen). Da über Metaprädikate wiederum Eigenschaften ausgesagt werden können, sind auch Meta-Meta-Prädikate möglich usf.

2.2 Grundfragen

Christa *Hauenschild* (TU Berlin/KIT-FAST, Eurotra-D- Begleitforschung) schilderte KI-relevante Phänomene in der maschinellen Übersetzung. Dazu teilte sie die MUE-Geschichte in die Computer-Phase, die computerlinguistische Phase und die Phase der KI-Modelle ein. Aus der KI-Sicht stellt sich Übersetzen als eine besondere Form von Problemlösen dar, wobei linguistisches Wissen eine Quelle u.a. ist. Der computerlinguistische Ansatz habe sich gegenüber der Computer-Phase durch eine Syntaxkomponente, die Trennung von Grammatik und Programm, die Spaltung in Analyse und Synthese sowie ein Transfer-Modell ausgezeichnet. Kritikpunkte seien mangelnde informatische Fundierung, Überschätzung der Syntax, Beschränkung auf Einzelsatzebene und Vernachlässigung der Semantik gewesen. Demgegenüber integriere der KI-Ansatz Syntax und Semantik, sprachliches und nicht-sprachliches Wissen. Er ziele auf Textorientierung und Generierungsorientierung. Probleme seien die praktische Ungeeignetheit des holistischen Ansatzes und der Verstehensbegriff in der KI. Wesentlich sei aber, daß der KI-Ansatz in der MUE plausible Lösungen produziere, während beim computerlinguistischen Ansatz meist alle zulässigen Lösungen ungewichtet produziert werden. Zentrales Problem der MUE ist die Auflösung von Ambiguitäten auf der Lexem-Ebene, da hier Mehrdeutigkeiten leicht zu einer multiplikativen Steigerung der Lesarten führen können. Hier würde die Verfolgung aller Lesarten große Probleme verursachen oder gar unmöglich sein. In der Abwägung beider Ansätze und ihrer jeweiligen Vor- und Nachteile plädiert *Hauenschild* letztlich für eine Integration beider Vorgehensweisen, wobei der holistische Ansatz der KI aufgegeben und durch den mehr auf Modularisierung des Verstehensprozesses abzielenden Ansatz ersetzt werden sollte.

Peter *Bosch* (Universität Nijmegen) behandelte das Problem der anaphorischen Verweise durch Pronomen, wobei geprüft wurde, welche Wissensquellen zu welcher Zeit zur Verfügung stehen. Grammatische Strukturen sind nur bei der Verarbeitung einzelner Sätze verfügbar. Morphosyntaktische Merkmale wie Genus sind so lange bekannt, wie ein Wort als Topic gilt. Pronominalisierung kann einmal dadurch begründet sein, daß durch syntaktische Agreements ohne Relevanz der Referenz ein Pronomen (syntactic pronoun) zu setzen ist: "Peter sagt Heinz, daß er gestern krank war". Andererseits werden Pronomen bekanntlich gesetzt, wenn gleiche Redegegenstände gemeint sind, wozu Weltwissen erforderlich ist (referential pronouns). Bei der Auflösung anaphorischer Referenzen (referential pronouns) entscheiden meist die semantisch gesteuerten Erwartungen, nur in

Einzelfällen sind zusätzliche Inferenzen mit Kombination mehrerer Auflösungsstrategien nötig. Die Steuerung von Erwartungen geschieht z.B. durch die Informationen über Genus und Numerus eingeführter Redegegenstände. Führt dies bezüglich des jeweils letzten passenden Redegegenstands zu semantischen Inkonsistenzen, gestaltet sich die Suche nach dem passenden Referenten aufwendiger.

2.3 Parsing

Manfred *Gehrke* (Ko-Autor Hans Haugeneder, Siemens AG) behandelte Parsing als Suchproblem. Sind beim Parsen mehrere Analysen möglich, stellt sich das Problem der Entscheidung zwischen den Lesarten. Bei einer BESTFIRST-Suche wird in der Analyse ständig ein Bewertungsschema angewendet (die plausibelste Lesart wird zuerst verfolgt). Bei der DEPTHFIRST-Suche wird immer die erste funktionierende Regel angewendet. Heuristische Kriterien zur Entscheidung sind Plausibilität von Grammatikregeln, Komplexität der aufgebauten Struktur und semantische Plausibilität. Solche heuristischen Faktoren können gewichtet und arithmetisch verknüpft werden. Vorangegangene Konstituenten können dynamisch die Plausibilität nachfolgender Einschätzungen verändern. Suchstrategien sollten deklarativ durch Meta-Regeln bestimmt werden können. Siemens hat eine Entwicklungsumgebung erstellt, die das Experimentieren mit verschiedenen Heuristiken und den arithmetischen Gewichtungen erlaubt. Dietrich PAULUS (Universität Erlangen-Nürnberg) stellte ein System zur morphologischen Analyse von Verben vor. Die korrekten morphologischen Veränderungen von Verben wurden in Rewriting-Rules gefaßt. Ihre Anwendung wird als Abfolge der Zustände eines endlichen Automaten interpretiert. Bei der lexikalischen Codierung der Verben wird der Stammvokal durch ein X ersetzt und an den Verbstamm angehängt, so daß für Verben wie WIEGEN, WAGEN, WAGEN nur ein Grundeintrag notwendig ist. Aus dem angehängten Stammvokal ergibt sich über eine Tabelle der korrekte Ablaut.

2.4 Generierung

Massimo *Poesio* (Universität Hamburg, vorher CSELT Turin) stellte ein taktisches Satzgenerierungssystem vor, wobei die Organisation lexikalischen Wissens betont wurde. Entgegen der häufigen Auffassung, die Probleme taktischer Generierung seien gelöst, weist *Poesio* auf entsprechende ungeklärte Fragen hin.Das taktische System GEOGRAPH "linearisiert" eine Information, die von einer strategischen Komponente in Form von Conceptual Graphs *(CG)* geliefert wird. Diese CG bestehen aus Konzepten und konzeptuellen Relationen, Konzepte können generisch (Referent variabel) oder individuell (Referent konstant) sein. GEOGRAPH benötigt zur Generierung Wissen über Morphologie, um die den Konzepten entsprechenden Wörter zu erzeugen; weiterhin ist Wissen notwendig, wie diese Wörter syntaktisch organisiert werden. Da Lexeme domänenabhängig bestimmte Merkmale aufweisen, wird kodiert, mit welchen Entitäten der Wissensbank sie korreliert sind. Neben morpho-syntaktischen Merkmalen werden auch Synonymie-Relationen im Lexikon abgelegt, wobei Synonymie bei gleicher Semantik vorliegt. Die Einträge sind so organisiert, daß Suchbegriffe sowohl Lexeme sind (Parsen), als auch Konzepte sein können (Generieren). Morphologische Regeln zur Flexion bei Kenntnis der Merkmale werden als" Morphologie Actors", syntaktische Regeln als "Syntactic Actors" bezeichnet. Ausgehend vom Head des CG werden jeweils Lexeme zugeordnet, denen dann wiederum die Informationen über möglichen Satzbau beigegeben sind. Verschiedene Prozeduren generieren dann die verschiedenen Satzteile. Endergebnis ist einmal die Liste der Oberflächen-Wörter, zum anderen eine syntaktische Struktur. Bei der Generierung kann Backtracking auftreten: Soll ein Passiv-Satz generiert werden, so muß ein transitives Verb gefunden werden; wenn diese Suche erfolglos ist, wird ein Aktiv-Satz erzeugt. Als aktuelles Hauptproblern wurde die Repräsentation syntaktischen Wissens hervorgehoben.

Jochen *Dörre/Stephan Momma* (Universität Stuttgart) stellten die Generierung aus f-Strukturen als strukturgesteuerte Ableitung dar. Die Leistung ihres Systems beschränkt sich darauf, aus den funktional erweiterten syntaktischen Strukturen auf der Grundlage einer LFG-Grammatik Oberflächenstrings zu erzeugen. Dabei gilt die übergebene f-Struktur als Zielstruktur, die ausgehend von der leeren f-Struktur und dem Start-Symbol der Grammatik als Start-c-Struktur aufgebaut werden soll. Bei dieser Rekonstruktion des Aufbaus der f-Struktur fällt die Satzoberfläche ab, deren Parsing genau die zugrunde gelegte Struktur ergeben würde.

Ebenfalls sehr oberflächennah setzt das von Martin *Eme/e* (Universität Stuttgart) geschilderte System FREG E auf. Auf der Grundlage von Kasusrahmen soll unter Berücksichtigung von Syntax und Morphologie ein Oberflächensatz erzeugt werden. Semantischen Rollen werden ziemlich unmittelbar grammatische Funktionen zugeordnet. Numerus und andere Merkmale werden in der syntaktischen Struktur nur kodiert, sofern sie von Defaults abweichen. Trägt eine der Konstituenten das Merkmal TOPIC, führt dies zu Struktur-Transformationen. Bezüglich der Wortstellung wird fehlende Markierung (Standardwortstellung)

von spezieller Markierung unterschieden.

In diesem Zusammenhang erläutert Ulrich *Heid* (Universität Stuttgart) die lexikalischen Grundlagen von SEMSYN. Hier wird ein semantischdeutsches Lexikon zur Zuordnung semantischer Symbole zu Lexemen und ein morphosyntaktisches Lexikon verwendet. Die semantische Repräsentation umfaßt Relationen, die sich in Funktionen und der Syntax niederschlagen, und semantische Symbole, die in Lexeme eingehen (erweiterte Kasusrahmen-Notation). Die Verbauswahl kann kontextsensitiv je nach Füllung der Tiefenkasus erfolgen. Wenn semantische Symbole nicht unmittelbar durch Lexeme repräsentiert werden können, werden ggf. regelhaft erzeugte Paraphrasen verwendet.

Die Trennung strategischer und taktischer Generierung im Zusammenhang mit SEMSYN stellte Dietmar *Rösner* (Universität Stuttgart) dar. Ursprünglich existierte, da im Rahmen eines Übersetzungssystems die semantische Basis Resultat der vorherigen Analyse des Quellsatzes ist, keine strategische Komponente. Diese wurde später als separates, auf einer Datenbank operierendes Modul hinzugefügt. Die Reihenfolge der Einzelsätze im Text wird schon in der strategischen Komponente festgelegt. Die strategische Komponente sorgt für eine kontextabhängige Realisierung von konzeptuellen Strukturen. Temporalbeziehungen werden durch Analyse der Datenbestände zu bestimmten Zeitpunkten generiert, wobei sich Änderungen oder Beständigkeiten ausmachen lassen. Zusätzlich lassen sich funktional-pragmatische Bestimmungen wie "emphasize negative" codieren, die dann in speziell ausgewählten Syntaxmustern ausgedrückt werden. Informationen über Topic und Fokus werden ebenfalls verarbeitet. Erweiterungen werden in Richtung auf Textsortenspezifikationen anvisiert.

Helmut *Horacek* (Universität Hamburg) behandelte das Problem der Trennung von Strategie und Taktik. Eindeutig der taktischen Generierung wurde die Bestimmung der Satzgrenzen und die Topikalisierung zugewiesen, während die Berücksichtigung stilistischer Aspekte strategischer Natur sei. Die Frage der Trennung beider Komponenten wurde in bezug auf bestehende Systeme gestellt: bei MUMBLE (McDonald) ist eine rein taktische Komponente realisiert, bei N AOS werden Ereignisse formal als Pattern-Muster dargestellt, die die Grundlage zur Verbauswahl bilden.
Melish bildet in seinem System Plan-Muster auf
Teile von f-Strukturen ab. In KDS (Mann, Moore) erfolgt eine formale Umformung, bevor eine taktische Generierung beginnt. TEXT (McKeown) ist hauptsächlich strategisch, die Wortwahl wird durch die Datenbankgrundlage bestimmt. Im System KAMP (Appelt) steht der integrative Aspekt im Vordergrund, die Ziele der Satzgenerierung werden verknüpft. Als bisherige Lücke in der Forschung und Entwicklung ist nach *H oracek* der Übergang zwischen Strategie und Taktik anzusehen. Hier ergibt sich nach seiner Meinung ggf. die Notwendigkeit einer Zwischenrepräsentation. Diskutiert wurde weiterhin der Gesichtspunkt, daß über die Güte verschiedener Resultate der Strategie evtl. erst das Ergebnis der taktischen Komponente entscheidet, so daß ein Backtraining in die Strategie hinein möglich sein sollte.

Norbert *Reithinger* (Universität Saarbrücken) stellt die Konzeption des im Rahmen des XTRA-Projektes zu entwickelnden Generierungssystems POPEL vor. Grundidee ist, daß die Wissensbasen sowohl für die Generierung, als auch für die Analyse verwendbar sind. In der WHAT-Komponente wird die Auswahl des Wissens und die Reihenfolge festgelegt, in HOW wird diese Semantik in deutschen Sätzen realisiert. Die Wahl der Semantik stellt im wesentlichen eine Auswahl des konzeptuellen Wissens dar. Sind die Informationen für HOW unzureichend, wird dies an WHAT zurückgemeldet. Die konzeptuelle Darstellung wird in eine oberflächennahe (Kasus-Rahmen-)Darstellung überführt, die Grundlage für einen Syntaxbaum ist. Die Überführung in Kasus-Rahmen ist notwendig, da die Semantik in verschiedenen, nicht generalisierten Formalismen übergeben wird. Textgenerierung wird von HOW nicht geleistet, vielmehr wird nur auf der Einzelsatzebene gearbeitet.

Elisabeth *Andre* (Ko-Autoren Thomas Rist, Gerd Herzog, Universität Saarbrücken) stellte die Generierung simultaner Beschreibungen von Fußball-Szenen im System SOCCER dar. Solche simultanen Beschreibungen stellen andere Anforderungen an eine Generierungskomponente als aposteriori-Beschreibungen. Prinzipien der Generierung unter Berücksichtigung des Benutzers sind Fokussierung, Selektion und Enkodierung. Wahrnehmung und Sprachproduktion müssen temporal aufeinander abgestimmt werden. ACTIONS werden durch Verben ausgedrückt, deren Oberflächen-Kasusrahmen frühzeitig zu bestimmen ist. Für die Morphologie und Syntax wird derzeit probeweise SUTRA-S eingesetzt.

Stephan *Busemann* (TU Berlin) arbeitet bei der Generierung mit einer GPSG-Grammatik, die erlaubt, sprachliche Universalien im Formalismus, einzelsprachliche Phänomene in der Grammatik
zu kodieren. Die Generierung beschränkt sich auf einzelne Sätze und setzt auf einer oberflächennahen semantischen Repräsentation auf. Die Grammatik verfügt über permissive ID-Regeln zur strukturellen Expansion und restriktive LP-Regeln zur Linearisierung. In der Semantik ist bereits festgelegt, daß z.B. ein Aussagesatz mit Verb-Zweit-Stellung erzeugt werden soll. Letztlich findet ein direkter Match der Eingabe-" Semantik",

in der lexikalische Knoten bereits spezifiziert sind, mit der genau spezifizierten Syntax statt (Einfüllen der Lexeme in die terminalen Slots). Regeln über Head Feature und Foot Feature betreffen prinzipiell den Transport der Instanziierungen der Agreement-Variablen von unten nach oben, von oben nach unten und von einem Knoten über den Mutterknoten zum Schwesterknoten.

2.5 Zusammenfassende Beurteilung

Die vorgetragenen Überlegungen und geschilderten Ansätze sind für unsere weitere Arbeit (besonders im Bereich Textgenerierung) von Wichtigkeit, zumal Generierung von natürlicher Sprache den Schwerpunkt der computerlinguistischen Vorträge darstellte. Für den Trend gilt weiterhin, daß aus dem Bereich der unifikationsbasierten Grammatikformalismen die LFG den Favoriten darstellt. Da die Verwendung dieses Formalismus erhebliche Probleme gerade im Bereich Generierung mit sich bringt (allein die im Gegensatz zur Relation Semantik-Syntax problemlos erscheinende Umsetzung der syntaktischen Struktur in einen Oberflächen-String erfordert in der LFG-Umgebung offensichtlich erhebliche Entwicklungsarbeit), stellt sich die Frage, warum andere unifikationsbasierte Formalismen mit großer Komfortabilität (wie der in unserem System verwendete DCG-Formalismus) relativ wenig Aufmerksamkeit genießen. Auch der Versuch der Anwendung einer GPSG zur Generierung zeigt, daß zwar auf der einen Seite die Trennung von ID-Formalismus ein Sache sprachliche Sachverhalte wie Merkmals-Kongruenzen im Numerus, Kasus etc. nur mit erheblichem Aufwand im Generierungsprozeß zu verarbeiten sind. Zwischen der Semantik einer zu generierenden Äußerung und der zu konstruierenden Syntax klafft (wie auch auf der Konferenz festgestellt wurde) eine Forschungslücke. Diese Lücke führt in den meisten Ansätzen dazu, daß die Generierung der Sätze auf einer sehr oberflächennahen "Semantik" aufsetzt, die bereits viele Oberflächenmerkmals spezifiziert. Die Herkunft dieser "Semantik" bleibt dann oft unklar. Die Grundlagen der oberflächennahen Spezifikationen dieser "Semantik" liegen in Entscheidungsprozessen der strategischen Komponente, so daß von einer deklarativen Zuordnung semantischer und syntaktischer Strukturen keine Rede sein kann. Die oberflächennahe "Semantik" führt damit mindestens bei sehr großer Mächtigkeit der zugrundeliegenden Grammatik zu Problemen der sinnvollen Zuordnung der verschiedenen grammatisch möglichen Sätze zu den angezielten Inhalten. Da andererseits die Forderung der Integration semantischer Aspekte explizit erhoben wird und sich auch eine Trennung strategischer und taktischer Generierung (wenn auch mit definierten Rückkopplungen) als akzeptiert erwies, kann der in WISBER/NUGGET gewählte Ansatz, taktische Generierung als semantisch fundierte Restriktion der freien Produktionen einer zugrundeliegenden Grammatik zu verstehen, in die erwähnte" Lücke" gut eingeführt werden. Diese Einschätzung wird dadurch bestärkt, daß in dem bislang wohl in Sachen Taktik fortgeschrittensten Textgenerierungssystem SEMSYN bezüglich der taktischen Komponente ähnliche Wege gegangen werden (wenn auch auf eine detaillierte syntaktische Beschreibung der zu generierenden Sätze zugunsten einer Kasusrahmen-Notation verzichtet wird). So wird z.B. streng zwischen Tokens der Wissensrepräsentation und zugeordneten Lexemen unterschieden; auch haben semantischpragmatische Informationen (z.B. über die Sichtweise, die Fokussierung bestimmter Themen) entsprechenden Einfluß auf die entsprechenden Texte. SEMSYN unterscheidet sich damit erheblich von Systemen, die dem HAM-ANS-Paradigma folgend (POPEL, SOCCER) die taktische Komponente mehr oder weniger auf "Oberflächentransformation" beschränken (Wortstellung, Morphologie). Beachtenswert war insgesamt, daß sehr oft Generierung noch auf die Satzebene beschränkt wird wie bei XTRA/POPEL-H oder GEOGRAPH. Erst auf der Textgenerierungsebene aber werden Forschungsergebnisse wie die von BOSCH zur anaphorischen Pronominalisierung formulierten Regelmäßigkeiten relevant. Zudem zeigt sich deutlich, daß die Probleme taktischer Generierung (wenn man wirklich von einer tiefen Semantik und nicht von einer Syntax-Semantik-Mischform ausgeht) alles andere als gelöst sind und kaum überwertet werden können.

Konrad Jablonski, Nixdorf Computer AG, Paderborn

Der Arbeitsplatz des Geisteswissenschaftlers
Ein Projekt sensibler Vermittlung zwischen Ergonomie und Informatik

Die condition sine qua non der Unterstützung geisteswissenschaftlicher Arbeit durch Methoden der Informatik liegt in einem "sanften" Design der Arbeitsumgebung. Das heißt: eine Überlagerung kognitiver Arbeit an Theorie und Text mit Verständnis- und Verfahrensfragen des Instrumentariums ist konsequent, wie es der Stand der Technik erlaubt, auszuschließen. Jeder Gedanke, der an ein *Wie* der Arbeit gerichtet wird, geht sofort für das *Was* verloren. Mit den bisher realisierten Systemen konnte oft eine krasse Zielverschiebung beobachtet werden: Geisteswissenschaftler wurden zu unprofessionellen Programmierern und verloren ihre ursprünglichen Forschungsziele länger desto mehr aus dem Blick.

Daher stellen sich entschieden höhere Anforderungen an die Ergonomie des Arbeitsplatzes wie z.B. bei Systemen der Bürokommunikation. Die Entwicklung des geisteswissenschaftlichen Arbeitsplatzes (GAP) im rein universitären Umfeld ist von vornherein zum Scheitern verurteilt, weil dort, mit wenigen Ausnahmen, langfristige alltägliche Arbeitspraxis mit professionellen Systemen nicht vorliegt und die Anwender selbst Phantasien vornehmlich aus unzulänglichen Methoden der Großrechner-Ara oder der PC-Welt schöpfen.

Hinzu kommt die Methodenfrage. Eine Methode dient der Lösung einer Fragestellung mithilfe eines implementierten Verfahrens. Sie muß zuverlässig, ihre Rahmenbedingungen bekannt sein und Schnittstellen zu anderen, inhaltlich verknüpften Methoden sollten problemlos benutzt werden können. Eine Sammlung solcher Methoden sollte dem Wissenschaftler am GAP in Form einer Methodenbank zur Verfügung stehen. Der aktuelle Stand der Softwareentwicklung hat keine Integration im vorgenannten Sinne bisher bewirkt: Die universitären Forschungs- und Förderungszyklen erlauben allenfalls Prototypen, die weder portabel, oft nicht reliabel und daher selten integrierbar sind.

Es wäre jedoch auch falsch, marktreife Produkte aus der Universität zu erwarten. In der Regel gewährleistet nur die industrielle Entwicklung Kontinuität über alle Phasen des Software-Lebenszyklus mit modernen Methoden des Software-Engineerings. Da aber der GAP im kommerziellen Sinne keinen Markt hat, der hohe Entwicklungskosten rechtfertigen würde, bedarf es einer dreiteiligen Infrastruktur: Dem Sponsoring großer Hardwarefirmen, dem Know-How und der Flexibilität kleiner Tech-Unternehmen, die der Universität "nahestehen", sowie der Definitionsgruppe auf universitärer Seite.

Während die Methodenentwicklung eine langfristige Aufgabe der Philologien und der computerlinguistisch-orientierten Grundlagenforschung bleibt, kann der GAP bei entsprechender Planung als integratives Rahmensystem angelegt werden, das unter vorrangiger Berücksichtigung des ergonomischen Designs zunächst realisierte Methoden integriert - etwa die Leistung der Systeme CLIO, LDVLIB, LEMMA, TUSTEP, ... - und in einer weiteren Phase neue Methoden "aufsaugt". Nur so kann einem überschaubaren Zeitrahmen - bei entsprechendem Vorgehen und angemessener Ausstattung - ein GAP geschaffen werden. Allerdings ist hierfür ein strategisches Konzept vonnöten, das im Zusammenwirken der Autoren realisierter Systeme mit kompetenten Informatikern und repräsentativen Anwendern innert kürzester Frist zu erarbeiten ist.

Das strategische Konzept beantwortet unter anderem folgende Fragen: Welche Leistungen werden am GAP direkt, welche über verteilte Systeme erbracht? Welche Wissens-

basis welcher Struktur ist der internen Welt des GAPs zugrunde zu legen? Wie werden Texte repräsentiert und inwieweit folgt man hierbei neueren Normen? Welche bereits existierende Hardware kommt unter Kosten- j N utzen-Gesichtspunkten in Frage, damit der GAP für den Anwender finanzierbar bleibt? Welche Rechte der Softwarehersteller müssen berücksichtigt werden? Wie kann die langfristige Weiterentwicklung des GAP institutionell gesichert und finanziert werden? Welche Schnittstellen sind zu schaffen? In welchem Maße werden Datenbankmethoden und KI-Techniken eingebunden? Welche sprachlichen Wissensquellen (Wörterbücher, Textarchive, Parser, ...) können einbezogen werden? Welcher Stil der Benutzerinteraktion soll rur eine verläufige7endgültige Version gewählt werden? Diese Liste bleibt hier zunächst ungeordnet und unvollständig.

Anschließend ist ein Reverse-Engineering der für die Erstversion ausgewählten Systeme vermutlich der kürzeste Weg, um zu greif baren Ergebnissen zu gelangen. Die Re-Implementierung der Methoden in einer modernen Programmiersprache, die Herstellerunabhängigkeit und damit Portabilität sowie die nötigen Strukturierungs- und Integrationswerkzeuge bereitstellt, stellt den nächsten Schritt dar. (Vermutlich erfüllt nur die Sprache C im Moment die genannten Bedingungen.) Daß hier aus Kosten- und Zeitgründen auf Standardwerkzeuge in der Systementwicklung zurückgegriffen werden muß, versteht sich von selbst.

Den strategischen und instrumentellen Skizzen soll nun aber - ebenso knapp die Anforderungen aus der Sicht der zwischen Text und Theorie operierenden Philologien/Linguistiken folgen. Hier bestehen sicher mehrere Schichten in der Arbeit am GAP, die sich aber in eine zeitlich entzerrte Folge bringen lassen:

Texte, Situationen und Wissensquellen stellen die Datenbasis dar. Ein quasi handwerklicher Prozeß besteht darin, hieraus Daten in einer Form zu erzeugen, die nicht nur historisch kritische Ausgaben, sondern auch statistische Auswertungen, Wörterbücher, Bibliographien, Glossare und dergleichen anfertigen lassen. Diese konventionellen Formen des textwissenschaftlichen Arbeitens müssen adäquat im GAP repräsentiert sein. Aus ihnen muß jederzeit ein publikationsreifes Manuskript zu erzeugen sein. Ergänzt wird diese materiale Stufe durch vorläufig strukturiertes Wissen, das wahlfrei mit den Methoden des Hyper- Text abgelegt wird. Der Begriff "Gedankeneditor" mag hier als kleine Provokation dienen, um einen möglichen Weg der Entwicklung anzudeuten.

Die folgende Stufe besteht aus Auswertungsmethoden, die es erlauben, mithilfe des Rechners Wissen in strukturierter Form aus der Datenbasis zu ermitteln. Solche operationalisierten Verfahren der Textanalyse stellen eine offene Menge von Methoden dar, die jederzeit, je nach Fortschritt der Grundlagenforschung, ergänzt werden kann. Dabei muß es auch am GAP eine "Werkzeugkiste" geben, die den Textwissenschaftler in die Lage versetzt, neue Methoden modular zusammenzubauen und damit zu experimentieren.

Der vorgenannten Stufe ist eine Erfassungskomponente zuzuordnen, die nicht operationalisiertes, intellektuell erarbeitetes Textwissen ergänzen hilft. Für die weitere Auswertung und Bearbeitung können neben NI- auch KI-Methoden versuchsweise eingesetzt werden. Hier ist zu beachten, daß damit der Stil konventionell text wissenschaftlicher Arbeit sicher verändert wird. Da diese Änderung lediglich instrumenteller Natur ist, sollte sie nicht mit einem "paradigmatischen Sprung" verwechselt werden. Mit Akzeptanzbarrieren ist sicher zu rechnen - warum auch nicht? Sie können Ausgangspunkt einer fruchtbaren Methodendiskussion sein.

Uber eine weitere Stufe reden, erscheint aus der derzeitigen Situation verfrüht, weil damit eine Realisierung in weite Zukunft rückt. Statt ausgiebiger Spekulation erscheint es sinnvoller, Zeit und Intelligenz darauf zu verwenden, das Werkzeug GAP im Sinne des eingangs gestellten Postulats soweit wie möglich zu verfeinern. Die Differenziertheit der Werkzeuge spricht, wenn die Kunst außer Betracht gelassen wird, für die Höhe einer Kultur. Dies folgt aus der Betrachtung der älteren Ge
schichte. Gerade die Geisteswissenschaftler haben die Chance, eine neue Werkzeugkultur zu fordern und mitzuformen.

Raimund Drewek, Stuttgart, Feb. 1988

Veranstaltungen

Veranstaltungskalender

Oktober 1989

28.9.88 – 1.10.88, PASSAU (D): Jahreskongreß der Gesellschaft für Angewandte Linguistik. Veranstalter: GAL , *Information: Bernd Spillner, Fachbereich Romanistik, Lotharstr. 65, D-4100 Duisburg*

5. – 7.10.88, CHARLES, ILL (USA): Workshop on Spatial Reasonning and Multi-Sensor Fusion. *Information: Si-Chin Chen, Dep. of Computer Science, University of North Carolina, Charlotte, NC 28223*

5. – 5.10.88, BERN (CH): Annual Conference on Artificial Intelligence . SGAICO'88 , *Information: Prof. H. Bunke, Institut für Informatik und angewandte Mathematik, Universität Bern, Länggasstr.51, CH-3012 Bern*

6. – 8.10.88, CLAUSTHAL (D): Arbeitstagung der SIG: Begriffsanalyse und Künstl. Intelligenz. Veranstalter: TU Clausthal , *Information: Prof. Dr. W. Lex, Institut für Informatik der TU Clausthal, Erzstr. 1, D-3392 Clausthal-Zellerfeld*

7. – 9.10.88, MELBECK (D): Interdisziplinäres Symposium über Erziehung – Lernen. Veranstalter: Arb.ber. Erziehung, Dt. Ges. für Semiotik , *Information: Gerd Jansen, Magdeburger Straße 50, D-2120 Lüneburg*

12. – 14.10.88, TOKIO (J): IAPR Workshop on Computer Vision. *Information: Mikio Takagi, Institute of Industrial Science, University of Tokyo, 7-22-1, Roppongi, Minato-ku, Tokyo 106, Japan*

17. – 19.10.88, HAMBURG (D): 18. Jahrestagung der Gesellschaft für Informatik / CONCURRENCY '88. Veranstalter: Univ. Hamburg, FB Informatik , *Information: Prof. Dr. R. Valk, Universität Hamburg, Rothenbaumchaussee 67-69, D-2000 Hamburg*

19. – 21.10.88, ATLANTA Georgia (USA): 6^{th} Symposium on Emperical Foundations of Inf. and Software Sciences. *Information: Pranas Zunde, School of Information and Computer Science, Georgia Institute of Technology, Atlanta, Georgia 30332, U.S.A.*

21. – 21.10.88, BASEL (CH): Informationssysteme im Unternehmen: Personal Computing, Corp. Comp.. (SI-Jahrestagung) , **SI'88** , Veranstalter: Schweizer Informatiker Gesellschaft (SI) , *Information: Schweizer Informatiker Gesellschaft, Geschäftsstelle, Postfach 570, CH-8027 Zürich*

25. – 28.10.88, STUTTGART (D): 3. Software-Ergonomie Herbstschule SEH'88. (Software-Ergonomie Herbstschule) , **SEH'88** , Veranstalter: GI Dt. Inf.-Akademie GmbH, GI-FG 2.3.1 , *Information: GI Deutsche Informatik-Akademie GmbH, Godesberger Allee 99, D-5300 Bonn 2*

27. – 29.10.88, BOCHUM (D): Kolloquium über Fernsehen: Medium und Spiegel. *Information: Walter A. Koch, Bochumer Semiotisches Colloquium, Postfach 102148, D-4630 Bochum 1*

November 1988

7.- 11.11.88, BANFF (CA): 3^{rd} Knowledge Acquisition for Knowledge-based Systems Workshop . *Information: John Boose, Adv. Techn. Center 7L-64, Boing Computer Services, Bldg. 33.07, 2760 160th Ave. SE, Bellevue, Washington, USA 98008*

7. – 9.11.88, DARMSTADT (D): Benutzerschnittstellen – Werkzeuge und Techniken zur Gestaltung interaktiver Systeme. (GI/GChACM-Fachtagung) , Veranstalter: GI-FG 2.1.2 (Interakt. Sys.); GChACM , *Information: Prof. Dr. H.-J. Hoffmann, TH Darmstadt, FB 20, Programmiersprachen und Übersetzer, Alexanderstraße 24, D-6100 Darmstadt*

14. – 19.11.88, HANNOVER (D): 5. Internationaler Leipniz-Kongreß. *Information: Institut für Philosophie, Universität Hannover, Postfach, D-3000 Hannover*

15. – 17.11.88, MELBOURNE (AU): 1^{st} Australian Knowledge Engineering Congress. *Information: Prof. B. Garner, Deakin University, Victoria 3217, Australia*

15. – 17.11.88, NIMES (F): Neuro-Nimes 88 – Neural Networks & Their Applications. (Neuro-Nimes) , Veranstalter: Association de recherche Cognitive (ARC) , *Information: EC2, 269-287, rue de la Garenne, 92000 Nanterre, France*

28.11.88 – 02.10.88, TOKIO (J): Fifth Generation Computer Systems 1988. FGCS '88 , Veranstalter: ICOT , *Information: Prof. Dr. H. Schweppe, FU Berlin, Institut für Informatik, Netorstraße 8-9, D-1000 Berlin 31*

29.11.88 – 01.12.88, BERLIN West (D): Elektronische Sprachdialogsysteme. BIGTECH , *Information: Prof. Dr. K. Fellbaum, Institut für Fernmeldetechnik, Technische Universität Berlin, Einsteinufer 25, D-1000 Berlin 10*

30.11.88 – 04.12.88, KASSEL (D): Symposium über Zeitzeichen, Zeitbewußtsein, Zeiterfahrung. *Information: Projekt "Zeit, Sprache, Geschichte", Universität GHS Kassel, Postfach 101380, D-3500 Kassel*

Dezember 1988

2. – 4.12.88, BERLIN (D): Dt.-franz. Symposium üb. semiotische ikonograph. u. kunstsoz. Aspekte. *Information: Michael Nerlich, Technische Universität Berlin, Sekr. TEL 3, Ernst-Reuter-Platz 7, D-1000 Berlin*

5. – 9.12.88, SANTA FE, New Mexico (USA): ACM–Conference on Document Processing Systems. *Information: Association for Computing Machinery, 11 West 42nd Street, New York, New York 10036, U.S.A.*

8. – 10.12.88, BERLIN West (D): Tagung über Symbole der Jugendkultur – Im Vergleich Berlin / Warschau. *Information: Roland Posner, Arbeitsstelle für Semiotik, Technische Universität Berlin, Sekr. TEL 6, Ernst-Reuter-Platz 7, D-1000 Berlin*

9. – 10.12.88, ANTWERPEN (B): Internationales Symposium über Sprachuniversalien. *Information: Johan van der Auwera, Universitaire Instelling Antwerpen, Departement Germaanse Filologie, Universiteitsplein 1, B-2610 Wilrijk*

Januar - März 1989

4. – 7.1.89, FT. LAUDERDALE, FL (USA): 2nd Int. Workshop on Artificial Intelligence and Statistics. *Information: W. Gale, AT&T Bell Laboratories 2C278, 600 Mountain Avenue, Murray Hill, NJ 07974, USA*

23. – 26.1.89, BERLIN West (D): Symposium über chin. Zeichenkonzeptionen u. westl. Zeichentheorien. Veranstalter: TU Berlin und Akad. für Ges.wiss. Peking , *Information: Roland Posner, Arbeitsstelle Semiotik, Technische Universität Berlin, Sekr. TEL 6, Ernst-Reuter-Platz 7, D-1000 Berlin*

1. – 3.3.89, ZÜRICH (CH): Datenbanksysteme in Büro, Technik und Wissenschaft . BTW'89 , Veranstalter: GI; SI , *Information: Tagung BTW 89, Prof. Dr. C. A. Zahnder, Institut für Informatik, ETH-Zentrum, CH-8092 Zürich*

8. – 10.3.89, ULM (D): Interaktion und Kommunikation mit dem Computer. (Jahrestagung 1989 der GLDV) , Veranstalter: **GLDV** , *Information:* **GLDV-89, c/o Dr. Dietmar Rösner, FAW, Postfach 2060, D-7900 Ulm**

15. – 18.3.89, PARIS (F): 4th Meeting of the International Society for Research on Emotion. (Meeting of the International Society for Research on Emotion) , Veranstalter: Int. Society for Research on Emotion , *Information: Matty Chiva, U.F.R. de Psychologie, Universite de Paris X, 200 Avenue de la Republique, F-92001 Nanterre*

28. – 31.3.89, INNSBRUCK (A): 5. Österreichische Artificial-Intelligence-Tagung 1989. Veranstalter: ÖGAI , *Information: Österreichische Gesellschaft für Artificial Intelligence, ÖGAI-Tagung 1989, Postfach 177, A-1014 Wien*

April – Juni 1989

2. – 0.4.89, PERPIGNAN (F): 4. Kongreß der Internationalen Vereinigung für Semiotik (IASS/A.I.S.). IASS/A.I.S , *Information: Gerad Deledalle, 15, Rout du Poussan, F-34140 Montbazin*

5. – 9.4.89, Xi'an (CHINA): Tagung über Textforschung in China und im Westen. *Information: Prof. Dr. H. Bluhme, Universitaire Instelling Antwerpen, Dep. Germ. Filologie, Universiteitsplein 1, B-2610 Wilrijk*

10. – 12.4.89, AUGSBURG (D): 13. Jahrestagung der GfK: Inhaltl. und numerische Analyse von Daten. *Information: Prof. Dr. O. Opitz, Lehrstuhl für Math. Methoden u. Wirtschaftswiss., Universität Augsburg, Memmingerstr. 14, D-8900 Augsburg*

29. – 31.5.89, GARMISCH-PARTENKIRCHEN (D): Wirtschaftlichkeit von Informationstechniken. (Int. Fachkonferenz der Dt. Ges. für Dokumentation e.V. (DGD)) , Veranstalter: DGD-KWID , *Information: Werner Schwuchow, GMD-Forschungsstelle für Informationswirtschaft, Schönhauserstraße 64, D-5000 Köln 51*

25. – 28.6.89, CAMBRIDGE, MASS (USA): 12th Int. Conference on Research and Development in Inf. Retrieval. SIGIR89 , Veranstalter: AICA-GLIR; BSC-IRSG; GI; INRIA , *Information: Prof. C. J. van Rijsbergen, Department of Computing Science, Glasgow University, Lilybank Gardens, Glasgow G12 8QQ, Scotland*

27. – 30.6.89, CHARLOTTESVILLE, VA (USA): 2^{nd} Conference of the Int. Federation of Classification Societies . (Conference of the Int. Federation of Classification Societies) , **IFCS'89** , *Information: IFCS-89, Department of Mathematics, Math.-Astronomy Building, University of Virginia, Charlottesville, VA 22903, USA*

Juli – September 1989

7. – 21.7.89, URBINO (I): Semiotisches Sommerinstitut. (Semiotisches Sommmerinstitut) , *Information: Christina Catani, Centro Internazinale di Semiotica e di Linguistica, Piazza del Rinascimento 7, I-61029 Urbino*

7. – 14.8.89, KIRCHBERG (A): Symposium über Wittgenstein: Versuch einer Neubewertung. *Information: Österreichische Ludwig-Wittgenstein-Gesellschaft, A-2280 Kirchberg am Wechsel*

21. – 27.8.89, EDINBURGH (GB): Special Session on Law and Semiotics. *Information: Roberta Kevelson, Center f. Sem. Research in Law, Governm. and Econ., Pennsylvania State Univ., Berks, Reading, PA 19610, U.S.A.*

28.8.89 – 1.9.89, SAN FRANCISCO, CA (USA): 11th IFIP Congress'89 - Better Tools for Professionals. (World Computer Congress IFIP Congress) , Veranstalter: IFIP , *Information: Herve Gallaire, ECRC, Arabellastraße 17, D-8000 München 81*

8. – 10.9.89, CAMBRIDGE, MASS (USA): Internationales Kolloquium zum 150. Geburtstag von Charles S. Peirce . *Information: Kenneth L. Ketner, Institute for Studies in Pragmaticism, Texas Tech University, Lubbock, TX 79409, U.S.A.*

11. – 14.9.89, MONTREAL (CA): 2^{nd} Int. Scientific Conference on Work with Display Units 89. Veranstalter: IRSST , *Information: General Secretary, Inst. de recherche en sante et en sec. travail 505, boul. de Maisonneuve Quest, Montreal, H3A 3C2, Kanada*

19. – 28.9.89, SLASKY DVUR (TS): 8th Summer School: Man-machine Interface in Scientific Environments. Veranstalter: Comp. Physics Group/Europ.Phys.Society , *Information: Dr. J. Nadrchal, Institute of Physics, 180 40 Praha 8, Tschechoslowakei*

20. – 25.9.89, DETROIT, MICH (USA): 11^{th} Int. Joint Conference on Artificial Intelligence. ijcai89 , *Information: Prof. Wolfgang Bibel, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada V6T 1W5*

26. – 28.9.89, PARIS (F): EUROSPEECH Conference. *Information: Jean-Pierre Tubach, Ecole Nationale Superieure des Telecommunications 46, rue Barrault, F-75634 Paris CEDEX 13*

Oktober – Dezember 1989

02.10.89 – 04.10.89, HAMBURG (D): 11. DAGM-Symposium: Mustererkennung. Veranstalter: DAGM in Zusammenarbeit mit Trägerges. , *Information: Prof. Dr. H. Burkhardt, Technische Informatik I, TU Hamburg-Harburg, Harburger Schloßstraße 20, D-2100 Hamburg 90*

17. – 20.10.89, BEIJING (CHINA): 9. Internationale Tagung über Mustererkennung. *Information: 9 ICPR Secretariat, Chinese Association of Automatisation, P.O. Box 2728, Beijing, VR China*

1990

15.04.90 – 21.04.90, SALONIKI (GR): 9. Weltkongreß für Angewandte Linguistik. *Information: Prof. S. Efstathiadis, P.O. Box 52, Aristoteles-Universität, GR-54006 Tessaloniki*

13. – 27.7.90, URBINO (I): Semiotisches Sommerinstitut. *Information: Christina Catani, Centro Internazinale di Semiotica e di Linguistica, Piazza del Rinascimento 7, I-61029 Urbino*

Gesellschaft für Linguistische Datenverarbeitung (GLDV)

Jahrestagung 1989

Universität Ulm, 8.3. - 10.3. 1989

Interaktion und Kommunikation mit dem Computer

Die Schnittstelle zwischen Mensch und Computer ist Gegenstand mehrerer Disziplinen. Die Tagung soll verschiedene Sichten auf diese Thematik zusammenführen.

Welche der möglichen Interaktionsformen (etwa Menutechnik, graphische Interaktion, natürliche Sprache, Kommandosprachen, Formalsprachen) dem Benutzer beim Arbeiten mit dem Computer die bestmögliche Unterstützung bietet, ist eine Forschungsfrage u.a. der Softwareergonomie. In der Linguistischen Datenverarbeitung (LDV) steht die natürliche Sprache als Medium im Vordergrund (z.B. natürlichsprachliche Zugangs- und Dialogsysteme). In der sprachorientierten KI-Forschung wird der Aspekt der Simulation menschlicher Kommunikation mit dem Rechner betont. Die Integration von natürlichsprachlicher Ein/Ausgabe mit anderen Interaktionsformen wird in den letzten Jahren immer stärker betrieben.

Mit ihrem Schwerpunktthema soll die Tagung einen Dialog zwischen den verschiedenen Forschungsperspektiven initiieren. Für die Jahrestagung 1989 der GLDV sind daher neben Beiträgen über aktuelle Forschungsergebnisse aus allen Gebieten der LDV insbesondere solche zum Schwerpunkt der Tagung erwünscht.

Dazu zählen vor allem:

Grundlagen, Modelle und interdisziplinäre Perspektiven: Kommunikationstheorie, Interaktionsmodelle, Bewertungskriterien...

Interaktionsformen: direkte Manipulation, graphische Interaktion, natürlichsprachliche Interaktion, multimediale Interaktion...

Kontextaspekte der Mensch-Computer-Interaktion: Probleme der Kontextinterpretation, adaptive Systeme, adaptierbare Systeme, Benutzermodelle, Dialogmodelle...

Benutzerunterstützung: Erklärungskomponenten, Hilfesysteme, tutorielle Systeme...

Empirie: Evaluationsstudien, Erfahrungsberichte (auch aus der industriellen Praxis)...

Die Beiträge werden in einem Tagungsband veröffentlicht, der zu Beginn der Tagung vorliegen soll.

Zeitplan:
Einreichen von begutachtungsfähigen Kurzfassungen (max. 4 Seiten, 4-fach) bis 15.10.1988

Benachrichtung über die Annahme bis 7.11.1988

Abgabe der druckfertigen Manuskripte bis 15.12.1988

Programmkomitee:
Brigitte Endres-Niggemeyer, FH Hannover
Thomas Herrmann, Universität Dortmund
Alfred Kobsa, Universität Saarbrücken
Dietmar Rösner, FAW Ulm

Papiere und Rückfragen bitte an:
GLDV-89
c/o
Dr. Dietmar Rösner
FAW
Postfach 2060
D-7900 Ulm
Tel. (0731) 142920/24
e-mail gldv-nl@ifistg.UUCP

Hintergründe & Begründungen

zur Konturierung des Faches Computerlinguistik

1 Berichtigung:

Im vorigen LDV-Forum (5, Nr. 2/3 (1987/88), S. 94f.) ist bei der Darstellung der Studiengänge für die Weiterbildung von Geisteswissenschaftlern der Begriff "Regelstudienzeit" fehl am Platze; es muß jeweils, wie auf S. 93 heißen: "vorgesehene Studienzeit" (zur Begründung S.u. Erläuterungen zu Kap.7).

2 Erläuterungen

zu den Kapiteln 5-7 von "Zur Konturierung des Faches Computerlinguistik"

Zu den Kapiteln 5 - 7, die bisher noch nicht veröffentlicht wurden, werden im folgenden Zusatzinformationen gegeben, die erklären, warum sich die *"Arbeitstreffen LD V-Studiengänge"* für die vorgelegte Textformulierung entschieden haben und die das Verständnis dieses Textes erleichtern sollen.

Kapitel 5 wurde in Anlehnung an "Empfehlungen zum Nebenfach-Studium der Informatik" (des Fakultätentages Informatik vom 11.11.1983) formuliert.

Kapitel 6 wurde deshalb notwendig, weil es in der Bundesrepublik die Konstruktion gibt, daß das Fach Computerlinguistik als ein Teil in einem anderen Fach studiert wird (von Einzelveranstaltungen wird hier abgesehen). Dabei kann es sein, daß der Spezialisierung auf CL eine beiden Fächern gemeinsame Grundlagenkomp0nente im Grundstudium vorausgeht (Studiengang Linguistik mit Studienrichtung Computerlinguistik/ Künstliche Intelligenz in Bielefeld). Ebenso ist möglich, daß die Fächer den Studenten im Grundstudium - wenn sie sich noch nicht für eine Spezialisierung entschieden haben - Lehrangebote machen (Studiengang Allgemeine Sprachwissenschaft mit Teilfach Linguistische Informationswissenschaft in Regensburg) . Und es gibt auch die Variante, daß fur die Studierenden aus einem Lernangebot beider Fächer und gemeinsamer Teile eine bestimmte Anzahl von Lehrveranstaltungen vorgeschrieben ist, so daß sich de facto eine Gewichtung zugunsten eines Faches ergeben kann, aber auch eine Gleichverteilung beider Fächer (Studiengang Kommunikationsforschung und Phonetik (mit den Teilen Phonetik und maschinelle Sprachverarbeitung)). Weiterhin kann zwischen mehreren Spezialisierungen gewählt werden wie in Bielefeld zwischen den Studienrichtungen Psycholinguistik oder Soziolinguistik oder Computerlinguistik o.a. und wie in Regensburg zwischen den Teilfächern Theoretische und Angewandte Sprachwissenschaft oder Linguistische Informationswissenschaft. In Bonn dagegen muß keiner der beiden Teile explizit gewählt werden, sondern die Studierenden können sich angesichts des Lehrangebots von Fall zu Fall entscheiden. Daraus ist leicht zu ersehen, daß diese Studiengänge mit "eingebetteter" CL auch untereinander nur bedingt vergleichbar sind. Denn nicht in jedem Fall ist es leicht möglich, eine Lehrveranstaltung von ihrem Titel her der CL zuzuordnen, so daß schon ein rein quantitativer Vergleich der SWS-Anzahl schwierig ist. Zudem hängt der Erfolg eines solchen Studiums sehr davon ab, ob zwischen den Lehrveranstaltungen beider Fächer eine inhaltliche Integration besteht, die die/die Studierende nachvollziehen kann, oder ob die Veranstaltungen beider Fächer nur zu einem Studienfach addiert werden. Wird ein solches "Kombinationsfach" im Magisterstudiengang vom Drei-Fächer-Typ als Nebenfach studiert, so reduzieren sich die SWS-Zahlen in der Regel auf die Hälfte, wobei wiederum beide Fächer zu gleichen Anteilen vertreten sein können oder (wie in Bonn) die Gewichtung auf ein Fach gelegt werden kann. Dazu ist zu bedenken, daß Studierende im Einzelfall weitere freiwillige Einarbeitung in das CL-Fach leisten mögen. Mit dem Rückverweis auf Kap.5 soll angesichts solcher Sachlage besonders darauf hingewiesen werden, daß keine *allgemeinen* Aussagen darüber möglich sind, wie sinnvoll ein solches *Nebenfach*-Studium ist. Erst recht unmöglich ist es, mit dem dann verbliebenen CL- Teil inhaltliche Anforderungen zu verknüpfen. (Ähnliche Bedenken haben in Regensburg dazu geführt, daß das sprachwissenschaftliche Fach

mit Linguistischer Informationswissenschaft ausschließlich als 1. oder 2. Hauptfach (im Zwei-Fächer-Typ) studiert werden kann.) Es ist anzumerken, daß der Diplom-Studiengang Informatik mit linguistischem Schwerpunkt in Koblenz hier außer Betracht bleiben kann, weil das Schwerpunktfach (mit 1/3 der Studienzeit) nicht den üblichen Nebenfächern in solchen Studiengängen entspricht.

Kapitel 7 stellte die Verfasser vor die Schwierigkeit, eine allgemeine Aussage treffen zu müssen, die mit den verschiedenen Auffassungen und Auslegungen des Hochschulrahmengesetzes der Bundesländer nicht kollidert. Deshalb wurde zunächst nur grundsätzlich zwischen den zwei Formen der akademischen Weiterbildung unterschieden:

- den Weiterbildungsstudiengängen und
- den übrigen weiterbildenden Studien (Kontaktstudien) .

Im Unterschied zu Kontaktstudien sind Weiterbildungsstudiengänge curricular aufgebaut und können durch Studienordnungen geregelt werden. Sie werden gemäß einer Prüfungsordnung mit einer Hochschulprüfung abgeschlossen, wobei je nach Typ ein Zertifikat oder ein Diplom- Zeugnis oder auch ein Promotionszeugnis vergeben wird. Als Studienvoraussetzung fordern Weiterbildungsstudiengänge vom Bewerber oder der Bewerberin in der Regel einen ersten berufsqualifizierenden Abschluß eines Hochschulstudiums (wie Magister, Diplom oder Staatsexamen). Dementsprechend sind Weiterbildungsstudiengänge dazu gedacht, das bereits erworbene Wissen der Hochschulabsolventen/innen zu vertiefen, durch Ergänzung weiterer Inhalte zu verbreitern oder auch in der Ausrichtung auf einen späteren Beruf besondere berufliche Qualifikationen zusätzlich zu vermitteln. (In einigen Auslegungen ist davon die Rede, daß sie geschaffen wurden, um Studiengänge, die zum ersten berufsqualifizierenden Abschluß führen, zu entlasten.) Weiterbildende Studiengänge werden in der Regel unmittelbar an das vorige Studium anschließen, sie setzen die ganztägige Präsenz des/der Studierenden voraus und sind nicht berufsbegleitend - wie etwa Abendstudien es wären. Manche Bundesländer erlauben weitere Bedingungen für die Zulassung zu solchen Studien, die die Hochschulen selbst für ihre Weiterbildungsstudiengänge festsetzen; andere Bundesländer lassen ausdrücklich die Möglichkeit offen, daß auch Bewerber/innen zugelassen werden, die ersatzweise rur den Hochschulabschluß eine vergleichbare (mehrjährige) Berufstätigkeit ("in verantwortlicher Position") vorweisen können. Jedenfalls aber ist der Zugang zu einem weiterbildenden Studiengang restriktiver gehandhabt als der zu einzelnen weiterbildenden Lehrveranstaltungen. Die Studierenden in einem Weiterbildungsstudium werden teilweise explizit als ordentliche Studierende bezeichnet - im Unterschied zu Gasthörern.

Von Studiengängen, die zu einem ersten berufsqualifzierenden Abschluß fuhren, unterscheiden sich die Weiterbildungsstudiengänge vor allem in zweierlei Hinsicht: Sie sind nicht von den generellen Zulassungsbeschränkungen, der Vergabe durch die ZVS, betroffen, und sie unterliegen nicht dem rechtlichen Rahmen, der mit der Regelstudienzeit gesetzt wird. Allerdings gelten auch für sie die Zeitvorgaben der Bundesländer, die regeln, daß 2 Semester nicht unterschritten und 4 Semester nicht überschritten werden sollen. Einige Bundesländer geben weitere Empfehlungen (z. T. in Erlassen), so z.B. daß Weiterbildungsstudiengänge nur solche Studieninhalte anbieten sollen, die zu dem vorangegangenen Studium bestimmte Bezüge aufweisen - jedenfalls soll ein Erststudium nicht durch beliebige Weiterbildungsstudiengänge ergänzt werden. Das bedeutet sowohl, daß sich ein Weiterbildungsstudiengang an Adressatenkreise mit bestimmter Vorbildung zu richten hat, als auch, daß bestimmte Ausbildungsziele verfolgt werden, die zur beruflichen Tätigkeit, etwa der eines Spezialisten führen (z.B. Experte/in für Tropenwasserwirtschaft). Naheliegend ist, daß Kultusministerien ein Interesse daran zeigen, bestimmte weiterbildende Studiengänge nur an solchen Universitäten einzurichten, die diese Studien in ihr bereits laufendes Lehrangebot grundständiger Studiengänge und ihre jeweilige Forschungsarbeit einbeziehen können. Diese Universitäten haben ihre Kompetenz für das in Frage stehende Fach ja bereits durch ihre grundständigen Studiengänge unter Beweis gestellt. - Mit "grundständigen" Studiengängen sind offensichtlich etablierte Studiengänge gemeint, die zu einem ersten berufsqualifizierenden Abschluß fuhren; ausdrücklich werden Studiengänge ausgenommen, die sich selbst erst im Aufbau befinden. - Diese Forderung kann man akzeptieren, auch wenn man die damit verknüpfte Überlegung ablehnt, daß auf diese Weise "kein zusätzlicher Sachmittel- und Personalbedarf entsteht". Zum Zweck einer fachgemäßen und bezüglich der Tätigkeitsfelder sinnvollen Ausbildung ist eine Absprache derer unbedingt ratsam, die weiterbildende Studiengänge im CL-Bereich eingerichtet haben, einrichten oder planen. Für das CL-Fach sollte das nicht nur auf Länderebene, sondern auf Bundesebene gelten.

3 Schlußbemerkung:

Die vorstehenden Zusatz-Informationen wurden gesammelt aus Gesetzestexten und Erlassen oder

Empfehlungen der Bundesländer sowie aus dem Hochschulrahmengesetz und Kommentaren dazu. Die gesamte Sachlage konnte hier nicht ausgeführt werden, was schon deswegen auch nicht zweckmäßig gewesen wäre, weil die Bundesländer über einzelne Fragen teils unentschieden sind, teils in Diskussion stehen. Für die Leserinnen und Leser, die sich genau informieren möchten, sind folgende Hinweise:

- zu "erster berufsqualifizierender Abschluß", "Regelstudienzeit": bes. §18(1), §10(1)-(4) HRG, nach Änderung durch das 3. Gesetz vom 14.11.1985 (in der Fassung der Bekanntmachung vom 9.4.1987 (BGBL.I, S.1170));

- zur Unterscheidung von "Weiterbildung" und "Aufbaustudium": §10(3) HRG;

- zu "Zusatz-, Ergänzungs- und Aufbaustudien": §10(5) HRG;

- zu "weiterbildendes Studium": §21 HRG;

Weiterhin zu "Weiterbildung und weiterbildenden Studiengängen, insbes. zum unterschiedlichen Gebrauch von "Aufbaustudiengang", "Ergänzungsstudiengang" und "Zusatzstudiengang":

- Baden-Württembergisches Universitätengesetz (30.10.1987): §48 und amtliche Begründung;
- Bayerisches Hochschulgesetz (19.2.1988): Art.50, anders im Entwurf für eine Gesetzesänderung;
- Berliner Hochschulgesetz (13.11.1986): §§25, 26;
- Hamburgisches Hochschulgesetz (22.5.1978): §§50, 51;
- Hessisches Hochschulgesetz (Okt.1987 §§48, 49;
- Niedersächsischer Hochschulgesamtplan: S. 74f.;
- Gesetz über die Wissenschaftlichen Hochschulen des Landes Nordrhein- Westfalen (20.11.1987): §§84, 87, 89 und Erlasse;
- Landesgesetz über die Wissenschaftlichen Hochschulen in Rheinland-Pfalz (9.9.1987): §§18, 31;
- Gesetz über die Hochschulen im Lande Schleswig-Holstein (1.5.1987): §§85a, 78;
- aus dem Saarland liegen mir nur schriftliche, aus Bremen z. Zt. nur mündliche Stellungnahmen vor. (Angegebene Daten beziehen sich auf die mir von den Kultusbehörden angegebenen Fassungen.)

AK Ausbildung u. Berufsperspektiven M. Lutz-Hensel LDV-

Kontaktadressen der GLDV-Arbeitskreise

AK "Textanalyse": *Prof Dr. B. Endres-Niggemeyer*, FH Hannover, FB BID, Hanomagstr. 8, D-3000 Hannover 91

AK "Spracherkennung, Sprachgenerierung *K.* G. *Schweisthal*, Institut für Phonetik und sprachliche Kommunkation der Universität München, Schellingstraße 3 V6, D-8000 München 40

AK "Maschinelle Übersetzung": *Dr. D. Rösner*, FAW, Postfach 2060, D-7900 Ulm

AK "Ausbildung und Berufsperspektiven": *M. Lutz-Hense1*, Institut für Kommunikationsforschung und Phonetik, Poppelsdorfer Allee 47, D-5300 Bonn 1

AK "Lexikographie": *J. Brustkern*, Institut für Angewandte Kommunikationsforschung und Phonetik, Poppelsdorfer Allee 47, D-5300 Bonn 1

A new series from Mouton de Gruyter

Cognitive Linguistics and Natural Language Processing

CALL FOR CONTRIBUTIONS

Managing editor: *Annely Rothkegel, Saarbrücken*
Editorial board: Hans-Jürgen Eikmeyer (Bielefeld), Maurice Gross (Paris), Walter von Hahn (Hamburg), James Kilbury (Düsseldorf), Bente Maegaard (Kopenhagen), Dieter Metzing (Bielefeld), Makoto Nagao (Kyoto), Helmut Schnelle (Bochum), Harold Somers (Manchester), Hans Uszkoreit (Saarbrücken), Antonio Zampolli (Pisa).

Consulting board: Christian Boitet (Grenoble), Lawrence Danlos (Paris), Giacomo Ferrari (Pisa), Dafydd Gibbon (Bielefeld), Christopher Habel (Hamburg), Christa Hauenschild (Berlin), Anna S. Hein (Göteborg), Ursula Klenk (Göttingen), Manfred Pinkal (Hamburg), St. G. Pulman (Cambridge, GB), Burkhart Rieger (Trier), Jun-ichi Tsujii (Kyoto).

The new series will serve to propagate theoretical linguistic investigations of language processes within the framework of cognitive models and computer implementations. To ensure linguistics the role it deserves in the domains of natural language understanding and linguistic aspects of artificial intelligence and cognitive science, it is preferable that such investigations regard language in terms of its function in the structure of discourse.

The series will be an international forum that offers studies representing recent research in Europe and seeks to further communication not only within the whole of Europe, but also with the United States, Japan, and other parts of the world. The studies will therefore appear in English.

Studies covering topics concerned with language processes are of primary interest. In relation to language use, the traditional domains such as syntax, morphology, and phonology as well as new developments in semantics and pragmatics will be focussed on. Within this domain, problems connected with the lexicon and the analysis of generation of sentences, texts, and dialogues are of interest. Results in these fields can be related to machine translation, question-answering, queries, speech production, text processing (analysis and generation), abstracting, information retrieval, and man-machine-communication in natural language.

The books of this series are meant to contribute to the theoretical foundation which is indispensable for the computer implementation of natural language systems. All theoretical approaches dealing with language processes are welcome. Indeed, the interdisciplinary orientation demands a broad theoretical spectrum. Restrictions are put, however, on the methods to be adopted in the series. Formal methods both constitute a common basis for the interdisciplinary aims and foster the implementation of the transferable results. The possibility of implementation must be ensured. On the other hand the models must be based on empirical studies. Emphasis will be given to linguistic investigations of actual texts and dialogues.

Inquiries regarding the submission of manuscripts can be sent to the Publisher or directly to:

Dr. Annely Rothkegel
Universität Saarbrücken
Computerlinguistik (5.5), Bau 4
D-6600 Saarbrücken
Federal Republic of Germany

mouton de gruyter
Berlin · New York · Amsterdam

Wörterbücher in sprachverarbeitenden Systemen, speziell maschineller Übersetzung

Ulrich Heid
Universität Stuttgart, Institut für masch. Sprachverarbeitung

Der Arbeitskreis Lexikographie der GLDV hat am 15. Juni 1988 ein Arbeitstreffen über Wörterbücher in sprachverarbeitenden Systemen, speziell maschineller Übersetzung veranstaltet.

Das vom Institut für maschinelle Sprachverarbeitung der Universität Stuttgart *(U. Heid)*, dem Wissenschaftlichen Zentrum Heidelberg der IBM Deutschland GmbH *(B. Barnett)* und EUROTRA-D, Saarbrücken *(U. Reuther, T.C. Gerhardt)* gemeinsam organisierte und sehr gut besuchte Treffen gab in drei Themenbereichen einen Überblick über derzeit laufende Forschungsarbeiten:

- Wörterbücher in maschinellen Übersetzungssystemen: allgemeine Fragen und Beispiele
  - Heinz-Dirk Luckhardt (Saarbrücken): Qualitative und quantitative Probleme des Lexikons in der maschinellen Übersetzung
  - Ursula Reuther und Erich Steiner (Saarbrücken): Über Verbeinträge in EUROTRA Transferwörterbüchern
  - Stephan Busemann und Christa Hauenschild (Berlin): Lexikalisches Wissen im Berliner GPSG-System
  - Susan Warwick (Genève): Grundlagen zum Aufbau des Wörterbuchs in einem experimentellen System
  - Wolfgang Hoeppner (Hamburg/Koblenz): Konzem oder Lexept? — Das ist nicht die Frage.
- Deskriptive Probleme aus der Sicht der Wörterbücher für maschinelle Sprachverarbeitungssysteme
  - Stephan Mehl (Koblenz): Dynamische semantische Netze
  - Ulrich Heid und Sybille Raab (Stuttgart): Kollokationen im Wörterbuch für multilinguale Generierung und maschinelle Übersetzung
  - Dieter Seelbach (Mainz/Paris): 'Noms composés' und mehrgliedrige Nominalgruppen in französischen Fachtexten
  - Folker Caroli (Bochum/Saarbrücken): Probleme der Klassifikation deutscher und französischer Lokalisierungsverben im Hinblick auf die Erstellung zweisprachiger Lexika für sprachverarbeitende Systeme
- Datenbanken und maschinelle Übersetzung
  - Lothar Lemnitzer (Heidelberg): Datenbanken im lexikalischen Prozeß
  - K.-H. Grigo und H.-J. Stellbrink (Essen): Das Ruhrgas-Terminologiedatenbanksystem XLT-1

Im Folgenden werden kurze Zusammenfassungen der Vorträge abgedruckt, die einen groben Eindruck von den behandelten Fragen vermitteln sollen. Die Autoren sind zum Teil dabei, an den Themen weiterzuarbeiten, so daß ausführlichere Arbeiten zu diesen Bereichen zu erwarten sind. Ebenso können Interessenten die Materialien zugänglich gemacht werden, die von den Autoren zur Vorbereitung des Treffens zusammengestellt wurden und allen Teilnehmern zugegangen sind.

Die zitierte Literatur ist am Schluß zusammengestellt.

1 Qualitative und quantitative Probleme des Lexikons in der maschinellen Übersetzung — Ein Fallbeispiel: Kompositaübersetzung – (Heinz-Dirk Luckhardt, Saarbrücken)

Bei der Konzipierung, Realisierung und Nutzung von Lexika in der maschinellen Übersetzung (MÜ) spielen zwei Sichtweisen eine Rolle:

- Die qualitative Sicht
  - Inhalt
  - Struktur
  - Rolle im Gesamtsystem
- Die quantitative Sicht (Bewältigung großer Datenmengen und daraus resultierende Fragen)

Beide Sichtweisen werden anhand der Rolle der Lexika bei der automatischen Kompositaübersetzung im Rahmen des Saarbrücker Translationsservice STS exemplifiziert. Zunächst einige Thesen:

1.1 Qualitative und Quantitative Aspekte

1.1.1 Die qualitative Sicht

- Lexika in der MÜlassen sich nur in ihrer jeweiligen Funktion für Analyse, Transfer oder Synthese verstehen, konzipieren und implementieren.
- Die Unterschiede zwischen der MÜund der intellektuellen Übersetzung sind auch und gerade im lexikalischen Bereich sehr groß.
- Die linguistische Grundlegung des MÜ–Systems wirkt sich auf die Formulierung der Regeln / Einträge und die Möglichkeiten der Vereindeutigung ambiger Begriffe aus.
- Die Komplexität der Einträge hängt ab von der Aufgabenverteilung zwischen Analyse, Transfer und Synthese, der Definition der Schnittstellen zwischen diesen Komponenten sowie der Konzipierung des MÜ-Systems als bi - oder multilingual.

1.1.2 Die quantitative Sicht

- Die Beschäftigung mit kleinen Lexika bereitet in keiner Weise auf die Probleme vor, die beim Einsatz großer Lexika im Rahmen der angewandten MÜvon Sachtexten auftreten.
- Die beim Erstellen großer Lexika neu auftretenden Probleme können änderungen oder Umgewichtungen auf der konzeptionellen Ebene erforderlich machen.
- Das Schwergewicht bei der Disambiguierung von Lesarten bzw. bei der automatischen Auswahl zielsprachlicher äquivalente geht von linguistischen Kriterien auf praktische Über (Benutzerpräferenz, Fachgebietscodes).

1.2 Die automatische Übersetzung von Komposita

Im Rahmen des Saarbrücker Translationsservice STS wurden bis Mitte 1988 ca. 1. Mio. laufende Wörter (Titel und Abstracts) 'computergestützt' üb ersetzt. Dabei erwies sich die Analyse und Übersetzung der äußerst zahlreichen nicht lexikalisierten Komposita als gravierendes Problem, für das die Computerlinguistik keine umfassende Lösung parat hat. In das eingesetzte MÜ-System SUSYSTS sind verschiedene Verfahren zur Kompositaanalyse und -übersetzung integriert, die anhand der laufenden Übersetzungen getestet und auf Schwachstellen untersucht wurden. Im Referat wurden verschiedene daraus resultierende Problemklassen unter dem Aspekt der Lexikalisierung von Komposita oder Simplizia (und von Informationen dazu) diskutiert. Als besonders kritisch erwies sich die Frage, ob überhaupt Komposita in Lexika aufgenommen werden und - wenn ja - in welchen Systemkomponenten (Analyse?, Transfer?). Außerdem bereitet natürlich auch hier das Problem der Disambiguierung Kopfzerbrechen (aber: vor allem von Nomina, wohingegen sich die Computerlinguistik am liebsten mit Verben beschäftigt). Insgesamt wurden folgende Probleme angesprochen:

- qualitative Sicht:
  - Lexikalisierung von komplexen Konzepten
  - Steuerung der Auswahl von Übersetzungsäquivalenten (Syntax, Semantik, praxisbezogene Kriterien)
  - Tiefe der syntaktisch-semantischen Beschreibung
  - lokale vs. globale Lösungen
  - Diskrepanz MÜ/intellektuelle Übersetzung

- quantitative Sicht
  - Bewältigung von großen Datenmengen
  - pos. und neg. Auswirkungen großer Lexika: Konsistenz, Deutigkeitenexplosion
  - Änderungen/Umgewichtungen auf konzeptioeneller Ebene (viel mehr Nomen- als Verbregeln erforderlich, Höhergewichtung praktischer gegenüber linguistischen Kriterien)
  - in der Praxis: Akzeptieren von 90%-Lösungen?
  - der ökonomische Aspekt (v.a. Arbeitszeit hochqualifizierter Mitarbeiter)
  - irreguläre Bildungen in laufenden Texten

2 Über Verbeinträge in EUROTRA-Transferwörterbüchern (Ursula Reuther und Erich Steiner, Saarbrücken)

Das Grundmodell eines Transferwörterbuchs in der Maschinellen Übersetzung ist zunächst einmal ein herkömmliches zweisprachiges Wörterbuch. Ein solches gibt an, welches alle möglichen Übersetzungen einer gegebenen ausgangssprachlichen Lesart sind. Nimmt man nun an, dass solche zweisprachigen Transferwörterbücher maschinenlesbar zur Verfügung stehen, so ergibt sich, wenn man ohne Einschränkungen vom Wortlaut ausgeht, ein VIELE ⟶ VIELE Transfer auf der Ebene des Wortlautes. In der maschinellen Übersetzung hat man für Analyse und Synthese Wörterbücher zu erstellen, in denen Lesarten von Einträgen nach bestimmten Kriterien unterschieden werden. Zu diesen Kriterien gehören syntaktische (Funktion und/oder Struktur von valenzgebundenen Gliedern) und semantische (semantische Merkmale, semantische Rollen). Erstellt man auf der Grundlage des Wissens in traditionellen zweisprachigen Wörterbüchern Transferwörterbücher mit den selben Kodierungen wie in Analyse- und Synthesewörterbüchern, so kann man die Zahl der möglichen Übersetzungen eines Lexems in sinnvoller Weise einschränken, indem man Übereinstimmung von Information verlangt.

Im folgenden soll nun näher auf die semantischen Informationen bei Verbeinträgen eigegangen werden, insbesondere auf semantische Rollen, die den Argumenten eines Verbs zugewiesen werden. Anhand eines Experiments (vgl. [ECKERT/HEID 1988]) zeigte sich, dass diese bei der Zuordnung von Übersetzungsäquivalenten im Transfer durchaus ein Selektionskriterium sein können. Im Verlauf des Experiments bestätigte sich die Annahme, daß für jedes System Semantischer Relationen sowie für jedes System zum lexikalischen Transfer die Phänomene Modale Hilfsverben, Halbhilfsverben ("raising") und Kontrollverben von theoretischem und praktischem Interesse sind. Ihre erfolgreiche und theoretisch motivierte Behandlung erscheint als unverzichtbare Voraussetzung für Transfer genauso wie für Analyse und Generierung. Deshalb sollen nun diese Phänomene im Zusammenhang mit den Semantischen Relationen bzw. Rollen kurz erläutert und im gegenwärtigen Formalismus von EUROTRA (vgl. EUROTRA Reference Manual 5.0) dargestellt werden.

2.1 Modale Hilfsverben

Im Moment werden nur deontische Verben bzw. deontische Lesarten als "governers" auf IS (=Interface Structur; die Ebene auf der der Transfer aufsetzt) dargestellt. Deshalb wird bei einem Satz wie *This may be true may* als feature repräsentiert und *be* durch folgenden Eintrag abgedeckt:

```
be = {lu=be,cat=v,ers_frame=subj_attrsubj,
      is_frame= arg1arg2,arg1=attribuant,arg2=classifier,
      role=gov}.
```

Die Werte des Attributs "arg_n" stellen die Semantischen Rollen dar; die Verwendung von "arg_n" ist nur deshalb notwendig, weil zur Zeit innerhalb EUROTRAs Kompatibilität zwischen zwei verschiedenen Ansätzen auf dem Gebiet der Semantischen Relationen gewahrt werden muß.

Für deontische Lesarten wie in *Every employee must turn up at 8 o'clock* sind Einträge wie der folgende notwendig:

```
must = {lu=must,cat=v,ers_frame=
        subj_xcomp,is_frame=arg1arg2,
        arg1=oblique, arg2=process,
        sctrl=y,role=gov}.
```

Must wird also wie ein Subjektskontrollverb behandelt. Es weist zwei SRs zu. Das Verb *turn up* im Konstituentensatz weist seinerseits eigene SRs zu. Durch das feature "sctrl=y" wird das leere Element, das auf der der IS vorangestellten Ebene (ERS) erzeugt wurde und das für das Subjekt des eingebetteten Satzes steht, lexikalisch gefüllt (vgl. auch Baumstrukturen in folgendem Abschnitt.)

2.2 Halbhilfsverben

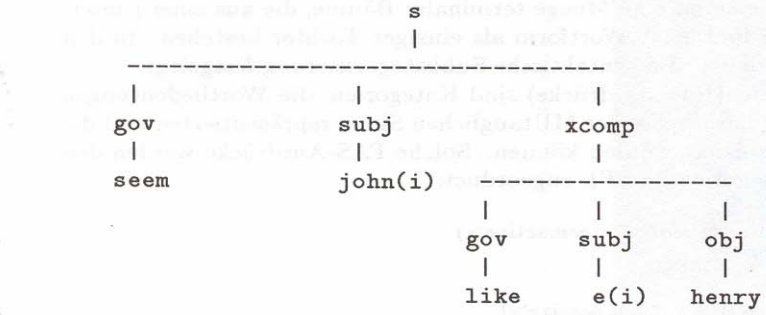
Beispiel: *John seems to like Henry.* Verbeintrag auf ERS:

```
seem = {cat=v,lex=seems,lu=seem,nb=sing,
       pers=3, tns=tensed,tense=pres,
       aux=no,ers_frame= subj_xcomp,
       rais=y,role=gov}.
```

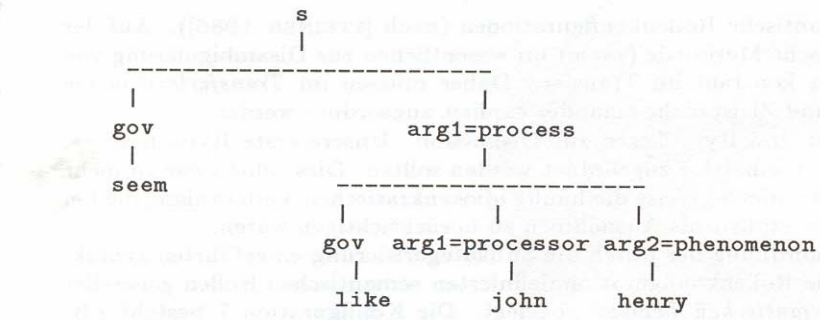
Verbeintrag auf IS:

```
seem = {lu=seem,cat=v,ers_frame=subj_xcomp,
        is_frame=arg1,arg1=process,role=gov}.
```

ERS-Repräsentation:



IS-Repräsentation:



Das Halbhilfsverb *seem* (raising) weist nur seinem propositionalen Argument die Platzhalter-SR PROCESS zu. In der ERS-IS Übersetzung wird das Subjekt von *seem* an seinen Platz als Subjekt des Konstituentensatzes zurückbewegt und erhält auch seine SR durch das Verb des Konstituentensatzes. Dies geschieht durch eine T-Regel folgender Form:

```
S: {cat=s} [GOV1: {cat=v,rais=y},
            SUBJ: {cat=np,sf=subj},
            XCOMP:{cat=s,sf=xcomp}
                  [GOV2:{cat=v},
                    ~ :{cat=np,sf=subj,np_type=e},
                   OBJ:{cat=np,sf=obj}]]
=> S <GOV1,XCOMP<GOV2,SUBJ,OBJ>>.
```

2.2.1 Kontrollverben

Die Behandlung von Kontrollverben ist ähnlich wie die bei "raising". Nur wird in der ERS-IS Übersetzung im Matrixsatz für das kontrollierende Subjekt und manchmal auch für das Objekt ein leeres Element mit der vom Verb des Matrixsatzes zugewiesenen Rolle bewahrt. Das raising Verb dagegen weist seinem Oberflächensubjekt keine Rolle zu.

3 Lexikalisches Wissen im Berliner GPSG-System (Stephan Busemann und Christa Hauenschild, Berlin)

Mit dem Berliner GPSG-Systeml ist maschinelle Übersetzung zwischen Deutsch und Englisch angestrebt. Das zugrundeliegende Übersetzungsmodell sieht Transfer u.a. auf einer oberflächennahen semantischen Ebene (Funktor-Argument-Strukturen; kurz FAS) vor. Die Abbildung in die und aus der Repräsentationssprache der F AS erfolgt im Rahmen der Theorie der generalisierten Phrasenstrukturgrammatiken (GPSG).

Für jede Sprache wird lexikalisches Wissen benötigt, das für Parsing und Generierung verwendet werden kann. Aus ökonomischen Gründen wird im Berliner GPSG-System ein Stammformenlexikon (SFL) verwendet. Das lexikalische Wissen gliedert sich wie folgt.

Für Lemmatisierung und Wortflexion ist im *morphologischen Lexikon* Information über Umlautung, Ablautung, Flexionsklassen und Partizipien zu explizieren.

Das für GPSG relevante lexikalische Wissen ist eine Menge terminaler Bäume, die aus einer (unterspezifizierten) GPSG-Kategorie als Mutter und einer Wortform als einziger Tochter bestehen. In den Kategorien wird durch Merkmal- Wert-Paare u.a. die syntaktische Subkategorisierung festgelegt.

Lexikalische Elemente auf der FAS-Ebene (Basisausdrücke) sind Kategorien, die Wortbedeutungen in einem eingeschränkten, für die Disambiguierung bei der MU tauglichen Sinne repräsentierten und die mit (semantischen) Merkmalen weiter spezifiziert werden können. Solche FAS-Ausdrücke werden den Wortstämmen mit ihren syntaktischen Kategorien im SFL zugeordnet:

(1) verabschied V *{<sub:6>}*
pred(*<conf.7>*, <psem:action>, *<gen:active»*

(1') Der Rat verabschiedet den Vorschlag.

(2) verabschied V{ *<sub:7>,<vf.pas>}*
pred(*<conf.7>,<psem:action>,<gen:passive»*

(2') Der Vorschlag wird vom Rat verabschiedet.

(3) verabschied - V{ *<sub:5>,<vf.pas>}*
- pred(*<conf.3>,<psem:action>,<gen:passive»*

(3') Der Vorschlag wird verabschiedet.

Das Merkmal *conf* beschreibt semantische Rollenkonfigurationen (nach [STEINER 1986]). Auf der F AS-Ebene dienen Rollen wie semantische Merkmale *(psem)* im wesentlichen zur Disambiguierung von Lesarten; sie sind jedoch nicht immer konstant im Transfer. Daher müssen im *Transferlexikon* die Rollenkonfigurationen für Ausgangs- und Zielsprache einander explizit zugeordnet werden.

Vor diesem Hintergrund stellen wir drei Hypothesen zur Diskussion. Unsere erste Hypothese ist, daß *sub* und conf lexemspezifisch im SFL einander zugeordnet werden sollten. Dies führt zwar zu mehr Redundanz im Lexikon, erfaßt aber auf einfache Weise die häufig idiosynkratischen Verhältnisse, die bei allgemeineren Zuordnungsmechanismen explizit als Ausnahmen zu berücksichtigen wären.

Zweitens scheint es möglich, die Zuordnung der durch die Subkategorisierung eingeführten syntaktischen Konstituenten zu den durch die Rollenkonfiguration definierten semantischen Rollen genereller zu halten; sie wird im *syntaktisch-semantischen Lexikon* abgelegt: Die Konfiguration 7 besteht z.B. aus AGENT und AFFECTED, die im Aktivfall bei allen Verben mit *sub:6* als NP{nom} und NP{acc} ausgedrückt wird. Somit kann bei der Zuordnung *'sub:6* zu *conf7'* (1) generell AGENT zu NP{nom} und AFFECTED zu NP{acc} zugeordnet werden. Im Passiv (2) liegt *sub:7* vor: bei *'sub:7* zu *conf7'* wird AGENT zu PP{von} und AFFECTED zu NP{nom} zugeordnet.

Anhand der Subkategorisierung des Verbs findet der Parser die entsprechende Rollenkonfiguration und kennt aufgrund der sub-conf-Relation die Rollennamen der zu analysierenden Konstituenten. Die Generierung kann ausgehend von F AS-Strukturen mit semantischen Relationen anhand des Verbgenus und der möglichen conf-sub-Zuordnungen der zu verwendenden Lexeme die syntaktischen Realisierungen (insbesondere Oberflächenkasus) finden.

Unsere dritte Hypothese besagt, daß alle Konstituenten und Rollen obligatorisch sein sollen. Auf der GPSG-Ebene spricht dafür, daß freie Ergänzungen eben nicht beliebig frei sind. Auf der FAS-Ebene ist

1 z.B. [BUSEMANN/HAUENSCHILD 1988], [HAUENSCHlLD/BUSEMANN 1988].

eine feste Stelligkeit für die aus theoretischen Gründen erwünschte Abbildbarkeit von FASen in intensionale Logik (und damit in ein Modell) nützlich. Schließlich ist die Zulassung optionaler Konstituenten für fakultative Mitspieler bei gleichzeitiger fester Stelligkeit von Rollenkonfigurationen problematisch mit der beschriebenen Relation zwischen *sub* und *conf,* weil mehr als ein Wert von *conf möglich* wird.

Diese sehr kurzen Thesen sind diskutabel, wenn nicht sogar fragwürdig. Wir freuen uns über Hinweise und Kritik und sind gerne zu einer ausführlicheren Diskussion bereit.

4 Building Up a Lexicon for NLP Applications (Susan Warwick, Genf)

This talk discussed criteria for building up a lexicon for NLP applications. The work in defining the lexical entries takes as a starting point that the necessary information cannot be defined apriori and therefore emphasis is placed on the definition language of the lexical data and the tools appropriate for experimentation. The sample lexical entries presented exemplify the concern that the information be concise and readable (for humans), interpretable (for a program), as well as modifiable and re-usable (as necessary for a given application and according to given linguistic concerns). The work described is situated in a project at ISSCO to develop tools for NLP applications (a unification based system) and to apply them in building a prototype Machine Translation system. The program which provides facilities for expanding these definitions into full lexical entries and which uses them in parsing natural language texts was demonstrated on grammar fragments of German and French.

An initial concern in designing the lexical entries was to define the information in such a way that it was not committed to a specific linguistic theory. The information was to be used for different grammars; in particular, one which maps texts to a semantic representation via a functional structure (LFG) and another which maps text directly to a situation schemata. It is hoped that the lexical entries can also be used in future experiment at ion with other linguistic theories.

Given the assumption that the entries will need modification and additions, another consideration was to abstract away from the specific values used by a given grammar (represented as attribute value pairs). By means of "lexical procedures", an integral part of the programming environment, the information is mapped into feature names and values used by the grammar. The "procedures" or expansion rules are defined by the linguist and can be modified for experimental purposes without changing the basic lexicon. Adding information to the basic lexicon implies adding new lexical expansion rules to interpret the information; modifications imply changing parts of the "procedures" or replacing so me entirely.

Since building a lexicon to be used in a computational application is a large and time-consuming task, the work must be defined such tImt a number of collaborators can work together. A further concern was therefore that the information be specified in a formalism that could be understood by a collaborator with perhaps little linguistic training to allow coding to continue somewhat independently of grammar development.

In light of these considerations, the strategy for defining lexical entries consists of two levels: an abstract definition of the properties of a given lexical item (or class of items) and lexical procedures which expand this information into full lexical entries according to the needs of a specific grammar.

The following sample entry of the German verb "warten" (to wait) is an illustration of the abstract definition which serves as the basis for making full lexical entries.

```
warten        v        !Subcat(np(nom),pp(auf,acc))
```

The entry can be read as "the string *warten* has category v(erb) and subcategorizes for a "nominative" NP and for a PP whose preposition is *auf* and which takes an "accusative" NP. An underlying assumption is that semantically the verb has two arguments corresponding to the two elements in the parentheses. A morphological component (coded as a finite state automata) associates the various forms of the verb to the base form.

The expansion of the lexical item is defined by the clause "!Subcat(00')'" the definition of which will vary according to its use in a given grammar. The programming language available to the linguist provides the means of stating generalizations over classes of lexical items in a modular and declarative way. The expansion introduced by! may have further expansions, ego "!Active(oo.)" and "!Passive(oo.)". The example given above could have many definitions for regularities which might be considered lexical, passing the information in the base entry as arguments to the procedures. The program allows passing of arguments as constants and variables as well as the use of primitive data types such as lists and terms.

The basic definition is neutral with respect to specific function names such *as* "subject"; these are defined in the "procedures" written by the linguist. For example, on the basis of the information *np(nom),* the verb could be assigned an active reading with a "subject" of category "np" and *case* "nominative", these names being assigned according to their use in a given grammar. An open quest ion is how much information can be stated in such a concise and abstract manner.

The basic definition is also not biased towards a specific application. The same information could be used to generate definitions for "human" dictionaries. In this case the information in the entry for warten would be expanded into a phrase *as* found in printed dictionaries, ego

jmd wartet auf etw (someone is waiting for something)

As the example suggests, the entry might need additional information to choose between "sorneone" and "something" . A feature such *as +hum* could be added to the definition given above, i.e. "np(nom,hum)".

This formalism and the tools provided for expressing the lexical information offer the power and flexibility necessary for building a lexicon for use in computational applications. The language adopted for coding the lexical data meet the criteria of simplicity, clarity and readability. The programming language provides the means of stating the information in a modular way and therefore a means of controlling the information (which may be quite complex and which will certainly be added to and modified). And finally, the program provides the environment for the necessary experimentation with the lexical component in a given application.

5 **Konzern oder Lexept? - Das ist nicht die Frage.** (Wolfgang Hoeppner, Hamburg und Koblenz)

Die Frage, die in diesem Beitrag erörtert wurde, ist, ob eine autonome lexikalische Semantik möglich ist, oder ob es eines Fundamentes für eine solche bedarf. Unter lexikalischer Semantik verstehe ich hier eine intralinguistische Anstrengung, durch verschiedene Methoden (Merkmale, Prozesse) die Bedeutung sprachlicher Einheiten in den Griff zu bekommen. Meine These hierzu ist, daß dies kein sinnvolles Unterfangen ist, sondern daß vielmehr der Bezug einer linguistischen Semantikbeschreibung auf eine ontologische notwendig ist.

Dies wurde durch die Unterscheidung zwischen konzeptuellem und assertionalem Wissen begründet: es ist etwas anderes, ob ich einen bestimmten Diskursbereich modelliere, oder ob ich die Art und Weise, wie über einen solchen Diskursbereich gesprochen wird, repräsentiere. Für die erste Aufgabe wird eine konsistente, zumindest auf einer Metaebene widerspruchsfreie Repräsentation benötigt. Für die zweite Aufgabe trifft dies nicht zu: es werden tagtäglich durchaus inkonsistente, widersprüchliche Äußerungen produziert und verstanden (z.B. *diese Windel ist naß, und doch schön trocken),* für die eine Repräsentation gesucht ist. Wichtig bei diesem Ansatz ist allerdings die Beziehung zwischen den beiden Modellierungsebenen, für die - wie in der Semantik üblich - Kompositionalität gefordert wird. Hierauf wurde an Hand konkreter Systeme eingegangen (WISBER, CMU-MT-PROJEKT, TRANSLATOR, JANUS).

Ein wesentlicher Punkt, der oft als Kritik für eine formale, letztendlich beweisbare Semantikebene ins Feld geführt wird, ist die Beschränktheit der Anwendungsmöglichkeit, d.h. die Beschränktheit der formal abgesicherten Prozesse, wie z.B. den Subsumptionsalgorithmus. Dominieren hier nicht formale Eigenschaften die Bedürfnisse der inhaltlichen Modelliering, ist eine Frage, die sich einem Systementwickler bei zahllosen Gelegenheiten aufdrängt. Eine Antwort darauf ist jedoch leichten Herzens kaum mit ja oder nein formulierbar. Wer wünscht nicht, daß allgemeine Prozess auch eine allgemeine Prozeßsemantik haben, sprich vollständig, korrekt und effizient sind? Wenn man sich derartige Prozesse daraufhin ansieht, dann wird man nicht sonderlich viele finden, die diesen Ansprüchen gerecht werden, und gleichzeitig die Leistungen erbringen, die beispielsweise für sprachverstehende Systeme gebraucht werden. Das Problem der Disjunktion in terminologischen Notationen sei hier nur genannt, stellvertretend für eine Vielzahl von Problemen, für die eine formal befriedigende Lösung nicht existiert.

Das Phänomen, dem ich hier nachgehe, ist jedoch auch nicht brandneu: wie allgemein sind meine Algorithmen, welche Konzessionen kosten mich wieviel, beschäftigt die Informatik schon seit einigen Jahrzehnten und, nebenbei bemerkt, auch die formale Linguistik, (z.B. bei Redundanzregeln). In der Künstlichen Intelligenz sind diese Themen insbesondere im Zusammenhang mit der default-Debatte und Logikerweiterungen diskutiert worden, und die Tatsache, daß diese Debatten auch heute noch aktuell sind und geführt werden, läßt die Vermutung nahelegen, daß es sich hier um Grundsätzliches handelt. Aus der (lesenswerten) Unzahl der Literatur zu diesem Thema sei hier nur auf ASHER 1983 und McDERMOTT 1986 verwiesen, und dies auch nur, weil diese Arbeiten mich kürzlich beschäftigt und durch ihre Explizitheit eingenommen haben.

Als Ergebnis möchte ich hier nur zusammenfassen: ein formal fundiertes Repräsentationsmodell ist eine Basis, die einem die Lösung von Problemen, z.B. auf dem Gebiet der lexikalischen Semantik teilweise abnimmt. Zu den verschiedenen mittlerweile existierenden T- und A-Boxen sind sicherlich noch eine Reihe anderer box-ähnlicher Konstrukte zu erfinden, deren Universalität sicher nicht so welt-umfassend ist, für die es sich aber lohnt, einen maximalen Allgemeinheitsgrad anzustreben.

Für eine linguistische Semantik liegen gerade bei notorischen Problemen Lösungsmöglichkeiten vor, die insbesondere Kompositionalität in solche Bereiche re-importiert haben, die sich dieser wünschenswerten Eigenschaft bislang hartnäckig zu verschließen schienen. Im Vortrag wurde dies an den Beispielen *Nominalkomposition, Metonymie* und *propositionale Einstellungen* gezeigt.2

6 **Dynamische semantische Netze** (Stephan Mehl, Koblenz)

In seinem Vortrag 'Dynamische semantische Netze' machte STEPHAN MEHL (EWH Koblenz) den Vorschlag, den aus der Kognitiven Psychologie stammenden Repräsentationsformalismus der Aktivationsausbreitungsnetze (COLLINS/LoFTUS 1975) für die maschinelle Lexikographie fruchtbar zu machen. Mit Hilfe solcher Netze ist es insbesondere möglich, die Disambiguierung von Homonymen und Homographen durch den sprachlichen Kontext als lexikalischen Zugriffsprozeß darzustellen.

Eines der zentralen Probleme der maschinellen Lexikographie besteht darin, Informationen aufzunehmen, die bei der Analyse von Texten eine Auflösung der semantischen Mehrdeutigkeiten erlauben. Üblicherweise wird so verfahren, daß für bestimmte, festgelegte syntaktische Positionen (z.B. die obligatorischen Ergänzungen eines Verbs) Lexemklassen als Selektionsrestriktionen angegeben werden. Diese Vorgehensweise hat zwei Nachteile:

1. Häufig fällt es schwer, brauchbare Lexemklassen zu spezifizieren.
2. Oft ist es nicht möglich, ein Homonym anhand der wenigen syntaktisch festgelegten Wörter im Kontext zu disambiguieren.

Beispiel:

Das Verb *abgeben* hat eine Reihe verschiedener Bedeutungen, u. a. solche, die durch die Quasi Synonyme *freisetzen, verkaufen* sowie *ausliefern* beschrieben werden können. Die verschiedenen Lesarten lassen aber zum Teil dieselben Ergänzungen zu: *Gas abgeben* kann einerseits *Gas freisetzen,* andererseits Gas *verkaufen* bedeuten; *Blumen abgeben* kann neben *verkaufen* auch *ausliefern* heißen.

Die Angabe von Selektionsrestriktionen für die valenzgebundenen Ergänzungen reicht daher nicht aus. In einem größeren Kontext wie dem folgenden wird die Lesart hingegen eindeutig bestimmt:

> *Aus den bituminösen Schichten des Jura wird infolge der Reibung aneinandergleitender Schollen eingeschlossenes Gas freigesetzt. Entlang einer* 25 *km langen Fläche können im Jahr 1.250.000* m3 *Gas abgegeben werden.*

Für die Bedeutungsbestimmung von *abgeben* zentral sind die vorangehenden Lexeme *eingeschlossen* und *freisetzen,* die beide semantisch mit *abgeben* verknüpft sind.

Die Grundidee des im Vortrag dargestellten Verfahrens besteht nun darin, semantische Relationen zwischen Lexemen, wie sie ohnehin in traditionellen einsprachigen Wörterbüchern kodifiziert sind (z.B. Synonymie, Hyperonymie; Kollokationen), als Kanten zwischen Lexemknoten in einem Netz darzustellen, und zwar in wesentlich größerem Umfang, als dies in gedruckten Lexika möglich ist. Verschiedenen Lesarten eines Wortes werden dabei in der Regel unterschiedliche Kantenbündel zugeordnet. Durch die im (satzübergreifenden) Kontext aufgefundenen Lexeme werden dann Teile des Netzes so aktiviert, daß die mit diesen Knoten verbundene Lesart eines Homonyms anschließend ausgewählt wird.

Durch die Verwendung individueller Relationen anstelle von globalen Klassen ist es gleichzeitig möglich, das Netz bei der Textgenerierung für die Unterscheidung von Quasisynonymen einzusetzen. Trotz ihrer sehr ähnlichen Bedeutung können z.B. *ausströmen* und *ausstrahlen* nicht völlig gleich verwendet werden: so kann man Wärme *ausströmen* oder *ausstrahlen, Helligkeit* nur *ausstrahlen, Gas* nur *ausströmen.* Diese Kollokationsphänomene sind einzelsprachlich verschieden und stellen daher in der Generierungsphase der maschinellen Übersetzung ein erhebliches Problem dar.

Die Erstellung eines solchen Netzes ist allerdings wegen der Vielzahl der aufzunehmenden Kanten relativ aufwendig, selbst wenn man davon ausgeht, daß in den meisten Fällen zwischen den Lexemen eines Kontexts keine direkte Verbindung besteht, sondern erst im Zuge der Aktivationsausbreitung ein vebindender Pfad erarbeitet werden muß. Zur Zeit wird in Koblenz im Rahmen eines Dissertationsvorhabens ein Netz mit einigen hundert Lexemen exemplarisch implementiert.

2 Ein Exemplar des Workshop-Papiers kann beim Autor angefordert werden. Eine ausführlichere Version ist in Arbeit.

7 Kollokationen im Wörterbuch für multilinguale Generierung und maschinelle Übersetzung (Ulrich Heid und Sybille Raab, Stuttgart)

In einem wissensbasierten Übersetzungs- oder Generierungssystems (wie z.B. [TOMITA/CARBONELL 1986]) zerfällt die Aufgabe der sprachlichen Realisierung von in semantischen Repräsentationen ausgedrückten Sachverhalten, vereinfacht gesagt, in zwei Teile: (a) die Wahl der syntaktischen Form der zu erzeugenden sprachlichen äußerung, (b) die Wahl geeigneter Lexikalisierungen für die in der semantischen Repräsentation verwendeten Konzepte.

Aus den Problemen, die sich bei der Lexikalisierung stellen, wird hier die Frage der Behandlung von Kollokationen herausgegriffen.

7.1 Begriffsbestimmung

Kollokationen sind polare Verbindungen von zwei Lexemen; eines ist semantisch autonom (in der Terminologie von [HAUSMANN 1984, 1985: *Basis*]: es kann unabhängig vom Kontext übersetzt werden; das andere Lexem determiniert semantisch die Basis und erhält selbst seine Bedeutung nur in der gegebenen Kollokation [HAUSMANN: *Kollokator*]: Übersetzungsäquivalente von Kollokatoren können nur bezogen auf einzelne Kollokationen bestimmt werden.[3]

Welche Basen welche Kollokatoren zulassen (vgl z. B. *des prix abordables, * des taux abordables*) ist durch die einzelsprachliche Norm festgelegt.

Zusammenfassend: Wissen über Kollokationsmöglichkeiten gehört zum (einzel-)sprachlichen Wissen, das bei der Lexikalisierung von Konzepten benötigt wird. Die "Basis" bestimmt die Wahl des "Kollokators".

7.2 Beschreibung im Wörterbuch

7.2.1 Beschreibung von Kollokationen mit Selektionsrestriktionen

Kollokationen werden in traditionellen Wörterbüchern in der Regel in Listen angegeben. Es erscheint nicht sinnvoll, Kollokationen in Wörterbüchern für sprachverarbeitende Systeme ausschließlich mit Selektionsrestriktionen beschreiben zu wollen, die für ein Verb oder Adjektiv angeben, Nomina welcher semantischen Klasse (⟶ semantische Merkmale)[4] als Ergänzungen oder als Bezugswort möglich sind. Semantische Merkmale beschreiben in gewisser Weise Weltwissen. Mit den Merkmalen [± DURATIV] kann z.B. beschrieben werden:

eine Anfrage, einen Befehl, den Nachrichtenempfang **abbrechen** ([- DURATIV]) ⟶ *annuler* vs. *das Drucken, die Operation* **abbrechen** ([+ DURATIV]) ⟶ *arrêter*.

Kollokationsmöglichkeiten erklären sich aber nicht ausschließlich durch Weltwissen, so daß sich zur Beschreibung von Kollokationsklassen die semantischen Merkmale als zu wenig trennscharf erweisen; zur Disambiguierung bei der Übersetzung müßte eine große Anzahl von schlecht motivierbaren, sehr detaillierten Merkmalen eingeführt werden.

7.2.2 Beschreibung von Kollokationen mit lexikalischen Funktionen (LF)

MEL'ČUK 1984 verwendet zur Beschreibung der Relationen, die Lexeme miteinander unterhalten, das Konzept der "lexikalischen Funktion". Durch LFen, d.h. die Anwendung eines semantisch vagen Operators auf ein Lexem können neue Lexeme erzeugt werden, "einwortige" oder Mehrwortlexeme; MEL'ČUK erfaßt damit Synonymie, Antonymie, Hyponomie/Hyperonomie und Konversion ebenso wie Derivationsprozesse und Kollokationsmöglichkeiten.

Die Analyse der in [MEL'ČUK 1984] beschriebenen Kollokationsmöglichkeiten von Nomina, die Gefühle bezeichnen[5], läßt die Annahme zu, daß für den Bereich von sehr allgemeinen LFen (z. B. der Ausdruck des Stattfindens, des Anfangs, Fortgangs bzw. Endes eines Vorgangs, Intensivierung, Kausativierung

[3] Dafür nachfolgend ein Beispiel: z.B. *dresser les barricades* (⟶ *errichten*), *un budget* (⟶ *erstellen*), *un bilan* (⟶ *aufsetzen*) etc.

[4] Generell zu semantischen Merkmalen z.B.: [CRUSE 1986], Kapitel 12.2. Vgl. auch [NAGAO/TSUJII 1986], [ZELINSKY-WIBBELT 1987] etc.

[5] Das Wörterbuch von [MEL'ČUK 1987] liegt uns noch nicht vor. In [MEL'ČUK 1984] sind Gefühlsbezeichner die einzige semantisch motivierbare Klasse, zu der mehr als 10 Lexeme beschrieben sind. Nach den Prinzipien von MEL'ČUK arbeitet aber auch BETTY COHEN in einem neuen Wörterbuch der Börsensprache, das uns ebenfalls noch nicht vorliegt. Kollokationen mit Gefühlsbezeichnungen sind auch in [HAUSMANN 1984] behandelt.

etc.) Standardannahmen hinsichtlich der Lexikalisierung durch Kollokationen formuliert werden können. Einzelne Ausnahmen müssen daneben separat behandelt werden:

Sentiment	⟶	{*admiration, colère, désespoir, enthousiasme, envie étonnement, haine, joie, mépris, respect*}
Default:		
OPER (Sentiment)	⟶	*ressentir* <SUBJ OBJ> *éprouver* <SUBJ OBJ> (OBJ PRED) ∈ {Sentiment}
Lexikalisierte Ausnahme:		
OPER (*admiration, haine*)	⟶	*nourrir* <SUBJ OBJ>

7.3 Zur Verwendung des Ansatzes von Mel'čuk im Wörterbuch eines sprachverarbeitenden Systems

Die Kollokationsbeschreibung mit LFen könnte für die Erstellung von Wörterbüchern für maschinelle Übersetzung und multilinguale Generierung Vorteile gegenüber bisherigen Ansätzen bringen: [6].

Vorausgesetzt, es lassen sich LFen à la [MEL'ČUK 1984] hinreichend klar definieren[7], so kann im Konzeptlexikon jede Konzeptklasse Slots für LFen aufweisen, wie bisher Slots für Argumente und für nichtregierte Umstandsangaben[8]

Wenn es möglich ist, Defaults für die Lexikalisierung der Anwendung bestimmter LFen auf die Lexeme anzugeben, so stünde damit ein flexibles Hilfsmittel zur Verfügung, das Problem der Kollokationsmöglichkeiten bei der Lexikalisierung von in semantischen Repräsentationen ausgedrückten Sachverhalten in den Griff zu bekommen. Zukünftige Arbeiten sollen es erlauben, diese Annahme zu erhärten.

8 Noms composés und mehrgliedrige Nominalgruppen in französischen Fachtexten (Dieter Seelbach, Mainz und Paris)

"Noms composés" sind zweigliedrige nominale Ausdrücke, vor allem vom Typ N de N, Adj N und N Adj, die in Fachtexten signifikant häufig auftreten, und bei deren Analyse oft voreilige Verallgemeinerungen vorgenommen wurden, die einer Überprüfung an größeren Datenmengen nicht standhalten. Viele traditionelle Sprachwissenschaftler sind sich darin einig, daß "noms composés" semantisch definierbar seien, und zwar als aus zwei Komponenten aufgebaute (neue) Bedeutungseinheiten. Zahlreiche Derivationen und nichtabgeleitete einfache Nomina sind aber semantisch ebenfalls als Einheit von zwei Bedeutungskomponenten zu definieren:

pédiatre	Kinderarzt
"médecin pour enfants"	
verger	Obstgarten
potager	Gemüsegarten
giboulées	Schnee-, Hagel-, oder Graupelschauer
accalmie	Wetterberuhigung
bruine	Nieselregen

Sie würden damit der semantischen Definition von "noms composés" entsprechen, wie *sonate pour piano, film pour enfants, chaussures pour enfants* oder *pluies verglaçantes* etc.

Die Synonymie von "*maître d'école*" und "*instituteur*" spricht ebenso gegen eine rein semantische Definition wie die kontrastive Tatsache, daß "noms composés" auch einfache Nomina als deutsche äquivalente haben, deren semantische Komponenten nicht ohne weiteres idenfizierbar sind:

[6] In [TOMITA/CARBONELL 1986] wird für die Wortwahl bei der Generierung ein Blackboard-Modell verwendet. Im Realisierungslexikon ist bei den einzelnen Lexemen ein Slot "collocates-with" vorgesehen, in dem bevorzugte Kollokationspartner aufgelistet werden. Vgl. [NIRENBURG/NIRENBURG 1988] und [NIRENBURG/NYBERG/KENSCHAFT 1988] und jetzt auch [NIRENBURG ET AL. 1988].

[7] vgl. z. B. die LFen [OPER], [INCEP], [CONT], [FIN], [CAUS], [PERM], [LIQU], [MAGN].

[8] vgl. z.B. links die Slots, rechts die syntaktische Standardrealisierung bei der Konzeptklasse ACTION:

[AGENT]	lo	SUBJ
[START], [END], [DURATION], [WHEN]	lo	temporaler ADJUNCT
[ASSOCIATED-ACTION]	lo	temporaler Nebensatz
[FREQUENCY]	lo	ADJUNCT

chaise à porteurs	Sänfte
cheval blanc	Schimmel
coup de feu	Schuß
pétrole lampant	Petroleum
temps lourd	Schwüle
marée haute	Flut
gelée blanche	Reif
brouillard léger	Dunst

Auch die eher syntaktische Definition der "noms composés" von MARTINET als "synthèmes", deren zweites Element nicht determiniert oder modifiziert sein darf, ("ne peut recevoir de détermination particulière"), ist voreilig:

milieux bien informés	gut unterrichtete Kreise
uranium faiblement enrichi	schwach angereichertes Uranium
front extrêmement froid	extreme Kaltfront
courant très perturbé	mit starken Störungen angereicherte (Luft)strömung

Eine sehr vorsichtige Vorgehensweise und einen brauchbaren Ansatz zur Definition von "noms composés" findet man in GROSS (1986). Wir haben diesen Ansatz erweitert und auf das Vokabular der Wetterberichte angewendet. Welche Tests sind es, die es erlauben, beispielsweise *températures matinales, fin de la semaine, front froid, petit matin als* "noms composés" zu definieren und entsprechend maschinell zu identifizieren und zu Übersetzen und *températures douces, fin de la pluie, vent froid, premier matin* nicht? Betrachten wir die Tests für die Erarbeitung von Lexikoneinträgen der Klasse N Adj.

"freies" Adjektiv in der Gruppe N Adj		Adjektiv im "nom composé" vom Typ N Adj
Bsp.: températures douces	versus	températures {maximales / matinales}
a) Prädikativität (Attributstellung) les températures sont douces		keine Prädikativität (Attributstellung) *les températures sont {maximales / matinales}
b) Nominalisierung la douceur des températures		keine Nominalisierung *la {maximalité / matinalité} des températures
c) Kein Bruch im Paradigma ("freie Distribution") températures agréable, hautes, basses, variables, modérées, élevées, fraiches, tropicales, polaires etc.		Bruch im Paradigma températures diurnes, nocturnes; minimales
d) Veränderlicher Numerus une/des température(s) douce(s)		keine Numerusveränderung *la/les températures {matinales / maximales} ↔ la t. m. à Strasbourg
e) Adverbeinschub températures extrêmement douces		kein Adverbeinschub *températures extrêmement {matinales / maximales}
f) Koordinierbarkeit températures douces et agréables		keine Koordinierbarkeit mit "freien" *températures {matinales / maximales} et douces
g) Weglaßbarkeit des Nomens(-) *les douces		Weglaßbarkeit des Nomens(+) les maximales
j) Austauschbarkeit mit (N) de un Modif N		Austauschbarkeit mit de (Det) N de Ø N — de le N
*températures de douceur *températures de la douceur températures d'une certaine douceur		températures du matin
i) *températures régnant à un moment donné de la douceur		Paraphrase mit N(-Adj) als Argument ≡ Relationales Adj températures régnant {le matin / à un moment donné de la matinée}
j) kein Einwortlexem als Synonym		Synonym: Einwortlexem les maxima
k) kein Element eines 'composé de composés' *températures douces saisonnières		Element eines"composé de composés" températures maximales saisonnières

Besonders "stark" sind die Kriterien h), i) und vor allem c).

Die formalen Klassen von 'noms composés' des Französischen geben keinen Hinweis auf die Formen

der deutschen Entsprechungen:

mer houleuse	rauhe See	mer du nord	Nordsee
mer intérieur	Binnenmeer	mer d' huile	glatte See
moyenne montagne	Mittelgebirge	petit matin	früher Morgen

Es bietet sich an, auch komplexe 'noms composés' als Einheiten ins Lexikon aufzunehmen, wenn mehr als zwei Nomina oder Adjektive im Spiel sind. Hier sind neue Klassen und Subklassen zu bilden, die als 'composés de composés' oder als *Expansionen* der Klassen A N, N A, N de N, N N etc. anzusehen sind.

Expansionen von NA:

1. A N / A une jeune fille prolongée	les hautes pressions méditerranéennes
2. N A / A la politique agricole commune	des passages nuageux matinaux
3. N de N / A du fil de fer barbelé	le régime d' ouest perturbé

Expansionen von N de N:

1. A N / de N le fin mot de l' histoire	la grande partie du pays
2. N A / de N une offre publique d' achat	les températures normales de saison
3. N / de A N une heure de grande écoute	un axe de hautes pressions
4. N / de N A un incendie d'origine criminelle	des masses d' air chaud
5. N / de N de N une station de sports d' hiver	un risque de chutes de neige

Es können Fälle von Ambiguität auftauchen wie in:

petites pluies locales (A / N A oder A N / A) (Expansion von A N oder von N A)
chutes de pluies verglaçantes (N / de N A oder N de N / A)

Über die Aufnahme von 'noms composés' ins Lexikon in Abhängigkeit vom Fachgebiet sind übrigens die klassischen Fälle von Ambiguität beim Parsing aufzulösen wie in:

une maîtresse de gymnastique amoureuse
(N de N / A oder N / de N A)

Es genügt hier, *maîtresse de gymnastique* im Lexikon zu haben, um die zweite Lösung auszufiltern.
Es wurden bisher etwa 150 'noms composés' vor allem vom Typ *Adj N, N Adj* und *N de N* sowie etwa 50 'composés de composés' (dreigliedrig) unterschiedlicher Subklassen in der Wetterberichtssprache ermittelt. Ein zweisprachiges Lexikon von 'noms composés' des Französischen und Deutschen, ohne das an eine automatische Übersetzung (besonders von Fachtexten) nicht zu denken ist, wird in einem deutsch-französischen Gemeinschaftsprojekt erarbeitet. – Über mehrgliedrige Nominalgruppen mit prädikativen Nomina als Köpfen wie in "orientation du vente á l'ouest" wurde hier – aus Platzgründen – nichts gesagt.

9 Probleme der vergleichenden Klassifikation deutscher und französischer Lokalisierungsverben im Hinblick auf die Erstellung zweisprachiger Lexika für sprachverarbeitende Systeme (Folker Caroli Bochum / Saarbrücken)

In dem hier vorgestellten Projekt wird eine vergleichende Klassifikation der deutschen und französischen Lokalisierungsverben erarbeitet, mit dem Ziel, alle syntaktischen und semantischen Informationen systematisch zu erfassen, die für die eindeutige Zuordnung korrespondierender Verwendungsweisen dieser Verben in Transferwörterbüchern nötig sind.

Bei der Klassifikation folgen wir dem im Laboratoire d'Automatique Documentaire et Linguistique (L.A.D.L) entwickelten Ansatz, der sich an Harris orientiert. In diesem Ansatz werden äußerungen als Verkettungen von Operatoren und Argumenten angesehen. Argumentstellen können mit elementaren oder komplexen Argumenten besetzt sein. Komplexe Argumente sind wiederum Operatoren, die ihrerseits weitere Argumente binden. Die Struktur der äußerungen mit komplexen Argumenten können auf eine Menge von äußerungen abgebildet werden, deren Operatoren nur elementare Argumente binden. Diese Projektion kann mit Hilfe von Paraphrasen vorgenommen werden, deren metasprachlichen Bildungsmittel in der zu beschreibenden Sprache selbst enthalten sind. (vgl.[GROSS 1981], [BOONS/GUILLET/LECLÈRE 1976])

Dieses Prinzip wird bei der Lösung verschiedener Probleme der Klassifikation in beiden Sprachen in äquivalenter Weise angewendet:

9.1 Abgrenzung der verbabhängigen Lokalergänzung durch Extraktion

Eine solche mögliche Projektion stellt die Geschehns-Paraphrase dar, die in der Verbklassifikation zur Abgrenzung verbabhängiger von satzabhängigen Präpositionalergänzungen herangezogen wird. Sie hat sich als hinreichend operationales Kriterium zur eindeutigen Bestimmung verbabhängiger Lokalergänzungen erwiesen. Da sie in der Literatur eingehend diskuiert ist, gehe ich auf sie hier nicht weiter ein. (vgl. die Darstellungen bei [BIERE 1976] und [EROMS 1981])

9.2 Explikation der lokalen Relation

Ausgehend von Verben mit präpositionaler Lokalergänzung läßt sich die allgemeine Struktur der Lokalisierungsverben mit Hilfe von allgemeinen lokalen Paraphrasen bestimmen:

(3) a Max stellt das Glas auf den Tisch
a' Max pose le verre sur la table
b Das Glas ist auf dem Tisch
b' le verre est sur la table
(Zustand t2 : Nach dem Prozeß der Lokalisierung)

(4) a Max zieht eine Zigarette aus der Schachtel
a' Max sort une cigarette du paquet
b Die Zigarette ist in der Schachtel
b' la cigarette est dans le paquet
(Zustand t1 : vor dem Prozeß der Lokalisierung)

(5) a Max geht zur Schule
a' Max va à l'école
b Max ist in der Schule
b' Max est à l'école
(Zustand t2 : Nach dem Prozeß der Lokalisierung)

(6) a Max kommt aus dem Zimmer
a' Max sort de la chambre
b Max ist in dem Zimmer
b' Max est dans la chambre
(Zustand t1 : vor dem Prozeß der Lokalisierung)

Wie die Paraphrasen b und b' in den Beispielen (3) bis (6) zeigen ist gemeinsames Merkmal aller Lokalisierungsverben die lokale Relation zwischen zwei Aktanten:

$$R_{[L]} (N_{[a]}, N_{[loc]})$$

Dabei bezeichnet $N_{[a]}$ das lokalisierte Argument, $N_{[loc]}$ den Ort der Lokalisierung. Bei den transitiven Verwendungsweisen ist das lokalisierte Argument das Akkusativobjekt, bei den intransitiven das Subjekt.

Solche elemenataren Operatoren, durch die sich diese Relation explizieren läßt, sind sein im Deutschen und *être* und *il y a* im Französischen. Der dynamische Aspekt der Lokalisierung kann durch Operatoren wie kommen, gelangen bzw. *aller, passer, venir* wiedergegeben werden.

9.3 Unterschiedliche Bezeichnung der lokalen Relation

Lokalisierungsverben lassen sich danach klassifizieren, durch welche Kombination von Aktanten die lokale Relation bezeichnet wird. Bei der Annahme eines maximalen Satzrahmens der lokalen Verwendungsweise eines Verbs mit Subjekt, Akkusativobjekt bzw. complément d'objet direct und präpositionaler Ergänzung sind theoretisch folgende Kombinationen denkbar:

		Ort		
		Subjekt	Objekt Cmp. dir.	Prp. Eg.
lokal. Argum.	Subjekt	1 –	2 +	3 +
	Objekt Cmp.dir.	4 +	5 –	6 +
	Prp. Eg.	7 +	8 +	9 ?

In der Tat sind folgende Kombinationen im Deutschen, wie im Französischen durch Beispiele belegt.

2 = Der Feind besetzt die Stadt
(Der Feind gelangt in die Stadt)
La fumée envahit la pièce
(La fumée passe dans la pièce

3 = Max tritt in das Zimmer
(Max kommt in das Zimmer)
Max entre dans la chambre
(Max va dans la chambre)

4 = Der Vulkan speit glühende Asche
(Glühende Asche kommt aus dem Vulkan)
Le volcan crache de la cendre brûlante
(De la cendre brûlante vient du volcan)

6 = Max legt das Buch auf den Tisch
(Das Buch gelangt auf den Tisch)
Max pose le livre sur la table
(Le livre (va + passe) sur la table)

7 = Der Käse wimmelt von Würmern
(Die Würmer sind in dem Käse)
Le fromage grouille de vers
(Les vers sont dans le fromage)

8 = Max belädt den Lastwagen mit Kisten
(Die Kisten (gelangen + kommen) auf den Lastwagen)
Max charge le camion de caisses
(Les caisses (vont + passent) sur le camion)

Jede dieser Kombinationen konstituiert im Deutschen wie im Französischen eine Klasse von Lokalisierungsverben. Zwischen den einzelnen Klassen bestehen bestimmte Ableitungsbeziehungen, d.h., es gibt Verben, die distinkte Verwendungsweisen haben, die zwei unterschiedlichen Klassen von Lokalisierungsverben zuzuordnen sind. Diese Beziehungen sind durch die Verbindungslinien zwischen den einzelnen Kombinationen skizziert.

Im Deutschen werden dies Ableitungen häufig durch Präfixe bzw. Partikel markiert. Im Französischen sind sie nur an der Umgruppierung der Aktanten ablesbar. Solche Ableitungsbeziehungen werden in der Klassifikation festgehalten, um Transfer auch dann zu ermöglichen, wenn es in einer Sprache zu einer Verwendungsweise der anderen Sprache kein äquivalent in der direkt korrespondierenden Klasse gibt.

9.4 Subklassifikation der Hauptklassen

Innerhalb dieser Hauptklassen werden Subklassen gebildet, je nachdem, welcher Typ von Lokalergänzung vorliegt. Die elementaren Typen von Lokalergänzungen werden festgelegt durch die Evaluation des Wahrheitswerts der elementaren lokalen Paraphrasen vor und nach dem vom Verb bezeichneten Prozeß. Dieses Testverfahren ist in den Beispielen (3) bis (6) bereits skizziert; es läßt sich in folgendem Schema zusammenfassen:

	VOR	NACH
TYP D. LE.		
Ausgangsp.	+	-
Zielpunkt	-	+
Statisch	+	+

Die Rollen des lokalisierte Arguments und des Orts der Lokalisierung können kumuliert werden mit anderen Rollen (z.B: lokalisierte Argument und effiziertes Objekt, Ort der Lokalisierung und aktives Subjekt, Ort der Lokalisierung und Instrument). Solche Kombinationen modifizieren die elementaren Typen der Lokalergänzung und konstituieren eigene Subklassen.

9.5 Weitere Trennung von Lesarten

Innerhalb der so konstiuierten Subklassen werden zur weiteren Trenung von Lesarten Selektionsbeschränkungen für die einzelnen syntaktischen Positionen herangezogen. Diese werden beschrieben einerseits durch semantische Merkmale andererseits durch syntaktische und semantische Klassifikatoren, die Klassen von Nomina definieren, die in bestimmten Positionen zugelassen sind. (Näheres hierzu sowie zum gesamten Verfahren der Klassifikation in [CAROLI 1988] und [GUILLET 1987]).

10 Probleme der vergleichenden Klassifikation deutscher und französischer Lokalisierungsverben im Hinblick auf die Erstellung zweisprachiger Lexika für sprachverarbeitende Systeme (Folker Caroli, Bochum/Saarbrücken)

In dem hier vorgestellten Projekt wird eine vergleichende Klassifikation der deutschen und französischen Lokalisierungsverben erarbeitet, mit dem Ziel, alle syntaktischen und semantischen Informationen systematisch zu erfassen, die für die eindeutige Zuordnung korrespondierender Verwendungsweisen dieser Verben in Transferwörterbüchern nötig sind.

Bei der Klassifikation folgen wir dem im Laboratoire d'Automatique Documentaire et Linguistique (L.A.D.L) entwickelten Ansatz, der sich an Harris orientiert. In diesem Ansatz werden äußerungen als Verkettungen von Operatoren und Argumenten angesehen. Argumentstellen können mit elementaren oder komplexen Argumenten besetzt sein. Komplexe Argumente sind wiederum Operatoren, die ihrerseits weitere Argumente binden. Die Struktur der äußerungen mit komplexen Argumenten können auf eine Menge von äußerungen abgebildet werden, deren Operatoren nur elementare Argumente binden. Diese Projektion kann mit Hilfe von Paraphrasen vorgenommen werden, deren metasprachlichen Bildungsmittel in der zu beschreibenden Sprache selbst enthalten sind. (vgl. [GROSS 1981], [BOONS/GUILLET/LECLÈRE 1976])

Dieses Prinzip wird bei der Lösung verschiedener Probleme der Klassifikation in beiden Sprachen in äquivalenter Weise angewendet:

10.1 Abgrenzung der verbabhängigen Lokalergänzung durch Extraktion

Eine solche mögliche Projektion stellt die Geschehens-Paraphrase dar, die in der Verbklassifikation zur Abgrenzung verbabhängiger von satzabhängigen Präpositionalergänzungen herangezogen wird. Sie hat sich als hinreichend operationales Kriterium zur eindeutigen Bestimmung verbabhängiger Lokalergänzungen erwiesen. Da sie in der Literatur eingehend diskuiert ist, gehe ich auf sie hier nicht weiter ein. (vgl. die Darstellungen bei [BIERE 1976] und [EROMS 1981].)

10.2 Explikation der lokalen Relation

Ausgehend von Verben mit präpositionaler Lokalergänzung läßt sich die allgemeine Struktur der Lokalisierungsverben mit Hilfe von allgemeinen lokalen Paraphrasen explizieren. Solche elementaren Operatoren sind *sein* im Deutschen und *être* und *il y a* im Französischen. Der dynamische Aspekt der Lokalisierung kann durch Operatoren wie kommen, gelangen bzw. *aller, passer, venir* wiedergegeben werden. Als gemeinsames Merkmal aller Lokalisierungsverben erweist dieses Verfahren der Explikation, daß durch den vom Verb bezeichneten Prozeß eine lokale Relation zwischen zwei Aktanten entsteht:

$$R_{[L]} (N_{[a]}, N_{[loc]})$$

Dabei bezeichnet $N_{[a]}$ das lokalisierte Argument, $N_{[loc]}$ den Ort der Lokalisierung.

10.3 Unterschiedliche Bezeichnung der lokalen Relation

Lokalisierungsverben lassen sich danach klassifizieren, durch welche Kombination von Aktanten die lokale Relation bezeichnet wird. Bei der Annahme eines maximalen Satzrahmens der lokalen Verwendungsweise eines Verbs mit Subjekt, Akkusativobjekt bzw. complément d'objet direct und präpositionaler Ergänzung sind theoretisch folgende Kombinationen denkbar:

		Ort		
		Subjekt	Objekt Cmp. dir.	Prp. Eg.
lokal. Argum.	Subjekt	1 –	2 +	3 +
	Objekt Cmp.dir.	4 +	5 –	6 +
	Prp. Eg.	7 +	8 +	9 ?

In der Tat sind folgende Kombinationen im Deutschen, wie im Französischen durch Beispiele belegt.

Jede dieser Kombinationen konstituiert im Deutschen wie im Französischen eine Klasse von Lokalisierungsverben. Zwischen den einzelnen Klassen bestehen bestimmte Ableitungsbeziehungen, d.h., es gibt Verben, die distinkte Verwendungsweisen haben, die zwei unterschiedlichen Klassen von Lokalisierungsverben zuzuordnen sind. Im Deutschen werden diese Ableitungen häufig durch Präfixe bzw. Partikel markiert. Im Französischen sind sie nur an der Umgruppierung der Aktanten ablesbar. Solche Ableitungsbeziehungen werden in der Klassifikation festgehalten, um Transfer auch dann zu ermöglichen, wenn es in einer Sprache zu einer Verwendungsweise der anderen Sprache kein äquivalent in der direkt korrespondierenden Klasse gibt.

10.4 Subklassifikation der Hauptklassen

Innerhalb dieser Hauptklassen werden Subklassen gebildet, je nachdem, welcher Typ von Lokalergänzung vorliegt. Die elementaren Typen von Lokalergänzungen werden festgelegt durch die Evaluation des Wahrheitswerts der elementaren lokalen Paraphrasen vor und nach dem vom Verb bezeichneten Prozeß. Dieses Testverfahren ist in den Beispielen (3) bis (6) bereits skizziert; es läßt sich in folgendem Schema zusammenfassen:

	VOR	NACH
TYP D. LE.		
Ausgangsp.	+	–
Zielpunkt	–	+
Statisch	+	+

Die Rolle des lokalisierten Arguments und des Orts der Lokalisierung können kummuliert werden mit anderen Rollen (z.B: lokalisiertes Argument und effiziertes Objekt, Ort der Lokalisierung und aktives Subjekt, Ort der Lokalisierung und Instrument). Solche Kombinationen modifizieren die elementaren Typen der Lokalergänzung und konstituieren eigene Subklassen.

10.5 Weitere Trennung von Lesarten

Innerhalb der so konstituierten Subklassen werden zur weiteren Trennung von Lesarten Selektionsbeschränkungen für die einzelnen syntaktischen Positionen herangezogen. Diese werden beschrieben einerseits durch semantische Merkmale andererseits durch syntaktische und semantische Klassifikatoren, die Klassen von Nomina definieren, die in bestimmten Positionen zugelassen sind. (Näheres hierzu sowie zum gesamten Verfahren der Klassifikation in [CAROLI 1988] und [GUILLET 1987]).

11 Datenbanken im lexikalischen Prozeß (Lothar Lemnitzer, Heidelberg)

'Datenbanken im lexikographischen Arbeitsprozeß' — das bedeutet die Anwendung eines der modernsten und in rascher Entwicklung sich befindenden Konzepte der Informatik auf eine traditionelle informationsverarbeitende Tätigkeit. Der Vergleich von acht in Hinsicht auf zwei- und mehrsprachige Lexikographie aufgebauten lexikographischen, lexikalischen und terminologischen Datenbanken, den ich in meinem Vortrag zog, bezieht sich auf die von mir im Rahmen des Heidelberger Projekts COLEX (COmputergestützte LEXikographie) gemachten Untersuchungen.

Ziel des von BLUMENTHAL im Rahmen des Heidelberger 'Forschungsschwerpunktes Lexikographie' von 1985 bis 1987 durchgeführten Projektes waren Entwurfsvorschläge für das Design eines computergestützten lexikographischen Arbeitsplatzes. Der Blick in die Werkzeugkiste des Informatikers ist dafür ebenso unerläßlich wie eine genaue Analyse des lexikographischen Arbeitsprozesses und seiner einzelnen Schritte. Im Rahmen dieses Projektes oblag mir die Aufgabe, die bisherigen Anwendungen von Datenbanktechnologie in lexikographischen Projekten zu untersuchen und zu vergleichen.

Man kann die verschiedenen Datenbankanwendungen in zweierlei Hinsicht klassifizieren. Zum einen danach, ob in ihnen genuin sprachliche, durch linguistische Analyse gewonnene Daten gespeichert sind, oder bereits lexikographisch aufbereitete Daten, also Wörterbuchartikel, Karteikarten o.ä., zum anderen danach, ob in ihnen die Ausgangsdaten des lexikographischen Prozesses gespeichert werden sollen oder Daten, die Ergebnis dieses Prozesses sind. Danach lassen sich unterscheiden: *Textdatenbanken* (sprachliche Ausgangsdaten), *lexikographische Datenbanken* (lexikographische Ausgangsdaten), *lexikalische Datenbanken* und *Terminologiedatenbanken* (sprachliche Ergebnisdaten), *Ergebnisdatenbanken* (lexikographische Ergebnisdaten).

Der Aufbau lexikographischer Datenbanken beruht auf der Analyse von bereits bestehenden Wörterbüchern, genauer: deren maschinenlesbarer Druckvorlagen. Die Übersetzung der Wörterbuchartikel in eine Datenbankstruktur ist also keine eigentlich linguistische oder lexikographische Arbeit. Dennoch gehört für die Übersetzung viel linguistisches und lexikographisches Know–how dazu. Das zeigt ein Beispiel aus der Bearbeitung des 'Longman Dictionary of Contemporary English' in Lüttich. Die satzgrammatischen Codes für die Verben bestehen aus der Angabe der Transitivität des Verbs und den möglichen distributionellen Umgebungen für dieses Verb. Das heißt, dieses lexikographische Textsegment ist zusammengesetzt. Informationen über die Kombination eines Transitivitäts'werts' mit einer bestimmten Satzumgebung lassen sich zuverlässig nur aus der Datenbasis gewinnen, wenn die Informationen aufgetrennt werden. Dafür ist über die Interpretation der Drucksteuerzeichen hinaus ein intellektueller Eingriff notwendig.

Ist die Beschreibung und Speicherung der Daten in lexikographischen Datenbanken nicht eigentlich theoriegeleitet, sondern pragmatisch orientiert, so steht bei lexikalischen Datenbanken und terminologischen Datenbanken die theoretische Erwägung am Anfang aller Beschreibung. Die Auswahl der lexikalischen Einheiten, die beschrieben werden sollen, und der Angaben zu diesen lexikalischen Einheiten obliegt dem Designer der Datenbank, sie sind nicht bereits durch irgendein anderes Medium vorgegeben. Bei der Modellierung lexikalischer Datenbanken kommt es auf die Auswahl der Beschreibungsbereiche (Morphologie, Syntax, Semantik etc.), auf die Wahl der Kategorien und deren Vokabular an. So orientiert sich die morphologische Beschreibung deutscher lexikalischer Einheiten in DOMENIGs 'Dedizierter Datenbank für Lexika' am morphologischen Beschreibungsverfahren von KOSKENNIEMI.

Das für die lexikographische Datenmodellierung geeignetste Datenmodell ist m.E. das relationale Modell. Relationale Datenbanken erweisen sich als flexibel gegenüber nachträglichen Veränderungen und Erweiterungen der Datenstruktur und ermöglichen bei Abfragen alle beliebigen Verknüpfungen zwischen Datenfeldern. Hierarchische und Netzwerkdatenbanken bedürfen der expliziten Verbindung einzelner Datensätze untereinander, was eben die Realisierung nicht vorgesehener Verbindungen bei Datenabfragen erschwert. Der entscheidende Nachteil relationaler Datenbanken liegt in dem Sachverhalt, daß die Verbindung der Datensätze untereinander nur virtuell ist, d.h. bei jeder konkreten Abfrage mittels mengenorientierter und relationaler Operationen neu hergestellt werden muß. Das kann bei großen Datenmengen zu einer nicht zu verantwortenden Verlängerung der Suchzeit führen. Außerdem sind bei sauber konzipierten relationalen konzeptuellen Schemata die Zusammenhänge zwischen den einzelnen Bestandteilen eines komplexen Objekts (z.B. eines Wörterbuchartikels) nicht mehr ohne weiteres zu rekonstruieren.

Im Sinne der Datenbanktheorie handelt es sich bei der Lexikographie um eine 'Non-Standard-Anwendung', eine Anwendung, deren Anforderungen über das übliche Maß hinausgehen. In letzter Zeit ist die Datenbankentwicklung dabei, den speziellen Bedürfnissen dieser Anwendungen entgegen zu kommen.

12 Das Ruhrgas-Terminologiedatenbanksystem XLT-1 (Hans J. Stellbrink und Karl H. Grigo, Essen)

12.1 Einführung

Im Bereich des Übersetzens verringert sich in der Bundesrepublik Deutschland die Zahl der Arbeitsplätze kontinuierlich. Neben wachsenden Fremdsprachenkenntnissen bei allen Akademikern sowie übersetzerischen Leistungen, die nicht dem tatsächlichen Bedarf entsprechen, sind die hohen Kosten der Übersetzung, die fast ausschließlich personalkostenabhängig sind, Ursache für diese Entwicklung. Nennenswerte Produktivitätssteigerungen lassen sich nur durch den Einsatz des Rechners erreichen. Zu nennen sind insbesondere Textverarbeitung, rechnergestützte Terminologiearbeit sowie Maschinenübersetzungen.

Den Durchbruch geschafft hat die Textverarbeitung unabhängig davon, wie die Übersetzung hergestellt wird (als Diktat oder direkt am Schreibsystem). Die rechnergestützte Terminologiearbeit hat mittlerweile auch die Schwelle der Wirtschaftlichkeit überschritten. Die Einführung von Terminologiesystemen ist allenthalben im Gespräch, jedoch werden immer wieder bei den Planern von Terminologiesystemen fehlende Marktübersicht, fehlende terminologische Fachkenntnis und generell Konzeptionslosigkeit bei der Einbindung von Terminologiesystemen in vorhandene DV-Umfelder, vorhandene Terminologienetzwerke und bestehende Übersetzungsabläufe deutlich.

12.2 Die Konzeption von XLT-1

Bei der Konzeption von XLT-1 wurde soweit wie möglich auf vorhandene Erfahrungen zurückgegriffen. Ausgangspunkt war die Grundüberlegung, daß Schreiben und Übersetzen zwei voneinander getrennte Abläufe sind, so daß eine Integration des Terminologiesystems in die Textverarbeitung keinen Sinn macht (zwei getrennte Arbeitsplätze). Aus Speicherkapazitätsgründen sowie der gewünschten Funktionsvielfalt kam deshalb nur ein Großrechnersystem in Frage. Zweite Grundüberlegung war, daß ein vorhandenes Software-System eingesetzt werden sollte, um einen Datenaustausch komplikationslos zu ermöglichen und von bereits vorhandenen Erfahrungen zu profitieren. Dieser Ansatz führte zur Entscheidung für die EURODICAUTOM-Software, die aber von Siemens auf IBM umgestellt werden mußte. Schließlich war weitere Maßgabe, daß das Terminologiesystem für spätere Anwendungen (z. B. Maschinenübersetzung) erweiterbar sein mußte.

12.3 Das Terminologiedatenbanksystem XLT-1

Die Terminologiedatenbank XLT-1, die auf einer IBM 3090 läuft, ist begriffsorientiert konzipiert. Jeder Eintrag in der Datenbank entspricht einem Begriff. Die Terminologielehre hat eindeutig nachgewiesen, daß nur so eine logische Zuordnung von Benennungen in vielen Sprachen, eine sachgerechte Behandlung von Synonymen, eine einmalige Speicherung von Zusatzinformationen und eine Verwaltung sehr großer Datenbestände möglich sind. Für XLT-1 sieht das Eintragsformat wie folgt aus:

Sprachenneutral:

1. Sachgebietsschlüssel

2. Autor
3. Pool
4. Hierarchische Zuordnung
5. Eintragsqualität

Sprachenbezogen:

1. Sprache 1
   a) Benennung Sprache 1
      α) Synonym Sprache 1
      β) Synonym Sprache 1
   b) Quelle Sprache 1
      α) Quelle Synonym 1 Sprache 1
      β) Quelle Synonym 2 Sprache 1
   c) Definition
   d) Beispielhafte Verwendung
   e) Zusatzbemerkungen (Sprachebene, Sprachgebiet, usw.)
2. Sprache 2 usw.

Von den einzelnen Eintragsfeldern ist lediglich der Eintragskopf (außer der hierarchischen Zuordnung) obligatorisch. Damit ist es z.B. möglich, für Begriffe in Sprachen, in denen keine Benennung vorliegt, das Feld frei zu lassen, oder lediglich Synonyme in einer Sprache zu sammeln.

Die Aufspeisung der Datenbank erfolgt sowohl durch Eingabe größerer Terminologien als auch bei der laufenden Übersetzungsarbeit. XLT-1 wird aber kaum zur systematischen Terminologiearbeit verwendet, denn hierfür steht die erforderliche Personalkapazität nicht zur Verfügung. Jeder Übersetzer ist eingabeberechtigt. Dies ist allerdings nur in einem kleineren Sprachendienst wie bei der Ruhrgas möglich, wo Vorgehensweisen leicht abgestimmt werden können. Die Eingabe erfolgt on-line, so daß jeder Eintrag auch sofort zur Abfrage zur Verfügung steht.

Die Datenbank ist in Datenpools organisiert. Neben dem allgemeinen Bestand gibt es Sonderpools für Projekte, wie z. B. das bei Ruhrgas erstellte Fachwörterbuch der Gasindustrie, oder Projekte der Ruhrgas oder ihrer Beteiligungsgesellschaften. Diese projektspezifische Terminologie wird nach der Abwicklung des Projekts wieder gelöscht.

Die Abfrage erfolgt im Regelfall on-line über die Bildschirme, mit denen alle Übersetzerarbeitsplätze ausgestattet sind, oder über die Terminals anderer Anwender in anderen Bereichen der Ruhrgas oder ihrer Beteiligungsgesellschaften. Es können jedoch auch beliebige Batch-Ausgaben z. B. mit Weitergabe zum automatischen Wörterbuchsatz oder als Liste für externe Übersetzungsbüros erfolgen.

12.4 Ziele

XLT-1 verfolgt verschiedene Ziele:

- Ersatz alter Handdateien zur Erhöhung der Produktivität des Übersetzungsdienstes
- Vereinheitlichung der Terminologie innerhalb des Ruhrgas-Konzerns
- Vorbereitung des Satzes des internationalen Fachwörterbuches der Gasindustrie
- Bereitstellung von Terminologie für externe Übersetzungsdienste zur Verbesserung der Qualität von außer Haus angefertigten Übersetzungen
- Verringerung des Terminologierecherchenaufwandes bei Ruhrgas durch Kooperation mit Dritten.
- Vorbereitung der Einführung der Maschinenübersetzung

Auf die beiden letzten Aspekte soll abschließend kurz noch eingegangen werden.

In keiner Wissensdisziplin scheint so viel Doppelarbeit stattzufinden wie im Bereich der Terminologie. Benennungen zu Begriffen werden von Normenausschüssen, Autoren und Übersetzungsdiensten immer wieder recherchiert. Eigentlich kann es sich aber auch der Berufsstand der Übersetzer nicht leisten, aufwendig in jedem Übersetzungsbüro die richtige englische Entsprechung für *Unterpulverschweißen, das Boyle-Mariott'sche Gesetz* oder eine *Rutschkupplung* zu suchen. Gerade die Vermeidung dieser Doppelarbeit durch Terminologieaustausch ist ein wesentliches Ziel von XLT-1.

Schließlich soll XLT-1 auch vorbereitend für die Maschinenübersetzung sein. Voraussetzung für jede maschinelle Übersetzung ist eine funktionierende hauseigene Vokabeldatei. Diese wird mit XLT-1 im Ansatz geschaffen, denn XLT-1 kann und soll um die zusätzlichen Funktionen, die in einem Wörterbuch für die Maschinenübersetzung notwendig sind, ergänzt werden.

Literaturverzeichnis

[Asher 1983] N. Asher: 'Understanding and Non-monotonic Reasoning', in: Proceedings of the Non-monotonic Reasoning Workshop. New Paltz, 1983, 1-20.

[Barnett/Lehmann/Zoeppritz 86] Brigitte Barnett, H. Lehmann, M. Zoeppritz: 'A Word Data Base for Natural Language Processing', in: Proceedings of COLING86, 1986.

[Biere 1976] Bernd Ulrich Biere: 'Ergänzungen und Angaben.', in: Helmut Schumacher (Hg.): 'Untersuchungen zur Verbvalenz. Eine Dokumentation über die Arbeit an einem deutschen Valenzlexikon' (= Forschungsberichte des Instituts für deutsche Sprache Mannheim 30), Tübingen, 1976, S.129-173.

[Boons/Guillet/Leclère 1976] Jean-Paul Boons, Alain Guillet, Christian Leclère: 'La structure des phrases simples en français. Constructions intransitives', (=Langue et Culture 8) Genf und Paris, 1976.

[Busemann/Hauenschild 1988] Stephan Busemann, Christa Hauenschild: 'A Constuctive View of GPSG or How to Make it Work', in: 12th Proceedings of COLING-88, Budapest, 1988.

[Caroli 1984] Folker Caroli: 'Les verbes transitifs à complément de lieu en allemand', in: Lingvisticae Investigationes, VIII,2, 1984, S. 225-267.

[Caroli 1988] Folker Caroli: 'Abschlußbericht zum Projekt: Vergleichende Klassifikation der deutschen und französischen transitiven Lokalisierungsverben', Bochum [masch.verf.], 1988.

[Collins/Loftus 1975] A.M. Collins, E.F. Loftus: 'A spreading activation theory of semantic processing', in: Psychological Review 82, 407-428.

[Cruse 1986] D.A. Cruse: 'Lexical Semantics', Cambridge University Press, Cambridge u.a., 1986.

[Czap/Galinski 1987] H. Czap, C. Galinski (Hg.): 'Terminology and Knowledge Engineering', Frankfurt, 1987.

[Domenig 1987] Marc Domenig: 'Entwurf eines dedizierten Datenbanksystems für Lexika. Problem, Analyse und Softwareentwicklung anhand eines Projekts für maschinelle Sprachübersetzung'. Beiträge zur philologischen und linguistischen Datenverarbeitung (Informatik und Informationswissenschaft Nr.17). Tübingen, 1987.

[Eckert/Heid 1988] U. Eckert, U. Heid: 'Semantic Relations in EUROTRA-D and Syntactic Functions in LFG: A comparison in the context of lexical transfer in MT', in: Steiner/Schmidt/ Zelinsky-Wibbelt (Hg.): From syntax to semantics - Insights from machine translation, London, 1988.

[Eroms 1981] Hans-Werner Eroms: 'Valenz, Kasus und Präpositionalphrase. Untersuchungen zur Syntax und Semantik präpositionaler Konstruktionen in der deutschen Gegenwartssprache', (=Monographien zur Sprachwissenschaft 11), Heidelberg, 1981.

[Eurotra 1988] EUROTRA Reference Manual, version 5.0., Juli 1988.

[Gross 1986] Gaston Gross: 'Typologie des noms composés: le lexique électronique des noms composés du français', Rapport ATP, CNRS, Paris, Université Paris 13, 1986.

[Gross 1981] Maurice Gross: 'Les bases empiriques de la notion de prédicat sémantique', in: Langages 53, 1981, S. 7-52.

[Guillet 1987] Alain Guillet: 'Constructions transitives à complément de lieu', Paris, 1987.

[Hauenschild/Busemann 1988] Christa Hauenschild, Stephan Busemann: 'A Constructive Version of GPSG for Machine Translation', in: E. Steiner, P. Schmidt, C. Zelinsky-Wibbelt (Hg.): 'From Syntax to Semantics – Insights From Machine Translation', London, 1988: 216-238.

[Hausmann 1984] Franz-Josef Hausmann: 'Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen', in: Praxis des neusprachlichen Unterrichts 31-4 (19840; 395-406.

[Hausmann 1985] Franz-Josef Hausmann: 'Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiel', in: Henning Bergenholtz und Joachim Mugdan (Hg.): 'Lexikographie und Grammatik', Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, Tübingen, 1985.

[Hutchins 1986] W.J. Hutchins: 'Machine Translation. Past, Present, Future', Chichester, 1986.

[Koskenniemi 1983] Kimmo Koskenniemi: 'Two-level Model for Morphological Analysis', in: Proceedings of Ijcai-83, 1983, 683-685.

[Koskenniemi 1983a] Kimmo Koskenniemi: 'Two-Level-Morphology: A General Computational model for Word-Form Recognition and Production', Dissertation, University of Helsinki, 1983.

[Lemnitzer 1988] Lothar Lemnitzer: 'Datenbanken im lexikographischen Arbeitsprozeß', Heidelberg, 1988 (Magisterarbeit, zu beziehen über den Autor).

[Luckhardt 1985] Heinz Dirk Luckhardt: 'Die SUSY-Lexika als linguistische Wissensbasis', CL-Report No. 7., Saarbrücken: Universität des Saarlandes: SFB100, 1985.

[Luckhardt 1987a] Heinz Dirk Luckhardt: 'Der Transfer in der maschinellen Sprachübersetzung', Sprache und Information Bd. 18, Tübingen, 1987.

[Luckhardt 1987b] Heinz Dirk Luckhardt: 'Probleme bei der automatischen Auswahl von Übersetzungsäquivalenten', in: Zimmermann/Kroupa/Luckhardt (Hg.), 1987.

[McDermott 1986] D.V. McDermott: 'A Critique of Pure Reason', Yale University, Research Report No.480, June 1986.

[Mel'čuk 1984] Igor A. Mel'čuk: 'Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-semantiques, I., (Les Presses de l'Université de Montréal) 1984.

[Nagao/Tsujii 1986] Makoto Nagao, Jun-ichi Tsujii: 'The transfer phase of the MU Machine Translation System', in: Proceedings of COLING 86, (Bonn: IKP), 1986: 97-103.

[Nirenburg et al. 1988] Sergei Nirenburg, Rita McCardell, Eric Nyberg, Scott Huffmann, Edward Kenschaft, Irene Nirenburg: 'Lexical Realization in Natural Language Generation', in: Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 12-14, 1988 (Pittsburgh, Pa.), Proceedings, (Pittsburgh, Pa.: CMU, ICMT), 1988

[Nirenburg/Nirenburg 1988] Irene Nirenburg, Sergei Nirenburg: 'Choosing Words Carefully', (Carnegie-Mellon University, Pittsburgh) 1988: Interner Bericht.

[Nirenburg/Nyberg/Kenschaft 1988] Sergei Nirenburg, Eric Nyberg, Edward Kenschaft: 'Inexact Frame Matching for Lexical Selection in Natural Language Generation', (Carnegie-Mellon University, Pittsburgh) 1988: Interner Bericht.

[Seelbach 1986] Dieter Seelbach: 'Zum propositionalen und textuellen Wissen für das Verstehen französischer Wetterberichte', LDV-Forum 4:1, 1986: S. 46-62.

[Snell 1983] B. Snell (Hg.): 'Term Banks for Tomorrow's World', London, 1983.

[Steiner 1986] Erich Steiner: 'Generating semantic Structures in EUROTRA-D', in: Proceedings of COLING 1986 (Bonn: IKP) 1986: 304-306.

[Thurmair 1986] Gregor Thurmair: 'Eine maschinelle morphologische Analyse des Deutschen', in: C. Schwarz, G. Thurmair (Hg.): 'Informationslinguistische Texterschließung', Hildesheim, 1986.

[Tomita/Carbonell 1986] Masaru Tomita und Jaime G. Carbonell: 'Another Stride Towards Knowledge-Based Machine Translation', in: COLING-86, Proceedings, Bonn, August 1986.

[Zelinsky-Wibbelt 1987] Cornelia Zelinsky-Wibbelt: 'Semantische Merkmale für die automatische Disambiguierung: ihre Generierung und ihre Verwendung', EUROTRA-D Working Paper Nr.4, Saarbrücken, 1987.

[Zimmermann/Kroupa/Luckhardt 1987] H.H. Zimmermann, E. Kroupa, H.-D. Luck hardt (Hg.): 'Das Saarbrücker Translationssystem STS - Eine Konzeption zur computergestützten Übersetzung', Veröffentlichungen der Fachrichtung Informationswissenschaft, Saarbrücken, Universität des Saarlandes, 1987.

Mitteilungen aus der GLDV

Die LDV-Forum–Redaktion bittet um Mithilfe

Folgende Mitglieder der GLDV waren beim letzten Versand des LDV-Forum unter ihrer uns bekannten Adresse nicht (mehr) erreichbar:

- H.-U. Block, Aachen, Rethelstraße 6
- M. Butt, Berlin, Sophie-Charlottenstr. 32
- G. Hoffmann, Trier, Am Weingarten 58
- M. Diestelmann, München, Winthirstr. 13a
- K. Dierks, Graz, Franckstr. 35
- G. Guckler, München, Herrmann-Vogel Str. 2
- S. Haubrich, Mainz, Frauenlobstr. 20
- W. Herrschbach, Uni Trier, Romanistik ¡- sei Jahren nicht mehr immatr.
- H.-D. Maas, Neunkirchen, Zweibrückerstr. 71
- A. Rau, Paderborn, Luethenweg 39
- V. Ripp(?), Berlin, Birkenstraße 12
- N. Ruge, Düsseldorf, Bilker Allee 200
- V. Schlögel, Münster, Sternstr. 39
- P. Schreiner, Gemaringen, Grundstraße 21
- B. Wesche, Stuttgart, Schönauerstr. 12a

Wer zur Aktualisierung unseres Adressbestandes beitragen kann, ist herzlich dazu aufgefordert, dies zu tun. Redaktion und GLDV bedanken sich sehr.

... Und sollte sich dereinst Ihre eigene Adresse ändern, so teilen Sie dies bitte umgehend der GLDV mit – der ansonsten notwendige Verwaltungsaufwand und die zusätzlichen Postgebühren sollten wirklich gespart werden.

Die Redaktion

Der Schatzmeister ...

Da folgende Mitglieder ihre Bankverbindungen gewechselt haben, sind ihre Mitgliedsbeiträge nicht mehr einziehbar und sie damit auch mit ihren Zahlungen im Verzug:

- Diestelmann, M.
- Gehrke, Manfred
- Guckler, Gudrun
- Kroupa, Edith
- May, Cornelia
- Rauschner, Hans-Dieter
- Ritzke, Johannes
- Rudolf, Klaus
- Sagawe, Helmuth
- Schreiner, Peter
- Steinkrüger, Werner

Die Genannten werden hiermit herzlich gebeten, mir möglichst umgehend ihre derzeit gültige Bankverbindung mitzuteilen.

Zur Konturierung des Faches Computerlinguistik

Udo Hahn, Matthias Heyn, Ludwig Hitzenberger
Bernhard Kelle, Hans-Dieter Lutz, Klaus Wothke

Die Zahl der Computerlinguistik2LStudiengänge nimmt ständig zu. Die Unterschiedlichkeit der fachlichen und institutionellen Einbindung und die Verschiedenartigkeit der Studiengangstypen3) mit entsprechend verschiedenen Abschlüssen erschwert die Überschaubarkeit und Vergleichbarkeit der Studienangebote für

- Lehrende
- Studienanfänger und fortgeschrittene Studierende
- Studienberater
- eventuelle Arbeitgeber.

Auch über Ausbildungsziele, Berufsqualifikation und Berufsprofil von CL-Absolventen bestehen bisher keine festumrissenen, sondern allenfalls vage und meist differierende Vorstellungen.

Auf Anregung der Arbeitskreise "Ausbildung und Berufsperspektiven" sowie "LDV und Nachbarn" haben daher Beirat und Vorstand der GLDV im März 1986 Arbeitstreffen interessierter Mitglieder initiiert, deren Ziel es war, etablierte und im Aufuau begriffene Studiengänge, deren Inhalte, Ausbildungsprofile sowie die sich daraus ergebende Art von Berufsqualifikation zu analysieren. Langfristige Zielvorstellungen für diese Arbeit waren,

- das Profil des Faches CL genauer zu konturieren und damit die Berufsqualifikation eines CL-Absolventen beschreibbar zu machen,
- inhaltliche Minimalvorstellungen (je Studiengangtypus) zu entwickeln,
- Lehr- und Lernziele einzelner Lehrveranstaltungen zu beschreiben,
- die "Durchlässigkeit" (Studienort-Wechsel) zwischen einzelnen Studiengängen zu erhöhen oder überhaupt erst zu ermöglichen,
- die Überschaubarkeit für die Studienberatung zu verbessern.

Für die gestellte Aufgabe erwies es sich als zweckmäßig, einen Diplomstudiengang Computerlinguistik (ohne eine institutionelle Einbindung) als *"tertium comparationis"* (s. Kap. 3) anzusetzen, zu dem die existierenden oder geplanten Studiengänge dann in Beziehung gesetzt werden können. Einig war man sich auch über die Fachbezeichnung "Computerlinguistik" (CL).

Empfehlungen [der GLDV] zur Konturierung des Faches Computerlinguistik:

1 Inhalte und Methoden des Faches Computerlinguistik

Grundlagen

Konstitutive Elemente der Computerlinguistik sind:

a) die Beschreibung natürlicher Sprache,

b) die Ausrichtung dieser Beschreibung an der Verarbeitung durch Computer,

c) die methodische Fundierung dieser Beschreibung durch den Bezug auf theoretische Prinzipien der Linguistik und Informatik.

Demzufolge bezieht die CL einige ihrer Methoden und Inhalte aus ihren Nachbardisziplinen Linguistik und Informatik, die sie einer eigenen Zielsetzung4) folgend in ihren Methoden- und Inhaltekanon integriert. In einer ersten Näherung sind dies:

"Vorfeld Linguistik"

1) Grundlegende Methoden empirischen Arbeitens und Kenntnis ihrer Ergebnisse im Bereich deskriptiver, struktureller, logischer oder quantitativer Linguistik

2) Modelle zur Phonologie, Morphologie, Syntax, Semantik, Pragmatik

3) Theoriegeleitete Beschreibung ausgewählter sprachlicher Phänomene (Nominalphrasen, Quantifikation, Koordination. . .)

4) Validierung von (Teil-)Theorien des Sprachsystems

"Vorfeld Informatik"

1) Mathematische Grundlagen (Algebra, Graphentheorie, formale Logik, Modelltheorie, Statistik); informatische Grundlagen (Rechnerarchitektur, Compilerbau, Betriebssysteme, Programmiersprachen, Software Engineering, Datenstrukturen, Algorithmentheorie)

2) Theorie formaler Grammatiken/Sprachen, Automatentheorie

3) DB-Systeme

4) Problemlösungsverfahren (Suchen, Planen, Lernen)

5) Testtheorie (Software- Validierung)

Zentrale CL-Bereiche

1) Grundlagen der Mensch-Computer-Interaktion (MCI) (Interaktionstheorie, Kommunikationsforschung, Kognitionswissenschaft; Dialogmodelle, Argumentationsmodelle, Benutzermodelle)

2) Parsing-Verfahren, Parsing-Strategien und grammatische Spezifikationssprachen5) 3)

Wissensrepräsentationsformalismen und -verarbeitungsprozeduren

4) Architekturen von natürlichsprachlichen Systemen

5) Systemevaluierung (Effektivität, Effizienz und empirische Adäquatheit implementierter natürlichsprachlicher Systeme, Technologiefolgenabschätzung)

2 Ausbildungsprofil und Fähigkeiten von Computerlinguisten6)

2.1 Ausbildungsprofil und Berufsprofil

Das Ausbildungsprofil eines Computerlinguisten setzt sich zusammen aus

a) Grundwissen bezüglich der Computerlinguistik mit ihren Vorfeldern Linguistik und Informatik und
b) Vertiefungswissen, das in der Regel durch einen Anwendungsbereich bestimmt wird, der wiederum die Hinzunahme von Anwendungswissen und Inhalten eines Nebengebietes erfordern kann. Dieses Vertiefungswissen ist auf mögliche Berufsfelder bezogen.

Grundwissen und Vertiefungswissen werden ergänzt durch praktisches Erfahrungswissen (interne und externe Praktika, Übungen, u.a.).

Vertiefungswissen kann in den folgenden Anwendungsbereichen (exemplarische, nicht vollständige Liste) erworben werden:

- Maschinelle Textverarbeitung / Bürokommunikation
- Maschinelle / Maschinengestützte Lexikographie
- Information und Dokumentation
- Maschinelle / Maschinengestützte Übersetzung
- Natürlichsprachliche Frage-Antwort-Systeme / Dialogschnittstellen
- Computerunterstützter Unterricht
- Entwicklung CL-spezifischer Tools

Entsprechend qualifiziert der hier vorgeschlagene Studiengang für Tätigkeiten im Bereich von Forschungseinrichtungen, von Fachinformations- und -dokumentationseinrichtungen, des Verlagswesens, in einschlägigen Softwarehäusern, im EDV-Ausbildungsbereich, u. a.

Grund- und Vertiefungswissen gemeinsam setzen den CL-Absolventen in den Stand, die Anforderungen des folgenden Fähigkeitenkataloges zu erfüllen.

2.2 Fähigkeiten des Computerlinguisten

Der Computerlinguist muß vorhandenes CL-Wissen adäquat einsetzen können. Das bedeutet:

2.2.1 Der Computerlinguist muß in der Lage sein zu entscheiden, ob ein sprachliches Problem auf dem gegenwärtigen Stand des Know-How mithilfe des Rechners sinnvoll zu lösen ist.

2.2.2 Der Computerlinguist muß die Fähigkeit haben, offene Problemfelder zu erkennen und neue Problemlösungen zu suchen.

2.2.3 Der Computerlinguist muß in der Lage sein, selbständig sprachliche Problemstellungen so aufzubereiten, daß sie sich mit dem Computer lösen lassen. Das bedeutet:

- Er muß mit fachspezifischen Verfahren eine Lösung finden und dabei gezielt das vorhandene sprachliche linguistische und informatische Wissen einsetzen (Problemanalyse) .
- In einem zweiten Schritt muß der Lösungsweg in einer für die maschinelle Verarbeitung geeigneten Form dargestellt werden (Spezifikation).
- Er muß den Lösungsweg in einem dritten Schritt in ein Programm(-System) umsetzen können (Implementation).

2.2.4 In interdisziplinären Softwareentwicklungsprojekten muß der Computerlinguist zwischen den verschiedenen Ansätzen der an der Lösung eines computerlinguistischen Problems beteiligten Fachrichtungen vermitteln können. Kritische Bereiche sind dabei die verschiedenen Problemlösungsstile, Bewertungsmuster und nicht zuletzt die Fachterminologien der daran beteiligten Disziplinen.

2.2.5 Ein Computerlinguist muß in der Lage sein, Lösungen in der CL zu bewerten und diese Bewertung explizit zu begründen. Das setzt die Kenntnis von spezifischen Eigenschaften sowie von V or- und Nachteilen unterschiedlicher Lösungswege voraus und die Kenntnis des innovativen Elementes bezüglich vorangegangener Lösungen.

2.2.6 Ein Computerlinguist muß weiterhin die Fähigkeit haben, die sprachlich fundierten Teilaspekte von Anwenderproblemen zu erkennen, in Kooperation mit dem Anwender zu beurteilen und gegebenenfalls einer Lösung zuzuführen.

Schließlich muß der Computerlinguist in der Lage sein, an der Etablierung eines in der Öffentlichkeit noch nicht geläufigen Berufsbildes mitzuarbeiten.

3 Computerlinguistik als Diplomstudiengang

Der Diplomstudiengang teilt sich in ein

- Grundstudium, das mit dem Vordiplom abgeschlossen wird, und
- ein Hauptstudium, das mit dem Hauptdiplom abgeschlossen wird.

Der Diplomstudiengang ist ein berufsqualifizierender Studiengang im üblichen Sinne.

Die Studiendauer beträgt 10 Semester + 1 Semester7) die Diplomarbeit.

Im Grundstudium und zu Beginn des Hauptstudiums liegen die Pflichtveranstaltungen; diese werden im Laufe des Hauptstudiums von Wahlpflichtveranstaltungen abgelöst.
Das Verhältnis zwischen Pflichtveranstaltungen, Wahlpflichtveranstaltungen und Wahlpflicht-veranstaltungen aus dem Vertiefungsbereich (Anwendungssysteme) soll stets 3 : 2 : 1 betragen.
Der hohe Anteil von Wahlpflichtveranstaltungen soll die Möglichkeit eröffnen, örtliche und institutionelle Besonderheiten an den einzelnen Studienorten berücksichtigen zu können.

Der Diplomstudiengang umfaßt 120 SWS; dabei entfallen auf Pflichtveranstaltungen 60 SWS, auf Wahlpflichtveranstaltungen ebenfalls 60 SWS (dabei wiederum auf Wahlpflichtveranstaltungen aus dem Vertiefungsbereich 20 SWS).

Zum Fach Computerlinguistik ist ein Nebenfach mit 24 - 40 SWS zu studieren. Als Nebenfächer kommen in Frage alle Fächer, die starke inhaltliche Bezüge zur Computerlinguistik haben (Informatik, Linguistik, Mathematik, Psychologie) und darüber hinaus alle Fächer, in denen CL-Bedarf im weitesten Sinne besteht. Lokale Gremien entscheiden über die Zulassung der Nebenfächer.

Im folgenden wird ein Rahmen für einen Diplomstudiengang angegeben, der sich auf den in Kapitel 1 angegebenen Methoden- und Inhaltekanon bezieht.
Dabei wird besonderer Wert darauf gelegt, daß die beiden Vorfelder Linguistik (als rekonstruktive Wissenschaft) und Informatik (als konstruktive Wissenschaft) nicht unvermittelt nebeneinander stehen, sondern bereits im ersten Semester durch eine eigene Veranstaltung miteinander verklammert werden, die dem interdisziplinären Charakter der CL entspricht.

Vorangestellt werden einige Erläuterungen zu den im Diplomstudiengang angesprochenen Veranstaltungsformen und Studienleistungen und einige generelle veranstaltungsunabhängige Empfehlungen.

Veranstaltungsformen

Alle Veranstaltungen, die keinen Vorlesungscharakter haben, verfolgen den Zweck, theoretisches Wissen durch praktische Fähigkeiten zu ergänzen bzw. in praktische Fertigkeiten umzusetzen; sie sind aber auch dazu gedacht, den Studierenden Gelegenheit zu geben, Ergebnisse ihrer Studien und Erfahrungen zu formulieren und auf diesem Wege auch ihre Kooperations- und sprachliche Ausdrucksfähigkeit auszubilden.

Vorlesung:

Vorlesungen dienen zur Einführung in größere Teilgebiete der Computerlinguistik, Linguistik und Informatik. In zusammenhängender Darstellung werden wissenschaftliches Grund- und Spezialwissen vermittelt und die zugehörigen Arbeitstechniken und Methoden exemplarisch vorgestellt.

Übung:

Übungen begleiten Vorlesungen. Dabei ist ihre primäre Aufgabe nicht darin zu sehen, den Stoff der Vorlesung zu wiederholen, sondern darin, daß sich die Studierenden mit dem Stoff intensiv und selbsttätig auseinandersetzen. Im Mittelpunkt steht die Vermittlung und Einübung praktischer Fertigkeiten. Von dieser Ausrichtung her ist es notwendig, Übungen in Gruppen (mit max. 20 Teilnehmern) abzuhalten.
Es werden Hausaufgaben gestellt, die selbständig schriftlich zu bearbeiten sind. In den Übungsstunden besteht die Möglichkeit, Aufgabenlösungen vorzutragen, sachliche Probleme und Verständnisschwierigkeiten im Zusammenhang mit dem Stoff der Vorlesung und/oder Übung unter fachkundiger Leitung zu diskutieren. Ubungen sind wesentliche und notwendige Ergänzung zu den Vorlesungen.

Proseminar:

Proseminare sind die Seminare des Grundstudiums. Dabei erhalten die Studierenden bereits im Grundstudium die Möglichkeit, selbständig ein eng begrenztes Thema zu erarbeiten, darzustellen und in einem kurzen Referat vorzutragen. Damit werden sowohl die in Übungen erworbenen methodischen Kenntnisse vertieft als auch die Fähigkeit geübt, Probleme und Sachverhalte darzustellen und zu vermitteln.

Seminar:

In den Seminaren werden Spezialbereiche bzw. spezielle Problemstellungen behandelt. Die Studierenden erhalten zur Aufgabe, ein komplexeres Thema anhand von Originalliteratur und weiterführender Literatur zu erarbeiten, kritisch-referierend darzustellen und in einem längeren Referat zur Diskussion zu stellen. Damit wird die Fähigkeit zur Darstellung und Vermittlung von Sachverhalten weiter vertieft.

Projektübung:

In Projektübungen soll eine umgrenzte, aber nicht-triviale Aufgabenstellung von kleinen Teams (3-4 Mitglieder) schrittweise und systematisch er- und bearbeitet werden: von der Definition der Problemstellung über die Zerlegung in Teilprobleme und die Abgrenzung und Bereitstellung des empirischen (sprachlichen) Materials und die Erarbeitung der notwendigen Beschreibungssprache(n)... bis hin zur Dokumentation von Stärken und Schwächen der vorgeschlagenen Problemlösung.

Projektseminar:

Die Projektseminare dienen sowohl dem Verständnis wichtiger neuerer Entwicklungen und aktueller Forschungsprobleme im Bereich der CL als auch der Vorbereitung auf die Berufspraxis. Hierbei arbeiten Gruppen von Studierenden an der Lösung umfangreicher CL-Probleme (von der linguistischen Analyse bis hin zur Implementierung) unter Anleitung von Dozenten. Neben dem Erwerb und der Festigung von CL-Kenntnissen soll die Arbeit in Projektseminaren vor allem folgende Themen berücksichtigen:

- psychologische und organisatorische Probleme der Teamarbeit,
- Rechnerunterstützung zur Organisation und Dokumentation der Arbeitsergebnisse,
- unterschiedliche Kommunikationsebenen und -formen bei der Software-Entwicklung und
-Konstruktion, wie Kommunikation innerhalb der Gruppe sowie mit Auftraggebern und Benutzern und Weiterentwicklern,
- Erfahrungen im Software-Entwurf, Abschätzung der Auswirkungen alternativer Entwurfsentscheidungen,

- Problematik der Aufspaltung in Teilaufgaben (Arbeitsteilung wie Modularisierung) und der Terminplanung,
- Test-, Evaluierungs- und Integrationsproblematik.

Studienleistungen

Studienarbeit:

Eine Studienarbeit steht bzgl. Komplexität und Zeitaufwand zwischen Seminararbeit und Diplomarbeit. Unter Betreuung eines Dozenten arbeiten die Studierenden ein Thema nach wissenschaftlichen Methoden auf, die im Studium bis zu diesem Zeitpunkt vermittelt worden sind. Dabei sollen nicht nur die theoretischen Kenntnisse, sondern auch Programmierkenntnisse und -fertigkeiten unter Beweis gestellt werden.

Diplomarbeit:

Die Diplomarbeit ist die unter Anleitung eines Hochschullehrers anzufertigende schriftliche Abschlußarbeit. Sie stellt den wesentlichen Studienschwerpunkt dar. Mit ihr sollen die Studierenden zeigen, daß sie in der Lage sind, eine nicht zu eng gewählte Aufgabenstellung innerhalb einer vorgegebenen Frist nach wissenschaftlichen Methoden selbständig zu bearbeiten und adäquat darzustellen.

Empfehlungen

1. Es wird empfohlen, ein Betriebspraktikum vorzusehen. Dieses Praktikum, das nicht zu den von der Hochschule anzubietenden Veranstaltungen gehört, soll dem/der Studierenden Einblicke in die Berufspraxis ermöglichen, die die Hochschule weder theoretisch noch praktisch vermitteln kann.
2. Die Durchführung von Übungen sollte durch Tutoren unterstützt werden.

Rahmen für einen Diplomstudiengang Computerlinguistik

Das Grundstudium umfaßt im:

1. Sem.	Einführung in die Computerlinguistik I, (V + Ü) Einführung in die Grundlage der Linguistik, (V + Ü) Mathematische und informatische Grundlagen **I,** (V+Ü) Programmierpraxis (V+Ü)	4SWS 4SWS 4SWS 4 SWS = 16 SWS
2. Sem.	Einführung in die Computerlinguistik 11 (V+Ü) Grundlegende Methoden empirischen Arbeitens (V + Ü) Mathematische und informatische Grundlagen 11 (V+Ü)	4SWS 4SWS 4 SWS = 12 SWS
3. Sem.	Theoriegeleitete Beschreibung ausgewählter sprachlicher Phänomene I (Projektübung) .. Programmiersprachen und Programmierkonzepte (V + U)	8 SWS 4 SWS = 12 SWS
4. Sem.	Theoriegeleitete Beschreibung ausgewählter sprachlicher Phänomene 11 (Proseminar) Datenbank-Systeme (V+Ü) Problemlösungsverfahren (V + Ü)	2 SWS 4SWS 4 SWS = 10 SWS
		: *50SWS*

Das **Hauptstudium** umfaßt 70 SWS:

- 34 SWS für Lehrveranstaltungen u.a. zu folgenden thematischen Bereichen:
  - Modellierung von Mensch-Computer-Interaktion
  - Parsingverfahren und -strategien
  - Wissensrepräsentationsverfahren
  - Systemarchitekturen
  - Evaluierung
- 20 SWS für Lehrveranstaltungen zu Vertiefungswissen aus einem Anwendungsbereich,
- 6 SWS für die Studienarbeit,
- 10 SWS für Projektseminar.

gesamt: 120 SWS

Inhalte der Veranstaltungen (Wesentliches in Stichworten)

EINFÜHRUNG IN DIE COMPUTERLINGUISTIK 1+ II

Einordnung der CL; CL und ihre Anwendungsbereiche; Basisalgorithmen der Sprachverarbeitung, Grundlagen des Parsing von natürlichen Sprachen; Grundlagen der Interaktionstheorie; zwischenmenschliche Kommunikation; Mensch-Computer- Interaktion (MCI); gesellschaftliche Auswirkungen.

Grundlagen von Informationssystemen: Informationsbegriff; Funktionsklassen von Informationssystemen (Referenzretrieval, Datenretrieval, inferenzbasierte Systeme), natürlichsprachliche Komponenten (Indexing, Klassifikation, Abstracting, Datenbanksynthese, Faktenextraktion, Maschinelle Übersetzung, Textgenerierung), Interaktionsklassen von Informationssystemen (Frage-Antwort-Systeme, Dialogsysteme); ...

EINFÜHRUNG IN DIE GRUNDLAGEN DER LINGUISTIK

Linguistik als empirische Wissenschaft; deskriptive Linguistik, formale Linguistik, kontrastive Linguistik, ...; Sprache und Welt; sprachliche Einheiten; linguistische Subsysteme (Phonologie, Morphologie, Syntax, Semantik, Pragmatik), ...

MATHEMATISCHE UND INFORMATISCHE GRUNDLAGEN I + II

Algebra; Graphentheorie; formale Logik (incl. Theorem-Beweis-Verfahren); Modelltheorie; Statistik; Algorithmen und Datenstrukturen; Theorie formaler Sprachen und Automatentheorie; ...

PROGRAMMIERPRAXIS

Grundlagen der strukturierten Programmierung anhand einer ersten Programmiersprache; praktischer Umgang mit Rechnern und Betriebssystemen;…

GRUNDLEGENDE METHODEN EMPIRISCHEN ARBEITENS

Beobachtung, Beschreibung, Erklärung; Datenerhebung, Transkription; Aufbau von Corpora, Corpusanalyse (Distributionsanalyse, IC-Analyse, H'); Klassifizierung, Hypothesen bildung, Abstraktion, Generalisierung, Theorienbildung, Theorienüberprüfung;

THEORIEGELEITETE BESCHREIBUNG

AUSGEWÄHLTER SPRACHLICHER PHÄNOMENE I+II

z.B.: Typisierung von Sprechakten nach AUSTIN, HABERMAS, SEARLE, ...;

z.B.: syntaktisch/semantische Beschreibung von Verben im Rahmen der Valenztheorie, der TG, der Kasusgrammatik, ...;
z.B.: Kennzeichnung und ihre sprachliche Realisierungsmöglichkeiten (Referenzsemantik, Syntax, Wortbildung, ...); ...

DB-SYSTEME

Grundlegende Konzepte von DB-Systemen; conceptual modelling; logische Datenorganisation; Datenmodelle; Sprachen für Datenmodelle (DDL,DML); Relationentheorie; physikalische Datenorganisation; Integrität; praktischer Umgang mit DB-Systemen; ...

PROGRAMMIERSPRACHEN UND PROGRAMMIERKONZEPTE

Über die erste Programmiersprache hinausgehende Konzepte, z.B. funktionaler Programmiersprachen (wie LISP), objektorientierter Programmiersprachen (wie SMALLTALK), logikorientierter Programmiersprachen (wie PROLOG), ...; Programmierumgebungen; Prototyping;...

PROBLEMLÖSUNGSVERFAHREN

Pattern-Matching; generate-and-test; means-end-Analysis; heuristische Suche ...; formale Schlußverfahren (Deduktion, Induktion, Abduktion); Lernverfahren; Planungsverfahren; ...

MODELLIERUNG VON MENSCH-COMPUTER-INTERAKTION

Kooperationsprinzipien; adaptive Systeme: Benutzermodellierung, Überzeugungssysteme; Sprechakt-Erkennung, Handlungsplangenerierung, Sprechaktgenerierung; Argumentationsmodelle, Dialogmodellierung; adaptierbare Systeme; ...

WISSENSREPRÄSENTATIONSVERFAHREN

Wissensrepräsenationsformalismen: Logiken; Semantische Netze, Frames, Scripts; Produktionssysteme; Wissensrepräsentationssprachen (FRL, KL-ONE, KRYPTON, OPS5, ...);

Beschreibungsformalismen für natürliche Sprachen: Dependenz, TG, GPSG, LFG, G & B, u.a.; grammatische Spezifikationssprachen, Unifikation;

PARSING-VERFAHREN UND -STRATEGIEN

Prinzipien des Compilerbaus; deterministisches, probabilistisches Parsing; top-down, bottom-up, island, chart, ...; unifikationsbasiertes Parsing; Parser-Generatoren; ...

SYSTEMARCHITEKTUREN

sequentielle, parallele, hierarchische, heterarchische, blackboard-Architektur; ...;

Architekturen von natürlichsprachlichen Systemen; ...

EVALUIERUNG

Leistungstests (Effizienz, Effektivität, Funktionalität) von natürlichsprachlichen Systemen; Prinzipien von Experimentaufbau und -durchführung bei der Entwicklung von natürlichsprachlichen Systemen...

4 Computerlinguistik als Hauptfach eines Magisterstudiengangs

Der Magisterstudiengang teilt sich in ein

- Grundstudium, das mit der Zwischenprüfung abgeschlossen wird, und
- ein Hauptstudium, das mit der Magisterprüfung abgeschlossen wird.

Der Magisterstudiengang mit CL als Hauptfach ist ein dem Diplomstudiengang entsprechend berufsqualifizierender Studiengang. Die Studiendauer beträgt mindestens 8 Semester + 1 Semester für die Magisterarbeit.
Für Magisterstudiengänge gibt es zwei verschiedene Typen:

- den Zwei-Fächer- Typ, in dem das 1. Hauptfach und das 2. Hauptfach jeweils ca. 60 SWS umfassen,
- den Drei-Fächer- Typ, in dem das Haupfach ca. 60 SWS umfaßt und jedes der beiden Nebenfächer ca. 30 SWS beansprucht.

Unter den folgenden Voraussetzungen ist ein Magisterstudium CL einem Diplomstudium CL inhaltlich vergleichbar:

Für den Zwei-Fächer-Typ werden die dem jeweiligen "Vorfeld" zuschreibbaren Veranstaltungen vom 2. Hauptfach übernommen. Als 2. Hauptfächer kommen in Frage Linguistik, Informatik und sprachwissenschaftlich ausgerichtete Einzelphilologien. Wie die Verteilung im einzelnen aussieht, wird vom jeweiligen Standort abhängen, und sie sollte ihren Niederschlag finden in einer von den beteiligten Fächern zu erarbeitenden Studien- und Prüfungsordnung.

Für den Drei-Fächer-Typ in seiner Ausprägung CL und zwei der Nebenfächer Informatik, Linguistik, sowie sprachwissenschaftlich ausgerichtete Einzelphilologie werden Teile der dem jeweiligen "Vorfeld" zuschreibbaren Veranstaltungen vom 1. bzw. 2. Nebenfach übernommen. Um aber eine vergleichbare Qualifikation zumindest prinzipiell sicherzustellen, sind an die Inhalte, die in diesen Nebenfächern vermittelt werden, bestimmte Forderungen von seiten des Hauptfaches zu stellen. Diese Inhalte sind in eine von den beteiligten Fächern zu erstellende Studien- und Prüfungsordnung einzuarbeiten.

In jedem Falle wird das Hauptfach CL die Aufgabe zu erfüllen haben, die Inhalte des/der jeweils anderen Faches/Fächer auf die Veranstaltungsinhalte der CL zu beziehen.
Weitere sinnvolle Kombinationen mit CL als Hauptfach sind denkbar, sind aber mit dem im Kapitel 3 skizzierten Diplomstudiengang nicht mehr vergleichbar. Dazu würden die Kombinationen von CL mit Psychologie, Informationswissenschaft, Formale Logik, u.a. zählen. Kombinationen mit Fächern ohne jeglichen inhaltlichen Bezug zur CL werden nicht empfohlen, können aber für Studierende im Einzelfall sinnvoll sein.

5 Computerlinguistik als Nebenfach

Computerlinguistik als Nebenfach sollte im Rahmen eines Diplomstudienganges oder im Rahmen eines Magisterstudienganges vom Drei- Fächer- Typ nur an solchen Universitäten angeboten werden, an denen CL auch als Hauptfach studiert werden kann oder aber an denen eine CL vorhanden ist, die personell, apparativ und mit Sachmitteln so ausgestattet ist, daß eine Forschung durchgeführt wird, die eine im Niveau akzeptable Lehre gewährleistet.
Welche Inhalte in welcher Form aus welchen Veranstaltungen des Rahmens für einen Diplomstudiengang (vgl. Kap. 3) gewählt werden, hängt einerseits ab von den örtlichen Gegebenheiten (vor allem von den Forschungsschwerpunkten) und andererseits von dem gewählten Hauptfach.
CL ist als Nebenfach zu einem Hauptfach Linguistik bzw. sprachwissenschaftlich ausgerichtete Einzelphilologie anders zu strukturieren als als Nebenfach zu einem Hauptfach Informatik.
Auf Grund des interdisziplinären Charakters des Faches CL erscheint es - von Einzelfällen abgesehen - wenig sinnvoll, CL mit Fächern zu kombinieren, die keinen Bezug zu CL haben.

6 Computerlinguistik als Teil in einem anderen Fach

Wenn Computerlinguistik nur als Teil (Studienrichtung oder Teilfach) eines Nebenfaches in einem Magisterstudiengang vom Drei-Fächer- Typ auftritt, dann ist zu bezweifeln, ob diese Kombination sinnvoll ist. Sie hätte nämlich bei den heutigen Gegebenheiten zur Konsequenz, daß CL mit nur ca. 15 SWS vertreten wäre. Es ist fraglich, was ein derart reduziertes Lehrangebot enthalten soll und welche Qualifikation damit vermittelt werden soll.

Dieser Zweifel gilt auch für den Fall, daß CL Teil eines Nebenfaches für einen Diplomstudiengang Informatik ist (zumal die Anzahl der SWS für Nebenfächer in einem Diplomstudiengang Informatik zwischen ca. 18 und 36 SWS differiert).
Tritt CL als Teil eines Hauptfaches auf, gilt das im ersten Abschnitt von Kap. 5 Gesagte.

7 Weiterbildungsstudiengänge in CL

Unter Weiterbildungsstudiengängen werden hier Studiengänge verstanden, die in den einzelnen Bundesländern als 'Aufbaustudiengang', 'Ergänzungsstudiengang' und 'Zusatzstudiengang' bezeichnet werden.
Unberücksichtigt bleiben in diesem Zusammenhang Zweitstudium und Weiterbildung im Rahmen von Einzelveranstaltungen (Kontaktstudium).
Studienvoraussetzung für einen Weiterbildungsstudiengang CL ist ein berufsqualifizierender Abschluß eines Hochschulstudiums (Mag., Dipl., Staatsex.); hierbei muß es sich um Fächer handeln, die in einem engeren inhaltlichen Zusammenhang zu CL stehen (wie z.B. Linguistik, Informatik, Einzelphilologien, Phonetik), oder um CL selbst.
Die Studiendauer beträgt in der Regel 4 Semester.
Der Studiengang wird durch eine Prüfungsordnung **und** eine Studienordnung bestimmt. Der Weiterbildungsstudiengang CL sollte nur an einer Hochschule angeboten werden, an der ein grundständiger CL-Studiengang existiert. Das Lehrangebot des Weiterbildungsstudiengangs ist institutionell und personell an das grundständige Lehrangebot zu binden.
Die Lehrveranstaltungen müssen sich inhaltlich nach den als Eingangsqualifikation geforderten Studienvoraussetzungen richten; ein Weiterbildungsstudiengang CL muß also bei vorangegangenem Studium der Informatik, Linguistik, Einzelphilologie, Phonetik oder CL jeweils verschiedene Lehrinhalte anbieten.
Eine Abstimmung zwischen den einzelnen Hochschulen bei der Einrichtung von Weiterbildungsstudiengängen CL ist unumgänglich.

Anmerkungen

(1) Mitgearbeitet und zur Entstehung dieses Papiers beigetragen haben:
I.S. Batori, R. Drewek, B. Endres-Niggemeyer, A. Franzke, T. Gardner, U. Hahn, P. Hellwig, M. Heyn, L. Hitzenberger, K. Jakob, B. Kelle, G. Knorz, W. Kreitmeier, W. Löser, H.-D. Lutz, M. Lutz-Hensel, P. Ovenhausen, H.P. Pütz, W. Putschke, U. Reyle, B. Rieger, B. Schaeder, K.-D. Schmitz, H. Schnelle, G. Schweisthal, G. Willee, K. Wothke

(2) Die Bezeichnung Computerlinguistik (CL) wird hier stellvertretend für die Vielzahl der zur Zeit gebräuchlichen Bezeichnungen des Faches verwendet.

(3) s. Studienführer Linguistische Datenverarbeitung (LDV) für die wissenschaftlichen Hochschulen der Bundesrepublik Deutschland, ermittelt und bearbeitet von Magdalene Lutz-Hensel, Stand Januar 1985.

(4) Diese Zielsetzung ist zur Zeit nicht einheitlich. V gl. **LDV** Forum 5 (1987), 76-78 mit einer Vorversion dieses Papiers, oder Bcitori, 1.; Krause, J.; Lutz, H.-D. (Hg.), Linguistische Datenverarbeitung. Versuch einer Standortbestimmung im Umfeld von Informationslinguistik und künstlicher Intelligenz. (= Sprache und Information. 4.) Tübingen: Niemeyer 1982.

(5) Dem liegt ein nicht allein syntaktisch geprägtes Verständnis von Parsing und Grammatik zugrunde.

(6) Die Bezeichnung "Computerlinguist" steht stellvertretend für "Computerlinguist" und "Computerlinguistin".

(7) Bedingt durch die Inhalte des Studiengangs wird von einer Studienzeit von 10 Semestern ausgegangen. Die formalen Bedingungen für die Studiendauer mögen aufgrund von Länderverordnungen o.ä. anders aussehen, d.h. ein achtsemestriges Studium fordern.