LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-Forum 13.1/2 (1996) Forum der Gesellschaft für Linguistische Datenverarbeitung e.V.

Herausgeber
Prof. Dr. Gerhard Knorz; Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)
Anschrift: Fachhochschule Darmstadt, Fachbereich Information und Dokumentation (IuD), Haardtring 100, D-64295 Darmstadt; Tel.: (06151) 16-8499; Fax: (06151) 168980; Email: knorz@fh-darmstadt.de

Redaktion
Gerhard Knorz, Ute Hauck

Wissenschaftlicher Beirat
Dr. Hans Billing, Harald Helsen, Prof. Dr. Wolf gang Hoeppner, Prof. Dr. Gerhard Knorz, Prof. Dr. Dietmar Rösner, Prof. Dr. Ulrich Schmitz, Dr. Matthias Wermke

Erscheinungsweise
Zwei Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober

Bezugsbedingungen
Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von DM 40,- (incl. Versand), Einzelexemplare zum Preis von DM 20,- (zuzügl. Versandkosten) bei der Redaktion bestellt werden.

Editorial

Mit dem LDV-Forum läuft das gegenwärtig so, daß es sich bei jeder Ausgabe um einen ganz speziellen Fall handelt. Bloß keine Routine also! Schön, daß es diesmal zumindest ein eingeplanter und angekündigter Sonderfall ist, und nicht einer, der sich gerade durch seinen Widerspruch zur Planung definiert. Die vorliegende Ausgabe ist eine Doppelnummer, die beide Hefte des Jahres 1996 zusammenfaßt und nach dem gegenwärtigen Stand der Arbeiten wohl einigermaßen rechtzeitig versendet werden kann. Als "gutgefüllt" war sie versprochen worden, und ich denke, das Versprechen kann als eingelöst gelten. Obwohl -, ja obwohl eine gewisse Schieflage, was das die Füllung der einzelnen Rubriken betrifft, kaum zu verbergen ist. Daß die Fachbeiträge das unbestrittene Zentrum dieser Ausgabe darstellen, ist nicht zu kritisieren. Daß aber Anregungen und redaktionelle Beiträge von den Mitgliedern der GLDV und aus den Arbeitskreisen so weitgehend ausbleiben, ist ein bedenkliches Zeichen. Es ist mehr als ein Indiz dafür, daß entweder die Gesellschaft zu wenige tragfähigen Initiativen entwickelt, oder aber daß diese Initiativen am LDV-Forum vorbeilaufen. Einfach zur Tagesordnung überzugehen bedeutet, die zugegebenermaßen noch unscharfe Diagnose zu bestätigen. Nachdenken, Diskussion und Handeln muß angesagt sein! Und ich denke, es gibt keinen Grund, systematisch an möglicherweise zentralen Ursachen vorbeizudenken. Um eine der sich aufdrängenden, kritischen Fragen explizit zu formulieren: Brauchen wir (noch) ein LDV-Forum in dieser Form? Mit dieser Qualität? Mit dem gegenwärtigen Spektrum an Beiträgen und Informationsleistungen. Ich bin mir sehr sicher, daß es eine Zeit gab, in der das LDV-Forum wesentlich zur Attraktivität der GLDV und auch zu einem Mitgliederzuwachs beigetragen hat. Diese Sicherheit (in Bezug auf die gegenwärtige Situation) ist mir abhanden gekommen. Äußern Sie sich! Üben Sie Kritik! Machen Sie Vorschläge! Oder gar Angebote! Damit die kleinen praktischen Hemmnisse Sie nicht von eventuellen guten Entschlüssen abhalten, hier direkt meine e-mail:

knorz@fh-darmstadt.de.

Weil gerade implizit das Stichwort "Internet" gefallen ist: Die Schnelllebigkeit dieses Mediums macht sicher einen Teil seiner Faszination aus, verursacht aber andererseits auch viele seiner Probleme. Ein Beispiel ist die URL, die Ihnen einen Blick in die Werkstadt des LDV-Forum ermöglichen sollte. Der Server des Fachbereichs Information und Dokumentation war seit Frühsommer 1995 am Netz und hat nun zum Wintersemester eine globale Umstrukturierung hinter sich gebracht. Die genannte URL führte dort noch eine ganze Zeitlang auf ein nicht (mehr) gewartetes Dokument, ehe auch diese Kulisse verschwand. Entschuldigung, sofern Sie enttäuscht einen

Versuch gewagt haben. Der GLDV-Server in Bonn teilt Ihnen seit geraumer Zeit die neue und richtige Adresse mit: Die redaktionelle Planung ist nunmehr in dem Informationsserver des Fachgebietes untergebracht, das ich hier an der Fachhochschule Darmstadt vertrete: in *WebSite Methodik.* Die URL des Einstiegsdokumentes für das LDV-Forum lautet:

http://www.iudjh-darmstadt.de/iud/wwwmeth/publl/dvforum/menuJ.htm

Vielleicht lassen Sie sich ja vom unbestreitbaren Reiz des interaktiven Mediums dazu verfuhren, Ihre Meinung oder gar eigenständige Beiträge in eine wieder etwas glänzendere Zukunft des LDV-Forums einzubringen. Im Übrigen greife ich eine Frage auf, die aus Koblenz an mich gerichtet wurde: Dürfen eigene oder fremde Forum-Beiträge in eigenen Web-Servern gespiegelt werden? Ich sehe keine andere Antwort als ein aufmunterndes "Ja"! Eine Quellenangabe ist, so denke ich, selbstverständlich, und eine kurze Mail an das Forum, die die Material-Übernahme mitteilt, wird erbeten.
Letztere Betrachtung enthüllt, daß es auf niedrigem Niveau eine Kommunikation mit dem LDV-Forum dennoch gibt, und dazu gehört dann auch die eine oder andere Reaktion auf Fehler oder Missverständliches im jeweils letzen Heft. So haben wir beispielsweise Wolfgang Höppner als Autor eines Veranstaltungsberichtes mehrdeutigerweise so abgedruckt, daß man ihn fälschlicherweise für den Veranstalter halten konnte. Ich komme gerne seiner Bitte nach und stelle fest, daß der Workshop über 'Discourse Markers' in Duisburg von Nils Lenke, Liesbeth Degand und Manfred Stede organisiert wurde. Und wenn wir damit bei den Namen von *Tüchtigen* sind, so schließe ich mit einem weiteren dieser Art, und hoffe daß es sich darüber hinaus um den Namen einer *Glücklichen* handelt: Ich schließe mit dem Auszug aus einem Zitat von Peter Hellwig, dessen *Kontext* (Nomen ist Omen) Sie im Heft nachlesen können: "Karin Haenelt hat ... am 8. Mai 1996 mit Bravour das Habilitationskolloquiuum bestanden". Herzlichen Glückwunsch zur *Venia Computerlinguistik* nochmals auf diesem Wege!

Und im Stil des alten Cato: Im Übrigen hoffe ich, von Ihnen zu hören Ihr

Gerhard Knorz

Titelgestaltung
Ute Hauck, Saarbrücken

Fachbeiträge
Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte (dreifach) sollten daher möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall zusätzlich auch noch auf Diskette *3W'* als ASCII oder LATEX-Datei übermittelt werden. Formatierungshilfen *(LDVForum.sty)* werden auf Wunsch zugesandt.

Rubriken
Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der Autoren wieder. Einreichungen sind - wie bei Fachbeiträgen - an die Redaktion zu übermitteln.

Redaktionsschluß
Für alle Rubriken mit Ausnahme der als Fachbeiträge eingereichten Manuskripte:
für Heft 14.1/97: 30. Apr. 1997 3
für Heft 14.2/97: 1. Aug. 1997

Herstellung W,
Saarbrücken

Druck
reha GmbH,
Saarbrücken
Auflage 400
Exemplare

Anzeigen
Preisliste und Informationen: Prof. Dr. Johann Haller, Institut für Angewandte Informationsforschung (W), Martin-Luther-Straße 14, 0-66111 Saarbrücken; Tel.: (0681) 38951-0; Fax: (0681) 38951-40; Email: hans@iai.uni-sb.de

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

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

HEADS UND MODIFIERS BEI DER KORPUSANALYSE

Gerda Ruge
TU-München
Arcisstr. 21
D-80290 München
ruge@informatik.tu-muenchen.de

Wilfried Brauer TU-München
Arcisstr.21 D-80290 München
brauer@informatik.tu-muenchen.de

Abstract: *Frühere Veröffentlichungen in der Computerlinguistik legen ihr Schwergewicht entweder auf linguistische Theorie, oder aber auf die empirisch Arbeit anhand von Korpora. Der folgende Artikel verbindet beides: Er befaßt sich mit dem Zusammenhang von bekannten linguistischen Theorien und der Extraktion von Heads und Modifiers aus Korpora. Drei theoretische Aspekte werden diskutiert: der Bezug zu Erkenntnissen aus der kognitiven Psychologie, der Bezug zur Merkmalssemantik und der Bezug ZU" modelltheoretischen Semantik. Zusammengenommen zeigt sich die Relevanz der Dependenzgrammatik in der Korpusanalyse. Diese ist allerdings nicht nur von theoretischem Interesse, sondern kann in der Praxis auch direkt bei verschiedenen Anwendungen eingesetzt werden.*

1 Einleitung

Zwei Sonderhefte *Computational Linguistics* 1993 zum Thema Korpusanalyse zeigen deutlich, daß das Interesse an diesem Gebiet im Aufwind ist. Die neueren Veröffentlichungen dazu sind jedoch überwiegend rein statistisch orientiert (z.B. /Smadja 93/ oder /Damerau 93/). Einige wenige Ansätze kombinieren statistische mit linguistischen Analyseverfahren (/Ruge 92/, /Hindie 90b/, /Grefenstette 94/, /Strzalkowski 95/). Allerdings steht hier die rein praxisorientierte Verbindung beider Methoden im Vordergrund. Bisher gibt es noch keine Veröffentlichungen zu theoretischen linguistischen Hintergründen der Korpusanalyse.

Die theoretische Linguistik hat eine lange Tradition, und sie hat viel Anerkennung gefunden. Derzeitige Veröffentlichungen zur Korpusanalyse erwecken allerdings den Eindruck, daß nun die Linguistik neu erfunden wird - und zwar auf statistischer Basis. Nie ist die Rede von irgendwelchen Zusammenhängen mit bekannten Theorien. Mit diesem Artikel wollen wir auf solche Zusammenhänge eingehen. Es fällt uns schwer, bei rein statistischen Arbeiten einen Zusammenhang mit linguistischen Theorien zu sehen. Dazu ist es schon nötig, linguistisch statistische Ansätze zu betrachten. Unseres Wissens nach beruhen alle linguistischen Korpusanalysesysteme zumindest teilweise auf der Head/Modifier-Relation, d.h. der Dependenzgrammatik.

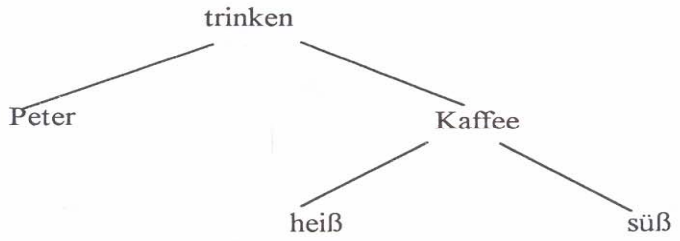
Die Theorie der Dependenzgrammatik geht ursprünglich auf Tesnierel zurück und wurde von /Hays 64/ formalisiert. Wir nennen hier alle Dependenzrelationen Head/Modifier-Relationen, auch beim Verbkomplement. Dependenzbäume unterscheiden sich von den Syntaxbäumen einer Phrasenstrukturgrammatik dadurch, daß ihre Struktur bereits von den einzelnen Syntaxregeln

1 laut /Kotschi 81, S.80/

einer Sprache abstrahiert ist, wenn Funktionswörter weggelassen werden. Beispielsatz 1 hat beispielsweise den gleichen Dependenzbaum wie Beispiel 2.

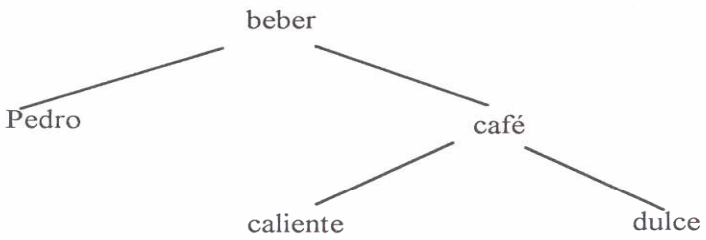
Beispiel 1 Peter trinkt einen süßen, heißen Kaffee.

Beispiel 2 Peter trinkt einen Kaffee, der süß und heiß ist.



Auch Sprachen mit unterschiedlichen Wortstellungsregeln haben bei äquivalenten Sätzen Dependenzbäume gleicher Struktur.

Beispiel 3 Pedro bebe un café dulce y caliente.



Laut /Hajicova 88/ sind Dependenzgrammatiken die einfachsten Grammatiken überhaupt. Diese Einfachheit prädestinieren die Dependenzanalyse zur Anwendung auf große Textmengen. Weltweit sind eine Reihe solcher Textanalyseverfahren entstanden, die alle mehr oder weniger auf den Ideen der Dependenzgrammatik beruhen: /Andreewsky, Fluhr 77/, /Berrut, Palmer 86/, /Dillon, Gray 83/, /Chiaramella et al. 86/, /Fagan 87/88; 89/, /Evans 90/, /Ruge et al. 91/ und /Smeaton 88/. Einige der Implementierungen der Dependenzanalyse sind wirklich auf große Textmengen anwendbar, da sie eine sehr schnelle Syntaxanalyse realisieren (/Hindle 90a/ und /Strzalkowski 92/). Der Parser von /Ruge et al. 91/ arbeitet darüber hinaus lexikonfrei, d.h. es gibt ein sehr kleines Lexikon mit Funktionswörtern und die Kategorien der Bedeutungswörter werden aus dem Kontext erkannt. Mit einem solchen Parser ist man dann auch in der Lage, bereichunabhängig zu arbeiten.

Leider gibt es in der Computerlinguistik kaum noch Veröffentlichungen zur Theorie der Dependenzanalyse; eine Ausnahme ist /Steinmann & Brzoska 95/. Hier wollen wir nun im Weiteren darstellen, welche Zusammenhänge zwischen der Head/Modifier-Extraktion aus Korpora und älteren linguistischen Theorien bestehen. Anschließend daran werden Einsatzmöglichkeiten für eine solche Extraktion behandelt. Es gibt zwei linguistische Theorien, zu denen wir einen starken Zusammenhang mit der Extraktion von Heads und Modifiers aus Korpora sehen: die strukturelle Semantik und die modelltheoretische Semantik. Es fällt auf, daß hier Semantik- und nicht Syntaxtheorien genannt werden. Dies liegt daran, daß die Dependenzanalyse keine reine Syntaxtheorie ist, in der lediglich die syntaktische Struktur eines Satzes beschrieben wird. Ein Dependenzbaum abstrahiert ja bereits von der syntaktischen Struktur und gibt statt dessen an, welches Wort sich auf welches bezieht. Und das führt uns gleich zu einem weiteren Aspekt von Head/Modifier-Relationen, nämlich ihrer kognitiven Relevanz.

2 Der kognitive Aspekt von Heads und Modifiers

Eine Besonderheit der Dependenzanalyse ist, daß sie sehr leicht intuitiv nachvollziehbar ist. Jeder Muttersprachler kann angeben, welches Wort sich auf welches in einem Satz bezieht, und das ist ja gerade das Wesentliche an der Dependenzrelation. Dependenzrelationen spiegeln deshalb die Muttersprachlerintuition bzgl. Syntax in einer natürlichen Art und Weise wider. Dies ist bereits das erste Argument dafür, daß Head/Modifier-Relationen eine besondere Position im menschlichen Gedächtnis einnehmen.

Ein weiteres Argument fußt auf einer Beobachtung von /Deese 64/ zur freien Assoziation. Bei der freien Assoziation wird einem Probanden ein Wort genannt, der Stimulus, und der Proband ist gehalten, mit dem Wort zu antworten, das ihm als erstes einfällt. Deese beobachtete, daß Assoziationen zu häufigen Adjektiven meistens deren Antonyme waren. Bei seltenen Adjektiven waren es meistens häufige Head-Nomen, so wie rot und Rose. /Wettier et al. 93/ stellten fest, daß menschliche Assoziationen im Wesentlichen auf häufiger Kookurrenz beruhen. Sie verglichen Kookurrenzstatistiken mit den Assoziationsantworten von 331 Probanden zu 100 Stimuli und stellen dabei eine starke Übereinstimmung fest. Die Stimulus/Assoziations-Paare waren entweder paradigmatische Relationen, Z.B. Tisch - Stuhl oder Mädchen - Junge, oder syntagmatische Relationen, wie Z.B. Schere - schneiden. Weitere Beispiele für syntagmatische Relationen sind: Kopf - Haare, Spinne - Netz, grün - Wiese oder Bett - schlafen. Nach Wettler et al. kommen solche Assoziationspaare häufig mit einem Abstand von wenigen laufenden Wörtern in Texten zusammen vor. Es ist anzunehmen, daß sie sich dann direkt aufeinander beziehen, z.B. die Haare auf dem Kopf, das Netz der Spinne, die grüne Witwe, im Bett schlafen. Diese Assoziationspaare wurden also aller Wahrscheinlichkeit nach häufig als Heads und Modifiers gebraucht.

/Strube 84/ untersuchte die sogenannte freie, fortgesetzte Assoziation. Hier gibt der Proband nach Nennung des Stimulus mehrere Assoziationsantworten ohne unterbrochen zu werden. Strube klassifizierte die Relation zwischen je zwei aufeinanderfolgenden Wörtern:

a) paradigmatische Relationen, im wesentlichen am Anfang der Assoziationskette, z.B. dreckig schmutzig
b) feststehende Begriffe, Z.B. blau - Himmel
c) syntagmatische Relationen, Z.B. Hammelfleisch - essen
d) gemeinsames Vorkommen in typischen Situationen, Z.B. Schweinebraten - Kartoffeln
e) persönliche 1 episodische Relationen: z.B. Woody Allen - weiß. Hierbei handelt es sich um eine nicht nachvollziehbare Produktion. Solche Assoziationen konnten durch Nachfragen beim Probanden als gemeinsam vorkommend in einer persönlichen Episode identifiziert werden.

Die Punkte b) und c) beschreiben wieder wahrscheinliche Head/Modifier-Relationen, wobei die stärker lexikalisierten Ausdrücke unter b) zu finden sind.

In diesem Abschnitt wurden nun zwei Punkte herausgearbeitet, die dafür sprechen, daß Head/Modifier-Relationen kognitiv existent sind: Einerseits die Einfachheit, mit der Menschen Head/Modifier-Relationen bestimmen können, und andererseits ihre Anwesenheit in menschlichen Assoziationen. Die nächsten Abschnitte befassen sich mit dem Zusammenhang zwischen der Dependenzrelation und semantischen Theorien.

3 Der Merkmalsaspekt von Heads und Modifiers

Die Idee, die Bedeutung eines Wortes mit Hilfe von semantischen Merkmalen zu beschreiben, geht ursprünglich auf die semantischen Felder von /Trier 32/ zurück, der verwandte Wörter zueinander in Beziehung setzte. Später benutzte man diese Verwandtschaft, um Wortbedeutungen

in Bedeutungseinheiten zu zerlegen (/Katz & Fordor 64/). Dabei entstehen Bedeutungscharakterisierungen für Wörter durch Merkmalslisten, etwa Mann als *{menschlich, männlich, erwachsen}*. Diese Merkmale eignen sich besonders zur Beschreibung von Kompatibilitäten. Beispielsweise ist der Satz Die Hälfte der Männer sind schwanger. semantisch nicht wohlgeformt. Dies läßt sich darauf zurückführen, daß Mann mit schwanger nicht kompatibel ist, weil dem Wort Mann das Merkmal *weiblich* fehlt. Semantische Merkmale stellen also ein adäquates Mittel dar, um Selektionsrestriktionen zu beschreiben.

Die Merkmalssemantik ist allerdings sehr umstritten, besonders wegen der folgenden Kritikpunkte:

- Es ist sehr fraglich, ob es eine universelle Merkmalsliste, also eine universelles Basissystem, gibt (/Lyons 71, S. 483f/).
- Es ist unwahrscheinlich, daß die Bedeutung von Wörtern umfassend durch Merkmale dargestellt werden kann (/Geeraerts 88/).
- Es ist nicht klar, wie Merkmale überhaupt interpretiert werden sollen (/Kastovsky 80/).

Auf diese Kritikpunkte werden wir später noch einmal zurückkommen, wenn wir ausblickartig darstellen, wie mit Hilfe der Head/Modifier-Relation künstliche semantische Merkmale aus Korpora gewonnen werden können. Zunächst wollen wir jedoch bewußt machen, daß es ausreicht, nur die Head/Modifier-Relationen zu berücksichtigen, um alle Selektionsrestriktionen eines Satzes zu erfassen. Dazu betrachten wir folgendes Beispiel:

Beispiel 4 Er überreichte eine Schachtel, die in dunkles Papier eingewickelt war.

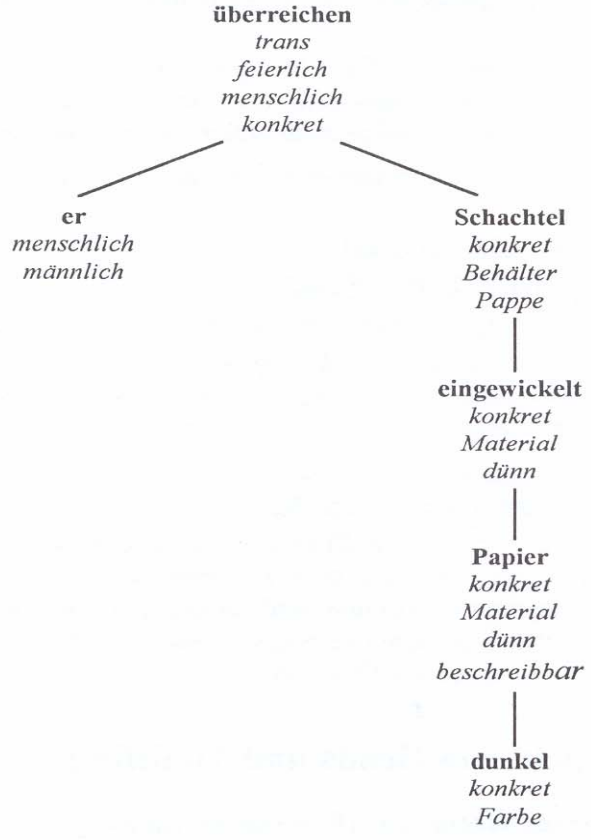


Abbildung 1 Dependenzbaum mit semantischen Merkmalen

Im folgenden wurde der Substitutionstest angewandt, um Kontrastpaare zu finden, mit denen semantisch wohlgeformte und nicht semantisch wohlgeformte Sätze verglichen werden können.

Beispiel 5 Er trank die Schachtel, die in dunkles Papier eingewickelt war.

Beispiel 6 Er trank die Schachtel, die mit dunklem Kaffee eingepinselt war.

Beispiel 7 Er trank den Kaffee, der in eine dunkle Tasse gegossen worden war.

Beispiel 8 Er trank die Farbe, die in eine dunkle Tasse gegossen worden war.

Beispiel 5 ist nicht wohlgeformt, weil trinken das Merkmal *flüssig* verlangt, welches Schachtel nicht besitzt. Auch wenn dieses Merkmal so wie bei Kaffee in Beispiel 6 vorkommt, bleibt der Satz doch nicht wohlgeformt, solange nicht der direkte Modifier von trinken dieses Merkmal besitzt (Beispiel 7). Als Modifier von trinken kann man beliebige Wörter mit dem Merkmal *flüssig* einsetzen, z.B. auch Öl oder Farbe, wie in Beispiel 8. Das gleiche Verhalten wie V erb/Objekt- Verbindungen zeigen auch Adjektiv/Nomen- Verbindungen:

Beispiel 9 Er gab ihr die senile Schachtel.

Beispiel 10 Er gab der senilen Frau die Schachtel.

Senil in Beispiel 9 verlangt das Merkmal *menschlich* und der Satz ist nicht wohlgeformt, obwohl er dieses Merkmal hat. Erst wenn senil ein Wort mit diesem Merkmal als direkten Head hat, ist der Satz korrekt (Beispiel 10).
Ein benötigtes Merkmal läßt sich auch nicht künstlich hinzufugen, so wie in Beispiel 11 und 12.

Beispiel 11 Er trank den Stein.

Beispiel 12 Er trank den flüssigen Stein.

Beispiel 11 ist nicht wohlgeformt, weil Stein das Merkmal *flüssig* fehlt. Es kann zwar wie in Beispiel 12 künstlich hinzugefugt werden, doch der Satz bleibt nicht wohlgeformt. Daß die Head/Modifier - Verbindung von trinken und Stein nicht möglich ist, ist nicht weiter beeinflußbar, da flüssig und trinken nicht direkt in einer Head/Modifier-Relation stehen.

4 Einsatz von Merkmalen

Die Eigenschaft von Selektionsrestriktionen, daß sie sich nur auf direkte Heads und Modifiers beziehen, läßt sich bekanntlich auch zum Disambiguieren von syntaktisch ambigen Sätzen einsetzen. Die Sätze 13 und 14 sind syntaktisch ambig. In einer Schachtel zu sein, verlangt das Merkmal *konkret.* Fehlt dieses Merkmal, so wie in Beispiel 14, so kann sich die Präpositionalphrase nur auf das Verb beziehen.

Beispiel 13 Er gab ihr die Schachtel mit Pralinen.

Beispiel 14 Er gab ihr die Schachtel mit Liebe.

Diese Disambiguierungsmethode wird in computerlinguistischen Systemen auch tatsächlich angewandt, etwa bei den Übersetzungssystemen von Siemens und IBM (/Hutchins 86, S.249/ und /Breidt 91/). Allerdings besitzt sie den Nachteil, daß die Merkmale manuell ins Lexikon eingetragen werden müssen, was sehr aufwendig ist und darüber hinaus sehr leicht zu Inkonsistenzen fuhrt.

Von /Hindle & Rooth 93/ wurde eine vollautomatische Disambiguierungsmethode vorgeschlagen. Sie beruht darauf, daß alle Head/Modifier-Relationen aus einem großen Korpus die Selektionsrestriktionen in diesem Korpus darstellen, auch wenn die Merkmale nicht explizit sind. Ihr Disambiguierungsansatz arbeitet folgendermaßen:

1) Extrahiere alle Head/Modifier-Relationen aus dem Korpus
2) Bestimme bei einer syntaktischen Ambiguität mögliche Bezüge in Form von Head/Modifier -Kandidaten.
3) Disambiguiere durch das Kandidatenpaar, das im Korpus auch vorkam.

Diese Methode hat allerdings zwei wesentliche Nachteilte: 1) Disambiguierung ist nur möglich, wenn genau der gleiche Head/Modifier-Link wie im aktuellen Satz auch im Korpus vorkam. 2) Die Methode ist sehr speicherintensiv, da nicht wenige Merkmale zu jedem Wort, sondern hunderte oder tausende von Heads und Modifiers gespeichert werden müssen.

Eine bessere Disambiguierungsmethode wäre also eine vollautomatische, die auf Merkmalen beruht. Eine solche Disambiguierungsmethode hat einer von uns bereits vorgeschlagen (/Ruge *95a1);* hier sollen nur die Grundzüge erläutert werden. Die Grundidee ist eine mathematische Transformation, mit der die Gesamtheit der Head/Modifier-Relationen eines Korpus, also die Selektionsrestriktionen aus dem Korpus, auf einige wenige (künstliche) Merkmale reduziert werden. Eine solche Reduktion ist in Abbildung 2 veranschaulicht. Zunächst werden die Wörter durch alle ihre Heads und Modifiers dargestellt (Punkte in Abbildung 2). Dann werden mit Hilfe einer Hauptkomponentenanalyse alle wesentlichen Gemeinsamkeiten - die Merkmale - herausgearbeitet (Pfeile im zweiten Teil von Abbildung 2). Zum Schluß werden die Wörter dann nur noch durch diese Merkmale dargestellt.

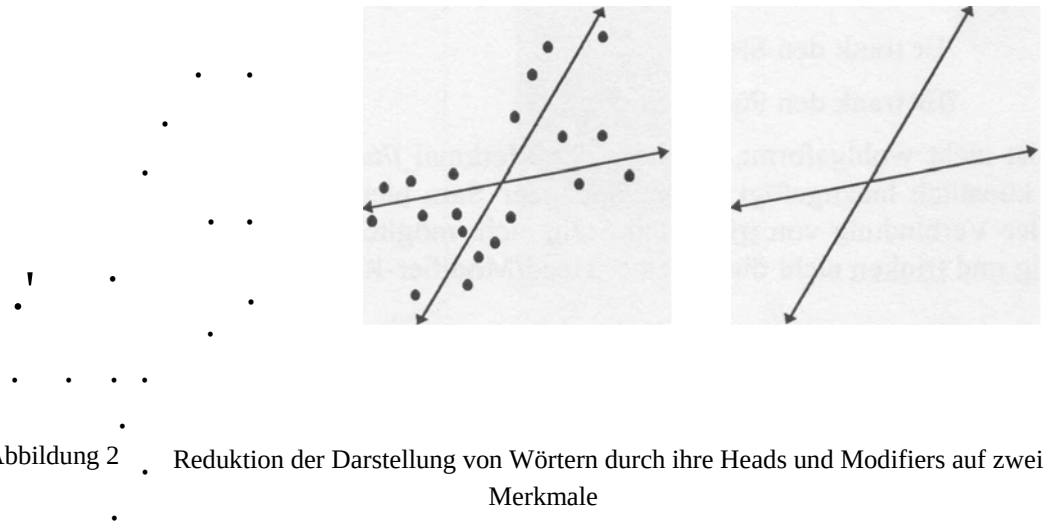


Abbildung 2 Reduktion der Darstellung von Wörtern durch ihre Heads und Modifiers auf zwei Merkmale

Die erzeugten Merkmale sind nicht mit den traditionellen Merkmalen wie *menschlich* oder *konkret* zu verwechseln. Sie entsprechen keinem Wort, sondern sind mathematische Konstrukte. Der Merkmalsvektor eines Wortes sieht z.B. folgendermaßen aus: (0.3, 1.2, 5.7, 0.6, 2.8, 4.3). Dabei gibt jede Zahl an, zu welchem Grad das entsprechende Merkmal auf das Wort zutrifft. Das erste Merkmal trifft mit 0.3 zu, das zweite mit 1.2 und so weiter. Was die einzelnen Merkmale bedeuten, läßt sich mit Hilfe der mathematischen Transformation nicht ablesen. Diese Merkmalsbeschreibung ist deshalb rein distinktiv: Sie beschreibt nicht, was die Wörter bedeuten, sondern nur, wie sich die Bedeutungen unterscheiden. Um diese automatisch generierten Merkmale von den normalerweise verwendeten abzugrenzen, nennen wir sie hier künstliche Merkmale oder Korpusmerkmale.

Zur Disambiguierung auf der Basis von künstlichen Merkmalen verfährt man folgendermaßen:

1) Extrahiere alle Head&-10difier-Relationen aus einem Korpus.

2) Bestimme zu jedem Korpuswort alle seine Heads und Modifiers und stelle die Korpuswörter durch Heads und Modifiers dar.

3) Wende die Hauptkomponentenanalyse darauf an. Die resultierenden Vektoren sind Merkmalsdarstellungen der Korpuswörter bzgl. des Korpus.
4) Bestimme bei einer syntaktischen Ambiguität mögliche Bezüge in Form von Head/Modifier -Kandidaten.
5) Disambiguiere durch das Kandidatenpaar, das am stärksten in seinen Merkmalen übereinstimmt.

Einen Nachteil hat allerdings auch diese Disambiguierungsmethode: Wenn zufällig semantisch verwandte Wörter Kandidaten zur Disambiguierung sind, wird falsch disambiguiert.

Beispiel 15 Peter beobachtet die Hand von seiner Tante mit den krummen Beinen.

Kandidaten, auf die sich mit den krummen Beinen beziehen kann, sind beobachtet, Hand und Tante. Mit der dargestellten Methode würde in diesem Fall allerdings nach Hand disambiguiert werden. Hand und Bein haben nämlich sehr viele gemeinsame Merkmale, weil sie semantisch verwandt sind. Wir erwarten allerdings, daß das zufällige Auftreten von semantisch verwandten Disambiguierungskandidaten sehr selten ist, so daß diese Disambiguierungsmethode normalerweise korrekt arbeitet. Derzeit wird an einer empirischen Evaluierung dazu gearbeitet.

Die verwendeten künstlichen semantischen Merkmale haben gegenüber den klassischen intellektuell zugeordneten Merkmalen nicht nur den Vorteil, daß sie vollautomatisch erzeugbar sind. Im Gegensatz zu den klassischen Merkmalen sind sie interpretierbar und zwar als Eigenschaften der Wörter bzgl. eines Korpus. Es stellt sich auch nicht mehr die Frage nach einem universellen Basissystem. Dies gibt es nicht, denn künstliche Merkmale werden fur jeden Korpus einzeln definiert.

Eine weitere Einsatzmöglichkeit der künstlichen semantischen Merkmale ist die lexikalische Disambiguierung bei automatisch übersetzten Texten. Bei der automatischen Übersetzung kommt es immer wieder vor, daß ein Wort mehrere Übersetzungen hat. Bei dem Übersetzungssystem MET AL der Fa. GMS wird dieses Problem zur Zeit durch Nacheditieren durch einen Übersetzer gelöst. IBM arbeitet an geeigneten Zusatzinformationen im Lexikon, die zu einer automatischen Disambiguierung fuhren (/Breidt 91/). Allerdings müßten die Lexikoneinträge manuell erzeugt werden. Durch Korpusmerkmale könnte dieses Problem vollautomatisch erledigt werden, und zwar folgendermaßen:

1) Extrahiere die Korpusmerkmale fur alle Korpuswörter aus einem Korpus in der Zielsprache. Dieser Korpus kann beispielsweise aus bereits nacheditierten Texten bestehen.
2) Gibt es dann ein Wort, das zwei mögliche Übersetzungen hat, so vergleiche die Merkmale beider Zielwörter mit denen aller übersetzten Wörter des ganzen Satzes. 3) WäWe das Zielwort als korrekte Übersetzung, dessen Merkmale am besten mit denen der Satzwörter übereinstimmen.

Korpusmerkmale lassen sich auch für die Erzeugung automatischer Hypertext-Links nutzen. Besonders interessant sind Links von Wort-Token zu der Definitionsstelle des Wortes im Text. Ausgehend von der folgenden Hypothese müßten sich diese Definitionsstellen finden lassen: Wörter werden definiert indem sie mit Wörtern umschrieben werden, die genau die Merkmale des definierten Wortes enthalten. Aufgrund eines Merkmalsvergleichs aller Wörter in den Textsätzen mit den Korpuswörtern müßten sich so Definitionsstellen, oder zumindest die wahrscheinlichsten Stellen für eine Definition finden lassen.

	CAR			
	Heads	208	Modifiers	182
1	BODY	109	ELEVATOR	142
2	SEAT	39	RAILROAD	98
3	DOOR	38	RAILWAY	60
4	SillE	33	TANK	46
5	POSITION	32	HOPPER	44
6	SPEED	31	RAIL	28
7	WHEEL	25	FREIGHT	27
8	TRUCK	22	TOY	26
9	TRA VEL	16	QUENCHING	25
10	STEREO	15	MOTOR	25
11	MOVEMENT	15	PAS SEN GER	21
12	END	15	SECOND	20
13	SYSTEM	14	LIFT	20
14	USE	13	LINKED	18
15	FRAME	13	LOADED	14
16	WASH	12	ADAPTED	12
17	HOLDING	12	GUIDE	11
18	COUPLER	12	MOVED	10
19	TOP	11	POSITION	9
20	MOUNTED	11	FLAT	9

Abbildung 3 Die häufigsten Heads und Modifiers des Wortes car aus Patentabstracts mit ihren Häufigkeiten

Beim Information Retrievallassen sich Head/Modifier-Relationen auch direkt einsetzen. Normalerweise werden alle bedeutungstragenden Wörter aus einem Text extrahiert, um ihn für das Volltext-Retrieval bereitzustellen. /Strzalkowski 941 nahm alle Head/Modifier-Paare aus dem Text hinzu und konnte so bessere Retrieval-Ergebnisse erzielen. Die Ergebnisse fallen besser aus, weil im Gegensatz zum herkömmlichen Retrieval unterschieden wird, ob zwei Wörter im Text sich aufeinander beziehen oder nicht. Head/Modifier-Relationen werden auch untersucht, um unter den Bedeutungswörtern in Texten die aussagefähigsten Phrasen zu finden, die das Textthema charakterisieren (/Smeaton 88/, /Fagan 87/88/).

All diese Anwendungen lassen sich realisieren, weil Head/Modifier-Relationen Träger von übereinstimmenden Merkmalen sind und durch die Head/Modifier-Relationen aus einem Korpus alle Selektionsrestriktionen in diesem Korpus erfaßt werden. Nach der Sicht der Merkmalssemantik betrachten wir Heads und Modifiers nun aus der Sicht der modelltheoretischen Semantik.

5 Der modelltheoretische Aspekt von Heads und Modifiers

Die modelltheoretische Semantik geht bekanntlich einerseits auf die Unterscheidung zwischen Extensionen und Intensionen zurück (/Frege 92/) und andererseits auf die Formalisierung von logischen Aussagen beginnend mit /Wittgenstein 21/. Für die grundlegenden Überlegungen im folgenden wollen wir uns auf Extensionen beschränken. Im Prinzip sind ähnliche Überlegungen auch für Intensionen möglich, allerdings sind Intensionen mathematisch etwas umständlicher zu handhaben. Die Extensionen vieler Wörter können durch Mengen von Objekten repräsentiert werden, z.B. Nomen, Adjektive und einstellige Verben. Andere werden durch Mengen von Tu-

pel repräsentiert, z.B. mehrstellige Verben, die sich wieder in einfache Mengen zerlegen lassen, z.B. in die Subjekt- und die Objekt-Mengen.

Man merkt schon, daß wir jetzt auf eine sehr stark vereinfachte Version der modelltheoretischen Semantik hinaus wollen, indem wir Extensionen einfach durch Mengen darstellen. Das ist nicht unbedingt im philosophischen Sinn adäquat. Um nur ein Beispiel zu diskutieren: Es lassen sich Adjektive nicht ohne weiteres als die Menge aller Objekte mit dieser Eigenschaft darstellen, ein leidenschaftlicher Biertrinker etwa braucht gar nicht leidenschaftlich zu sein. Dies und vieles mehr sind Vereinfachungen, die durch zwei Punkte gerechtfertigt werden: 1) Bei den späteren Anwendungen werden wir nur Wortarten betrachten, die im wesentlichen durch Mengen repräsentiert werden können, z.B. Nomen und Adjektive. 2) Die im folgenden dargestellten theoretischen Überlegungen sollen auf Korpora angewendet werden. Einzelne Gegenbeispiele und Ausnahmen verhindern nicht, daß grundlegende Aussagen die Gesamtheit der linguistischen Strukturen in einem Korpus korrekt charakterisieren.

Zwischen den Extensionen (als Mengen) können unabhängig von irgendwelchen Bedeutungen zwei verschiedene Relationen unterschieden werden: Der Schnitt der Menge ist nicht leer und die Mengen sind Teilmengen voneinander. Ersteres ist z.B. der Fall, wenn ein Element zwei Eigenschaften hat, wie etwa durch der grüne Apfel ausgedrückt wird. Hier wird ein Element beschrieben, das sowohl grün ist, wie auch ein Apfel. Gibt es ein solches Element, dann ist der Schnitt von ˇapfel und ˇgrün nicht leer. Äpfel können demnach grün sein, jedoch nicht alle Äpfel sind grün, deshalb kann ˇgrün als mögliche Eigenschaft von ˇapfel angesehen werden. Umgekehrt ist ein Apfel zu sein aber auch eine mögliche Eigenschaft von grünen Gegenständen. Wir nennen deshalb die Relation zwischen zwei Extensionen **externe Eigenschaft**, wenn der Schnitt zwischen den Mengen nicht leer ist. Ist eine Extension Obermenge einer anderen, so nennen wir dies eine **interne Eigenschaft**. Obermengen können mit Merkmalen gleichgesetzt werden. Die Bedeutung der Obermenge ist bereits in der Bedeutung des Wortes enthalten, so wie bei dem berühmten weißen Schimmel. Eine wesentliche Eigenschaft von internen Eigenschaften ist, daß sie in normalen Texten nicht erwähnt werden, das wäre ja doppelt gemoppelt.

Interne Eigenschaften sind Obermengen, müssen aber nicht unbedingt Oberbegriffe sein. Weiß ist z.B. kein Oberbegriff von Schimmel. Um zwischen Oberbegriffen und anderen internen Eigenschaften unterscheiden zu können führen wir die Begriffe *echte Eigenschaft* und *echter Oberbegriff* ein. **Echte Oberbegriffe** sind genau das, was der übliche Sprachgebrauch mit Oberbegriffen meint: Es gibt eine Menge von definierenden Eigenschaften und alles, was diese Eigenschaften hat, fällt unter den echten Oberbegriff. **Echte** (interne) **Eigenschaften** sind Eigenschaften, die jedes Element einer Menge hat. Im Gegensatz zu echten Oberbegriffen lassen sich echte Eigenschaften auf mögliche Eigenschaften zurückführen.

Die eben beschriebenen Definitionen sollten jetzt noch einmal formal dargestellt werden:

Zwei Mengen sind Eigenschaften voneinander, wenn sie ein gemeinsames Element haben.

Definition 1: **Eigenschaft**

$$egs(A,B) \leftrightarrow \exists x \, (A(x) \wedge B(x))$$

Eine Menge A ist interne Eigenschaft von einer anderen B, wenn B in A enthalten ist. Beispielsweise ist ˇgelb eine interne Eigenschaft von ˇbanane.

Definition 2 **interne Eigenschaft**

$$int(A,B) \leftrightarrow egs(A,B) \wedge \forall x \, (B(x) \rightarrow A(x))$$

Eine Menge A ist externe Eigenschaft von der Menge B, wenn sie Eigenschaft ist, jedoch nicht interne Eigenschaft. Beispielsweise ist ˇgrün eine externe Eigenschaft von ˇapfel. Es gibt grüne Äpfel, doch nicht alle Äpfel sind grün.

Definition 3 **externe Eigenschaft**

$$ext(A,B) \leftrightarrow egs(A,B) \wedge \neg\, int(A,B)$$

P ist eine echte (interne) Eigenschaft der Menge A, wenn es eine externe Eigenschaft Q von A gibt und Q eine Teilmenge von P ist. D.h. eine echte Eigenschaft läßt sich auf eine externe Eigenschaft, die nicht jedes Element von A hat, zurückführen. Insbesondere heißt das, daß P kein bloßer Oberbegriff von A ist, sondern wirklich eine Eigenschaft darstellt. ˇFarbe ist eine echte Eigenschaft von ˇobst.

Definition 4 **echte Eigenschaft**

$$ee(P,A) \leftrightarrow \exists Q\ (\ ext(Q,A) \wedge int(P,Q)\ )$$

P ist ein echter Oberbegriff von A, wenn es eine Menge von Eigenschaften Q_i gibt, die P definieren: Für genau die Extensionen, für die alle Q_i gelten, gilt auch P. ˇObst ist beispielsweise ein echter Oberbegriff für ˇbanane.

Definition 5 **echter Oberbegriff**

$$eo(P,A) \leftrightarrow \exists Q_1..\exists Q_n(\ \forall B\ (\ ee(Q_1,B)\wedge..\wedge ee(Q_n,B) \rightarrow eo(P,B) \wedge int(P,B)\ ) \wedge ee(Q_1,A)\wedge..\wedge ee(Q_n,A)\ )$$

Jetzt soll der Zusammenhang zwischen Extensionen und der Head/Modifier-Relation in Korpora betrachtet werden. Im Gegensatz zu frei erfundenen Beispielen können für reale Korpora bestimmte Annahmen getroffen werden: Aussagen in realen Korpora sind meistens wahr, sinnvoll und relevant.

Der einfachste Fall eines Head/Modifier-Links ist der zwischen Adjektiven und Nomen, im Beispiel mit extensionaler Übersetzung angegeben:

Beispiel 16 fast car

$$\lambda x\ (\ ^{\vee}car(x) \wedge\ ^{\vee}fast(x)\ )$$

Wenn es ein schnelles Auto gibt - was der Fall ist, wenn die Phrase in einer wahren relevanten Aussage vorkommt - dann haben ˇfast und ˇcar ein gemeinsames Element, und beide Extensionen sind Eigenschaften voneinander. Selbst wenn die Aussage negiert vorliegt, also etwa

Beispiel 17 Indeed it was not a fast car.

$$\exists x\ (\ ^{\vee}car(x) \wedge \neg\ ^{\vee}fast(x)\ )$$

muß es schnelle Autos geben. Gäbe es nämlich keine schnellen Autos, so würde die Aussage den Eindruck erwecken, daß es sie doch gäbe; sie wäre dann nicht relevant. Unter der Annahme der Relevanz von Beispiel 16 und 17 kann ihren Übersetzungen also folgendes hinzufügt werden:

$$\exists y\ (\ ^{\vee}fast(y) \wedge\ ^{\vee}car(y)\ )$$

Die Frage ist jetzt, ob das Adjektiv und das Nomen externe oder interne Eigenschaften voneinander sind. Sicher kann man sich Beispiele vorstellen, bei denen es sich um die Relation der internen Eigenschaft handelt, wie beim weißen Schimmel. Es ist jedoch unwahrscheinlich, daß diese Relation - abgesehen von wenigen Definitionen - in realen Texten gefunden wird. Normalerweise ist es nämlich nicht relevant, interne Eigenschaften zu erwähnen. Deshalb muß man davon ausgehen, daß hier normalerweise die Relation der externen Eigenschaft vorliegt.

Das gleiche Muster wie bei Adjektiven findet sich bei Nomen und einstelligen Verben.

Beispiel 18 The disc rotates.

$\exists x\ (^{\vee}\text{disc}(x) \wedge {}^{\vee}\text{rotate}(x))$

Mehrstellige Verben denotieren keine Mengen von Objekten, sondern je nach Stelligkeit Mengen von Paaren oder Tripeln von Objekten. Betrachten wir hier die Mengen die durch Projektionen dieser Tupelmengen entstehen, also Subjekt-Mengen und Objekt-Mengen. Solche Projektionen können auch durch Gerundien und Partizipien ausgedrückt werden:

Beispiel 19 The bottom is fixed by a screw.

$\exists x\ (^{\vee}\text{screw}(x) \wedge \exists y(\ ^{\vee}\text{bottom}(y) \wedge {}^{\vee}\text{fix}(x,y))\ )$
$\exists x\ (^{\vee}\text{screw}(x) \wedge {}^{\vee}\text{fix}^1(x)\ );\ {}^{\vee}\text{fix}^1 := \{x: \exists y\ {}^{\vee}\text{fix}(x,y)\ \}$
$\exists y\ (^{\vee}\text{bottom}(y) \wedge {}^{\vee}\text{fix}^2(y)\ );\ {}^{\vee}\text{fix}^2 := \{y: \exists x\ {}^{\vee}\text{fix}(x,y)\ \}$

Beispiel 20 The fixing screw is made of steel.

Beispiel 21 The fixed bottom cannot be removed.

Auch hier gilt wieder, daß sich Head/Modifier-Relationen in externen Eigenschaften zwischen Mengen, die sich aus den Denotaten ergeben, manifestieren. Das gilt auch für Head/Modifier-Kombinationen weiterer Wortarten, soll aber hier nicht weiter ausgeführt werden.

Da Head/Modifier-Relationen der Relation der externen Eigenschaft zwischen den Denotaten der Wörter entsprechen, charakterisiert die Gesamtheit aller Head/Modifier-Relationen in einem Korpus weitgehend die Relation der externen Eigenschaft zwischen den Denotaten. Dies gilt abgesehen von Synonymie und Polysemie. Im Fall von Synonymie stehen verschiedene Head/Modifier-Relationen für eine Relation der externen Eigenschaft zwischen den Denotaten. Im Fall von Polysemie steht eine Head/Modifier-Relation für verschiedene externe Eigenschaften. Die Vollständigkeit einer solchen Head/Modifier-Charakteristik hängt entscheidend vom Umfang des Korpus ab, sowie von seiner Abdeckung eines Bereichs. Im folgenden soll gezeigt werden, wie aufgrund der Struktur von externen Eigenschaften der Denotate auf die Synonymierelation zwischen Wörtern geschlossen werden kann. Da die Head/Modifier-Relationen im Wesentlichen den externen Eigenschaften entsprechen, kann mit einer gewissen Genauigkeit von der Gesamtheit der Head/Modifier-Relationen in einem Korpus auf die Synonymierelation zwischen den Wörtern bzw. auf ihre semantische Ähnlichkeit geschlossen werden.

6 Synonymie auf der Basis von externen Eigenschaften

Bei Extensionen handelt es sich nicht um irgendwelche Mengen, sondern um Mengen mit denen Begriffe erfaßt werden. Jetzt soll ein Axiomensystem aufgestellt werden, das Extensionen charakterisiert, und zwar in Termen von externen und internen Eigenschaften.

Wenn sich zwei Extensionen unterscheiden, so unterscheiden sie sich in mindestens einem Element. Dies kann mit Hilfe von externen oder internen Eigenschaften ausgedrückt werden. Man kann z.B. sagen: "Alle Bananen sind gelb, aber Äpfel nicht unbedingt." oder "Äpfel können sauer sein, aber Bananen sind es nie."

Axiom 1 **Ausdrückbarkeit**

$\neg(A{=}B) \rightarrow \exists P(\ \neg(\ \text{int}(P,A) \leftrightarrow \text{int}(P,B)\ ) \vee \neg(\ \text{ext}(P,A) \leftrightarrow \text{ext}(P,B)\ )\ )$

Um den Zusammenhang zwischen internen und externen Eigenschaften einer Extension zu diskutieren, soll zunächst ein Zitat von /Wittgenstein 21, § 2.0123/ betrachtet werden.

> "Wenn ich den Gegenstand kenne, so kenne ich auch sämtliche Möglichkeiten seines Vorkommens in Sachverhalten. (Jede solche Möglichkeit muß in der Natur des Gegenstandes liegen.) Es kann nicht nachträglich eine neue Möglichkeit gefunden werden."

D.h., daß mögliche Eigenschaften durch die internen Eigenschaften vorbestimmt sind. Insbesondere sind externe Eigenschaften mögliche Eigenschaften. Wenn eine externe Eigenschaft vorliegt - somit eine Eigenschaft möglich ist - dann gibt es eine interne Eigenschaft, die diese Möglichkeit eröffnet.

Axiom 2 **Vorbestimmtheit der externen Eigenschaften**

$$ext(A,B) \rightarrow \exists P\ (int(P,A) \wedge int(P,B))$$

Externe Eigenschaften implizieren also interne Eigenschaften. "Orange" impliziert "farbig", das impliziert "sichtbar" und das wiederum "materiell". Man kann sich Mengen vorstellen, für die das nicht gilt, etwa eine Menge, die als Elemente einen orangen Ball und die Idee zu Woody Allens letztem Film enthält. Zwar impliziert eine externe Eigenschaft des Balls, seine Farbe, daß er sichtbar ist, doch gilt das nicht für die Idee. Extensionen sind aber nicht beliebige Mengen, sondern Mengen, die mit einem Wort oder einer Phrase beschrieben werden können. Wir können deshalb voraussetzen, daß Extensionen in dem Sinne homogen sind, daß ihre Elemente gemeinsame oder ähnliche Eigenschaften haben.

Das nächste Axiom postuliert, daß interne Eigenschaften entweder echte Eigenschaften sind, oder aber echte Oberbegriffe. Entweder geht also eine interne Eigenschaft auf externe Eigenschaften zurück, oder es handelt sich um einen definierten Oberbegriff. Es gibt keine Extension, in der willkürlich irgendwelche Elemente auftreten.

Axiom 3 **Homogenität der internen Eigenschaften**

$$int(P,A) \rightarrow (\ eo(P,A) \vee ee(P,A)\ )$$

Mit den eben dargestellten Axiomen läßt sich der folgende Satz zeigen: Wenn alle externen Eigenschaften von zwei Extensionen übereinstimmen, dann sind sie gleich. Den vollständigen Beweis zeigt /Ruge 95b, S. 54-64/. Hier sei nur die Skizze gegeben:

Aus Definition 1 folgt, daß bei Gleichheit von externen und internen Eigenschaften zweier Extensionen auch die Gleichheit der Extensionen folgt.

Satz 1 $$(\forall P(\ int(P,A) \leftrightarrow int(P,B)\ ) \wedge \forall P(\ ext(P,A) \leftrightarrow ext(P,B)\ )\ ) \rightarrow A=B$$

Aus der Übereinstimmung der externen Eigenschaften zweier Extensionen kann mit Definition 4 die Übereinstimmung aller ihrer echten Eigenschaften gezeigt werden.

Satz 2 $$\forall P(\ ext(P,A) \leftrightarrow ext(P,B)\ ) \rightarrow \forall P(\ ee(P,A) \leftrightarrow ee(P,B)\ )$$

Aus der Übereinstimmung der externen Eigenschaften zweier Extensionen folgt auch die Übereinstimmung ihrer echten Oberbegriffe. Das kann mit Hilfe von Satz 2 und Definition 5 gezeigt werden.

Satz 3 $$\forall P(\ ext(P,A) \leftrightarrow ext(P,B)\ ) \rightarrow \forall P(\ eo(P,A) \leftrightarrow eo(P,B)\ )$$

Mit den Sätzen 2 und 3 und Axiom 3 erhält man schließlich, daß aus der Übereinstimmung der externen Eigenschaften die Übereinstimmung der internen folgt.

Satz 4 $$\forall P(\ ext(P,A) \leftrightarrow ext(P,B)\ ) \rightarrow \forall P(\ int(P,A) \leftrightarrow int(P,B)\ )$$

und Verb/Objekt-Relation, wie auch Head/Modifier-Links in Nominalphrasen. Strzalkowski verwendet zusätzlich Gewichte, die Worthäufigkeiten im gesamten Korpus einbeziehen.

Eine Synonymgenerierung aus Korpora kann nicht nur als Hilfsmittel bei der Thesauruserzeugung eingesetzt werden. Die generierten Synonyme können auch zur Suchfrageerweiterung verwendet werden. Hier wird ein typisches Problem im Information Retrieval angegangen: Wenn ein Benutzer eines Retrieval-Systems eine Suchfrage absetzt, werden nur die Dokumente gefunden, die die benutzen Suchwörter enthalten. Dokumente mit Synonymen oder Oberbegriffen werden nicht gefunden, auch wenn sie genau dem charakterisierten Thema entsprechen. Grefenstette (S. 137ft) und Strzalkowski untersuchten die automatische Suchfragenerweiterung; hier wird die ursprünglich vom Benutzer formulierte Frage um automatisch generierte Synonyme der Suchwörter erweitert und die Retrieval-Ergebnisse zu dieser erweiterten Frage werden bestimmt. Beide fanden eine Verbesserung der Retrieval-Ergebnisse, was auf eine hohe Güte der generierten Synonyme hindeutet. Suchfrageerweiterungen verbessern die Retrieval-Ergebnisse nämlich nur, wenn Synonyme einbezogen werden, aber nicht, wenn sonstige Assoziationen zu den Suchwörtern einbezogen werden.

8 Anwendbarkeit von Head/Modifier-Ansätzen bei der Korpusanalyse

Damit Head/Modifier-Ansätze für echte Anwendungen eingesetzt werden können, ist es eigentlich nur nötig, eine funktionierende, schnelle Head/Modifier-Extraktion zu implementieren. Dies ist allerdings schon in mehreren Projekten gelungen. /Hindie 90al berichtet über eine Frei-Text-Syntaxanalyse mit anschließender Head/Modifier-Extraktion, die eine Million Wörter über Nacht bearbeitet. /Strzalkowki 92/ gibt eine Rate von 0,5 Sekunden pro Satz bei seiner Dependenzanalyse an. Damit können eine Million Wörter in etwa 8 Stunden bearbeitet werden. Die gleiche Geschwindigkeit wie bei Strzalkowski und Hindie geben auch /Ruge et al. 91/ an, mit 19 KB pro CPU-Sekunde. Ihre Dependenzanalyse von Nominalphrasen wurde auch bzgl. ihrer Fehlerrate evaluiert. Hier wurden 85% der Head/Modifier-Token korrekt erkannt und 14% nicht vorhandene Links irrtümlich eingeführt. Diese Fehlerraten sind zu verkraften, weil die Weiterverarbeitung wie oben dargestellt sehr robust ist, was die evaluierten Anwendungen in Abschnitt 7 zeigen.

9 Schlußfolgerungen

In diesem Artikel wurde dargestellt, daß Head/Modifier-Relationen aus drei verschiedenen Gründen relevante linguistische Einheiten sind: zum ersten sind sie kognitiv relevant, zum zweiten spiegeln sie Selektionsrestriktionen und damit Merkmalsverteilungen und zum dritten entsprechen sie weitgehend der Relation der externen Eigenschaft zwischen den Extensionen der Wörter. Diese theoretischen Eigenschaften von Heads und Modifiers führen zu verschiedenen, sinnvollen Anwendungen. Solche Anwendungen sind wirklich einsetzbar, da die Head/Modifier-Extraktion sehr einfach ist und deshalb effizient und robust implementiert werden kann.

Danksagungen: Wir danken Johannes Ritzke für seine Anregungen zu diesem Artikel.

Literatur

Andreewsky, A.; Fluhr, C.: "Computational Learning of Semantic Lexical Relations for the Generation and Automatical Analysis of Content", Proc. Of IFIP Congess, 1977,667-672

Berrut, C.; Palmer, P.: "Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing", Proc. of ACM Conf. on Research and Development in IR. 1986, 123-130

Breidt, E.: "Die Behandlung von mehrdeutigen Verben in der maschinellen Übersetzung", IWBS Report 158, ISSN 0938-1864, IBM, Wissenschaftliches Zentrum, 1991

Chiaramella, Y.; Defude, B.; Bruandet, M.; Kerkouba, D.: "IOTA: A Full Text Information Retrieval System", Proc. of ACM ICRDIR 1986,207-213

Damerau, F.: "Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts", Inf. Proc. & Management 29(4), 1993,433-447

Deese, J.: "The Associative Structure of some Common English Adjectives", 1. of Verbal Learning and Verbal Behavior 3(5), 1964,347-357

Dillon, M.; Gray, A.: "FASIT: A Fully Automatic Syntactically Based Indexing System", JASIS 34(2), 1983, 99-108

Evans, D.: "Conceptual Management in Text via Natural-Language Processing: The CLARIT Approach", Working Notes of the 1990 AAAI Symposium on "Text based Intelligent Systems", Stanford, 93-95

Fagan, J. (87/88): "Experiments in Automatic Phrase Indexing of Document Retrieval: A Comparison of Syn tactic and Non-Syntactic Methods", Dissertation, Correll University, Ithaca, New York, 1987/88

Fagan, J. (89): The Effectiveness of a Non-Syntactic Approach to Automatic Phrase Indexing for Document Retrieval", JASIS 40(2), 1989 115-132 Frege, G.: "Über Sinn und Bedeutung", Zeitschrift für Philosophie und philosophische Kritik 100, 1892,25-50 Geeraerts, D.: "Cognitive Grammar and the History of Lexical Sematnics" in Rudzka-Brygida (Hrg.): "Topics in Cognitive Linguistics", Current Issues in Linguistic Theory 50, Amsterdam, 1988,647-677 Hays, D.: "Dependency Theory: A Formalism and some Observations", Language(40)4, 1964,511-525

Hindie, D. (9Oa): "A Parser for Text Corpora", Technical Memorandum, AT&T Bell Laboratories, Dec. 1990, auch erschienen in Atkins, A.; Zampolli, A.: "Computational Approaches to the Lexicon", Oxford University Press, 1993

Hindie, D.(90b): "Noun Classification from Predicate-Argument Structure", Proc. of 28th ACL Conf., Pitts burg, 1990,268-275

Hindie, D.; Rooth, M.: "Structural Ambiguity and Lexical Relations", Comp. Ling. 19(1), 1993, 103-120 Hutchins, W.: "Machine Translation: Past, Present, Future", John Wiley & Sons, 1986

Kastovsky, D.: "Zur Situation der lexikalischen Semantik" in Kastovsky, D. (ed.): "Perspektiven der lexikalischen Semantik", Bouvier Verlag Herbert Grundmann, Bonn, 1980

Katz, J.; Fordor, J.: "The Structur of Semantic Theory" in Fordor, 1.; Katz, 1.: "The Structure of Language: Readings in the Philosophy of Language", Englewood Cliffs, NJ, Prentice Hall, 1964,479-518

Kotschi, T.: "Einleitendes zum Begriff der Valenz" in Kotschi, T.: "Beiträge zur Linguistik des Französischen", Gunter Narr Verlag, Tübingen, 1981

Lyons, J.: "Einführung in die moderne Linguistik", Beck'sche Verlagsbuchhandlung, München, 5. unveränderte Auflage, 1971

Ruge, G. (92): "Experiments on Linguistically Based Term Associations", Inf. Proc. & Mangement 28(3), 1992,317-332

Ruge, G.; Schwarz, C.; Warner, A. : "Effectiveness and Efficiency in Natural Language Processing for Large Amounts of Text", JASIS 42(6), 1991,450-456

Ruge, G. (95a): "Die automatische Extraktion semantischen Wissens aus großen Korpora", in Hitzenberger, L.: "Angewandte Computerlinguistik" (Proceedings der GLDV-Jahrestagung Regensburg 1995), Georg Olms Verlag, Hildesheim 1995, 179-193

Ruge, G. (95b): "Wortbedeutung und Termsassoziation", Reihe "Sprache und Computer" Band 14, Georg Olms Verlag, Hildesheim, Zürich, New York, 1995

Smadja, F.: "Retrieving Collocations from Text: Xtract", Computational Linguistics 19(1), 1993, 143-177

Smeaton, A. (88): "Using Parsing of Natural Language as Part of Document Retrieval", Dissertation, University Of Glasgow, Research Report CSC/88/RI, 1988

Steinmann, F.; Brzoska, *C.:* "Dependency Unification Grammar for PROLOG", Computational Linguistics 21(1), 1995,95-102

Strobe, G.: "Assoziation: Der Prozeß des Erinnerns und die Struktur des Gedächtnisses", Springer Verlag, 1984

Strzalokowski, T. (92): "TIP: A Fast and Robust Parser for Natural Language", Proceedings of COLING 92, 198-204

Strzalkowski, T. (94): "Robust Text Processing in Automated Information Retrieval", Proceedings of 4th Conference on Applied NLP, Stuttgart, 1994, 168-173

Strzalkowski, T. (95): "Natural Language Information Retrieval", Inf. Proc. & Management 31(3), 1995, 397 417

Trier, J.: "Sprachliche Felder", Zeitschrift für Deutsche Bildung 8(9), 1932,417-427

Wettler, M,; Rapp, R; Ferber, R: "Freie Assoziation und Kontiguitäten von Wörtern in Texten", Zeitschrift für Psychologie 201, 1993, 103-112

Wittgenstein, L.: "Tractatus logico-philosophicus", 1921, Suhrkamp Verlag, Ausgabe 12, (7. Auflage, 1969)

MASCHINELLE WERKZEUGE FÜR DIE FACHTEXTÜBERSETZUNG

Konzepte, Produkte und ihr wirtschaftlicher Nutzen

Uta Seewald Universität Hannover
seewald@mbox.rose.uni-hannover.de

1 Der Fachübersetzungsmarkt

Die Fachtextübersetzung hat im Zusammenhang mit der immer stärker werdenden internationalen Verquickung der Handelsbeziehungen und der Eröffnung neuer Absatzmärkte inzwischen eine ökonomische Schlüsselfunktion erhalten. Allein in Europa rechnet man am Ende dieses Jahrtausends mit einem Übersetzungsvolumen von etwa 1,8 Milliarden Seiten pro Jahr,1 was einem Markt von mehr als 70 Milliarden US Dollar entspricht. Nicht zuletzt auch bedingt durch die europäische Norm, die die Übersetzung von Bedienungsanleitungen für Geräte in die jeweilige Sprache des Absatzmarktes vorschreibt, bilden Dokumente aus dem Bereich der technischen Dokumentation, wie Bedienungsanleitungen, Reparaturanleitungen und Handbücher, den größten Anteil am Übersetzungsvolumen. Aber auch bei der Kontaktaufnahme mit ausländischen Handelspartnern, bei Verhandlungen und im Rahmen der Produktwerbung ist sprachliches Know-how erforderlich, das über das Englische, das heute als internationale Verkehrssprache dient, hinausgeht. Aufgrund des marktbedingten wachsenden Bedarfs an Fachübersetzungen ist verständlich, daß gegenwärtig ein starkes Interesse an Maschineller Übersetzung (MÜ) bzw. an effizienten Werkzeugen zur Unterstützung des Übersetzungsprozesses besteht.

Seit Beginn der fünfziger Jahre sind umfangreiche Fördermittel in Projekte zur MÜ geflossen. Die in die sechziger Jahre zurückreichenden Anfänge der großen Übersetzungssysteme SYSTRAN, METAL und LOGOS, die sich nach jahrzehntelanger Entwicklung noch heute im Einsatz befinden, sind ebenso wie das von 1982 bis 1994 von der Europäischen Gemeinschaft geförderte System EUROTRA, das Übersetzungen zwischen den Sprachen der Mitgliedsländer der Gemeinschaft berücksichtigt, Ergebnis dieser Förderungen.

Inzwischen investieren auch zahlreiche Unternehmen in die MÜ. Vor allem Firmen, die technische Dokumentationen erstellen, die für den internationalen Vertrieb ihrer Produkte in die Sprachen der Exportländer übersetzt werden müssen, sehen hier immense Einsparmöglichkeiten. So setzt zum Beispiel Microsoft insgesamt 300 Lizenzen des in den vergangenen Jahren von der Siemens-Tochter Sietec vertriebenen EUROLANG- Übersetzungsspeichers ein, der Baumaschinenproduzent Caterpillar das an der Carnegie-Mellon Universität in Pittsburgh entwickelte Übersetzungssystem KANT. Die japanischen Firmen Fujitsu und Hitachi haben internationale firmeneigene Abteilungen für MÜ, deren Hauptaufgabe die Übersetzung von Handbüchern ist. Firmen wie Systran, CompuServe, Globallink oder die britische Firma Translation

1 Schrnitt, Peter A. (1993): Der Translationsbedarf in Deutschland: Ergebnisse einer Umfrage. In: Mitteilungsblatt für Übersetzer und Dolmetscher 39/5, 3-10.

Experts vertreiben nicht nur die von ihnen entwickelten Übersetzungssysteme, sondern bieten unter Einsatz ihrer Systeme sowie zahlreicher Humanübersetzer selbst auch *online* Übersetzungsleistungen an.

Während die in den vergangenen dreißig Jahren entwickelten maschinellen Übersetzungssysteme in der Regel besondere Hardwareerfordernisse stellten, die aufgrund der Anschaffungskosten für ein Übersetzungsbüro oder einen Privatanwender kaum finanzierbar waren, werden seit der Kommerzialisierung der MÜ-Systeme in den achtziger Jahren inzwischen auch Systeme für Workstations oder Rechner der PC-Klasse angeboten.

Neben den eigentlichen Übersetzungssystemen bieten sich als effiziente Hilfsmittel für die Fachtextübersetzung vor allem Übersetzungsspeicher und Terminologieverwaltungsprogramme an. Nachfolgend sollen einige Charakteristika solcher Systeme herausgegriffen werden, um die Leistungsfähigkeit verfügbarer maschineller Werkzeuge skizzieren zu können.

2 Merkmale maschineller Übersetzungssysteme

2.1 Lexikonkomponente

Auf dem Sprachsoftwaremarkt werden schon für etwa 50 DM Produkte angeboten, die als "Übersetzungssystem" oder "Translator" firmieren. Wie bei dem von der südafrikanischen Firma Pink Software vertriebenen Produkt handelt es sich hierbei häufig um Programme, die auf einer reinen Wort-zu- Wort-Übersetzung bzw. Substitution von Wörterbucheinträgen basieren. Die Wörterbücher enthalten vielfach lediglich Eins-zu-eins-Entsprechungen von ausgangssprachlichen und zielsprachlichen Einträgen.

Sowohl ein dem Übersetzer als Nachschlagwerk dienendes elektronisches Wörterbuch als auch ein einem maschinellen Übersetzungssystem zugrunde liegendes Wörterbuch muß jedoch Informationen enthalten, die dem Übersetzer bzw. dem System bei der Übersetzung eines Textes die Selektion des zutreffenden zielsprachigen Ausdrucks erlauben. Insbesondere bei großem Wortschatz stellt die Auswahl eines Übersetzungsäquivalentes ein erhebliches Problem dar, da mit wachsendem Wortschatz auch die Zahl der Homonyme steigt. Auf dem Markt befindliche Systeme, wie das von Intergraph vertriebene System TRANSCEND, der von der Firma v. Rheinbaben und Busch vertriebene Personal Translator oder Tl von Langenscheidt, bei deren Entwicklung diese Problematik reflektiert wurde, ermöglichen daher, bei der Übersetzung eines Textes bestimmte fachspezifische Teillexika als Filter auszuwählen, um die Zahl der Homonyme zu reduzieren. Wo dies nicht gelingt, werden dem Benutzer mehrere Alternativen, teilweise mit der dazugehörigen Wörterbuchdefinition, zur Auswahl angeboten.

2.2 Übersetzungsverfahren

Übersetzungssysteme unterscheiden sich danach, auf welchem Wege die Übersetzung von einer Sprache in eine andere realisiert wird, ob und in welchem Maße der menschliche Übersetzer in den Übersetzungsprozeß eingreifen muß und, wo dies erforderlich ist, an welcher Stelle im Übersetzungsprozeß dies geschieht. Für den praktischen Einsatz der Systeme ist schließlich von Bedeutung, welche Sprachpaare verfügbar sind und welche Übersetzungsrichtungen unterstützt werden.

Die meisten der heute verfügbaren leistungsfähigen Systeme basieren auf dem Transferansatz. Gegenüber dem direkten Übersetzungsmodell (Kern des SYSTRAN-Systems), das ausgehend von einer Quellsprache lediglich im Hinblick auf die Besonderheiten einer Zielsprache hin konzipiert ist, zeichnet sich der Transferansatz durch größere Flexibilität bei der Entwicklung von Modulen verschiedener Sprachpaare aus, gegenüber dem Interlingua-Modell (z.B. KANT) umgeht er die Schwierigkeiten einer sprachunabhängigen Interlingua-Repräsentation (siehe Abb.

1). Transfermodelle werden danach unterschieden, ob sie ausschließlich eine syntaktische Analyse vornehmen, auf deren Basis der Transfer in eine formale syntaktische Repräsentation der Zielsprache erfolgt (z.B. das System SHALT von IBM-Japan), ob zusätzlich ein semantischer Transfer vorgenommen wird (z.B. METAL, EUROTRA und das System HICAT der Firma Hitachi), wie dies etwa in Systemen der Fall ist, die eine Unifikationsgrammatik zur syntaktischsemantischen Beschreibung verwenden, oder ob die Übersetzung gar ausschließlich auf einem semantischen Transfer beruht. In neueren Ansätzen, wie dem von Whitelock als Shake-and-Bake bezeichneten lexikonbasierten Transfer, werden Vorteile des direkten Ansatzes mit jenen des syntaktisch-semantischen Transfers kombiniert (Grundlage von Entwicklungen der Firma SHARP).2

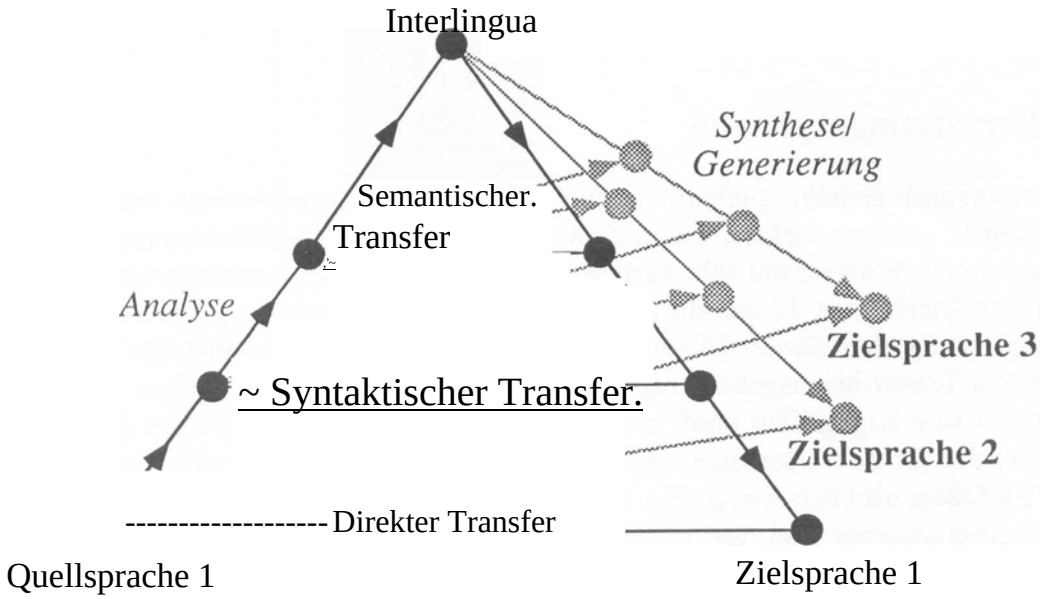


Abb. 1: Das Vauquoi'sche Dreieck zur Darstellung der verschiedenen Transferansätze zur maschinellen Übersetzung.

In einigen auf dem Markt verfügbaren Systemen kommen auch statistische Verfahren zur Anwendung, wie sie z.B. im Zusammenhang mit der Nutzbarmachung großer bilingualer Korpora für maschinelle Übersetzungssysteme entwickelt worden sind (vgl. das IBM-System CANDI DE).J

Hinsichtlich der Art der Beteiligung des menschlichen Übersetzers, d.h. der Stelle im Übersetzungsprozeß, an der der Benutzer eingreifen muß, werden Systeme unterschieden, die eine Präedition, d.h. eine vorbereitende Anpassung des zu übersetzenden Textes in bezug auf die Lexik und die vom System erkannten syntaktischen Strukturen erfordern, solche, die eine Postedition, d.h. eine nachträgliche Überarbeitung des maschinell übersetzten Textes durch einen menschlichen Übersetzer erfordern und interaktive Systeme, bei denen der Benutzer während des Übersetzungsprozesses aufgefordert wird, Ambiguitäten des Eingabetextes aufzulösen (z.B. das am GET A in Grenoble entwickelte dialogbasierte System LIDlA).

Die meisten der heute kommerziell verfügbaren Systeme erfordern eine Postedition des maschinell übersetzten Textes. Wie hoch der Aufwand der Überarbeitung durch den Übersetzer ist, hängt von der Qualität der vom System erstellten Übersetzung - auch als Rohübersetzung bezeichnet - ab.

2 Whitelock, Pete (1992): Shake-and-bake translation. In: Proceedings of the 14th International Conference on Cornputationa1 Linguistics (COLING-92), Nantes.
J Brown, P.F. et al. (1990): A statistical approach to Machine Translation. In Computational Linguistics, 16(2),79-85.

Bezogen auf den Automatisierungsgrad kann Übersetzung unter Zuhilfenahme der gegenwärtig auf dem Markt erhältlichen Systeme entweder als *Machine-Aided Human Translation* oder als *Human-Aided Machine Translation* qualifiziert werden. Von einer vollautomatischen maschinellen Übersetzung ohne Beteiligung eines menschlichen Übersetzers ist man heute weitgehend abgerückt. Der Grund hierfür ist in erster Linie in der geringen Qualität der maschinell erstellten Übersetzungen zu sehen. Diese hat ihre Ursachen in der in natürlichen Sprachen auftretenden Homonymie sowie in Problemen, die sich bei der Analyse äußerst komplexer syntaktischer Strukturen stellen. Im fachsprachlichen Bereich, wo in der Regel ein hohes Maß an terminologischer Normierung und zum Teil auch eine geringere Varianz im Aufbau von Satzstrukturen bzw. syntaktischen Konstruktionen besteht, können maschinelle Übersetzungssysteme deshalb am erfolgreichsten eingesetzt werden. Ein Ausweis hierfür ist das bereits in den siebziger Jahren in Kanada entwickelte Übersetzungssystem METEO, das bis heute Wetterberichte vom Englischen ins Französische übersetzt.

3 Übersetzungssysteme

Vollautomatisch erstellte qualitativ hochwertige Übersetzungen können mit den bisherigen Systemen nicht geleistet werden. Die auf dem Markt befindlichen Übersetzungssysteme, die in ihrer Funktionalität insgesamt sehr variieren, erfordern alle eine anschließende Überarbeitung des maschinell übersetzten Textes durch den menschlichen Übersetzer, jedenfalls wenn der Übersetzungsauftrag über eine Rohübersetzung zum Zweck der Beurteilung des Inhalts eines fremdsprachigen Textes hinausgeht. Aufgrund sehr beschränkter Wörterbücher, auf die beim Übersetzungsprozeß zugegriffen wird, und zum Teil nur äußerst rudimentärer syntaktischer Analysen kann der für den Übersetzer anfallende Aufwand für die Überarbeitung unter Umständen höher ausfallen, als übersetze er den Text von vornherein ohne Zuhilfenahme eines maschinellen Übersetzungssystems (vgl. die Übersetzungsergebnisse in Abb. 2). Sinnvoll wird der Einsatz eines solchen Systems erst dann, wenn dieses dem Benutzer beim Übersetzen eine zeitliche Ersparnis garantiert. Auch bei Systemen mit einer sehr weitgehenden linguistischen Analyse des Ausgangstextes setzt der ökonomische Nutzen häufig erst dann ein, wenn das dem System zugrunde liegende Basiswörterbuch durch ein benutzerdefiniertes Wörterbuch auf die individuellen Übersetzungserfordernisse abgestimmt wird. Eine solche, einer Trainingsphase vergleichbare Anlaufzeit muß in der Regel stets einkalkuliert werden, und zwar auch bei den die Preisskala anführenden Produkten wie SYSTRAN oder LOGOS (ca. 45.000 DM), die das Ergebnis jahrelanger Forschungs- und Entwicklungstätigkeit sind.

Deutscher Quelltext	Englische Übersetzung
Geben Sie bitte den Erläuterungstext für den WB-Abschnitt ein	Give please the @@Erläuterungstext for the @@Wb-paragraph a
Bitte ENTWEDER :1: ankreuzen ODER :2: eingeben.	PIe ase EITHER : 1: mark OR :2: enter.
Überleitung U-Nr. :1: nicht veraniaßt, da keine Bestätigung	Transition @@U-No. :1: does not cause, there no confmnation
Löschung im allgemeinen BP nicht erlaubt	Deletion in the general @@BP does not allow
Löschen der markierten Position: 1: zulässig	Delete the marked position: 1: permissibly
Löschung nicht ausgeführt, da keine Bestätigung	Deletion not implement, there no confirmation
Keine Löschung, nur Kostenaus-schluß bei :1: :2: möglich	No deletion, only cost exclusion at : 1: :2: possibly

Abb. 2: Maschinelle Übersetzung von Fehlermeldungen eines Immobilienfinanzierungsprogramms mit dem Power Translator von Globalink. (Von "@@" angeführt werden vom System nicht erkannte Zeichenketten.)

Nicht für alle Systeme stehen so viele Sprachpaare zur Übersetzung zur Verfügung wie bei SYSTRAN (derzeit 9 Sprachpaare, weitere in Vorbereitung) bzw. den inzwischen nicht mehr weiterentwickelten Sun-Versionen von METAL (ebenfalls 9 Sprachpaare) und LOGOS (7 Sprachpaare) oder dem von Intergraph bzw. HEI-SOFT und TRADOS vertriebenen System TRANSCEND (8 Sprachpaare). So ist das für PCs angepaßte METAL-System, das von der Firma GMS in München entwickelt und seit kurzem unter dem Namen Tl bei Langenscheidt vertrieben wird, derzeit lediglich für die Sprachpaare Deutsch-Englisch und Englisch-Deutsch erhältlich. Dasselbe gilt z.B. auch für den Personal Translator PT bzw. PT/plus, der von der Firma v. Rheinbaben & Busch Electronic Publishing in München vertrieben wird. Während alle Übersetzungssysteme in der Regel mindestens mit Englisch als Quell- oder als Zielsprache arbeiten, gegebenenfalls auch weiteren westeuropäischen Sprachen, sind etwa Russisch oder Japanisch nur äußerst selten verfügbar. Für Rechner der PC-Klasse wird Russisch mit Deutsch oder Englisch als Sprachpaar gegenwärtig nur von dem in Rußland entwickelten und von der Firma HEl-SOFT vertriebenen Übersetzungssystem STYLUS bearbeitet.4

Für Anwender komfortabel sind solche Übersetzungssysteme, die sich wie STYLUS oder Tl vom (gewohnten) Textverarbeitungsprogramm aus aufrufen lassen, da sie auf vertrauten Funktionen der Arbeitsumgebung aufbauen (siehe Abb. 3).

4 Es sei angemerkt, daß die Mehrzahl der auf dem Markt befindlichen Übersetzungssysteme auf PCs unter Windows laufen, während nur wenige wie STYLUS auch für Macintosh-Rechner erhältlich sind.

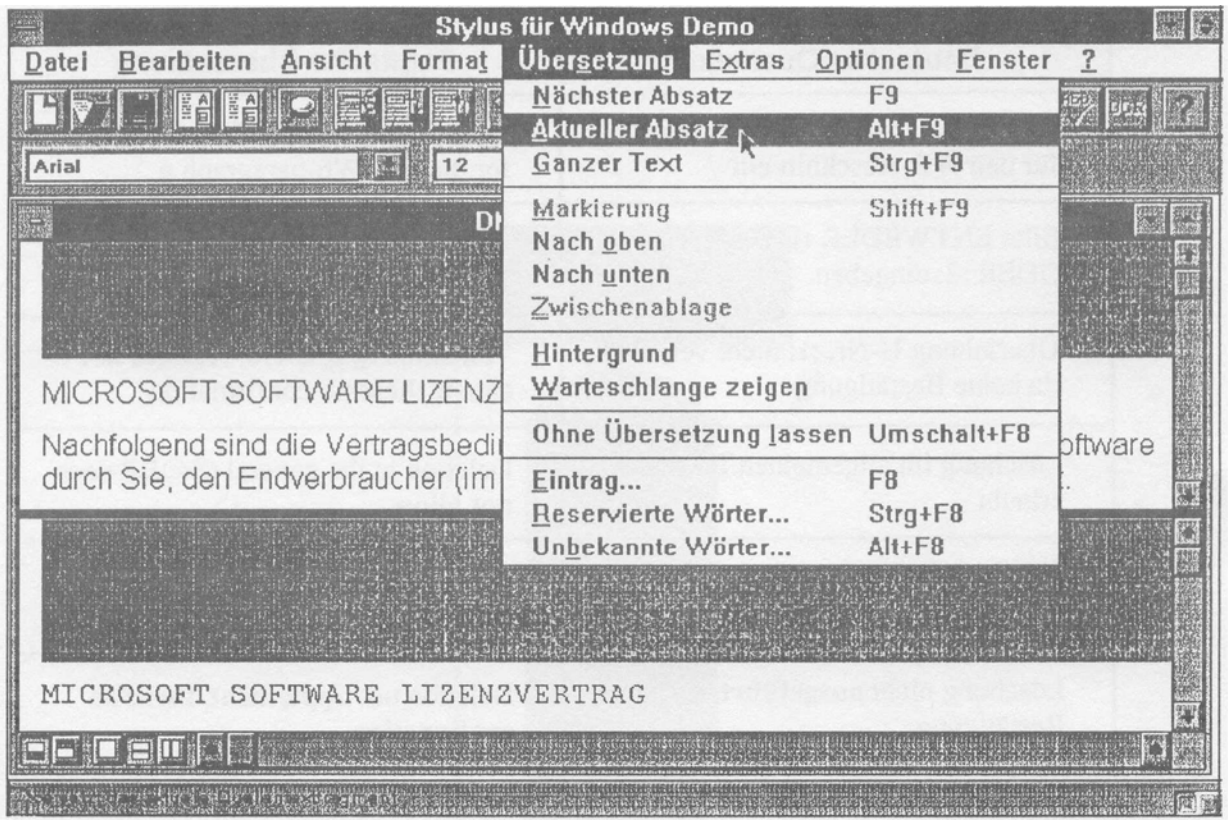


Abb. 3: Oberfläche des Übersetzungssystems STYLUS.

4 Übersetzungsspeicher

Insbesondere bei der Übersetzung großer Datenmengen oder einer Vielzahl von Texten derselben Textsorte haben sich Übersetzungswerkzeuge anderen Typs, sogenannte Übersetzungsspeicher oder *Translation Memories*, bewährt. Dabei handelt es sich um Programme, die in Ausgangstexten bereits übersetzter Dokumente nach identischen oder ähnlichen Textsegmenten suchen und die entsprechenden zielsprachigen Übersetzungen bei der Bearbeitung eines neuen Dokuments als Übersetzungsvorschlag anbieten. Die Systeme bedienen sich zur Ermittlung identischer bzw. ähnlicher Sätze unterschiedlicher Parsingalgorithmen sowie der *Fuzzy Logic* (Unscharfe Logik), wobei der Benutzer z.T. den Grad der Übereinstimmung von Sätzen, die bei der Übersetzung berücksichtigt werden sollen, selbst angeben kann. Die derzeit auf dem Markt befindlichen Translation Memories sind Entwicklungen von STAR in Böblingen (Transit), TRADOS in Stuttgart (Translator's Workbench), IBM in Stuttgart (Translation Manager) und Sietec in München (EUROLANG Optimizer). Die Systeme bestehen im wesentlichen aus drei Komponenten: dem eigentlichen Translation Memory, einer linguistischen Datenbank, in der einander zugeordnete ausgangssprachliche und zielsprachliche Übersetzungseinheiten gespeichert werden (Translator's Workbench, Translation Manager, EUROLANG Optimizer) bzw. Textsegmente verschiedener Dateien über Indizes einander zugeordnet werden (Transit), um sie für spätere Übersetzungen zur Verfügung zu haben; einer Wörterbuchkomponente, die eine vom Benutzer zu erstellende terminologische Datenbank enthält, deren Einträge mit dem ausgangssprachlichen Text verglichen werden und deren jeweilige Übersetzungen durch Betätigung von Symbolfunktionen mit der Maus in den Zieltext übernommen werden können; sowie einem Editor, der den Quelltext und den übersetzten Text in zwei verschiedenen Fenstern bzw. unterschiedlich markierten Bereichen am Bildschirm anzeigt (siehe Abb. 4), wobei Formatinfor-

mationen des Quelltextes in der Regel automatisch in den Zieltext übernommen werden. Während die Formatinformation bei Transit vor der Übersetzung extrahiert wird, so daß die Formatierung dem Übersetzer bei der Übersetzung nicht unmittelbar visuell zur Verfügung steht, werden mit Word oder WordPerfect erstellte Ausgangstexte beim Arbeiten mit der Translator's Workbench in ihrem Originallayout im Editor angezeigt. In den Text inserierte graphische Elemente, Textbausteine oder Referenzfelder erscheinen im Quelltexteditor der Translator's Workbench als besonders gekennzeichnete Symbole, die frei positionierbar per Mausklick in den Zieltext eingesetzt werden können, wodurch der Übersetzer auch bei der oft sehr zeitintensiven Layoutgestaltung der Übersetzung vom System unterstützt wird.

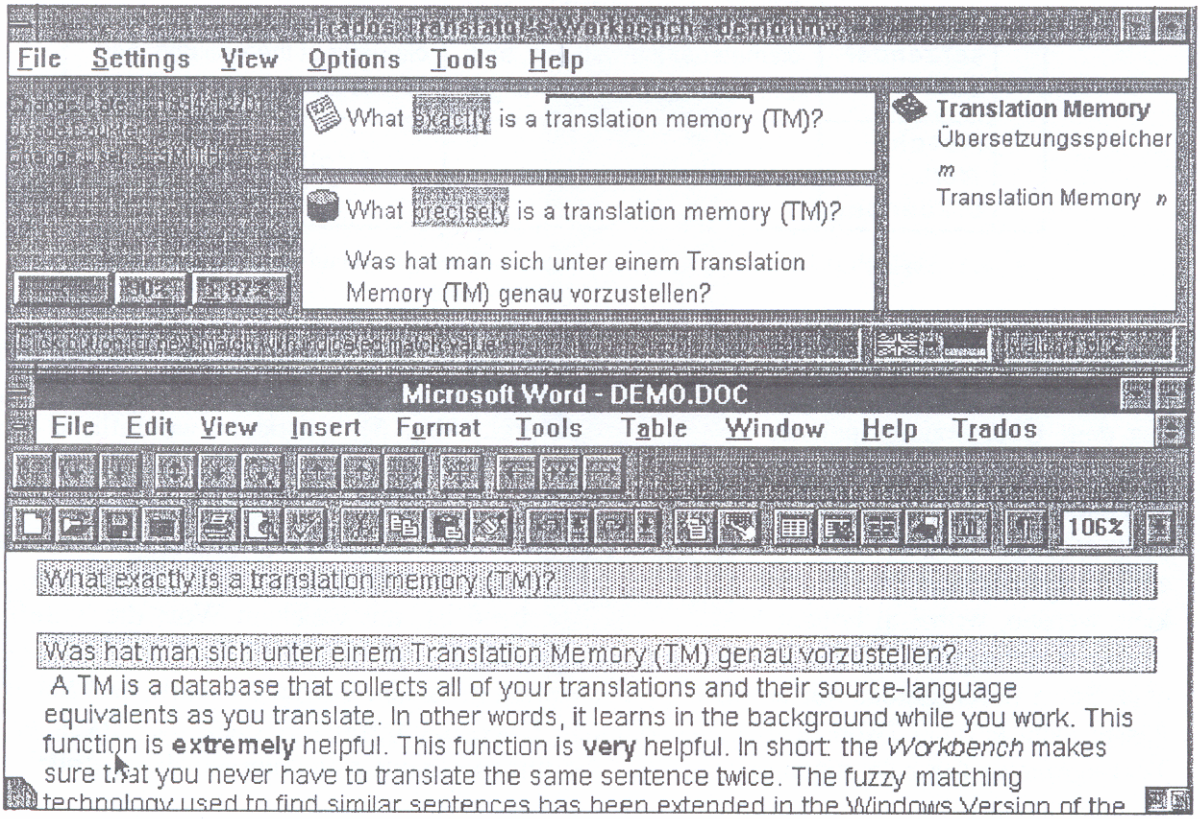


Abb. 4: Oberfläche der Translator's Workbench.

Übersetzungsspeicher erleichtern vor allem dort die Arbeit, wo ein wesentlicher Bestandteil der Übersetzungsarbeit darin besteht, beispielsweise Software-Handbücher zu übersetzen, die beim Updating der Software in der Regel überarbeitet bzw. neu übersetzt werden müssen.

Ein Vergleich verschiedener Systeme zeigt,[5] daß auch bei Übersetzungsspeichern zahlreiche Leistungsunterschiede bestehen, die bei der Ermittlung nur partiell übereinstimmender Sätze, bei sogenannten *Fuzzy Matches*, besonders deutlich werden. Grund dafür ist die unterschiedliche Berücksichtigung der Position bestimmter Wörter im Satz, insbesondere der Satzanfänge, die für den Abgleich zwischen dem zu übersetzenden Text und dem Dokument im Übersetzungsspeicher herangezogen werden. Entsprechend fallen die Markierungen von Auslassungen, Umstellungen, Austauschungen oder Einfügungen im Text im Vergleich zum Eintrag im Translation Memory von Übersetzungsspeicher zu Übersetzungsspeicher oft sehr unterschiedlich aus. Einige Beispiele mögen das anhand der Übersetzung mittels des Translation Ma-

[5] Nach Reineke, Uwe (1994): Zur Leistungsfähigkeit integrierter Übersetzungssysteme. In: Lebende Sprachen 3, 97-104.

nager/2 von IBM auf der einen und mittels der Translator's Workbench von TRADOS auf der anderen Seite verdeutlichen (Abb. 5 und 6).6

		Translator's Workbench	Translation Manager
1	Über-setzungs-speicher	The hoses are 100 mm in diameter at the wall end, and extend up to 200 cm in length.	The hoses are 100 mm in diameter at the wall end, and extend up to 200 cm in length.
	Zu über-setzender Satz	The hoses are 95 mm in diameter at the wall end, and extend up to 180 cm in length.	The hoses are 95 mm in diameter at the wall end, and extend up to 180 cm in length.
	Match-Wert / Erkennung	100 %	(Vollständige) Erkennung

Abb. 5: Abgleich zwischen einem zu übersetzenden Textsegment und einem Textsegment des Übersetzungsspeichers.

Während unterschiedliche Zahlenangaben in ansonsten identischen Textsegmenten von beiden hier gegenübergestellten Systemen erkannt und durch die entsprechenden Angaben des Satzes aus dem Übersetzungsspeicher ersetzt werden (vgl. Abb. 5), existieren wesentliche Unterschiede zwischen den Systemen bei der Erkennung von Sätzen, bei denen beispielsweise der Satzanfang modifiziert oder zentrale Wortgruppen gegenüber dem Textsegment im Übersetzungsspeicher erweitert worden sind (vgl. Abb. 6). Bei der Behandlung von Umstellungen ganzer Wortgruppen zeigt sich, daß diese von den Systemen unterschiedlich behandelt bzw. markiert werden, wobei in hohem Maße ausschlaggebend ist, mit welchem Wort die betreffende Wortgruppe beginnt. Da die Systeme keine syntaktische Analyse der Eingabesätze im eigentlichen Sinne durchfuhren, werden auch Ähnlichkeiten zwischen Sätzen im Aktiv und Sätzen im Passiv nicht erkannt (Abb. 6, Bsp. 4). Für den Übersetzungsprozeß spielt das Erkennen von Textsegmenten des zuletzt genannten Typs allerdings auch nur eine untergeordnete Rolle.

Beispielmaterial aus Reineke, Uwe, a.a.O.

		Translator's Workbench	Translation Manager
2	Über-setzungs-speicher	Insert the monitor's power cable into a suitable electrical outlet.	Insert the monitor's power cable into a suitable electrical outlet.
	Zu über-setzender Satz	Connect the monitor's power cable to an earthed electrical socket.	Connect the monitor's power cable to an earthed electrical socket.
	Match-Wert / Erkennung	SS %	Keine Ähnlichkeit erkannt
3	Über-setzungs-speicher	The recorder includes a multi-system tuner.	The recorder includes a multi-system tuner.
	Zu über-setzender Satz	The HR-D637MS video cassette recorder includes a multi-system tuner.	The HR-D637MS video cassette recorder includes a multi-system tuner.
	Match-Wert / Erkennung	74%	Keine Ähnlichkeit erkannt
4	Über-setzungs-speicher	In June 1984 she was sentenced to tour months' imprisonment by the judge.	In June 1984 she was sentenced to tour months' imprisonment by the judge.
	Zu über-setzender Satz	In June 1984 the judge sentenced her to four months' imprisonment.	In June 1984 the judge sentenced her to tour months' imprisonment.
	Match-Wert / Erkennung	Keine Ähnlichkeit erkannt	Keine Ähnlichkeit erkannt

Abb. 6: Verschiedene Erkennungsgrade beim Abgleich zwischen zu übersetzenden Textsegmenten und Textsegmenten des Übersetzungsspeichers.

Die Möglichkeit, den Wert des Fuzzy-Matches vom Benutzer festlegen zu lassen, gestattet, auch bei geringeren Übereinstimmungen zwischen dem zu übersetzenden Text und dem Text im Übersetzungsspeicher noch Übersetzungsvorschläge zu bekommen, wobei dann allerdings häufig aufgrund zu großer struktureller oder inhaltlicher Abweichungen Vorschläge vom Benutzer verworfen werden müssen. Im Fall der Translator's Workbench besteht außerdem die Möglichkeit, sich eine bilinguale Konkordanz von Textsegmenten mit einem vom Benutzer vorgegebenen Wort ausgeben zu lassen. Dies erlaubt, in bereits übersetzten Texten nach unterschiedlichen Übersetzungen eines Wortes in verschiedenen Kontexten zu suchen (vgl. Abb. 7).

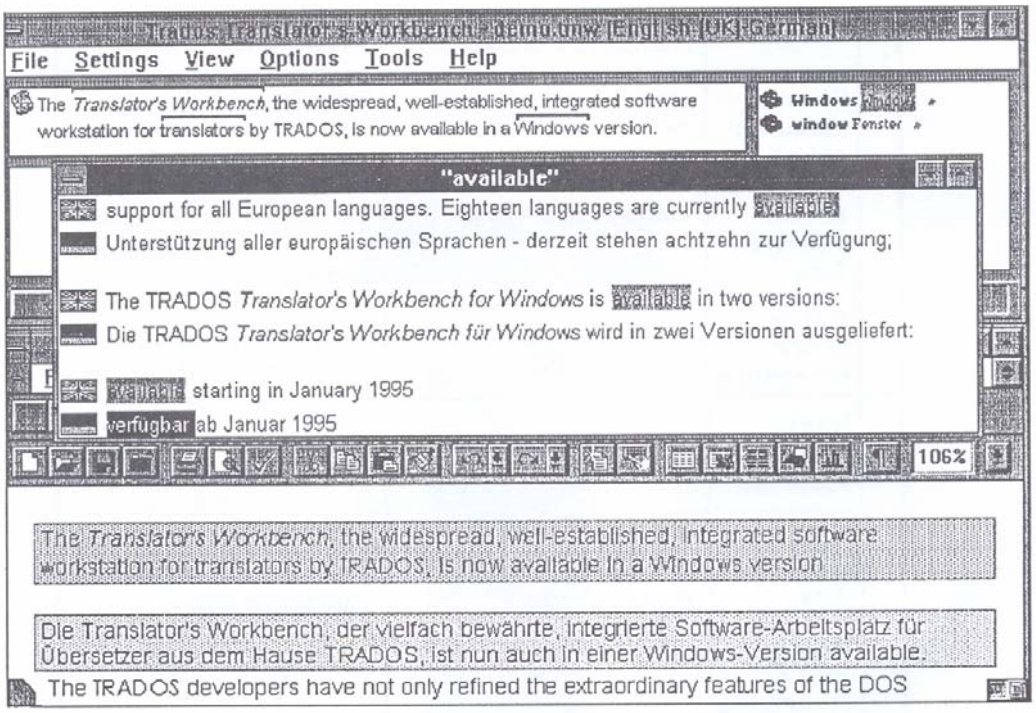


Abb. 7: Aus dem Übersetzungsspeicher erstellte Konkordanz mit dem Wort "available".

Eine besondere Leistungsfähigkeit erreichen die Systeme, wenn sie mit einem komfortablen Terminologieverwaltungssystem kombiniert werden, wie dies beispielsweise bei Transit mit der Terminologiedatenbank TermStar oder bei der Translator's Workbench mit MultiTerm der Fall ist. Diese Terminologiesysteme sind sowohl gekoppelt an das Translation Memory einsetzbar, wo die in den einzelnen Textsegmenten auftretende Terminologie jeweils in einem eigenen Fenster angezeigt und auch im Text markiert wird, oder wo Wörter aus dem Quelltext über Tastendruck oder Menüsteuerung in die Terminologiedatenbank übernommen werden können. Sie können darüber hinaus auch eigenständig bzw. von einem Textverarbeitungsprogramm aus aufgerufen werden und eignen sich für die Erstellung multilingualer Glossare, wobei neben Textdaten auch graphische oder Tondokumente als Einträge in die Datenbank aufgenommen werden können. Sowohl TermStar als auch MultiTerm unterstützen die Suche mit Teilzeichenketten, MultiTerm darüber hinaus auch Fuzzy Matching zwischen Suchbegriff und Einträgen in der Terminologiedatenbank. Derartige Funktionen tragen dazu bei, die Übersetzung von Wortableitungen und Komposita bzw. der darin enthaltenen Konstituenten bei der Übersetzung großer Textmengen konsistent zu halten.

Für Fälle, in denen in einem Translation Memory keine ausreichenden Übereinstimmungen zwischen zu übersetzendem Text oder Satz und dem entsprechenden Dokument im Übersetzungsspeicher bestehen, bietet sich der kombinierte Einsatz von Übersetzungsspeichern und maschinellen Übersetzungssystemen an (Abb. 8). Schnittstellen zwischen verschiedenen Systemen existieren bereits: die Translator's Workbench ist mit dem Übersetzungssystem TRANSCEND von Intergraph kombinierbar, das Übersetzungssystem der Firma LOGOS verfügt über eine Schnittstelle zum Übersetzungsspeicher EUROLANG Optimizer, SYSTRAN erlaubt die Einbindung des IBM Translation Memory. Einige Produkte haben Übersetzungsspeicher und maschinelles Übersetzungssystem inzwischen als Module eines Systems integriert (so z.B. das Übersetzungssystem Personal Translator PTplus von der Firma v. Rheinbaben & Busch in

München, das in Kooperation mit IBM entstanden ist), andere werden in dieser Richtung weiterentwickelt (so das aus METAL hervorgegangene Übersetzungssystem T1, das bis Ende 1996 auch mit einem Übersetzungsspeicher ausgestattet sein soll).

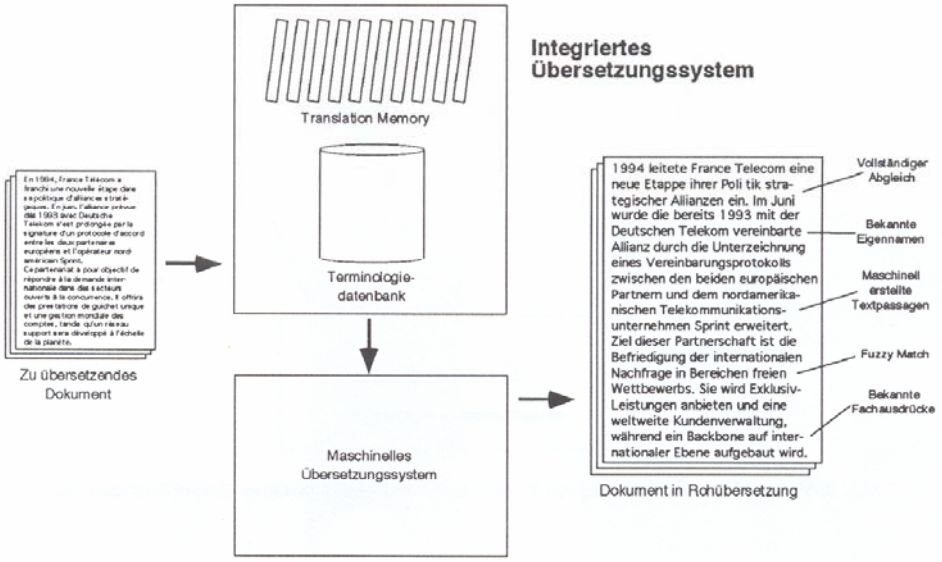


Abb. 8: Integriertes Übersetzungssystem.

5 Derzeitige Entwicklungen

Während die derzeit auf dem Markt befindlichen Systeme ausschließlich zur Übersetzung geschriebener Texte entwickelt worden sind, werden im Bereich der Forschung bereits Systeme konzipiert, die auch die Übersetzung gesprochener Sprache berücksichtigen. Japanische und amerikanische Projekte sowie das vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) geförderte Projekt Verbmobil sind hier anzusiedeln. Ziel des Verbmobil-Projektes ist die Entwicklung eines mobilen Systems, ähnlich einem Diktiersystem oder einem Laptop, mit dem Übersetzungen von Verhandlungsdialogen in direkten Gesprächssituationen möglich werden sollen. Das Verbmobil-System ist ein System zur Übersetzung gesprochener Dialoge, das Deutsch und künftig auch Japanisch als Ausgangssprache verarbeitet und die betreffenden Dialogsequenzen in das Englische übersetzt. Grundannahme dieser Vorgehensweise ist, daß Handelspartner, hier Deutsche und Japaner, sich heute in der Regel des Englischen als gemeinsamer Verhandlungssprache bedienen. An den Stellen, an denen die Dialogpartner jedoch nicht auf ihre Englischkompetenz vertrauen wollen, sollen sie das Verbmobil zur Übersetzung der betreffenden Dialogsequenz einsetzen können. Das Verbmobil-System erhält also gleichsam die Funktion eines Vermittlers zwischen zwei Dialogpartnern (Abb. 9). Allerdings hat das System — abgesehen von den Schwierigkeiten bei der Verarbeitung von gesprochenen Dialogen — die als Ziel für das Endprodukt formulierte Miniaturisierung noch nicht erreicht, denn zum gegenwärtigen Zeitpunkt läuft Verbmobil auf einer Sparc-Workstation mit Spezialprozessoren und 256 MB RAM. Das erstaunt jedoch nicht, wenn man berücksichtigt, daß hier nahezu alle Bereiche der Verarbeitung natürlicher Sprache und ihrer technologischen Umsetzung angesprochen sind.

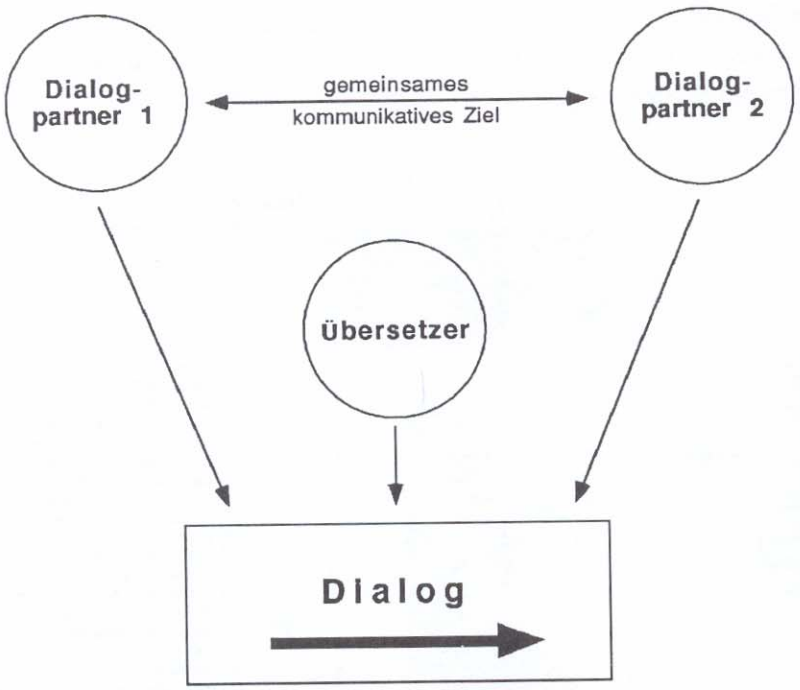


Abb. 9: Verb*mobil* — Übersetzung von Verhandlungsdialogen in direkten Gesprächssituationen.

Literatur

Alexandersson, Jan (1995): Plan Recognition in VERBMOBIL. Verbmobil-Report 81.

Boitet, Christian/Blanchon, Hervé (1994): Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup. In: Machine Translation 9, 99-132.

Brown, P.F. et al. (1990): A statistical approach to Machine Translation. In Computational Linguistics, 16(2), 79-85.

Minnis, Stephen (1994): A Simple and Practical Method for Evaluating Machine Translation Quality. In: Machine Translation 9, 133-149.

Reineke, Uwe (1994): Zur Leistungsfähigkeit integrierter Übersetzungssysteme. In: Lebende Sprachen 3, 97-104.

Schmitt, Peter A. (1993): Der Translationsbedarf in Deutschland: Ergebnisse einer Umfrage. In: Mitteilungsblatt für Übersetzer und Dolmetscher 39/5, 3-10.

Schwanke, Martina (1991): Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis. Berlin [u.a.]: Springer.

Seewald, Uta (1995): Antibabylonisch. Maschinelle Übersetzung — Marktübersicht: Kommerzielle Systeme und Werkzeuge. In: iX 12, 88-103.

Whitelock, Pete (1992): Shake-and-bake translation. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), Nantes.

DIE KONSTANZER LFG-UMGEBUNG

Bruce Mayo FG
Sprachwissenschaft Universität
Konstanz Postfach 5560 D
177 78434 Konstanz
bruce.mayo@uni-konstanz.de

Zusammenfassung: *Seit Mitte der 80er Jahre werden an der Universität Konstanz DCG-, PATR- und LFG-Grammatiken maschinell implementiert. Die Erfahrung hat gezeigt, daß die Umsetzung eines Grammatikentwurfs per Hand in maschinell interpretierbaren Code, z.B. Prolog, schwierig und fehleranfällig ist. Ein jetzt laufender Versuch, Grammatikentwürfe von Studierenden und Projekmitarbeiterinnen in einem abstrakten Formalismus direkt am Computer formulieren und testen zu lassen, verspricht einen besseren Erfolg. Einige Überlegungen zum Design des Systems, ihre Hintergründe und ihre Implementierung in unserem neuen System werden besprochen und mit anderen Implementierungen verglichen. Eine Besonderheit des beschriebenen Systems ist die Einbeziehung synchronischer und diachronischer Wortbildungsprozesse in ein modularisiertes Gesamtmodell der Sprachverarbeitung.*

Abstract: *Starting in the mid-1980s, grammars in the DCG, PATR and LFG formats have been implemented at the University of Constance. Experience has shown that translating a grammar specification into machine interrretable code, e.g., Prolog,* is *difficult and error-prone. A current attempt to let students and research co-workers formulate and test grammar sketches in an abstract formalism directly on the computer promises greater success. Some of the design considerations, background and the implementation in our new system are discussed and compared to other implementations. A special feature 01 the described system* is *its modularized processing model, which takes synchronic and diachronic derivation al processes into account.*

1 Hintergründe

Während Fortschritte in der Computergraphik dazu geführt haben, daß für Architekten Reißbrett und Lineal schon weitgehend zur Vergangenheit gehören, läßt eine vergleichbare "Computerisierung " der deskriptiven Linguistik auf sich warten, obwohl sie schon vor einem Jahrzehnt angekündigt wurde (z.B. in Shieber 1985). Für viele deskriptiv orientierte Linguisten bleibt dieser Schritt noch eine Chimäre, denn größere Lexika und Syntaxwerkzeuge, die für einsatzfähige Implementierungen notwendig wären, hat die Computerlinguistik bislang in nur wenigen Fällen hervorgebracht. Noch letztes Jahr hat eine Internet-Umfrage nach großen Implementierungen im Rahmen der Lexikalisch-Funktionalen-Grammatik (LFG) jeweils nur eine größere Grammatik für Englisch, Französisch und Deutsch ergeben, wobei die deutsche Grammatik noch als "in progress" beschrieben wurde, und entsprechend vollständige Lexika noch nicht vorhanden waren. [1]

[1] Die großen kommerziellen Systeme zur automatischen Textübersetzung und -indizierung scheinen dieser Behauptung zu widersprechen, aber sie werden in der Forschungsliteratur seltsamerweise kaum wahrgenommen. Vgl. Seewald 1995.

Der ausbleibende praktische Erfolg computerlinguistischer Werkzeuge hat sicherlich nicht mit einem Mangel an implementierbarem Wissen über einzelne Sprachen zu tun. Für die weit verbreiteten europäischen Sprachen gibt es eine beträchtliche Fülle von sehr detailliertem Beschreibungsmaterial in Form von Grammatiken, Lexika und deskriptiven Detailuntersuchungen, und wahrscheinlich ist sehr viel mehr verfügbar, als in irgendeiner Implementierung bislang untergebracht wurde. Was immer noch fehlt, ist eine benutzerfreundliche Brücke vom Wissen der deskriptiv orientierten Linguistik zu den bereitgestellten, aber bislang wenig benutzten Werkzeug en der Computerlinguisten. So bleiben Textverarbeitungs- und Datenbankprogramme wohl die einzigen Hilfsmittel, die die Arbeit der meisten deskriptiven Linguisten wirklich weiterbringen.

Einige Erfahrungen mit linguistischen Implementierungen in Projekten an der Universität Konstanz unterstützen die Ansicht, daß für die Linguistik viel getan wäre, wenn der mühsame Schritt von intuitiver linguistischer Beschreibung zur Formalisierung gekürzt oder so weit wie möglich eliminiert werden könnte - wenn der Linguist (öfter die Linguistin) direkt *am* Computer Beschreibungen entwickeln könnte, etwa wie ein Architekt, der seine Pläne direkt an der Workstation entwirft, ändert und testet. Die Nachteile einer Computerlinguistik ohne solche Hilfsmittel zeigte ein Projekt der 80er Jahre in unserem Institut. Ein nicht-triviales Fragment des Französischen wurde von Linguisten beschrieben und anschließend von Programmierern in erweiterte Definite-Clause-(DCG) Regeln umgesetzt (Mayo 1995). Es wird nicht überraschen, daß die implementierte Grammatik manchmal nicht die Analysen lieferte, die die Linguisten erwartet haben, aber sie war inzwischen auf eine Größe gewachsen (32 Kbyte unkommentierten Prolog-Code), die einen sicheren Vergleich mit der linguistischen Analyse nicht mehr zuließ. Es war kaum mehr möglich festzustellen, ob die ursprüngliche linguistische Arbeit subtile Fehler bzw. Inkonsistenzen enthielt, oder ob lediglich die Umsetzung in DCG inkorrekt war, und folglich blieb die erhoffte Verifizierung der linguistischen Analyse aus. Daraufhin wurde beschlossen, solche Beschreibungen nicht mehr per Hand, sondern nur über einen Compiler zu übersetzen, so daß eine Beschreibung möglichst interaktiv mit unmittelbarer Beteiligung der Linguisten entwickelt werden konnte.

Das damit begonnene Projekt,2 das aus einer früheren Zusammenarbeit mit dem Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart hervorging (Eiseie & Dörre 1986), wurde auch von der "Grammar Writer's Workbench" (fortan GWB) inspiriert, die seit mehreren Jahren am Xerox P ARC entwickelt wird und jetzt auch für andere Forscher freigegeben worden ist (Kaplan & Maxwell 1993). Die GWB setzt einen mächtigeren Rechner voraus, als den, den wir zur Verfügung hatten, und legt andere Annahmen über die Gesamtstruktur der '"Sprachverarbeitung fest, als wir. Die GWB ist aber äußerst flexibel, komfortabel und ausgereift, und sie wird in naher Zukunft bestimmt einen Standard für computerunterstützte Linguistik setzten. Im Hinblick darauf werde ich abschließend die Beschreibung unseres Systems durch einige Hinweise auf die GWB ergänzen.

2 Was sollte eine linguistische Werkbank leisten?

Der heute stattfindende Einzug des Computers in die diversesten Fachgebiete geht mit der Entwicklung von interaktiven, graphisch-orientierten Oberflächen einher, die den meisten Computerbenutzern heute beinahe selbstverständlich sind. Es liegt auf der Hand, daß auch ein Hilfsmittel für Linguisten das Gedächtnis möglichst wenig mit abstrusen Kommandos taxieren sollte. Ferner sollte es sprachbeschreibende Informationen möglichst in der Form aufnehmen, in der die Benutzer sie bereits formuliert, und es sollte die daraus folgenden Analysen möglichst

2 Das Computer-Programm ist Teil eines von der Deutschen Forschungsgemeinschaft finanzierten Projekts, "Semantik der Derivation". Für einige Vorschläge bin ich dem Leiter des Projekts, Christoph Schwarze dankbar; verbleibende Mißverständnisse gehen zu meinen Lasten.

schnell und bildhaft darstellen. Dabei bleibt noch offen, welche Informationen ein Benutzer in das System einbringen sollte: Sind nur Lexikoneinträge, Wort- und Satzbauregeln erforderlich, oder sollte eine Werkbank auch so flexibel sein, daß der Benutzer auch diverse Strategien der Sprachverarbeitung und -organisation spezifizieren kann? Auf diese Fragen werde ich später zurückkommen.

Die wichtigste Überlegung für eine linguistische Werkbank ist aber die Wahl des Darstellungsformalismus, denn der Formalismus muß den Weg von den oft inexakten, intuitiven Analysen des deskriptiven Linguisten zu einem lauffähigen Computerprogramm leicht begehbar machen. Einige Veröffentlichungen des letzten Jahrzehnts – nehmen wir stellvertretend Gazdar & Mellish 1989 – erwecken den Eindruck, daß die Beschreibung natürlicher Sprachen am Computer Vertrautheit mit Programmiersprachen verlangt. "Implementierte Linguistik" b rau c h t aber ebensowenig Kenntnis informatischer Fundamente wie workstation-basierte Architektur. Sprachverarbeitende Systeme der Zukunft werden wahrscheinlich weder in Prolog noch in LISP noch in PATR-II erstellt werden, sondern Formalismen einsetzen, die auf kompakte und klare Weise die Fakten darstellen, die deskriptive Linguisten für eine Sprache schnell und zuverlässig ermitteln können. Zur Zeit gelten wohl HPSG und LFG als wichtige Schritte in Richtung solcher Formalismen, obwohl selbst diese nicht leicht zu beherrschen sind. Ich vermute, daß große Systeme der Zukunft erst mit Hilfe von Dialogschnittstellen entstehen werden, die über Abfragen zur Klassifizierung, Sinnrelationen, Kontextbeispielen usw. sprachliches Wissen induzieren können. Einige Werkzeuge dieser Art werden bereits in den großen Systemen zur maschinellen Übersetzung (z.B. Logos) eingesetzt. Eine ähnliche Entwicklung kann man heute mit den sehr ausgereiften Datenbankprogrammen beobachten, die durch raffinierte Benutzerschnittstellen die Datenverwaltung und Geschäftsabläufe eines Betriebs übernehmen können, ohne daß je ein Programmierer zu Rat geholt werden muß.

3 Syntax und Lexikon in der Konstanzer LFG-Werkbank (KLU)

Solchen Überlegungen folgend, wurde die Konstanzer LFG-Umgebung (KLU) als ein Werkzeug konzipiert, das Linguisten und Linguistinnen helfen sollte, Beschreibungen kleinerer Sprachfragmente am Rechner zu entwerfen und zu testen, frei von den Schwierigkeiten und Tücken der unmittelbaren Programmierung. Die Werkbank sollte nicht mehr an Computerkenntnissen voraussetzen, als ihre Benutzer für den Umgang mit Textverarbeitungsprogrammen meistens schon besitzen. Als linguistisches Beschreibungsmittel hat sich der LFG-Formalismus nach Bresnan 1982 angeboten, deckt er sich doch gut mit den Begrifflichkeiten wie Konstituenz- und Dependenzanalyse, die auch weniger formal orientierte Philologen in einer Einführung in die Syntax kennenlernen. Außerdem läßt sich der Kern dieses Formalismus nach einer gut bekannten Strategie (Eisele & Dörre 1986) mit wenigen Schritten in einen erweiterten DCG-Formalismus übersetzen, der in Prolog als Parser sofort ausführbar ist.

Zur Vergegenwärtigung und für Leser, die den LFG-Formalismus nicht bereits kennen: Die Konstituenzfakten werden in LFG als kontextfreie Produktionsregeln und die Dependenzrelationen mit zusätzlichen Gleichungen notiert, die den DCG-Regeln als zusätzliche Prolog-Klauseln hinzugefügt werden. Stellt ein Linguist bei der strukturellen Untersuchung einer Sprache Z.B. fest:

(1) eine Nominalphrase besteht syntaktisch aus
einem Determinator,
einer optionalen Adjektivphrase, die eine modifzierende
Funktion in bezug auf die NP hat, und
einem Nomen ohne die Eigenschaft "Eigenname"

so läßt sich diese intuitive Beschreibung fast eins-zu-eins in die LFG-Notation von (2) umsetzen:

(2)
$$\text{NP} \sim \underset{t=,\}}{\text{Det}} \quad \left(\underset{(t\text{MOD})=,\}}{\text{AdjP}} \right) \quad \underset{\substack{t=,\} \\ -,(\ ,\}\ \text{PROPER})}}{\text{N}}$$

wobei die Gleichung $t = ,\}$ besagt, daß die Konstituenten Det und N als Köpfe der Phrase fungieren, d.h., daß alle ihre Eigenschaften in die übergeordnete NP eingehen. In unserer, wie in anderen LFG-Werkbanksystemen, werden die Reihen des LFG-Formats gegen Spalten getauscht, und die Sonderzeichen durch Zeichen eines üblichen Schriftsatzes ersetzt, so daß Z.B. im Konstanzer Format (2) wie in (3) geschrieben wird.

(3) NP->
```
Det     A = V
[ AdjP (A MOD) = v ]
N       A = V -(v PROPER) .
```

Im Konstanzer System wird eine Grammatik als Textdatei mit einem eingebauten (oder externen) Texteditor erstellt, und sie wird durch Auswahl einer Menüfunktion der Werkbankumgebung compiliert. Aus der LFG-Regel (3) erzeugt unser Compiler den Prolog lI-Code von (4):

(4)
```
Stx:NP(s1, s99, f, _done) ->
        lex_insertn(:SRCH_CTX, "Det",s1, s2, f1, _done)
          merge(f, f1)
          (AdjP(s2, s3, f2, _done) % fakultative Konstituente
                AND merge(f, (MOD.w1).r1) AND merge(f2, w1)
                OR 'e'(s2,s3,_,-> )        % leere Konstituente
        lex_insertn(:SRCH_CTX, "'N"',s3, s4, f3, _done)
          merge(f, f3)
          neg_constraint(f3, PROPER, v585, _done)
        eq(s99, s4)
        pars_protk(" Stx:NP ", s1, s99, f); %ENDE NP
```

Natürlich lassen sich nicht alle Beobachtungen über eine Sprache so leicht formalisieren. Es wird aber sicherlich einleuchten, daß für einen Linguisten die Übersetzung der deskriptiven Information von (1) zu (3) viel einfacher und zuverlässiger ist, als sie von (1) zu (4) wäre.3

Zusätzlich zu den Syntaxregeln enthält eine formale Grammatik gewöhnlich auch ein Lexikon, in dem die syntaktische Kategorie, die Morphologie, die Bedeutung usw. der einzelnen Wörter angegeben sind. In unserem (nicht-standard gemäßen) Format hat der Lexikoneintrag eine semantische Beschreibung, die für ein Verb wie *geben* folgendermaßen aussehen kann:

3 Das KLU-System ist im Prolog 11+ von Prolog/A, Marseille, implementiert. Aus historischen Gründen wird noch Prolog 11- statt ISO-Syntax erzeugt. Unterschiedliche Teile der Sprachbeschreibung (z.B. Syntax, Wortbildung, Wissenskonzepte) werden auf hierarchisch geordnete Module verteilt, und alle Symbole gehören implizit dem ranghöchsten Modul des Sprachverarbeitungsmodells an, in dem sie vorkommen. Regeln aber gehören explizit zu dem Modul, in dem sie deklariert werden ("Stx:NP" z.B. weist die Regel NP dem Syntax-Modul zu). Terminale Symbole (lexikalische Kategorien) werden nicht direkt, sondern durch die Regel lex Jnsertn!6 eingefügt, die eine evtl. notwendige morphologische Analyse außerhalb der Satzsyntax durchführen kann. oe' ist die vorvereinbarte leere Konstituente, die als Alternative zu AdjP die Optionalität „[AdjP]" realisiert. Um Benutzerprotokolle möglichst übersichtlich zu halten, werden Unifikation (mergel2; vlg. unifY/2 in Gazdar & Mellisch 1989:235) und Existenzforderungen (constraintl4, neg_constraintl4) so weit wie möglich parallel zur Konstituentenanalyse durchgeführt. Die GWB verfolgt die entgegengesetzte Strategie: Gleichungen werden erst nach der Konstituentenanalyse gelöst. pars -protk/4 ist für ein einschaltbares Ablaufprotokoll.

```
(5) § Geben(a,t,z) => ereignis(e,TRANSFER(a,t,z)
                        & agens(e,a)
                        & thema(e,t)
                        & ziel(e,z) ).
```

Die Beschreibung besagt, daß *geben* ein dreistelliges Transfer-Prädikat mit den Rollen Agens, Ziel, und Thema ist; ferner, daß es als temporales Ereignis instanziiert wird. Für ein Verb des gleichen lexikalischen Feldes wie *erhalten* ist der Eintrag erwartungsgemäß ähnlich, außer daß die Argumente der TRANSFER-Handlung anders realisiert sind, wie in:

```
(6) § Erhalten(z,t,q) => ereignis(e, TRANSFER(q,t,z) &
                                    ziel(e,z)
                                    & thema(e,t)
                                    & quelle(e,q) ) .
```

Die semantische Beschreibung kann beliebig detailliert sein, sie muß aber auf jeden Fall im Regelkopf so viele Argumentstellen spezifizieren, wie sie in die Syntax projezieren soll.

Es wäre unklug, den Werkbank-Formalismus so zu gestalten, daß die Lexikoneinträge nur semantische Informationen, wie Rollen, Bedeutungspostulate, usw. tragen könnten, wie bei (5) und (6). Sicherlich ist es in der generativen Literatur üblich, die thematischen Rollen unmittelbar in die Satzstruktur einzusetzen (es kommen allerdings zusätzliche Markierungen hinzu, die die späteren Oberflächenstellungen steuern sollen). Nach der lexikalisch-funktionalen Theorie wird die Einsetzung der thematischen Rollen nicht in der Syntax, sondern durch sogenannte lexikalische Regeln vermittelt, die einen von der Satzsyntax getrennten, eigenen Bereich der Grammatik bilden. Jedoch kann uns im Moment keine der bekannten Grammatiktheorien sehr viel über diesen Projektionsschritt sagen, und es ist evident, daß neben regelmäßigen, syntaxähnlichen Prinzipien auch willkürliche Entscheidungen der Lexikalisierung eine Rolle spielen: Bekanntlich kann in (5) das Ziel-Argument *z* von engl. *give* als NP realisiert werden *(give the schoo/ a microscope),* für das fast synonyme *donate* nicht, sondern nur in einer PP mit *to (donate a microscope to the schoo!).*

Um diese Lücke der heutigen Theorien zu berücksichtigen, ist es unumgänglich, zusätzliche Information ins Lexikon einzuführen, um die nicht vorhersagbaren syntaktischen Eigenschaften eines Wortes festzulegen. Der LFG-Formalismus kommt dieser Forderung mit einer expliziten syntaktischen Repräsentation nach; im KLU-Format sieht der syntaktische Teil des Lexikons folgendermaßen aus:

```
(7) /gEb/ V, Kea, (A PRED) = "Geben«ASUBJ),(AOBJ),(AOBJ2»"
                   (AOBJ2 CASE) =c DA T .
```

Hier erscheinen phonologische und morphologische Beschreibungen des Stamm-Morphems /gEb/ mit der Angabe eines Konjugationsparadigmas "Kea". Die Großschreibung mit "E" in /gEb/ spezifiziert, daß Umlaut stattfindet. Außerdem wird festgehalten, daß die Argumente von (5) auf entsprechende grammatische Funktionen (SUBJ[ekt], OBJ[ekt], usw.) stellenweise abgebildet werden.

Je nach theoretischer Überzeugung kann man diese grammatischen Funktionen als sprachliche Universalien oder schlicht als eine Berechnungshilfe für die Einsetzung der thematischen

Argumente betrachten.4 Für eine spätere Ausbaustufe unseres Systems ist vorgesehen, daß der syntaktisch-morphologische Teil des Eintrags (7) auf (7') reduziert werden kann:

(7') *lgebl* V, Kea, (A PRED) = "Geben<>" .

In dieser Form werden die morphologischen Daten und die Flexionsklasse des Eintrags wie in (7) festgehalten, aber eine explizite Spezifizierung der syntaktischen Argumente (z.B. (ASUBJ»), der Kasus-Markierungen (AOBJ2 CASE) =c DA T) und der Kontrollrelationen verschwinden. Es bleibt nur ein leerer Verweis auf das semantische Prädikat. Angenommen, daß die syntaktischen Informationen, die in (7) erscheinen, nach Regeln einer "mapping "-Theorie ableitbar sind, wird der Lexikon-Compiler dafür sorgen, daß diese zur Compile-Zeit automatisch erzeugt werden, d.h., daß (7') zu (7) kompiliert wird. Explizite Angaben im syntaktischen Eintrag haben dennoch Priorität vor den Regeln der Mapping-Theorie. Sind also Teile der syntaktischen Beschreibung unregelmäßig, d.h., nicht ableitbar (Blockierung des OBJ2-Argumentes bei engl. *donate),* muß man die Unregelmäßigkeiten im syntaktischen Teil des Eintrags explizit angeben, wie in (8), um eine reguläre Zuweisung von grammatischen Funktionen zu verhindern.

(8) *Idonatel* V, Reg, (A PRED) = "Give«ASUBJ),(AOBJ),(AOBL-TO»"

Die Benutzer unseres Systems müssen also ihr intuitives, deskriptives Wissen in den Formalismen annotierter Produktionsregeln wie (3), horn-klausel-ähnliche Logik wie (5) und syntaktisch-lexikalischer Einträge wie (7) umsetzen. Hinzu kommen Regel-Formate für lexikalische Ableitungen (7) zu (7') und allomorphische Variationen, die aber noch in der Entwicklung sind. Über diese rein deklarativen Spezifikationen hinaus verlangt KLU auch eine modulare Unterteilung des sprachlichen Wissens, und in diesem Punkt unterscheidet sie sich am stärksten von der GWB und anderen mir bekannten Systemen. s

4 Die modulare Struktur von KLU

Ein oft gepriesenes Ideal rein deklarativer Formalismen wie PATR-II, HPSG und der heutigen LFG ist, daß sie ihren Benutzern erlauben, ihr linguistisches Wissen möglichst ohne Angaben des Wo und Wann für die Verarbeitung dieses Wissens zu spezifizieren. Unsere Konstanzer Werkbank rückt etwas von diesem Ideal ab, indem sie ihre Benutzer zwingt, durch Modularisierung gewisse prozeßbezogene Fragen der Sprachverarbeitung und der Sprachentwicklung in die Sprachbeschreibung einzubeziehen. Zu diesem Schritt führte nicht ein einziger, entscheidender Grund, sondern eine Vielzahl von Überlegungen unterschiedlicher Art. Aus der Sicht des Software-Engineerings ist es immer erstrebenswert, ein größeres Programm in funktionale Module aufzuteilen, um Namenskonflikte zu vermeiden und funktionale Dependenzen lokal zu halten. Aus linguistischer Sicht schien es wichtig, viele Begriffe nach Verwendungsbereich zu differenzieren, wie z.B. grammatisches und semantisches Genus bzw. semantische und kognitive Prädikate. Die Modularisierung hat uns schließlich dazu verholfen, die Grammatik nicht bloß als statischen, mathematischen Formalismus, sondern auch als eine Darstellung sprachlicher Prozesse in bezug auf "Mikro- und Makrodiachronie" zu sehen (mehr dazu im Abschnitt 5).

4 Ursprünglich wurden die grammatischen Funktionen als Primitiva der Universalgrammatik postuliert, aber viele neuere Arbeiten, die der LFG zuzurechnen sind, betrachten sie eher als Kürzel für einen Komplex aus syntaktisch-semantischen Merkmalen. Vgl. Kiss 1993.

5 Weitere Informationen übers Internet zu LFG-Implementierungen:
Charon-System des IMS, Stuttgart: ftp.ims.uni-stuttgart.de, Verzeichnis /pub/Charon
System von A Andrews: Avery.AndreWS@anu.edu.au
Allgemeiner Überblick: http://clwww.essex.ac.uk/LFG/systemsl

Die Modulstruktur ist aber auch ein wesentlicher Aspekt der Benutzeroberfläche, die helfen soll, den umfangreichen Datenbestand einer Sprachbeschreibung zu organisieren. Aus der Sicht des Benutzers von KLU besteht die Beschreibung einer Sprache aus sieben Modulen, wovon sechs als Textdateien erstellt werden und eine explizite Darstellung im Menü „Module" (vgl. Untermenü „Selection" von Abb. 1) haben. Das siebte, nicht dargestellte Modul ist das „System", zu dem einige Funktionen gehören, die als universale Bestandteile der Sprachverarbeitung vom Benutzer nie spezifiziert und nie geändert werden; dieses siebte Modul enthält solche Regeln und Symbole, die z.B. mit dem Parsen, mit der Einfügung lexikalischer Einheiten, oder mit der semantischen Auswertung von Fragen und Behauptungen zu tun haben.

Die Module sind in drei Bereiche aufgeteilt: Zum ersten Bereich gehört eine sprachunabhängige aber vom Benutzer erstellte Wissensrepräsentation als Modul „Cog[nitive]. concepts". Im zweiten Bereich (unter der ersten gestrichelten Linie) liegen Module, die mit der Satzanalyse zu tun haben. Drei davon, „Lex[ical]. Semantics", „Lex. Operatives" und „Syntax", werden vom Benutzer als Teile der Sprachbeschreibung erstellt, während zwei, „Discourse rep[resentation]." und „Temp[orary] concepts" erst im Laufe der Satz- bzw. Textanalyse Inhalte bekommen, die aber für die Verarbeitung weiterer Sätze Bedeutung haben können. Der dritte Bereich „Morphotactics" hat ausschließlich mit Wortbildung zu tun. Er enthält Wortbildungsregeln und den semantisch leeren Teil des Lexikons, der aus Flexionsmorphemen besteht. Das letzte Modul „Allomorphy" beschreibt Unregelmäßigkeiten der Orthographie, die z.B. mit Umlaut, Klitisierung usw. zu tun haben können.[6]

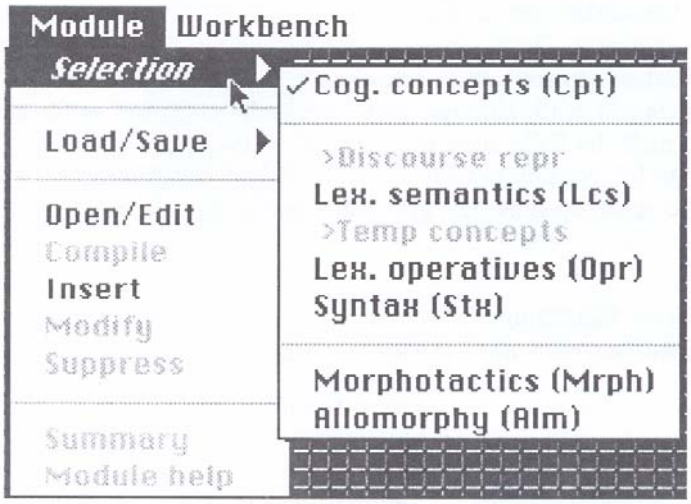


Abbildung 1. Das Modulmenü

4.1 Modul Kognitive Konzepte „Cog. concepts"

Im leeren Zustand, d.h., bevor das Werkbank-System eine Sprachbeschreibung compiliert bzw. geladen hat, erscheint in dem Untermenü „Selection" von Abb. 1 nur „Cog. concepts" (für Kognitive Konzepte) unschattiert, also verfügbar. Somit wird der Benutzer, bevor er anfängt, gezwungen irgendeine S p r a c h e zu beschreiben, wenigstens als Skelett ein Modul zu schreiben, das die W e l t des sprachlichen Diskurses beschreibt - eine kleine, sprachneutrale Wissensbasis. Zum Glück haben wir für unsere Lehrgrammatiken die Vereinbarung, daß unsere

[6] Allophonische, allomorphische und orthographische Angelegenheiten werden alle salopp unter der Bezeichnung „Allomorphy" behandelt. Klitisierbare Lexeme werden im Modul „Morphotactics" spezifiziert und der Satzsyntax zur Analyse angeboten.

Linguistenwelt - ob für Spanisch, Italienisch oder Französisch - nur eine kleine Märchenwelt von Feen, Zwergen, Schlössern und Rittern ist, so daß dieses Modul immer klein und immer das gleiche ist. Wir benutzen die Wissensbasis z. Zt. hauptsächlich, um konzeptuelle und Rollen-Attribute abzufragen; da sie aber in Prolog erstellt wird, kann sie beliebig detailliert werden.

Soll die Semantik eines Verbs etwa verlangen, daß ein Ritter als Argument agentive Eigenschaften haben soll, kann die Wissensbasis die Eignung zum selbständigen "agentiven" Handeln über solche Fakten und Regeln wie in (9) feststellen:

(9) a. agens(X):- concepCprops(X, P) & member(caput, P).
b. concepCprops('EQUES', [caput, masculus, fortis]).

Diese besagen, eine Entität X ist agentiv, falls zu der Menge ihrer Eigenschaften die Eigenschaft Person (lat. *caput)* gehört (9a). Da einem Ritter (lat. *eques)* diese Eigenschaft in concept_props/2 zugeschrieben wird (9b), kann ein Ritter als Agens fungieren.

4.2 Lexikalische Semantik "Lex. semantics"

Im Gegensatz zum Weltwissen soll das Modul Lexikalische Semantik nur Informationen enthalten, die durch das Erlernen einer Sprache erworben werden, und die von Sprache zu Sprache auch unterschiedlich sein können. Die Grenze zwischen sprachunabhängigem und sprachgebundenem Wissen ist natürlich schwer zu ziehen, aber praktische Systeme werden es sich nicht leisten können, die Gesamtheit des für die Sprachverarbeitung relevanten Wissens für jede Sprache neu zu implementieren. Durch die explizite Trennung in unserem System sollen die Benutzer angehalten werden, Semantik so zu schreiben, daß zwar möglichst viel in dem Modul "Cog. concepts" untergebracht wird, aber nur unter der Bedingung, daß es für andere Sprachen nie geändert werden muß. Im Falle eines einfachen Nomens wie frz. *chevalier* 'Ritter' ist fast die ganze verzeichnete Information in (9b) enthalten, so daß der Eintrag im Modul Lexikalischer Semantik nur die sprachspezifischen grammatischen Eigenschaften hinzufügen muß, wie in (10).

(10) § Chevalier => EQUES(u).

```
/chevalier/ Nom, nil,     (" PRED ) = "Chevalier"
                          (" GEN) = MAS
                                  (" NUM) = SING . ("
/chevaliers/ Nom, nil,    PRED ) = "Chevalier"
                          (" GEN) = MAS
                                  (" NUM) = PLURAL.
```

Weil bei der Kompilierung dieses Moduls die übergeordneten Kognitiven Konzepte bereits vorhanden sein müssen, wird das Prädikat EQUES aus der Wissensbasis geerbt. Natürlich muß die lexikalische Semantik nicht immer so einfach sein; man kann im Semantischen Lexikon auch eine sehr detaillierte dekompositionale Semantik, z. B. für Verben angeben.

Während das Lexikon semantische mit syntaktischer (funktionaler) Information zusammenmischt, besteht die Satzverarbeitung auf einer strengen Trennung zwischen funktionaler Satzstruktur und Satzbedeutung. Die Satzbedeutung hat in KLU einen besonderen Statu s als Wert einer kognitiv verankerten, deshalb vordefinierten Sprechaktfunktion. Diese Funktion wird stets in der äußersten Ebene der f-Struktur, als Wert des Attributs FORCE spezifiziert, wie in Abb.2.

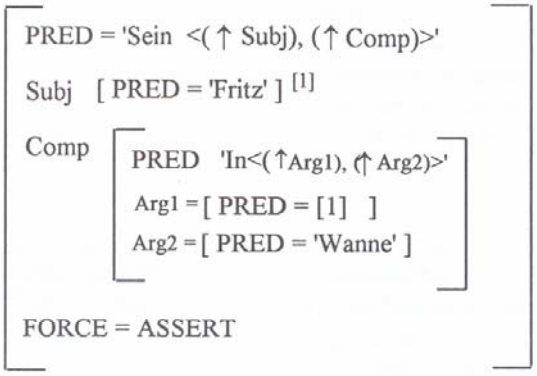


Abbildung 2. F-Struktur für *Fritz ist in der Badewanne.*

Die Motivation für diese Konstruktion war, die diskursbezogene Satzbedeutung von der formalen Semantik des obersten PRED[icates] der f-Struktur zu trennen. Die Satzbedeutung wird sowohl durch dieses PRED als auch durch den Aussagetyp des Satzes (z.B. Behauptung, Befehl, Frage) und Merkmale des Verbs (z.B. Konjunktiv) aufgebaut. Dies entspricht ungefähr der Intuition, daß *Fritz in der Badewanne* eine formale Semantik (IN(Fritz, Wanne)) aber keine Aussagekraft hat. Im Gegensatz dazu ändert *Fritz ist in der Badewanne* den globalen Diskurs-Kontext, indem es ihn um diese Relation erweitert, während ein anderer Satz *Ist Fritz in der Badewanne?* dieselbe semantische Relation enthält, aber den Kontext nicht ändert. Die Anwesenheit einer FORCE-Funktion ist eine Wohlgeformtheitsbedingung des Satzes, und diese Funktion muß normalerweise in den Regeln vorkommen, die die Grundformen der Sätze (Fragesatz, Ausruf, Behauptungssatz, usw.) definieren. Solange diese Funktion keinen Wert bekommen hat, kann aus einer f-Struktur keine Bedeutungsstruktur erzeugt werden.

Bislang wurde nur eine Sprechaktfunktion für ASSERT voll implementiert, aber die Werte INTERROG und IMPERAT sind auch vorgesehen.[7] INTERROG soll die aktuelle Datenbasis des Diskurs-Kontextes abfragen, und IMPERAT soll nützliche Anweisungen ausführen. „Bringe mir einen Kaffee!" wird wohl nicht realisiert, aber „Zeige die Regel für NP! " könnte als IMPERATive Satzform ohne großen Aufwand vom Rechner ausgeführt werden. Dieses Verfahren geht einen anderen Weg als viele neuere und in der GWB implementierte Entwürfe zur Semantik in LFG, die die semantische als eine von vielen möglichen, gleichwertigen, aus Lexikon und K-Struktur projizierten Strukturen betrachten.

4.3 Lexikalische Operatoren „Lex. operatives"

Wenn einmal Kognitive Konzepte und Lexikalische Semantik vorhanden sind, kann ein Modul der Lexikalischen Operatoren („Lex. operatives") erstellt werden. Dieses Modul hat zwei entscheidende Eigenschaften gegen über der Lexikalischen Semantik: Seine Einträge enthalten manchmal pronominale bzw. deiktische aber, keine konzeptuellen Prädikate ((↑ PRED) = 'PRO' aber nicht 'EQUES'), und es kann durch Derivation nicht erweitert werden. Es entspricht also den traditionell als „geschlossen" und „funktional" bezeichneten Klassen der Determinatoren, Pronomina, Konjunktionen; außerdem kommen in KLU die Derivationsmorpheme hinzu (ich komme im nächsten Kapitel auf diese zurück). In vielen Fällen handelt es sich hier um Wörter, die sprachgeschichtlich durch „semantic bleaching" gegenüber den Einträgen der

[7] Van Eijck & Alshawi 1992 sehen einen Modus-Operator vor, der auf oberster Ebene imp für Befehle, whq für WH-Fragen, ynq für Ja-Nein-Fragen, und dcl für Behauptungssätze angibt (S. 20).

übergeordneten Lexikalischen Semantik ihre eigene, strukturierte Semantik verloren haben. (Beispiele sind Auxiliar-Verben und Präpositionen wie ttz. *a inpenser a qc* .) Da die Operatoren nicht durch produktive Derivation entstehen, können sie als mono-morphemisch gelten. Die Grenzen zur Lexikalischen Semantik sind sicherlich eher fließend, und die Zugehörigkeit zu diesem Modul hängt oft stark von der spezifischen linguistischen Analyse ab.

4.4 "Syntax"

Die syntaktische Beschreibung einer Sprache beschränkt sich in den Unifikationsgrammatiken wie LFG zunächst auf Oberflächenstrukturen, die nach den üblichen Verfahren der Konstituentenanalyse zu ermitteln sind und bereits in (1) bis (3) angedeutet wurden. Die Konstituenzrelationen werden in Phrasenstrukturregeln zusammengefaßt, wie in (3). Die Relationen, die die generative Tradition durch Theta-Rahmen und Tiefenstruktur darstellt, haben ungefähre Entsprechungen in der LFG in den Argumentstrukturen der Prädikate (vgl. (5) bis (7')) und in funktionalen Gleichungen, die als Annotationen zu den Phrasenstrukturregeln und Lexikoneinträgen erscheinen dürfen. Folglich wird die generative Analyse des Satzes

(11) Paul verspricht zu singen.

wonach der Agens von *versprechen* in die Stellung des externen Arguments von *singen* "bewegt" wird, in der LFG-Analyse durch folgende, getrennte Bedingungen erfaßt: Erstens wurden - in der Entstehungsgeschichte des Lexikons - die agentiven Argumentstellen von *versprechen* und von *singen* auf die grammatische Funktion Subjekt abgebildet. Eine Gleichung im Lexikoneintrag von *versprechen,* die dieses Verb als "equi" charakterisiert, bewirkt, daß das SUBJ seiner Verbalergänzung mit dem eigenen Subjekt gleichgesetzt wird.

Der entscheidende Unterschied zur generativen Analyse liegt aber darin, daß die Abbildung der thematischen Rolle auf eine grammatische Funktion n ich t in der Syntax, sondern als sog. "mapping"-Regel innerhalb des Lexikons erfaßt wird. So entstehen z.B. Aktiv-, Passiv- und Partizipialformen eines Verbs nicht durch eine Transformation in der Syntax, sondern über zwei unterschiedliche Realisierungen eines semantischen Konzepts im Lexikon, die zu unterschiedlichen syntaktischen Beschreibungen fuhren, etwa einmal mit dem thematischen Agens als SUBJ, einmal als eine OBL[ique] Ergänzung. Nach einem häufig anzutreffenden Mißverständnis muß diese Abbildung stets "off-line" und in völliger Trennung von der Syntax stattfinden, so daß das Lexikon eine begrenzte, endliche Größe haben muß. In KLU geschieht dieser Schritt in der Regel zur Zeit der Lexikon-Compilierung. Bresnan & Kaplan 1982:xxxii-xxxiii suggerieren aber, daß die Abbildung von Argument-Struktur zur funktionalen Struktur in LFG lediglich formal in einen getrennten Bereich herausfaktorisiert wird; die Bezeichnung dieses Bereichs als Lexikon schließt nicht aus, daß die Abbildung der Argumente und der damit verbundenen Wortbildungsprozesse als Teil der Satzanalyse bzw. Produktion ablaufen könnte. Kommen rekursive Wortbildungsregeln bei der Satzanalyse zur Geltung, ist die Anzahl möglicher Wörter ebenso uneingeschränkt wie die Anzahl möglicher Sätze, trotz eines endlichen Lexikons.8 Die sprachlich gegebene Möglichkeit eines dynamisch erweiterbaren Lexikons ist bislang im formalen Apparat der LFG und anderer Unifikationsgrammatiken bislang kaum berücksichtigt worden, aber sie erfordert lediglich ein paar Erweiterungen, die im Kapitel 5 genannt werden.

8 Baayen 1991 gibt einen Überblick über die statistische und psycholinguistische Evidenz für diese Annahme und stellt fest, daß Wortbildung durchaus simultan mit der syntaktischen Analyse geschehen muß: „such formations [produktiv ableitbare Derivationen] will generally be stored in memory but. . . storage in memory is not obligatory as long as all properties of the types are fully predictable by rule. When not available from memory. . . low-trequency formations can be parsed or generated by the relevant word formation rules of the language .. (S. 126).

4.5 Wortbildung und Flexion: „Morphotactics"

Ohne die Möglichkeit Flexionsprozesse darzustellen, kann das Vollformlexikon einer stark flektierenden Sprache enorme Ausmaße annehmen. Ein Verfahren des „affix-stripping", um Flexionsmaterial getrennt vom Stamm eines flektierten Wortes behandeln zu können, wurde in Bresnan 1982:18-20 angedeutet, und in neueren LFG-Arbeiten kommt zusätzlich zur syntaktischen auch eine Flexionsklasse im Lexikon hinzu. In KLU wird dieser Vorschlag genutzt, um das Lexikon zu komprimieren. Beim Lesen des Eingabe-Strings versucht KLU stets, jedes volle Wort zunächst als Symbol im Lexikon zu finden. Falls dies nicht gelingt, versucht KLU innerhalb des Wortes möglichst große Segmente zu finden, die als Symbole im Lexikon erscheinen, indem Morpheme, die zu einer wortsyntaktischen Kategorie gehören (z.B. Nominal-Suffix) als Substrings von rechts und von links und von außen nach innen gesucht werden. Dadurch wird erreicht, daß die Segmentierung automatisch durch die Einträge des Lexikons gesteuert wird, und daß lexikalisierte Segmente immer Vorrang vor möglichen Dekompositionen haben.

In (10) haben wir gesehen, wie ein frz. Nomen im Singular und im Plural realisiert werden kann - die zwei Varianten sind im Lexikon explizit angegeben. Nach einer möglichen Analyse besteht das Nomen aus einem Stamm und einem Flexionsmorphem, das entweder als leeres Morphem (Ø) für Singular oder als *-s* für Plural erscheinen darf, wie in (12).

(12) $N \rightarrow$ NomStamm N-Sufx
$\uparrow = \downarrow \quad \uparrow = \downarrow$

Diese Wortbildungsregel erlaubt uns, den Eintrag (10) auf (13) zu kürzen, mit Hinzunahme der Flexionsmorpheme des Paradigmas NDs (14):

(13) /chevalier/ NomStamm, NDs, (↑ PRED) = "Chevalier"

(14) a. /Ø/ N-Sufx, NDs, (↑ NUM) = SING
b. /s/ N-Sufx, NDs, (↑ NUM) = PLURAL

Solange (10) im Lexikon von KLU bestehen bleibt, hat (13) keine Wirkung, weil *chevalier* und *chevaliers* nicht segmentiert werden müssen. Falls die Einträge von (10) nicht vorkommen, wird die Wortsegmentierung versuchen, den String „chevaliers" in größtmögliche lexikalisch verzeichnete Teile aufzuteilen, nämlich „chevalier" + „s". Dann kann der Wortparser die Regel (12) benutzen, um ein N[omen] in die Satzstruktur einzufügen.

Um die Regel (12) für sg. *chevalier* zu benutzen, muß das Flexionsparadigma für reguläre Nomen ein „leeres" bzw. Null-Morphem postulieren, das als die Wortkonstituente N-Sufx erscheint und die Information SING[ular] trägt, wie (14a). KLU kennt ein vordefiniertes Morphem /Ø/, das durch einen leeren String akzeptiert wird. Da die Wortsegmentierung aber noch nichts von der Wortbildung wissen kann, könnte sie gezwungen sein, unzählige Segmentierungen zu unternehmen, um alle möglichen, unsichtbaren Null-Morpheme für den Wortparser vorzuschlagen. Daher wird /Ø/ eigentlich nur an solchen Stellen im Wort akzeptiert, wo eine Art abstraktes Pseudo-Morphem erscheint, wie in (15).

(15) /chevalierF/ NomStamm, NDs, (↑ PRED) = "Chevalier" .

Die Wortsegmentierung sorgt dafür, daß das Pseudo-Morphem /F/ als getrenntes Terminalsymbol für die Wortanalyse erscheint, und die Allomorphie (s. u.) erlaubt die Expansion von /F/

entweder zu /Ø / oder zu /s/. Die Information über Numerus kommt jeweils aus einem entsprechenden Eintrag des Paradigmas für Pluralbildung in *-s*.

Normalerweise ist die Information, die ein Flexionsmorphem trägt, eine andere als im Stamm bereits vorhanden ist, und sie kann dem Stamm einfach durch Unifikation hinzugefügt werden. Bei Derivationsmorphemen hingegen ist dies manchmal nicht der Fall: Haben wir für dt. *Zähl + ung* im Lexikon

(17) a. /zählF/ V, (^ PRED) = „Zählen".
b. /ung/ V-Suffix, (^ PRED) = „EVENT-OF".

wird es nicht möglich sein, die zwei Prädikate „Zählen" und „EVENT-OF" zu einem einzigen Prädikat für *Zählung* zu unifizieren. Mehr muß geschehen - aber das, was geschehen muß, liegt außerhalb dessen, was im Rahmen der Unifikationsgrammatik vorgesehen ist. Wie zwischen f-Struktur und Satzbedeutung eine Modulgrenze gezogen wurde, so muß auch zwischen Wortstruktur und Wortbedeutung eine bislang in der LFG nicht explizit gezogene Schnittstelle ausgearbeitet werden. Die Details möchte ich aber auf Kapitel 5 verschieben.

4.6 „Allomorphy"

Eine einigermaßen vollständige Lemmatisierung der Formen, die in einem schriftlichen Text erscheinen können, ist bekanntlich sehr aufwendig (Vgl. Kaplan & Kay 1994, Lenders 1994). KLU bietet lediglich ein paar einfache Möglichkeiten, die orthographische Variation zu berücksichtigen, um das Lexikon klein zu halten und morphologische Segmente ohne großen Aufwand zu identifizieren.[9]

Wie in Kapitel 4.5 schon angedeutet, unterscheidet KLU zwischen Oberflächen-Strings in der schriftlichen Eingabe und lexikalischen Symbolen. Damit soll der inzwischen verbreiteten Annahme Rechnung getragen werden, daß das phonologische Lexikon nur unterspezifizierte und abstrakte Darstellungen enthählt. Diese Annahme läßt sich wohl auf das schriftliche Lexikon nicht ohne weiteres übertragen, und die Zitierformen unseres Lexikons sind auch keine phonologischen Darstellungen, aber sie sind auch nicht notwendigerweise identisch mit den Eingabe-Strings, die sie akzeptieren sollen. Um einfache, regelmäßige Variation zu berücksichtigen, können abstrakte 'Pseudophoneme' in den Zitierformen als Großbuchstaben erscheinen. Das Modul „Allomorphy" definiert eine Menge von orthographischen Ersetzungsregeln, die solche Großbuchstaben au f Strings, und diese evtl. auf lexikalische Morpheme bzw. Stämme abbilden können. Um die Pluralform von frz. *chevalier* zu bilden, hat das Pseudomorphem /F/ in /chevalierF/ (15) die String-Realisierungen „s" und leer („"), nach den Ersetzungsregeln (18):

(18) a. /...F/ → „s" → /.../ + /s/
b. /...F/ → „"→ /.../ + /Ø/

Diese Strings gehen in die Wortstruktur als die Morpheme /s/ bzw. /Ø/ ein. Die Regel (18a) bewirkt zunächst, daß das Pseudo-Phonem /F/ durch den Oberflächen-String „s" akzeptiert wird. Nach der Segmentierung geht „s" in die Wortstruktur als ein getrenntes mophologisches Segment ein, nämlich /s/. So wird aus dem String „chevaliers" die Segmentkette /chevalier/ + /s/, während aus „chevalier" /chevalier/ + /Ø/ entsteht.

[9] Die kommerzielle Entwicklung der „2-level morphology"zu „industrial strength linguistics" (in den Worten Lauri Karttunens) läßt das Problem der Lemmatisierung in vieler Hinsicht als gelöst gelten; auf einen eigenen effizienten Lösungsversuch haben wir verzichtet.

Falls in der Allomorphie-Regel die zweite Ersetzung fehlt, wird der String akzeptiert, ohne daß ein weiteres Morphem-Segment der Wortstruktur hinzukommt. (Z. Zt. gibt es keinen Compiler für diese Regeln, aber die entsprechenden Prolog-Regeln sind leicht zu formulieren.)

4.7 Arbeiten mit den Modulen

Wie die meisten modernen Entwicklungsumgebungen, bietet unser Prolog (PrologIA 1995) viele eingebaute Schnittstellen zu den Graphik- und Menüfunktionen des Betriebsystems, so daß eine relativ benutzerfreundliche Umgebung ohne großen Aufwand programmiert werden konnte. Der Benutzer braucht also keine Prolog-Kenntnisse und nur soviel Wissen über die Rechnerumgebung, wie für ein Textverarbeitungsprogramm erforderlich ist. Nach einer kurzen Einführung in die spezifische Arbeitsweise des Werkbank-Programms kann er bzw. sie sich ausschließlich dem Erlernen des LFG-Formalismus und der Lösung spezifischer linguistischer Problemstellungen widmen. Eine umfangreiche und detaillierte Fehlerbehandlung hat sich allerdings als unabdingbar erwiesen.

Benutzer werden vom System gezwungen, eine Sprachbeschreibung in sechs Schritten zu entwickeln, entsprechend den sechs schwarzgedruckten Modulnamen im Untermenü "Selection" von Abb. 1. Der erste Schritt ist jeweils, über den Menüpunkt "Open/Edit" eine Textdatei zu ändern bzw. zu erstellen, welche die deskriptiven Daten des jeweiligen Moduls enthält. Abb. 3 zeigt als Beispiel einen Auszug aus dem Modul "Lex. concepts" für eine kleine französische Grammatik.

```
# Description français.
# Governed functions are.
SUBJ { & } ,
OBJ { &} .

# Non governed functions are.
NUMBER {SG, PL} , TENSE
{PRES, IMPF, FUT} .

# Lexical categories are.
Degr,
Adv.

# Inflectional paradigms are.
PARADIGM {NDs, VK1 } .

# Lexical concepts are.
§ Penser(s) ->
        ereignis(e, COGITARE(s»
        agens(s) .
      /pens/ VStamm, VK1, (A PRED) = "Penser«ASUBJ»" .

# End franyais.
```

Abbildung 3. Quelltext-Format des Moduls "Lex. concepts"

Ein Klick mit der Maus auf "Compile" übersetzt die Datei in das interne Format des Systems. Danach sollte man die Daten mit einem entsprechenden Testmodus überprüfen (siehe Kapitel 4.8), bevor andere Module entwickelt werden. Einen Überblick über den geladenen Inhalt des

jeweils gewählten Moduls gibt der Menüpunkt „Summary", so wie Abb. 4 einen Ausschnitt eines Lexikons für Franz ösisch zeigt.

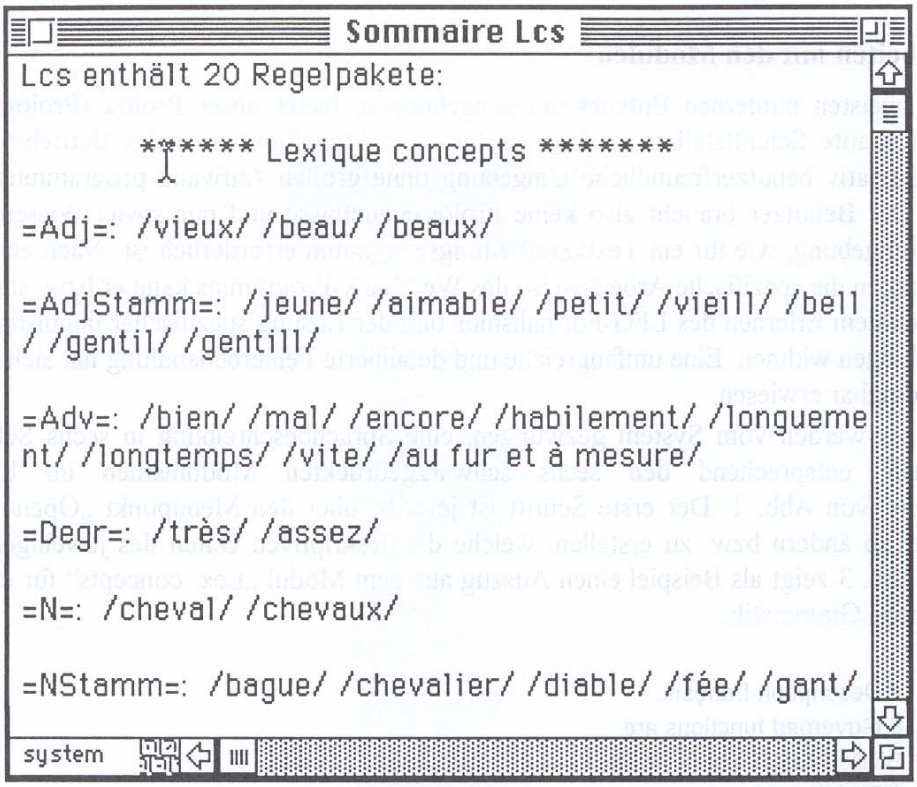


Abbildung 4. Liste der Einträge im Modul „Lex. concepts" (Lcs)

Der Menüpunkt „Modify" bietet die Möglichkeit, einen einzelnen Lexikon-Eintrag bzw. eine Syntax-Regel zu ändern, ohne die ganze Datei neu zu compilieren (was je nach Rechner und Dateigröße schon etliche Minuten dauern kann).

4.8 Überprüfung der Sprachbeschreibung

Der Weg vom intuitiven Wissen zur Formalisierung scheint immer da am längsten zu werden, wo die formale Beschreibung Fehler und widersprüchliches Verhalten hervorruft. Sowohl bei Anfängern als auch bei erfahrenen Linguisten kann man aber immer wieder beobachten, daß der Prozeß, in dem der erste formale Entwurf einer Grammatik korrigiert, geglättet und überarbeitet wird , oft mehr an tiefer Erkenntnis weckt, als die vorangegangene, unimplementierte Formalisierung. Ohne Übertreibung kann man also sagen, daß der Wert einer linguistischen Werkbank vor allem in ihren Hilfsmitteln für die Fehlersuche zu finden ist. Diese Hilfsmittel sind naturgemäß viel stärker an die gewählten Verarbeitungsalgorithmen und Datenformaten gebunden, als der Formalismus selber, und verlangen von den Benutzern des Systems eine Gewöhnung, die für eine abstrakte Verwendung entbehrlich wäre. Jedoch führt die Fehlersuche am Computer meistens zum vertieften Verständnis des Formalismus und der linguistischen Analyse. Im Einklang mit dem Ziel, eine möglichst flexible Umgebung für die uneingeschränkte Benutzung des LFG-Formalismus anzubieten, stellt die GWB zu diesem Zweck sehr vollständige und vielfältige graphische Mittel zur Verfügung, die den Parsbaum und die f-Struktur eines Satzes Stück für Stück aufbauen und diese in verschiedenen Formaten und mit wählbarer Detailtiefe anzeigen können. Außerdem können funktionale Gleichungen (die getrennt von der Konstitu-

entenanalyse gelöst werden) und Bedingungen (constraints), die mit Konstituenten verbunden sind, getrennt angeschaut und ihre Lösung bzw. ihr Scheitern protokolliert werden.

Im viel bescheideneren Rahmen unseres Systems sind die Hilfsmittel eingeschränkter, aber auch etwas anders konzipiert als in der GWB. Die strenge Modularisierung der Sprachbeschreibung, die KLU verlangt, bietet eine (noch nicht ganz implementierte) Möglichkeit, statische Konsistenztests durchzuführen, die manchem Laufzeitfehler vorbeugen können. Weil alle Symbole (Konstituenten, grammatische Funktionen und Werte) einer Sprachbeschreibung im jeweiligen Modul explizit deklariert werden müssen, können vor allem im Lexikon inkonsistente Angaben zur Compile-Zeit abgefangen werden.

Die Hilfsmittel für die Fehlersuche zur Laufzeit bestehen wesentlich in Einschränkungen der Grammatik und in Ablaufprotokollen. Um einen Fehler zu lokalisieren, ist es sehr nützlich, keine vollständigen Testsätze eingeben zu müssen, sondern nur einzelne Konstituenten bzw. Wörter, deren Beschreibungen vermutete Fehlerquellen enthalten. Durch das Untermenü von Abb. 5 kann der KLU-Benutzer die Analyse auf einen engen Bereich wie „Word " beschränken, in diesem Fall Wörter der Kategorie Prep[osition]. Die Grammatik ist dann so eingeschränkt, daß einzelne Präpositionen akzeptiert werden, als ob sie Sätze wären. Sie führen zu einer f-Beschreibung, die in diesem Fall den Lexikon-Eintrag der Preposition wiedergibt. Phrasale Kategorien können mit „Phrase" gewählt werden.

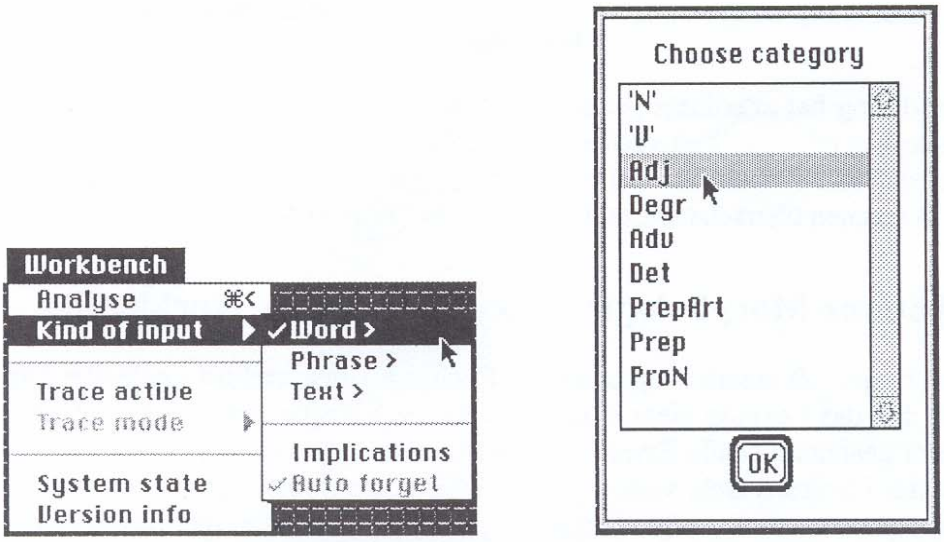


Abbildung 5. Beschränkung der Analyse auf die Kategorie „Prep"

Eine weitere Hilfe bei der Fehlersuche besteht in der Möglichkeit, Zwischenergebnisse der Morphologie- und Satzbearbeitung schrittweise anzuschauen. Diese Ablaufsprotokolle werden über ein weiteres Menü (Abb. 6) ausgewählt und können in toto ein- und ausgeschaltet werden über den Menüpunkt „Trace active".

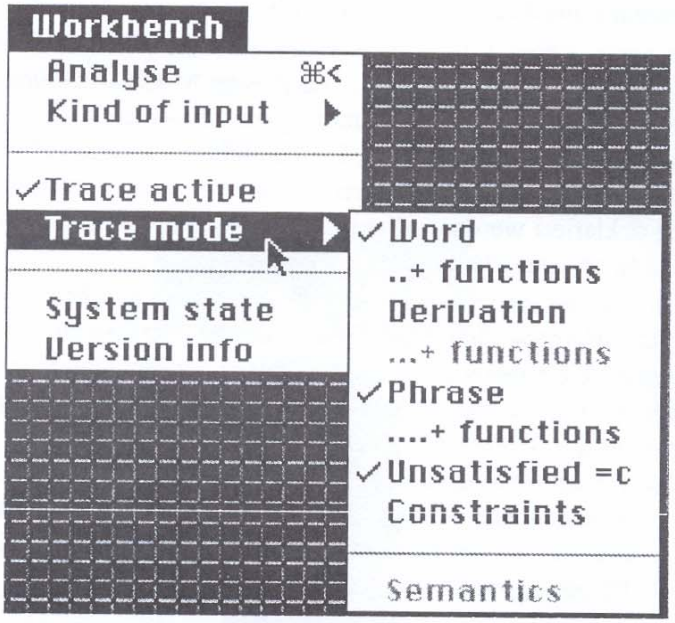


Abbildung 6. Aktivierung der Protokolle für jedes akzeptiertes Wort u. jede Phrase, mit Anzeige aller nicht gelösten fordernden Gleichungen.

Unsere Erfahrung hat allerdings gezeigt, daß Anfänger mit den von KLU erzeugten, ausführlichen Protokollen oft mehr Zeit verlieren als sie dadurch sparen sollten. Dies liegt sicherlich zum Teil daran, daß alle Protokolle im Line-Modus nacheinander in einem einzigen Fenster erscheinen, statt in eigenen überschaubaren Fenstern, wie in der GWB.

5 Derivations-Morphologie in der Konstanzer Werkbank

Jedem, der länger mit einem Programm zur Rechtschreibekorrektur gearbeitet hat, ist wohl klar geworden, daß das Lexikon einer real existierenden Sprache nichts statisches ist, sondern stets und in kaum geahntem Maße Erweiterungen (und Verluste) erlebt. Diese oft unterschätzte Generativität des Lexikons zieht wieder großes Interesse auf sich, aber in formalen Modellen wurde sie bislang oft vernachlässigt. Für frühe generative Modelle, die derivationale Vorgänge in der Syntax angesiedelt hatten, war dieser Aspekt der Wortbildung nicht problematisch. Jedoch wollten Bresnan & Kaplan 1982 solche Prozesse von der Syntax trennen, um die Komplexitätskosten zu vermeiden, die im transformationellen Modell mit wortnahen generativen Prozessen verbunden waren. Um die LFG vom transformationellen Modell abzugrenzen, haben sie den statischen Charakter wortnaher Prozesse hervorgehoben und sie von der syntaktischen Komponente nicht nur getrennt, sondern - in manchen Formulierungen - auch isoliert. Andererseits wurde in der Einleitung zu Bresnan 1982 eigentlich nur behauptet (S. xxxi+), daß die Ergebnisse lexikalischer Regeln zwischengespeichert, dann von der syntaktischen Komponente in einer einzigen Operation aus diesem Speicher geholt würden, um damit eine komplexe „on-line"-Bearbeitung durch Transformationen zu umgehen. Den Zwischenspeicher haben Bresnan & Kaplan aber pauschal mit dem Lexikon identifiziert, was wohl zu einigen Mißverständnissen geführt und bei manchem Leser den Eindruck erweckt hat, daß die LFG spontane Wortbildungsprozesse nicht formal erfassen könnte (z.B. bei Pulman 1992:76).

Eine lexikalische Ausrichtung eines Verarbeitungsmodells wie in der LFG verlangt aber keineswegs, daß Wortbildung von der Syntax völlig isoliert wird. Die vermeintlichen Schwierigkeiten solcher Modelle mit derivationalen Prozessen können beseitigt werden, wenn man das

Lexikon differenzierter sieht, nämlich als hierarchisch und mit statischen und dynamischen Unterbereichen versehen. In KLU wird mit den vier oben beschriebenen Unterbereichen "Lex. concepts", "Temp. concepts", "Lex. operators", "Morphotactics" eine solche lexikalische Organisation zur Erprobung angeboten. Die Dynamik des Lexikons ergibt sich daraus, daß die Wortbildung ("Morphotactics") mit der Syntax so zusammenarbeiten kann, daß Derivationsmorpheme (im Modul "Lex. operators") aus Einträgen des statischen Lexikons ("Lex. concepts") neue Lexikoneinträge während der Satzverarbeitung bilden können.

Wie bereits in Kapitel 3 bemerkt, wird die Projektion von lexikalischer Argumentstruktur zur Syntax ("mapping" bzw. "linking") in der transformational orientierten Literatur oft so betrachtet, als ob sie in der Syntax geschähe, dagegen wird sie in der LFG als eine lexikalische Operation beschrieben, getrennt von den Operationen der syntaktischen Satzanalyse bzw. -generierung. Abgesehen von den Verarbeitungskosten dürften die zwei Sichtweisen formal weitgehend äquivalent sein, aber in bestimmten Fällen kommt es doch darauf an, ob "mapping" sozusagen "on-line", also als Teil der Satzanalyse, oder "off-line" im Lexikon stattfindet. Solche Fälle findet man bei morphologisch komplexen Wörtern, deren Semantik in einer Lesart transparent aus der Wortstruktur ersichtlich, aber in einer anderen Lesart völlig opak ist. Bei morphologisch transparenten Wörtern ist es plausibel, daß das Wort segmentiert und Morphem-für-Morphem in die (Tiefen)struktur des Satzes eingefugt wird - dies wird durch die Beobachtung unterstützt, daß noch nie gehörte Neologismen auf Anhieb verstanden werden, obwohl ihre genaue Bedeutung manchmal erst durch den Satzkontext klar wird. Bei semantisch opaken Wörtern andererseits kann die Verarbeitung nicht morphemweise vorgehen, weil die in die Satzstruktur eingehende Bedeutung des Wortes aus seinen Bestandteilen nicht systematisch gewonnen werden kann.

Zur Illustration: aus dt. *verscheiden* kann man ein adjektivisches Partizip *verschieden* 'gestorben' über einen regelmäßigen Flexions- bzw. Derivationsprozeß gewinnen. Die Bedeutung 'gestorben' hat aber wenig mit der häufigeren Bedeutung 'unterschiedlich' zu tun, und Sätze wie *Anna und Maria sind verschieden* werden normalerweise nach dieser zweiten Bedeutung verstanden. Wir möchten aber annehmen, daß Hörer, die für partizipiales *verschieden* die Bedeutung 'tot' in ihrem mentalen Lexikon nicht präsent haben, sie aus *verscheiden* ableiten können, und daß diese Bedeutung nicht durch die bereits vorhandene Bedeutung 'unterschiedlich' ausgeschlossen wird.

In KLU wird ein solcher Konflikt dadurch vermieden, daß das Ergebnis der systematischen Ableitung ('tot') in einen sekundären Bereich "Temp concepts" abgelegt wird, wovon es erst dann in einen Satz eingefugt wird, wenn der fest lexikalisierte Eintrag ('unterschiedlich') aus "Lex. concepts" sich als unpassend erweist. Hier bleibt es auch eine Weile, bis der Diskurskontext, für den es erfunden wurde, vergessen ist (in der Praxis muß der Benutzer von KLU den Menüpunkt "Auto forget" von Abb. 5 anklicken). Solche neu erzeugten, zwischengespeicherten Derivate können aber durch wiederholte Verwendung über eine weitere Schnittstelle von "Temp. concepts" zu "Lex. concepts" gelangen. Die Funktion, die sie in "Lex. concepts" fixiert, kann eine semantische Verschiebung oder Einschränkung der unmittelbar abgeleiteten Bedeutung bewirken, wie sie eben in der geschichtlichen Entwicklung eines Derivats oft geschieht. (Die Funktion ist formal beschreibbar, aber in KLU nur sehr rudimentär implementiert, weil sie eine komplexe Wissensbasis und diverse Interaktionen mit dem Gesamtlexikon voraussetzen muß.)

Die dynamische, "on-line"-Derivation wird durch eine Elaborierung der lexikalischen Einfugung realisiert. Wegen der komplexen Struktur des Lexikons können präterminale Symbole der Syntax (lexikalische Kategorien) prinzipiell nicht in einer einzigen, ungeteilten Operation aus dem Lexikon substituiert werden, denn eine mögliche Substitution könnte mehrmals (in ver-

10 V gl. Grimm 1956; 1064 "dann ist durch eine besondere Sympathie. . . jede halbverschiedene Zärtlichkeit wieder auf einmal lebendig" (Goethe).

schiedenen Untermodulen) oder gar nicht (bei noch erforderlicher Derivation) vorkommen. Die lexikalische Einfügung in KLU sucht terminale Symbole in der Modul-Hierarchie von Abb. 1 (von oben nach unten), und der Parser fügt versuchsweise immer die längsten auffindbaren lexikalischen Segmente zuerst in den Baum ein, bevor er eine Zerlegung eines Segments untersucht. So haben fixe Kollokationen Vorrang vor ihren Bestandteilen, und bei Wörtern haben Vollformeinträge Vorrang vor ihren morphologischen Zerlegungen. Allerdings geschieht die Einfügung von nicht-lexikalisierten Derivaten n i c h t durch die Einfügung einzelner Morpheme in den Parsbaum des Satzes, sondern nach einem Schema, das die Wortgrammatik - im Einklang mit der lexikalistischen Sichtweise - von der Satzgrammatik streng getrennt hält, so daß das Derivat nur als Ganzes in den Satz eingefügt wird, als ob es eine bereits lexikalisierte Form wäre; die Verarbeitungskosten bleiben nach der ersten Derivation dadurch streng begrenzt. Wie dies möglich wird, ist in Abb. 7 und 8 angedeutet.

Wie in Kapitel 4 erwähnt, geht die Wortsegmentierung in unserem System so vor, daß 1) Segmente von den unteren Lexikonmodulen zuerst (also zunächst Flexions-, dann Derivationsmorpheme), 2) vom Wortende und -anfang zur Wortmitte[11] unter 3) Beibehaltung größtmöglicher innerer Segmente gesucht werden.[12] Erst nachdem die Einfügung einer gefundenen Segmentierung scheitert (durch syntaktische, semantische, bzw. kontextbedingte Unverträglichkeit) kann versucht werden, das Wort weiter in Wort-, Stamm-, oder sonstige Morphemsegmente zu zerlegen.

So wird z.B. „Zählung" sicherlich als Stamm im Lexikon vorliegen, die Pluralform wahrscheinlich nicht, so daß der String „Zählungen" die Segmentierung (19) bekommt.

(19) /zählung/ + /en/

Weil die Attribute von /zählung/ und /en/ sich unifizieren lassen (vgl. (17)), kann die Syntax die Morphemkette ohne weiteres als Kategorie N nach der Wortbildungsregel (16a) akzeptieren. Dies entspricht dem oberen, normalen Pfad der lexikalischen Einfügung in Abb. 8.

Nehmen wir an, ein Text führe die Neubildung *Melkungen* ein. Nachdem sie das Pluralmorphem abgeschnitten hat, würde die Segmentierung „Melkung" im Lexikon suchen aber nicht finden. Ein String entsprechend dem Derivations-Morphem /ung/ könnte aber abgeschnitten werden, und der Rest „Melk"würde auch im Lexikon (als Verbalstamm) vorliegen. Jetzt aber kann die Morphem-Kette /melk/+/ung/+/en/ nicht unmittelbar in den Satzbaum eingefügt werden, weil der Lexikoneintrag für /ung/ (20) ein besonderes Derivationsprädikat (DPRED) enthält, ähnlich dem semantischen PRED(ikat) eines Eintrags im Konzeptlexikon. Anders als PRED aber darf ein DPRED nicht in den Satzbaum eingefügt werden: Diese Bedingung verwirklicht die Barriere zwischen Satz- und Wortsyntax.

(20) /ung/ NSfx, (↑ DPRED) = 'EVENT-OF<(↑ ARG)>'

(Formal kann man sich vorstellen, daß der Kopf jeder syntaktischen Regel implizit eine Gleichung (↑ DPRED) = Ø trägt, die die Unifikation mit (20) blockiert. Dies entspricht der Beobachtung, daß Derivationsmorpheme nicht als eigenständige Wörter erscheinen können.) Obwohl

[11] Wichtig war, die Wortbildung mit der modularen Struktur des KLU-Lexikons zu verzahne n, da einige Studien daraufhindeuten, daß Segmente am Anfang und am Ende eines komplexen Wortes anders lexikalisiert sind als die inneren Segmente. Nach dem in Kiparsky 1982 beschriebenen und von Clahsen 1996 experimentel bestä tigten Modell, sind die äußeren Segmente meistens semantisch leer (wie Flexionsmorpheme) aber frei kombinierbar, also wie in den KLU-Modulen „Morphotactics" und „Lex. Operators", wä hrend innere Segmente zunehmend semantisch und morphologisch fixiert sind („Lex. concepts"). Wie sonst i n KLU wurde deshalb hier versucht, dem Benutzer in erster Linie eine in bezug zum Gesamtmodell linguistisch plausible Darstellung zu bieten.

[12] Baayen 1991: 131 erwä hnt einige psycholinguistische Untersuchungen, die tendenziell diese Annahme rechtfertigen.

die Unifikation scheitert, gibt es für diese besondere Art des Scheiterns eine Art Fehlerbehandlung, die den unteren Pfad von Abb. 8. einleitet.

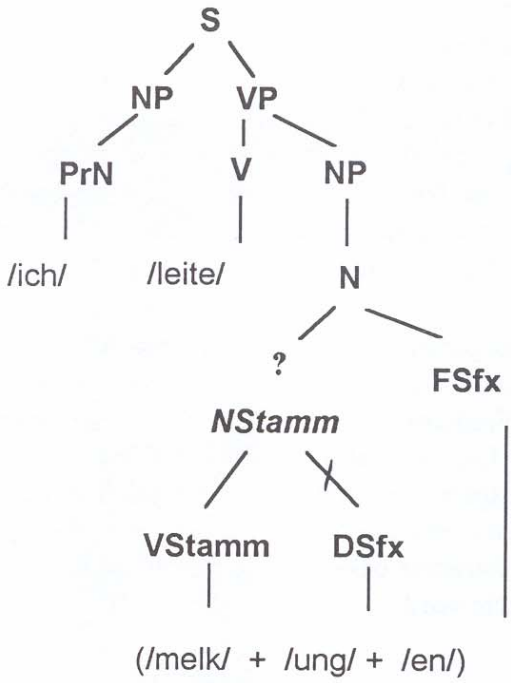


Abbildung 7. Analyse von „Ich leite Melkungen."

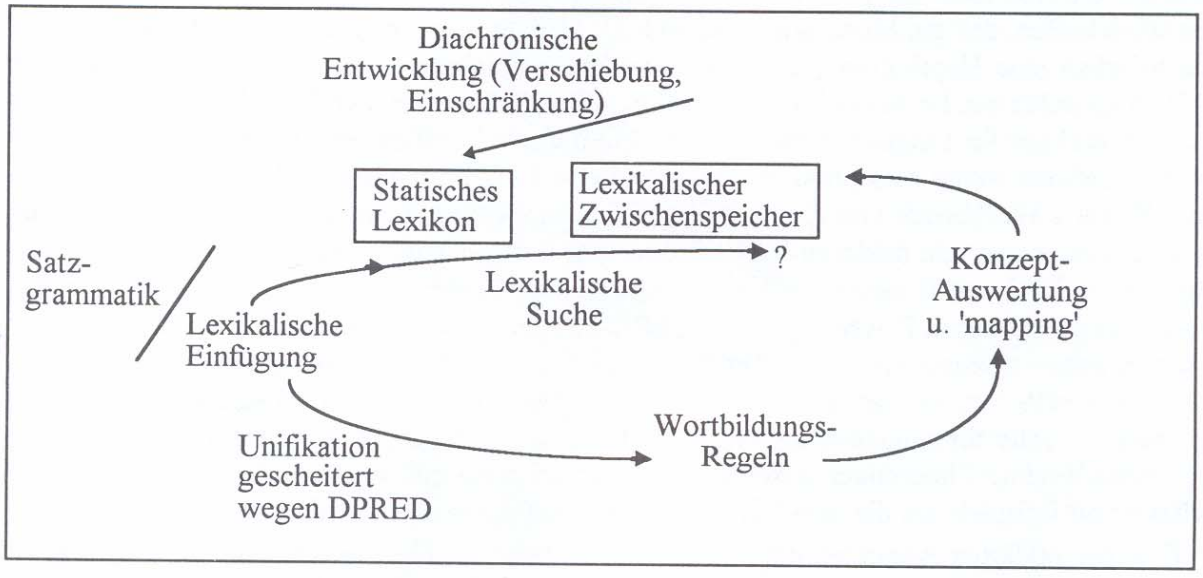


Abbildung 8. Lexikalische Einfügung und diachronische Entwicklung.

Mit dem semantischen Wert des DPREDs startet die Fehlerbehandlung einen Vorgang, der versucht, das semantische Prädikat EVENT-OF(Melken) pragmatisch bzw. konzeptuell auszuwerten. Wie dieser Prozeß genau aussieht, wäre eine eigene Untersuchung wert - mögliche Vorstellungen in bezug auf z.T. ähnliche ital. Derivationsprozesse wurden von meinen Kollegen in Mayo et al. 1995 entwickelt. Nehmen wir *Zählung* als ein denkbares Modell für diese Deri-

vation mit der Semantik EVENT -OF(Zählen), dann könnte EVENT -OF(Melken) das Ereignis sein, das dann eintritt, wenn in einem irgendwie abgeschlossenen Vorgang etwas gemolken wird. Ein solches Konzept muß allerdings in einer Wissensrepräsentation ("Cog. concepts", Abb. 1) in der Form eines Handlungsprädikats mit definierten Aktanten gefunden werden, deren Eigenschaften (agentiv, empfindend, usw.) bekannt sind. Als nächster Schritt muß dieses so gefundene Konzept nach den Regeln einer bislang nur in Umrissen bekannten Mapping-Theorie (rechts in Abb. 8) neu lexikalisiert werden, und das neue Lexem kommt in eben den Zwischenspeicher, der in der LFG fur die Ergebnisse lexikalischer Regeln schon immer vorgesehen war. Die Fehlerbehandlung endet damit, daß die Einfugung der Konstituente NStamm (Abb. 7) neu gestartet wird. Jetzt aber hat sich der Zustand des Systems geändert: Der Stamm /melkung/ liegt im Lexikon vor, wie /zählung/, nur noch nicht fest lexikalisiert sondern in "Temp. concepts". Jetzt kann die Neubildung eingefugt und die Analyse des Satzes zu Ende geführt werden.

Viele Details des Vorgangs "Auswertung und mapping" von Abb. 8 sind noch geklärt, und es wäre deshalb nicht sinnvoll, für sie einen eigenen Formalismus in KLU anzubieten. In Prolog haben wir aber bereits Funktionen implementiert, die aus einer semantischen Derivationsformel einen entsprechenden Lexikoneintrag erzeugen. Dieser Bereich unseres Systems gilt noch als projekt-spezifisches Experimentierfeld, und es wird dem normalen Verbraucher nicht zugänglich gemacht. Jedoch möchte ich behaupten, daß manche Spekulation über derivationale Prozesse einen anderen Charakter bekommt, wenn sie im Rahmen eines konkreten Verarbeitungsmodells zu Ende gedacht wird.

6 Ausblick: Die Grammar Writer's Workbench von Xerox PARC

Durch ihre Modularisierung der Sprachverarbeitung und ihre Festlegung von prozeßrelevanten Schritten in der Sprachverarbeitung und -entwicklung, verfolgt KLU eine Strategie, die sie vom rein deklarativen Ideal der Unifikationsgrammatiken wie HPSG und LFG absetzt. Wir möchten natürlich hoffen, daß die Modularisierung in KLU sich als notwendig zeigen wird, aber z. Zt. ist sie lediglich eine Hypthothese, die ansonsten kaum Spielraum für andersdenkende Benutzer läßt. Angesichts des heute noch unvollständigen Stands linguistischer Theorie sollte eine Computer-Werkbank für Linguisten aber flexibel sein und die Form der erlaubten deskriptiven Analysen möglichst wenig einschränken. Der erweiterte LFG-Formalismus, der jetzt in der Grammar Writer's Workbench von Xerox PARC implementiert ist, kommt diesem Wunsch gut entgegen, weil er den rein deklarativen Charakter des Formalismus weitgehend behalten hat. Leider hat die LFG seit Bresnan 1982 keine umfassende Darstellung mehr erlebt, und die inzwischen vorgenommenen Erweiterungen des LFG-Formalismus sind teilweise in schwer zugänglichen Arbeiten dokumentiert.13 Der Formalismus wird immer noch zu Unrecht stark mit der Theorie identifiziert; es wäre aber meiner Meinung nach richtiger, ihn als eine allgemeine Programmiersprache für linguistische Theoriebildung zu verstehen, denn die Analysen von Government-Binding-Theoretiker lassen sich mit ihm oft auch gut erfassen. Zur Illustration führe ich ein paar Beispiele an, die dem Handbuch der GWB entnommen sind.

Eine der erklärten Absichten der lexikalisch-funktionalen Theorie war (und ist noch), Konstituenz ohne leere Kategorien und ohne Bewegung zu beschreiben. Bewegung von Konstituenten kann im LFG-Formalismus nicht explizit als Transformationen von Baumstrukturen dargestellt werden, aber Bewegungs-Relationen können meistens mit Unifikationsgleichungen er-

13 Seit der Einrichtung eines FTP-Servers bei Xerox P ARC und einer WWW-Seite an der Universität Essex in England hat sich diese Situation etwas geändert. Der FTP-Server wird als parcftp.xerox.com erreicht, einige wichtige Dokumente sind im Directory /pub/nl. Der WWW-Server ist http://clwww.essex.ac.uk/LFG/; eine LFG-Bibliographie ist zugänglich über http://clwww.essex.ac.uk/search/. Siehe auch Dalrymple et al. 1995 für einige der wichtigsten neueren Arbeiten.

faßt werden. „Bewegungen", die über mehrere Baum-Knoten hinweg stattfinden („long-distance dependencies") werden mit einer Variante dieser Notation ausgedrückt, in der Bedingungen der Bewegung als Pfad durch eine Kette von syntaktischen Funktionen in der f-Struktur beschrieben werden. Zum Beispiel:

(21) VP
(↑ SUBJ) = (↓ COMP COMP* SUBJ)

besagt, daß das Subjekt der VP mit dem Subjekt irgend eines regierten Komplements identisch ist. Der Pfad kann durch noch komplexere reguläre Ausdrücke spezifiziert werden, die Bedingungen wie „government " und „barriers" entsprechen können. Da er aber nicht in bezug auf einen Tiefenstrukturbaum, sondern die f-Struktur definiert wird, entspricht er nicht ganz der transformationalen Analyse.

Transformationale Analysen mit Leerstellen („gapping") können durch das Hinzufügen von leeren Konstituenten notiert werden. Dies ist bei LFG-Theoretikern verpönt, der Formalismus verbietet es aber nicht: Will man z.B. für das häufig fehlende Subjekt im Italienischen eine leere NP-Konstituente postulieren, kann man schreiben

(22) S → NP VP
(↑ SUBJ) = ↓

S → e VP
(↑ SUBJ) = 'PRO'

was bedeutet, das eine fehlende NP an Subjektstelle als PRO evaluiert wird. (Sowohl dies als auch ein leeres Ø-Morphem in der Wortstruktur werden auch in KLU akzeptiert).

Die Faktorisierung von Dominanz und Präzedenz in Generalized Phrase Structure Grammar (GPSG) hat auch ein Darstellungsidiom in LFG bekommen. Die Regel

(23) S → [NP (↑ SUBJ) = ↓ , VP]

besagt, daß S aus NP und VP in beliebiger Reihenfolge besteht, während

(24) NP < VP

bestimmt, daß NP vor VP erscheinen muß.

Viele Programmiersprachen bieten die Möglichkeit, durch sogenannte Macros einen Programmtext zu „parametrisieren". Durch einen ähnlichen Mechanismus erlaubt die GWB die Einführung von „Parametern" in eine Sprachbeschreibung, d.h. sehr allgemeine, abstrakte Bedingungen, die sich durch den ganzen Regelapparat auswirken können. Wird z.B. eine Sprache als „kopf-initial"gesehen, muß der Grammatik-Schreiber normalerweise dafür sorgen, daß für jede Kategorie die Expansionsregel ihre minimale Projektion am linken Ende enthält. Ein explizites „Setzen" des Parameters ist nicht möglich. Mit einer Macro-Deklaration kann man aber explizit dokumentieren, das für eine Menge betroffener Kategorien die Minimal-Projektion v o r der entsprechenden Phrasal-Kategorie erscheinen soll:

(25) KopfTyp = { V | A | N | P} < { VP | AP | NP | PP}

Alle Phrasenstrukturregeln, die dieser Bedingung unterliegen sollen, brauchen nur Dominanz-Information zu spezifizeren, mit der Zusatzbedingung „& @KopfTyp", wie in

(26) VP → [V, NP (↑ OBJ) = ↓ , NP (↑ OBJ2) = ↓] & @KopfTyp

Da die Konstituenten der VP zwischen eckigen Klammern stehen, wird hier nichts über ihre lineare Präzedenz gesagt; die weitere Bedingung, daß V vor VP erscheinen muß, wird durch „@KopfTyp" eingeführt. Durch die logische Konjunktion („&") mit der durch „@KopfTyp" eingeführten Präzedenz-Relationen entsteht dann die Bedingung „V < VP" (V kommt vor VP). Sollte der Linguist entdecken, daß Nominalphrasen keine Kopfbedingungen haben, reicht es, N und NP aus dem Macro zu streichen, ohne die NP-Regeln selber zu ändern. Um die ganze Grammatik auf „kopf-letzt" umzustellen, muß man nur die Bedingung „< „ in (25) durch „>„ ersetzen.

Obwohl die GWB den LFG-Formalismus fast zu einer allgemeinen und daher theorieneutralen linguistischen Programmiersprache erhebt, ist die Gebundenheit dieses Formalismus an eine inzwischen sehr umfangreiche linguistische Forschungsliteratur nicht zu unterschätzen. Der Weg, der von manch anderem Projekt, wie der Core Language Engine (Alshawi 1992), eingeschlagen wird, einen eigenen, neuen Formalismus mit eigenen, neuen Beschreibungsparadigmen zu spezifizieren, wirft dieselben Probleme auf, die sonst aus der Einführung einer neuen Programmiersprache bekannt sind: Man kann zwar in dem neuen Formalismus auch alles ausdrükken, was man vorher konnte, gelegentlich auch schöner - aber die meisten Linguisten werden es leider nicht tun. Die Verbindung zu einer Gemeinschaft von Forschern, die stabile und gut begründete linguistische Analysen in dem neuen Formalismus umsetzen und weiterentwickeln können, wird nicht auf Anhieb gelingen, und es wird schwierig sein, von qualifizierten Linguisten Rat und Änderungsvorschläge zu den unvermeidlichen Problemen eines großen Systems zu holen, wenn sie den neuen Formalismus nicht oder nur halb beherrschen.

7 Schlußbemerkung

Die Ausbreitung linguistischer Werkzeuge wie der Grammar Writer's Workbench könnte das Selbstverständnis der deskriptiven Linguistik grundsätzlich ändern. Statt über vermutete Prozesse und Beziehungen in einer Sprache zu spekulieren, wäre Linguisten und Linguistinnen die Möglichkeit gegeben, bislang unüberschaubare Faktenmengen zu erfassen und deskriptive Hypothesen empirisch zu testen. Ein Studium, das diese Möglichkeit berücksichtigt, würde möglicherweise für ihre Absolventen auch einen Weg in den langsam aber unaufhaltsam sich herausbildenden Industrie-Zweig des „linguistic engineering" öffnen.

Literatur

Alshawi, Biyan 1992: The Core Language Engine. Cambridge, Ma.: MIT Press. Baayen, Barald 1991: Quantitative aspects of morphological productivity. In: Yearbook of Morphology 1991, Hrsg. Geert Booij & Jaap van Marle. Dordrecht: Kluwer, S. 109-149.

Bresnan, Joan, Hrsg. 1982: *The Mental Representation 01 Grammatical Relations.* Cambridge, Ma.: MIT Press.

Bresnan, Joan & Ronald Kaplan 1982: Introduction: Grammars as Mental Representations of Lauguage. In Bresnan 1982, S. xvii-liii.

Clahsen, Barald., et al. 1996: Compounding and inflection in German child language. In: Yearbook of Morphology 1995, Hrsg. Geert Booij und Jaap van Marle. Dordrecht: Kluwer.

Dalrymple, Mary, Rondald M. Kaplan, John T. Maxwell m & Annie Zaenen 1995: Formal Issues in Le xical-Functional Grammar. Stanford, CA: CSLI.

Eiseie, Andreas & Jochen Dörre 1986: A Lexical Functional Grammar System in Prolog. In: Proceedings of Coling '86, 11th International Conference on Computational Linguistics, 551-553. Bonn: Institut für angewandte Kommunikations- und Sprachforschung e. Y.

Gazdar, Gerald & Chris Mellish 1989: Natural Language Processing in Prolog: An Introduction to Computational Linguistics. Wokingham: Addison-Wesley. Grimm, J. u. W. 1956: Deutsches Wörterbuch. Leipzig: Hirze11956. Kaplan, Ronald M. & Martin Kay 1994: Regular Models of Phonological Rule Systems. Computational Lin guistics 20:3, S.332-378.

Kaplan, Ronald M. & John T. Maxwell **In** 1993: Grammar Writer's Workbench. Internal report, Xerox Corporation. Verfügbar als "lfgmanual.ps" (postScript-Datei) per anonymes FTP von parcftp.xerox.com, im Directory /pub/nI.

Kiparsky, Pau11982: Word-formation and the lexicon. In: Proceedings ofthe1982 Mid-America Linguistics Conference, Lawrence, KS, Hrsg., F. Ingemann. S. 3-29.

Kiss, Tibor 1993: Lexical-Functional Grammar. In: Syntax: Ein internationales Handbuch der zeitgenössischen Forschung, Hrsg. Joachim Jacobs et al., S. 581~01. Berlin: de Gruyter = Handbücher zur Sprach und Kommunikationswissenschaft 9.1, Hrsg. Steger, H.& H. E. Wiegand.

Lenders, Winfried 1994: Morpholympics - Ein Unternehmen der GLDY. In: LDV-Forum 11:1, S. 5 (vgl. weitere Beiträge dieser Ausgabe).

Mayo, Broce 1995: Describing Verbs of Motion in Prolog. In: Lexical Knowledge in the Organization of Lauguage, eds. Urs Egli, Peter E. Pause, Christoph Schwarze, Amim v. Stechow and Götz Wienold, S. 203243. Amsterdam: Benjamins = Amsterdam Studies in the Theory and History of Linguistic Science, Series IV, 114.

Mayo, Broce, Marie Theres Schepping, Christoph Schwarze & Angela Zaffanella 1995: Semantics in the derivational morphology of Italian: implications for the structure of the lexicon. Linguistics 33, S. 883-938.

Pulman, Stephen G. 1992: Unification-Based Syntactic Analysis. In: Alshawi 1992, S. 61-82.

Prolog A, 1995. Prolog 11+, Version 4.2. Marseille.

Seewald, Uta 1995: Antibabylonisch: iX Multiuser-Multitasking-Magazin (12) S. 88-103.

Shieber, Stuart 1985: Criteria for designing computer facilities for linguistic analysis. Linguistics 23: 189 211.

van Eijck, Jan & Hiyan Alshawi 1992: Logical Forms. In: Alshawi 1992, S. 11-39.

MASCHINELLE ÜBERSETZUNG AUF DER BASIS DER LOGOTECHNIKI

Maria Theresia Rolland

Zusammenfassung: *Zunächst wird eingegangen auf die Sprache als Gegenstand der Übersetzung. Danach werden die zentralen Prinzipien zur Ermittlung der Sprachstruktur am Beispiel des Deutschen aufgezeigt. Auf der Basis dieser Erkenntnisse lassen sich die Grundlagen eines Computersystems erstellen, in dessen Zentrum die "Relationsbasis" steht, die eine automatische Wissensakquisition ermöglicht. Zum Aufbau eines Übersetzungssystems massen nun, gesondert für die einzelnen zu übersetzenden Sprachen, nach den angegebenen Prinzipien die jeweiligen Relationsbasen festgestellt werden. Der entscheidende Schritt ist dann der, die Relationsbasen parallel zu setzen. Jetzt ist der Computer in der Lage, zu "übersetzen". Er muß in den Relationsbasen lediglich das auffinden, was schon vorher als äquivalent definiert ist.* Ggf. *sind noch Wortstellungsregeln bzw. sprachspezifische Spezialregeln zu beachten. Es werden Übersetzungsbeispiele aus dem Deutschen und Englischen angeführt.*

Summary: *Machine Translation based on Logotechnique*
First, we will *present language as a subject of translation. Then we* will *describe the essential principles of identifying the language structure using the German language as an example. These results are the basis for constructing a natural language system. This focuses on the "relation base", which enables an automated knowledge acquisition. When bui/ding a translation system it* is *necessary to establish the relation bases according to the given principles for each of the particular languages to be translated. The most important step then* is *to create the link between the relation bases. The computer* is *now able "to translate". It only must look up in the relation bases, what has already defined to be equivalent before. Additionally, rules of word-order or language-oriented special rules may have to be taken into account. We will give translation examples for German and English.*

1 Sprache als Gegenstand der Übersetzung

Wenn man die Sprache mit dem Computer verarbeiten will und das Ziel sogar die Verarbeitung mehrerer Sprachen mit dem Ergebnis einer maschinellen Übersetzung sein soll [vgl. Copeland et al., 1991; EC, 1995], erfordert dies zunächst eine Besinnung auf das, was Sprache ist; denn die Anforderungen, die der Computer stellt, sind eindeutig: Er kann nur Formalisiertes verarbeiten. "Diese Aufgabe, die natürliche Sprache formal und maschinengerecht darzustellen, gestaltet sich äußerst schwierig: sie zu bewältigen, ist ein heute noch unerreichtes Forschungsziel. Bislang ist auch nicht geklärt, ob dieses Ziel prinzipiell erreichbar ist" [SCS, 1990: 97f.], was teilweise offensichtlich bezweifelt wird, da zumindest Siemens [Spiegel, 1995: 97], die Weiterentwicklung des Übersetzungssystems :MET AL [vgl. z.B.: SCS, 1990: 47ff.] aufgegeben hat

1 *Die Lösung des Sprachenproblems ist eine grundlegende Voraussetzung für die Einigung, die wirtschaftliche Leistungsfähigkeit und die politische Stabilität Europas.* (Aus: Wege zu einer Europäischen Sprachintrastruktur. Bericht an die Kommission der Europäischen Gemeinschaften, 31.03.1992)

es wird jetzt weitergeführt von der Gesellschaft für multilinguale Systeme mbH (GMS), Berlin. Im Gegensatz dazu bemüht man sich bei VERBMOBIL [z.B.: DFKI et al., 1991; Reuse, 1994: 50f; BMBF, 1995; FAZ, 1995], einem Projekt, das von 1992 an auf 10 Jahre konzipiert ist, ein mobiles Dolmetschgerät zu schaffen, wobei zusätzlich zur Übersetzung auch noch die gesprochene Sprache einbezogen wird. Allerdings sind Ergebnisse nur für kleinere Anwendungsbereiche geplant. Methodisch geht man dabei hinsichtlich der Sprachanalyse immer noch von einer Trennung von syntaktischer und semantischer Analyse aus.

Wenn man nun neu beginnt und sich zunächst darüber klar wird, was es mit dem zu verarbeitenden Gegenstand, der Sprache, auf sich hat, wie eine Sprache prinzipiell aufgebaut ist, d.h. nach welchen Regeln und Gesetzmäßigkeiten sie "funktioniert", dann ist es auch möglich, die einzelne Sprache adäquat zu verarbeiten und damit auch der angestrebten maschinellen Übersetzung zugänglich zu machen; d.h. das "heute noch unerreichte Forschungsziel" ist erreichbar. Die Aussage: "Ein vollautomatisches maschinelles Übersetzungssystem hoher Qualität (Fully Automated High Quality Translation System = F AHQT -System) gibt es derzeit noch nicht und wird es auch mittelfristig nicht geben" [SeS, 1990: 93], kann nunmehr dahingehend korrigiert werden, daß auf Grund neuer Erkenntnisse und Einsichten in den Aufbau der Sprache die Grundlagen auch für den Aufbau vollautomatischer natürlichsprachlicher Übersetzungssysteme gelegt sind und ihre Realisierung realistisch erscheint [Rolland, 1994; 1995a].

Die Sprache nimmt im Leben des Menschen und im Zusammenleben der Völker eine zentrale Stelle ein. "Unter dem Erbe, das die Menschen von Generation zu Generation weitergegeben haben, ist die Sprache ... eines der wertvollsten kulturellen Instrumente. ... Die Sprache ist ... ein unentbehrliches Denkwerkzeug. Sie ermöglicht" Äußerungen vielfältigster Art "über Vergangenheit, Gegenwart und Zukunft... Und eben dadurch wird die Sprache zum Herrschaftsinstrument" [Danzin et al., 1992: 10].

Wenn man Sprachen übersetzen will, macht man sich Gedanken über den möglichen Zusammenhang der Sprachen und damit auch über ihren Ursprung, den man jedoch nicht wird enträtseln können [vgl. die Bibliographie von Hewes, 1975]. Wilhelm von Humboldt [1820: 14f] sagt dazu folgendes: "Die Sprache muß zwar, meiner vollsten Überzeugung nach, als unmittelbar in den Menschen gelegt angesehen werden; denn als Werk seines Verstandes in der Klarheit des Bewußtseins ist sie durchaus unerklärbar. ... So natürlich die Annahme allmählicher Ausbildung der Sprache ist, so konnte die Erfindung nur mit ein e m Schlage geschehen. Der Mensch ist nur Mensch durch Sprache; um aber die Sprache zu erfinden, müßte er schon Mensch sein.
... Sie geht notwendig aus ihm selbst hervor ..., aber so, daß ... mithin das erste Wort schon die ganze Sprache antönt und voraussetzt. ... Die wahre (Natur) der Spracherfindung liegt nicht sowohl in der Aneinanderreihung und Unterordnung einer Menge sich aufeinander beziehender Verhältnisse, als vielmehr in der unergründlichen Tiefe der einfachen Verstandeshandlung, die überhaupt zum Verstehen und Hervorbringen der Sprache auch in einem einzigen ihrer Elemente gehört. Ist dies gegeben, so folgt alles übrige von selbst, und es kann nicht erlernt werden, muß ursprünglich im Menschen vorhanden sein." Die Sprache wird also als integrierender Bestandteil des Menschseins gesehen.
st es ein Faktum, daß die Menschheit lückenlos in Sprachgemeinschaften gegliedert ist r, 1964]. "Durch die gegenseitige Abhängigkeit des Gedankens und des Wortes voneinander klar ein, daß die Sprachen nicht eigentlich Mittel sind, die schon erkannte Wahrheit darzustellen, it mehr, die vorher unerkannte zu entdecken. Ihre Verschiedenheit ist nicht eine von Schällen n, sondern eine Verschiedenheit der Weltansichten selbst." [Humboldt, 1820: 27]. Jede Sprache ıls spezifischer "Zugriff' auf die Welt ihr eigenes "Weltbild" auf [Weisgerber, 1962].

Eine Verschiedenheit des Weltbildes verhindert nun keineswegs prinzipiell eine Übersetzung [vgl.: ses, 1990: 16], noch wird diese durch eine den verschiedenen Sprachen gemeinsame so

genannte "Tiefenstruktur" [Chomsky, 1973; vgl. Rolland, 1994: 10; 551f.] ermöglicht - denn jede Sprache hat ja gerade ihr *eigenes Weltbild* -, sondern es verhält sich nach Humboldt [1820:16f.] - wie auch die Übersetzungspraxis zeigt - so: "Die Erfahrung bei Übersetzungen aus sehr verschiedenen Sprachen ... zeigt ..., daß sich, wenn auch mit großen Verschiedenheiten des Gelingens, in jeder jede Ideenreihe ausdrücken läßt. Dies aber ist bloß eine Folge der allgemeinen Verwandtschaft aller und der Biegsamkeit der Begriffe und ihrer Zeichen", d.h. [1820: 28f.]:

"Vergleicht man in mehreren Sprachen die Ausdrücke für unsinnliche Gegenstände, so wird man nur diejenigen gleichbedeutend finden, die, weil sie rein konstruierbar sind, nicht mehr und nichts anderes enthalten können, als in sie gelegt worden ist. Alle übrigen... enthalten weniger und mehr, andere und andere Bestimmungen. Die Ausdrücke sinnlicher Gegenstände sind wohl insofern gleichbedeutend, als bei allen derselbe Gegenstand gedacht wird; aber da sie die bestimmte Art, ihn vorzustellen, ausdrücken, so geht ihre Bedeutung darin gleichfalls auseinander. ... Denn da die Sprache zugleich Abbild und Zeichen (ist, ist sie) nicht ganz Produkt des Eindrucks der Gegenstände, und nicht ganz Erzeugnis der Willkür des Redenden."

Hiernach ist es verständlich, daß und wieso es durchaus in gewisser Weise Entsprechungen zwischen verschiedenen Sprachen geben kann, die sich u. a. jedoch in ihrer Ausprägung und den

damit vorliegenden sprachspezifischen Charakteristika möglicherweise völlig unterscheiden (z.B. kann dem Wort in der einen Sprache ein Nebensatz in der anderen Sprache entsprechen: die *semantikorientierte* Vorgehensweise - the approach *based on semantics);* auch können andere "Bilder" gebraucht werden *(Himmel und Erde in Bewegung setzen - to leave no stone unturned);* ggf. sind *keine* Entsprechungen vorhanden, so ist z.B. *Kindergarten* in englischen Texten in Deutsch beizubehalten.

Wenn man Sprachen übersetzen will, von denen jede in einer Fülle von Wörtern und den damit bedingten Spracheigentümlichkeiten, wie z.B. ggf. Wortart, Flexion und Konstruktionsmöglichkeiten vorliegt, dann ist es zunächst erforderlich, jede Sprache in ihrer spezifischen Beziehungsstruktur zu untersuchen und diese aufzudecken und dann erst die Entsprechungen zu denjenigen Sprachen herzustellen, in die die Ausgangsprache übersetzt werden soll. Auch wenn die einzelne Sprache in sich spezifisch strukturiert ist, so erscheinen die Prinzipien ihres jeweiligen Aufbaus gleichartig, so daß man grundsätzlich die gleichen Verfahren verwenden kann, um die einzelne Sprache zu erforschen. Auf welche Weise dies realisierbar ist, ist am Beispiel der deutschen Sprache detailliert dargelegt [Rolland, 1994]. An dieser Stelle sollen daher nur die zum Verständnis notwendigen allgemeinen Grundlagen über die Sprachstruktur wiederholt werden. Im übrigen sei auf die genannte Darstellung verwiesen. Im Anschluß daran wird im folgenden gezeigt, und zwar an Hand von Beispielen aus den Sprachen: Deutsch und Englisch, wie man sich die maschinelle Übersetzung vorzustellen hat.

2 Der Aufbau der Sprache

Die Sprache ist - wie in Kapitel 1 betont - gemäß Humboldt [vgl.: 1830-35: 46] ihrem *Wesen* nach eine *"wirkende Kraft",* die dazu dient, "die Welt in das Eigentum des Geistes umzuschaffen", die also in der Auseinandersetzung mit der Welt diese anverwandelt und dadurch ein spezifisches *Weltbild* schafft. Ihrer *Daseinsform* nach ist die Sprache, zu der neben der Lautform entscheidend der Inhalt gehört, eine *Wirklichkeit* [Weisgerber, 1962: 15], so daß es möglich ist, die *Inhalte* in dieser Existenzform zu erforschen - man ist also bei der Sprachuntersuchung nicht auf die *Sachwelt* angewiesen, sondern kann rein sprachimmanent verfahren. Ihrer Struktur nach ist die Sprache ein wohlgeordnetes System [Saussure, 1967: 27], ein *Beziehungsgefüge,* das nach erkennbaren Regeln und Gesetzmäßigkeiten aufgebaut ist, die es explizit zu machen gilt.

Das Kernelement der Sprache, aus dem die größeren Einheiten: Syntagmen, Sätze und Texte aufgebaut werden, ist das Wort. Das Wort ist der Baustein aller Zusammenhänge und der Schlüssel zur Erkenntnis. Daher heißt die Methode, auf der die Sprachuntersuchung beruht: *Logotechnik* ("Worthandhabung"). So ist im folgenden einzugehen auf das *Wort* als Ganzheit, die *Wortart* - eng gekoppelt mit den *Satzgliedern,* auf den komplexen *Wortinhalt* und seine Einzelkomponenten, die ihrerseits die Satzstruktur steuern, so daß man schließlich in den *Bauplänen* das *potentielle Relationsgefüge* einer Sprache erhält, d.h. die möglichen Vorgaben für die Realisierung jedweder sprachlichen Äußerung.

Dieses Relationsgefüge ist sprachspezifisch für die einzelnen Sprachen zu ermitteln. Dann sind diese Relationsgefüge für die Übersetzung parallel zu setzen - eine zeitaufwendige und schwierige Aufgabe für Übersetzungsexperten. Aus diesen potentiellen Relationen und ihren Entsprechungen lassen sich dann alle Syntagmen, Sätze und Texte identifizieren und übersetzen, da jede sprachliche Äußerung einen Extrakt aus diesen möglichen Vorgaben darstellt. Der Computer kann dann richtig "übersetzen", aber nicht, weil er auf einmal den Textinhalt "verstehen" würde, sondern weil er "gemäß seinen Fähigkeiten" lediglich die sprachlichen Pendants auffinden muß, wie sie im Vorfeld als mögliche Entsprechungen von den Übersetzungsexperten angegeben sind.

2.1 Das Wort

Das Wort ist eine Ganzheit aus Lautform und Inhalt, wobei der Inhalt maßgebend ist tUr die Bestimmtheit des Sprachmittels [Weisgerber, 1962: 206ff].

Abb. 1 Struktur des Wortes

Diese Wortdefinition hat ungeheure Konsequenzen; denn sie besagt, daß nicht die Lautform, sondern der Inhalt das Wesentliche darstellt, daß also ganz entscheidend Homonyme zu klären sind. Es gibt demnach Z.B. nicht *ein* Wort: *Zug* mit vielen Bedeutungen, sondern viele verschiedene Wörter mit der gleichen Bezeichnung, z.B.: *Zug1* (Eisenbahn), *Zug2* (Schriftzug), *Zug3* (Charakterzug), *Zug4* (Durchzug), *Zug5* (Gesichtszug) usw.

Auf diese Weise wird deutlich, daß jedes Wort nur einen, wenn auch komplexen Inhalt hat - die vielzitierte Mehrdeutigkeit und Ungenauigkeit löst sich auf. Auf Grund seines Inhalts bestimmt jedes Wort seine potentielle Wortumgebung selbst, die aber erst dann eindeutig feststellbar wird, wenn neben der Homonymenklärung auch ermittelt ist, welche Wortarten es in einer Sprache wirklich gibt, welche Satzglieder mit ihnen korrespondieren, welche Funktion die Flexionsausprägung hat und wie es sich mit der relationalen Abhängigkeit zwischen den Wörtern verhält. An dieser Stelle soll nicht noch einmal eine Auseinandersetzung mit ähnlich scheinenden vorhandenen Ansätzen, wie Valenztheorie, Kasustheorie, Parsingverfahren usw. erfolgen [vgl. dazu: Rolland, 1994: 8ff; Rolland 1995b]. Es wird hier lediglich auf zentrale, *neu* erkannte Faktoren im Bereich der Sprachstruktur hingewiesen, so daß allein schon dadurch verdeutlicht wird, daß alles, was darauf aufbaut, dementsprechend auch anders ist als bei den bisherigen Verfahren.

Es geht hier zunächst also um die *Wortidentifikation,* d.h. um die Feststellung, welche unterschiedlichen Wörter es gibt. So bewirkt z.B. die Homonymenklärung, die in allen Wortarten durchzuführen ist, auch große Veränderungen beim Verb, also im Konjugationssystem, wobei hiervon an dieser Stelle aus der Fülle der zu klärenden Wörter nur drei Problemkreise herausgegriffen werden sollen: "Subjektkategorien", "Partizipien" und "Hilfsverben":

Subjektkategorien:
Es sind nicht nur die bekannten 3 Personen (vereinfacht statt: Lebewesen) im Singular und Plural zu unterscheiden, sondern es muß außerdem auch als eigenständig und nicht unter das Personensubjekt subsumiert [Duden, 1995: §324] das Sachsubjekt im Singular und Plural berücksichtigt werden, so daß man inhaltbezogen 8 *Subjektkategorien* erhält (man kann ja jetzt nicht mehr von *Personen* sprechen): ich, du, er-sie-es (person), er-sie-es (Sache), wir, ihr, sie (Person), sie (Sache), z.B.: *"Es* (das Kind) schläft. - *Es* (das Buch) hat mir gut gefallen." Es gibt dementsprechend Verben, die sich in allen Formen von Aktiv und Passiv realisieren, wie z.B. *ziehen ("der Vater* zieht das Kind" und Passiv, *"das Lotsenboot* zieht das Schiff' und Passiv) oder auch solche, die ein Personensubjekt im Aktiv, ein Sachsubjekt im Passiv aufweisen, Z.B.: *mitteilen (er* teilte mir das Ergebnis mit, *das Ergebnis* wurde mir mitgeteilt) vgl. ferner: *versenden, andeuten* usw.
Die Unterscheidung von Person und Sache ist eine entscheidende Komponente innerhalb der Struktur der deutschen Sprache, z.B.:

Person	Sache
Lehrer	Tisch
anschwärzen	blühen
geizige	nichtige
gern	saftig
seitens	anläßlich

Abb. 2 Person-/ Sachbezug

Partizipien:
Es gibt nicht nur 2 Partizipien [vgl. Duden, 1995: §208], sondern 4, da man unterscheiden muß zwischen den Partizipien: Aktiv Präsens und Perfekt und Passiv Präsens und Perfekt: *ziehend, gezogen (habend), gezogen (werdend), gezogen (worden seiend),* auch wenn normalerweise die Klammerausdrücke: *habend, werdend, worden seiend* nicht im Text stehen.

Hilfsverben:
Es sind nicht nur, wie bisher üblich, 3 Hilfsverben zu unterscheiden, sondern 4; d.h. neben den entsprechenden Vollverben *(sein* - existieren, *sein* - sich befinden, *sein* - gehören; *haben* - besitzen; *werden* - sich verwirklichen und den Kopulaverben: *sein:* groß sein; *haben:* Glück haben; *werden:* fertig werden) gibt es die Hilfsverben: *sein* und *haben* zur Bildung der Zeiten des Perfektstammes sowie *werden-Futur* (kommen werden) und *werden-Passiv* (gerufen werden). Es lassen sich eindeutig die Verbformen der beiden Hilfsverben: *werden* voneinander unterscheiden [vgl. Rolland, 1994: 60].

Eine weitere einschneidende Änderung wird durch die Unterscheidung von *Adjektiv* und *Adverb* bedingt, z.B.: Adjektiv: das *schöne* Buch; Adverb: das Buch ist *schön.* Es handelt sich nicht [vgl. Rolland, 1994: 95ff.], wie bisher meist angenommen [Duden, 1995: §447], um ein "prädikativ" gebrauchtes Adjektiv, sondern um 2 verschiedene Wärter, die auch anders gesteigert werden, Z.B.: Adjektiv: (das) *schöne, schönere, schönste* (Buch): Adverb: *schön, schöner, am schönsten* [vgl. Weinrich, 1976: 230; Bergenholtz/ Schaeder, 1977: 108].

Von besonderer Bedeutung ist auch die Homonymenklärung bei Präpositionen. So sind in haltbezogen u. a. eine ganze Reihe von Wärtern: *aus* zu unterscheiden, z.B.:

aus 1:	aus dem Haus	(von welchem Ort?)	
aus2:	aus Gold	(aus welchem Material?)	
aus3:	aus Haß	(aus welchem Grund?)	
aus4:	aus der Tasse	(aus welchem Gefäß?)	
aus5:	aus dem 3 . Jahrhundert	(aus welcher Zeit?)	usw.

Erst wenn die Homonyme geklärt sind, stehen die Wörter fest und damit das Ausgangswortmaterial für die Computerverarbeitung; denn es ist das Wortgut, das auch der Mensch verwendet, wenn er spricht.

2.2 Wortarten und Satzglieder

Ohne an dieser Stelle auf die Begründung für den Aufweis der Wortarten im einzelnen eingehen zu wollen [vgl. Rolland, 1994: 53ff], soll hier nur das grundsätzliche Ergebnis vorgestellt werden:

Im Deutschen sind 6 in sich in vierfacher Weise gegliederte Wortarten zu unterscheiden, zu denen bestimmte Satzglieder in Wechselbeziehung stehen. Zur Erläuterung der Wortart *Konjunktion* sei folgendes angemerkt: Konjunktionen verbinden gleichartige Satzglieder. Sie differieren gemäß ihrem Gebrauch. Es gibt *spezielle Konjunktionen,* wie Z.B.: *weil, damit, obwohl* usw., die 2 Prädikate verknüpfen, wobei sich aus der Verbindung von Konjunktion und Prädikat bestimmte *Nebensätze* ergeben (z.B.: Es freut mich, *daß* du *kommst.* - Wir gehen spazieren, *obwohl* es *regnet.).* Außerdem gibt es die *generellen Konjunktionen,* wie z.B.: *und, sowie, oder* usw., die den ursprünglichen Satz um ein gleichartiges Satzglied erweitern (z.B.: Er kaufte Brot *und Butter),* wobei Konjunktion und "Ursprungssatzglied" ein neues Satzglied ergeben, das terminologisch als *Konjunktionalbestimmung* gefaßt ist. Neben den Ursprungssatzgliedem: Subjekt, Prädikat, Objekt, Umstandsbestimmung, Attribut gibt es also auch noch die Konjunktionalbestimmung. Somit stellen alle Wörter im Satz, allein oder in Kombination, auch Satzglieder dar. Danach sind - ohne auf die Untergliederungen und alle Nebensatzmöglichkeiten im einzelnen einzugehen [vgl. Rolland, 1994: 117ff] - zu unterscheiden:

Wortarten	Satzglieder
Verben	Prädikat
Substantive (einschließlich Artikel und Pronomina)	Subjekt, Objekt, Umstandsbestimmung
Adjektive	Attribut (so definiert)
Adverbien (einschließlich Interjektionen)	Adverbsubjekt Adverbobjekt Adverbumstandsbestimmung
Präpositionen	Präpositionalobjekt präpositionale Umstandsbestimmung
Konjunktionen: - spezielle - generelle	 bestimmte Subjekt-, Objekt-, Umstandssätze Konjunktionalbestimmungen

Abb.3 Wortarten und Satzglieder

Erst durch die detaillierte Einsicht in Art, Anzahl und Ausprägung der Wortarten und den zu ihnen in Wechselbeziehung stehenden Satzgliedern ist die Grundlage zum Durchschauen der Satzstruktur gelegt. Da der Mensch sich sprachlich in diesen Einheiten bewegt, müssen sie so genau erforscht werden, wenn sie einer Computerverarbeitung zugänglich sein sollen.

2.3 Der Wortinhalt

Der Wortinhalt ist ein komplexes Ganzes [vgl. Rolland, 1994: 121ff]. Als solcher steht er selbst eine Struktur aus Inhaltzügen - im Verhältnis wechselseitiger Abhängigkeit zu der Struktur, die den von ihm aus aufweisbaren Wörtern zugrunde liegt [vgl. Rolland, 1994: 129]. Indem man diese Struktur feststellt, erhält man eine linguistisch exakte Grundlage für eine

Deutung des Inhalts; denn Inhalte lassen sich als geistige Ganzheiten nur beschreibend umfassen [Weisgerber, 1962]. Im folgenden wird gezeigt, welches diese Strukturen prinzipiell sind.

Der Wortinhalt besteht aus 2 Teilinhalten, dem *speziellen Inhalt,* der nur dem einen Wort eignet (z.B. meint: *beschaffen* immer *beschaffen,* gleichgültig, in welcher Form das Wort ausgeprägt ist: *er beschaffte, sie werden beschaffen, ihr hättet beschafft* usw.) und dem *generellen Inhalt,* den das Wort mit Wörtern der gleichen Wortart teilt.

Der 1. Teilinhalt, der spezielle Inhalt jedes Wortes, besteht aus 4 Teilzügen, wobei es so ist, daß inhaltlich ähnliche Wörter, je nach dem Grad ihrer inhaltlichen Ähnlichkeit, in ein, zwei oder drei Teilzügen übereinstimmen, im vierten unterscheiden sie sich nur noch in Nuancen, z.B.: *genehmigen, erlauben, die Erlaubnis geben, die Erlaubnis erteilen, gestatten* u. a. haben inhaltlich gemeinsam, daß jemand die Meinung äußert, *daß* (etwas geschehen) *dürfe.* Die Wörter: *erlauben, die Erlaubnis geben, die Erlaubnis erteilen* haben darüber hinaus gemeinsam, daß diese Äußerung auf Grund der *Autoritätsstellung* erfolgt - sie differieren darin, daß *erlauben* "verantwortungsbewußt" ist, *die Erlaubnis geben* größeres Gewicht hat und "besonders verantwortungsbewußt" ist, *die Erlaubnis erteilen* noch größeres Gewicht hat und "zutiefst verantwortungsbewußt" ist [vgl. Rolland, 1969, 291f].

Im generellen Inhalt sind 2 Inhaltzüge zu unterscheiden:

- der 1. *generelle Zug,* der die *Flexionsausrichtung* meint, also die gemäß Tempora, Kasus bzw. Komparation bestehende Inhaltkomponente. Im 1. generellen Zug erfolgt die *Verknüpfung zum Satzglied,* was formal zum Teil wortimmanent, zum Teil in mehreren Wörtern ausgeprägt ist, Z.B.:

- Substantiv:	Subjekt::	der Computer (funktioniert)
- Verb:	Prädikat:	(er) hat gesagt, sag-te, ging
- Adjektiv:	Genitivattribut:	(des) schöne-n (Buches)
- Adverb:	Umstandsbestimmung:	am schönsten
- Präposition:	Präp. Umstandsbestimmung:	auf den Tisch, auf dem Tisch
- Konjunktion:	Umstandssatz:	(Sie freute sich,) obwohl er lachte.
	Objektsatz:	(Er verriet,) daß sie fertig war.
	Konjunktionalbestimmung:	(Sie verkauft Bücher) und Hefte.

  Das heißt: Jedes Wort jeder Wortart hat solch eine Flexionsausrichtung, auch wenn die "Beugung" nicht am Wort direkt greifbar ist, sondern, wie bei den Präpositionen, am Kasus der abhängigen Wörter sichtbar wird (bei Substantiven) bzw. in einer Frage (bei Adverbien), z.B.: *bis jetzt* (bis wann, bis zu welchem Zeitpunkt?). Außerdem trifft dies entsprechend auch auf die Konjunktionen zu, die ja gleichartige Satzglieder verknüpfen und dadurch ihren Flexionsbezug erhalten.

- der 2. generelle Zug, der die *Konstruktionsausrichtung* meint, also die Relationen des Ausgangswortes zu den abhängigen Wörtern. Daher erfolgt im 2. generellen Zug die *Verknüpfung zwischen Satzgliedern.,* z.B.:

Beispiel Relation die Rede *des*		*Satzglieder*
Direktors gefiel	wessen, die wer äußert?	Subjekt - *Genitivobjekt*
Er verschenkte das Buch *des Direktors*	wessen, das wer besitzt?	Akkusativobjekt - *Genitivobjekt*
er kam *gestern*	wann, während welchen Zeitraums?	Prädikat - *Umstandsbestimmung*

Zu beachten ist, daß die Konstruktionsausrichtung, also die Relationen, in *spezifizierten Fragewörtern* greifbar sind, die *Konstruktionselemente* genannt werden, wobei sich die Fragewör-

ter gleichsam als Pendant zu der jeweiligen *Feststellung* erweisen. Die Relationen ihrerseits stellen eine Komponente im 2. generellen Zug des jeweiligen Ausgangswortes dar. So hat Z.B. nicht der Genitiv eine *Funktion* (Genitivius subjektivus), sondern das Ausgangswort hat eine *Beziehung,* da sich das inhaltliche Spezifikum durch die Relation des Ausgangswortes ergibt wie die angeführten Beispiele verdeutlichen. Das Faktum des Satzgliedes dagegen, im Beispiel: das *Genitivobjekt,* ist in vielen Fällen gleich. Die Strukturformel des Wortinhaltes lautet danach:

<table>
<tr><td colspan="3">Gesamtinhalt</td></tr>
<tr><td colspan="3">Teilinhalte</td></tr>
<tr><td>l. Teilinhalt</td><td colspan="2">2. Teilinhalt</td></tr>
<tr><td>spezieller Inhalt</td><td colspan="2">genereller Inhalt</td></tr>
<tr><td></td><td>1.
genereller Zog</td><td>2.
genereller Zog</td></tr>
</table>

Abb. 4 Die Strukturformel des Wortinhalts

Wenn auch zur *Feststellung* der Teilinhalte eine solche Aufgliederung in Einzelzüge vorgenommen wird, so muß der Wortinhalt selbst, da er eine eigenständige Ganzheit darstellt, auch immer als solche betrachtet werden. Bei der Rede von sogenannten Überschneidung oder Überlagerung von Inhalten handelt es sich um eine unangemessene Betrachtungsweise des Inhalts, bei der man inhaltliche Einzelzüge herausgreift, *verselbständigt* und vergleicht. Im Gegensatz dazu zielt die Rede von der Einmaligkeit eines Inhalts auf den Gesamtinhalt, der in sich gegliedert ist und die verschiedensten Einzelzüge umfaßt, der aber als eigenständiges Ganzes einem anderen eigenständigen Ganzen gegenübersteht. Indem man nun zunächst linguistisch exakte Vorgaben ermittelt und diese dann zum Ausgangspunkt einer Interpretation macht, erhält man in dieser die Inhaltbeschreibung des Wortes [Beispiele: Rolland, 1969: 272ff].

2.4 Gliederung des Wortschatzes

Die inhaltliche Ähnlichkeit der Wörter sowie ihre Ausrichtung in der Flexion (Verknüpfung zum Satzglied) und in der Konstruktion (Verknüpfung zwischen Satzgliedern), d.h. ihre *Inhaltkomponenten,* bilden die Kriterien für die Gliederung des Wortschatzes. Im einzelnen ergibt sich folgende Differenzierung:

Gemäß ihrer inhaltlichen Ähnlichkeit gliedern sich die Wörter in sogenannte *Sinnklassen,* d.h. Gruppen gerade noch ähnlicher Wörter [zum Sinnklassenkriterium vgl. Rolland, 1969: 115]. Innerhalb dieser Sinnklassen ergeben sich Gruppen inhaltlich ähnlicher Wörter gemäß den genannten 4 Teilzügen des *speziellen Inhalts,* so daß man insgesamt als *Wortklassen* (nicht zu verwechseln mit Wortarten!) erhält: *Großklassen, Kernklassen* und die *Kleinklassen* mit den *Einzelwörtern,* z.B.:

Sinnklasse	"Verben der sprachlichen Äußerung"
: Großklasse:	"Verben der Meinungsäußerung"
Kernklasse:	"erlauben"
Kleinklasse:	erlauben, die Erlaubnis geben, die Erlaubnis erteilen
	gestatten
	genehmigen, die Genehmigung geben, die Genehmigung erteilen
	stattgeben usw.

Die einzelnen Klassen erhalten zum leichteren Auffinden eine *Wortklassennummer.* So heißt Z.B.: "WK6: 3.4.21.1" = Wortklasse 6 (Verben), Sinnklasse 3, Großklasse 4, Kernklasse 21, Kleinklasse 1. Eine Kleinklasse enthält meist nur sehr wenige Wörter, oder auch nur ein einziges, so daß es sich erübrigt, auch noch die Einzelwörter zu nummerieren.

Gemäß ihrer jeweiligen generellen Inhaltzüge ergliedern sich die Wörter in folgender Weise:

- Nach dem 1. generellen Zug (der spezifischen Flexionsausrichtung) ergeben sich pro Wortart jeweils 10 Flexionsklassen, so daß man insgesamt 60 Flexionsklassen für das Deutsche erhält. Jedes Wort hat Anteil an einer dieser Klassen [vgl. Rolland, 1994: 164 ff]
- Nach dem 2. generellen Zug (der spezifischen Konstruktionsausrichtung) ergibt sich eine noch festzustellende Anzahl bestimmter Konstruktionsklassen (so gehören Z.B. alle die Verben in eine spezielle Klasse, die die gleiche Art und Anzahl von Relationen haben, die in den entsprechenden Fragewörtern greifbar sind, wie: *wer, welche Person?; was, welche Sache?; wann, zu welchem Zeitpunkt?; wann, während welchen Zeitraums?; wann, innerhalb welchen Zeitraums?; wann, nach Ab/auf welcher Zeit?* usw.)

Insofern in den generellen Zügen die *allgemeinen wortartlich geprägten Einheiten* zum Ausdruck kommen und jeweils bestimmte Wörter eine Flexionsklasse und eine Konstruktionsklasse miteinander teilen, ergeben sich von daher *Kategorialklassen.*

Jedes Wort jeder Wortart wird gemäß seiner 3 Inhaltkomponenten, dem speziellen Inhalt, dem 1. generellen Zug und dem 2. generellen Zug, in dreifacher Weise geordnet: in eine Sinnklasse, eine Flexionsklasse und eine Konstruktionsklasse.

Wortarten	**Wortschatzgliederung**	
Verb/ Substantiv/ Adjektiv/ Adverb/ Präposition! Konjunktion	Sinnklassen	
Verb/ Substantiv/ Adjektiv/ Adverb/ Präposition! Konjunktion	Flexions-klassen	}Kategorial-
Verb/ Substantiv/ Adjektiv/ Adverb/ Präposition! Konjunktion	Konstruktions-klassen	} klassen

Abb. 5 Struktur des Wortschatzes

Das bedeutet, daß der Wortinhalt selbst die semantische Gliederung des Wortschatzes steuert und daß damit eine von der Sprache vorgegebene Ordnung des Wortschatzes vorliegt, in der jedes Wort seine Stellung im Ganzen hat, und zwar jeweils *einmal* in der Sinnklasse, der Flexionsklasse und der Konstruktionsklasse. Wenn diese Ordnung ermittelt ist, kann man je nach Wunsch und Erfordernis aus diesen Vorgaben das gewünschte Wortgut auswählen und sich die notwendigen Wörter zusammenstellen. Angenommen, man will das Wortgut für bestimmte Sachgebiete und für bestimmte Systeme zusammenstellen, dann wählt man aus den Sinnklassen, in denen das Wortgut nach der inhaltlichen Ähnlichkeit gegliedert ist, die betreffenden Wörter aus und erhält sie jeweils in der vorgegebenen Ordnung. Damit sind alle Systeme in diesem wichtigen Punkt kompatibel. Das gleiche Wort, das nur *einmal* innerhalb der Sinnklassen vorkommt, kann also mehrfach, d.h. so oft, wie benötigt, in den Sachgruppen angerührt werden. Sach- und Sprachordnung stehen also gewissermaßen quer zueinander.

2.5 Baupläne - Relationsgefüge

Wie funktioniert nun das Miteinander der Wörter? Jedes Wort ist auf Grund seiner Inhaltkomponenten so ausgerichtet, daß es einerseits andere Wörter bedingen kann (neue <- *Bücher)und* andererseits selbst von anderen Wörtern bedingt werden kann (kaufen -> *Bücher).* Es ist nun so: In den Relationen sind die Abhängigkeiten der Konstruktionsausrichtung des Ausgangswortes greifbar. Insofern jedes Wort einen nur ihm eignenden Inhalt hat - selbst wenn dieser sich aus Einzelzügen zusammensetzt, die das Wort mit anderen Wörtern teilt -, auf Grund dieser spezifischen inhaltlichen Ausprägung ist die Gesamtheit der abhängigen Wörter ausgangswortspezifisch; d.h. es meint:

beschaffen:	
wer?	Beschaffende Instanz (Händler, Kaufmann, ...)
was?	Beschaffungsobjekt (Möbel, Bücher, Job, ...)
wann, zu welchem Zeitpunkt?	Beschaffungszeitpunkt (am 15.5.95, ...)
wann, während welchen Zeitraums?	Beschaffungszeitraum? (gestern, heute, ...) usw.
bestellen:	
wer?	Bestellende Instanz (Kunde, Bedarfsträger, ...)
was?	Bestellungsobjekt (Möbel, Bücher, Menü, ...)
wann, zu welchem Zeitpunkt?	Bestellungszeitpunkt (am 15.5.95, ...)
wann, während welchen Zeitraums?	Bestellungszeitraum? (gestern, heute, demnächst, ...)
usw.	

Es gibt also durchaus bei den abhängigen Wörtern Überschneidungen, aber ihre Gesamtkonstellation ist pro Ausgangswort verschieden. Außerdem ist bei Zeitangaben zu berücksichtigen, daß - da der Verbalprozeß in Zeiten der Gegenwart, Vergangenheit und Zukunft abläuft - Verbform und Umstandsbestimmung zusammenpassen müssen, z.B.: "Er *beschaffte* das Buch *gestern",* nicht möglich: * "Er *wird* das Buch *gestern beschaffen".*

Die Gesamtheit der von einem Ausgangswort abhängigen Relationen, die die zugehörigen Konkretisierungen implizieren, heißt *Bauplan.* Jedes Wort jeder Wortart hat also, gemäß seiner spezifischen inhaltlichen Ausprägung, seinen eigenen Bauplan. Stimmen mehrere Wörter in Art und Anzahl der die Relationen kennzeichnenden Fragewörter überein, so gehören sie lediglich dem gleichen *Bauplantyp* an. Der Terminus Bauplan meint also, der Intention der *inhaltbezogenen Grammatik* gemäß [Weisgerber, 1962] die geistige Konzeption möglicher Abhängigkeiten *eines* Wortes *(beschaffen:* wer? Händler usw., was? Möbel usw.), nicht aber, wie bisher üblicherweise angenommen, den Plan zum Bau von Sätzen bzw. Syntagmen *vieler verschiedener* Wörter *(beschaffen, kaufen, holen:* wer, was, wem, wann? usw.) - und dann meist auch noch der kleinsten Einheiten, wie sie valenzmäßig festgelegt sind. Jede Äußerung ist somit ein Extrakt aus den Vorgaben, wie sie im Bauplan vorliegen. An Arten von Bauplänen sind zu unterscheiden:

Baupläne	**Wortarten**
Satzbaupläne	auf der Basis des Verbs
Syntagmenbaupläne	auf der Basis der übrigen Wortarten
erweiterte Baupläne	auf der Basis der generellen Konjunktionen

Abb. 6 Baupläne und Wortarten

Die Gesamtheit aller Baupläne ergibt das mögliche Beziehungsgefüge oder das *potentielle Relationsgefüge* der Sprache. Auf dieses Relationsgefüge greift der Mensch zu, wenn er spricht. Nur geschieht das so selbstverständlich und unreflektiert korrekt, daß keinem die Struktur dieses Relationsgefüges bewußt ist.

Durch den speziellen Einblick in den Aufbau der Sprache [Rolland, 1969: 29fI], der besagt, daß jedes Wort die Bedingungen für seine Geltung selbst enthält und daß diese Bedingungen in der Struktur der von einem Ausgangswort aus aufweisbaren Wörter greifbar ist, ist es nunmehr möglich, dieses Relationsgefüge explizit zu machen. Man muß dabei beachten, daß die Relationen, die die Konstruktionsausrichtung ausmachen, zu unterscheiden sind von den Bedingungen, die für den Aufweis des speziellen Inhalts notwendig sind. Selbst wenn man also bei einem Verb wie z.B. *lügen* feststellt:

1. *genereller Zug:* Flexionsformen mit Personensubjekt im Aktiv,
2. *genereller Zug:* Relationen: *wer, wen, warum, wie oft, wozu, wann?* und dgl. - dann hat man noch nicht die Angaben, die man für den Aufweis des *speziellen Inhalts* benötigt. Bei einem "Sprachverb" wie: *lügen* ginge es dabei um die Feststellung des Äußerungsinhalts, der bei

lügen, belügen, vorlügen; schwindeln, vorschwindeln usw. in einigen inhaltlichen Einzelzügen ähnlich ist, in anderen gerade differiert [Details s. Rolland 1969; 1994: 154].

Ferner ist die übliche Aussage von den flektierten und nicht flektierten Wortarten nicht haltbar, wie oben dargelegt ist [anders: Duden, 1995: §121ff].

Wortarten	Relationsgefüge
Verben Flexion	Satzbaupläne: Relationen -> Konkretisierungen aus den Wortklassen ->Flexion
Substantive	Syntagmenbaupläne: Relationen -> Konkretisierungen aus den W ortklassen ->Flexion
Adjektive	Syntagmenbaupläne: Relationen -> Konkretisierungen aus den Wortklassen ->Flexion
Adverbien	Syntagmenbaupläne: Relationen -> Konkretisierungen aus den Wortklassen ->Flexion
Präpositionen	Syntagmenbaupläne: Relationen -> Konkretisierungen aus den Wortklassen ->Flexion
Konjunktionen spezielle, generelle	Syntagmenbaupläne: Relationen -> Konkretisierungen aus den Wortklassen ->Flexion
generelle Konjunktionen	erweiterte Baupläne: - erweitert um das Satzglied: Prädikat -> Satzbauplan - erweitert um das Satzglied: Subjekt! Objekt! Umstandsbestimmung! Attribut: -> Syntagmenbauplan - erweitert um das komplexe Satzglied: Subjekt-/ Objekt-/ Umstands-/Attributsatz: -> Satzbauplan

Abb. 7 Die Sprache als Relationsgefüge

3 Grundlagen eines Computersystems

Die Sprache ist also, wie eingangs betont, ihrer Struktur nach ein Beziehungsgefüge - das aber nunmehr in seinem Aufbau durchschaut werden kann. Jede Sprache stellt auf Grund ihres nur ihr eignenden Weltbildes solch ein spezifisches Relationsgefüge dar. Indem man das Relationsgefüge mit Kennungen, den sogenannten *Kategorieneinträgen* versieht [Rolland, 1994: 352ff], die sich beziehen auf: Wortart, Genus, Kasus, Numerus, Substantivart (Person! Sache) usw. und ganz entscheidend auch auf die Relationen in Form von Fragewörtern, also die Konstruktionselemente, erhält man die sogenannte *Relationsbasis.* Auf der Grundlage der Relationsbasis wird die *Wissensbasis* automatisch aufgebaut. Im folgenden gehen wir daher auf diese beiden Komponenten näher ein.

3.1 Relationsbasis

Die zentralen Elemente der Relationsbasis stellen die Baupläne dar, also Satzbaupläne, Syntagmenbaupläne und die erweiterten Baupläne. In den *Bauplänen* ist angegeben: Ausgangswort, Relation, Wortklasse (z.B.: *kaufen: was? -> Geräte WK1.2.25),* in der *Wortklasse* (WK1.2.25): Bezugswort und Relation mit Hinweis auf die Flexion, terminologisch gefaßt als Formgruppe (z.B.: *Computer, was?, Formgruppe X).* In der *Formgruppe* erfolgt der Vergleich mit dem Be

zugswort (z.B.: *Computer*), das dann als relevant identifiziert werden kann. Ist das geschehen, werden die bei der Formgruppe ermittelten Kategorieneinträge in das *Deskriptionsschema* eingetragen und zusammengestellt, bis alle Wörter des Satzes entsprechend gekennzeichnet sind. In diesen Kennungen erhält man u.a. in der Gesamtheit der festgestellten Relationen, die ja in Form von Fragewörtern vorliegen, als Pendant zum Feststellungssatz einen Fragesatz. Dieser bzw. Teile des Fragesatzes bilden dann später die Ausgangselemente der *Frage*, die der Benutzer an das System stellt. *Antwort* sind dann die Wörter des Feststellungssatzes.

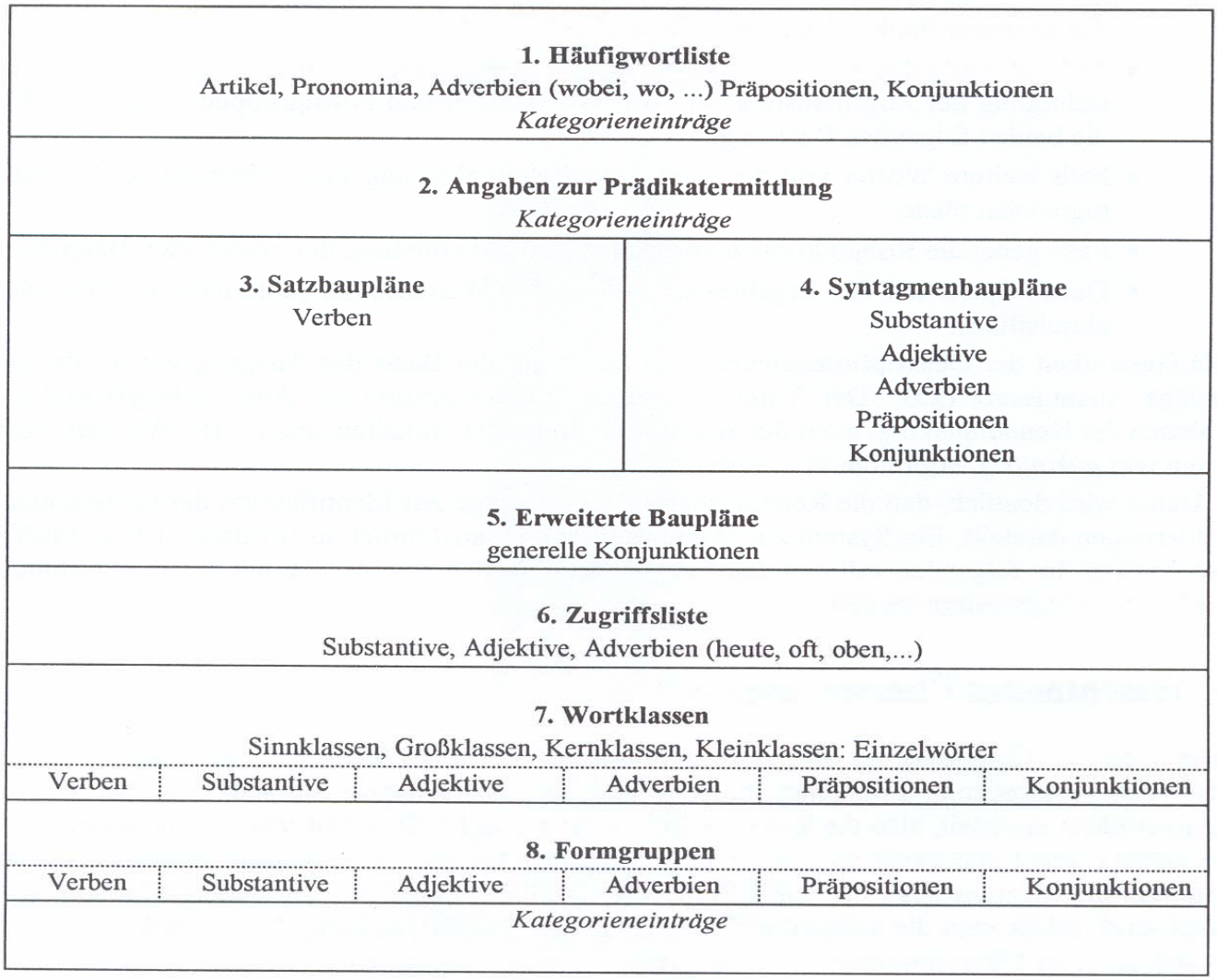


Abb. 8 Struktur der Relationsbasis

Entscheidend ist, da bei der Ermittlung der Satzstruktur alles auf dem Verb beruht, die *Ermittlung des Prädikats*. Um dieses möglichst schnell feststellen zu können, werden zunächst mit Hilfe einer *Häufigwortliste* die häufig vorkommenden Wörter wie Artikel usw. durch entsprechende Kategorieneinträge gekennzeichnet, um die Zahl der für die Prädikatermittlung infrage kommenden Wörter zu reduzieren. Ist das Prädikat (mit Hilfe entsprechender Fakten und Regeln) festgestellt, dann wird das Ergebnis der Prädikatermittlung in Form von Kategorieneinträgen im Deskriptionsschema festgehalten. Ferner ist eine alphabetische *Zugriffsliste* der Formen der Nicht-Verben und der Nicht-häufigen-Wörter aufgestellt, mit deren Hilfe über die Wortklassennummer die Zuordnung des Wortes zum betreffenden Konstruktionselement schnellstmöglich erfolgen kann. Die Struktur der Relationsbasis besteht also aus: Häufigwortliste, Angaben zur Prädikatermittlung, Satzbauplänen, Syntagmenbauplänen, erweiterten Bauplänen, Zugriffsliste, Wortklassen und Formgruppen.

3.2 Wissensbasis

Die *Wissenbasis* wird dadurch aufgebaut, daß auf der Relationsbasis operiert wird und das Ergebnis, also das jeweilige Deskriptionsschema, als Einheit der Wissensbasis gespeichert wird. Auf diese Weise ergibt sich eine automatische Wissensakquisition. Die einzelnen durchzuführenden Schritte (vgl. die Struktur der Relationsbasis) sind folgende:

- Einlesen eines Satzes und Kennzeichnung der Position der Wörter.
- Vergleich der Wörter mit der Häufigwortliste und Kennzeichnung der Treffer.
- Als zentraler Punkt: Prädikatermittlung.
- Auf der Basis des Prädikats: Abarbeitung des Satzbauplans, und zwar unter Berücksichtigung der Zugriffsliste sowie von Wortklassen und Formgruppen - was auch für die beiden folgenden Punkte gilt.
- Falls weitere Wörter von den direkt ermittelten abhängig sind: Abarbeitung der Syntagmenbaupläne. . Falls generelle Konjunktionen vorhanden sind: Abarbeitung der erweiterten Baupläne. . Durch Speichern der Ergebnisse: Aufbau der Wissensbasis (automatische Wissensakquisition) .

Die Gesamtheit der Deskriptionsschemata, in denen auf der Basis des Ausgangswortes der jeweilige Aussagesatz (z.B.: Der Vater *hat* seiner Tochter gestern ein Auto *gekauft)* und im Rahmen der Kennungen u.a. auch der zugehörige Fragesatz enthalten sind (z.B.: Wer *hat* wem wann was *gekauft?),* ergibt die Wissensbasis.

Damit wird deutlich, daß die Relationsbasis die Grundlage zur Identifikation der eingelesenen Äußerungen darstellt. Ein System zur Wissensabfrage ist ausführlich in [Rolland, 1994: 346ff] beschrieben. Im folgenden soll nun deutlich gemacht werden, wie es sich mit einem maschinellen Übersetzungssystem verhält.

4 Maschinelles Übersetzungssystem

Indem man im Gegensatz zu vorhandenen Ansätzen [vgl. z.B.: Slocum, 1989; Fabricz, 1989] nach den angegebenen Prinzipien zunächst jede zu übersetzende Sprache in ihrer Beziehungsstruktur ermittelt, also die jeweilige Relationsbasis aufstellt, erhält man die Strukturen der Sprachen explizit und damit so, wie sie dem Menschen bei seinem Sprachtun vorliegen. Indem man nun die Relationsbasis der einen Sprache und die Relationsbasis der anderen Sprache parallel setzt, erhält man die adäquaten Entsprechungen. Die Herstellung der Parallelität ist eine *einmalige,* von Übersetzungsexperten durchzuführende Aufgabe und gilt dann für alle Äußerungen, da jede Äußerung einen Extrakt aus diesen Vorgaben darstellt. Somit kann jeder Satz, jedes Syntagma im Rahmen dieses Wortgutes identifiziert und daher auch automatisch übersetzt werden; denn der Computer sucht ja nur die innerhalb der Vorgaben existierenden Entsprechungen auf

Welche Sprache Ausgangssprache, welche Zielsprache ist, bleibt sich gleich, da ja eine Parallelität zwischen den relevanten Relationen der Sprachen und ihren Konkretisierungen, d.h. den abhängigen Wörtern in den entsprechenden Flexionsformen besteht. Hinzu kommen bei den jeweiligen Sprachen natürlich noch die *Wortstellungsregeln,* mit deren Hilfe dann aus dem "Parallel-Gesetzten" der eigentliche Satz in der korrekten Wortabfolge in der einzelnen Sprache gemacht wird. Je nach Spracheigentümlichkeit einer Sprache sind ggf weitere Besonderheiten zu berücksichtigen, was per Regel geschehen kann.

Prinzipiell verhält es sich dann so: In der Ausgangssprache wird ein Satz eingelesen und gemäß den Vorgaben in der Relationsbasis identifiziert und mit den entsprechenden Kennungen

versehen. Von diesen Kennungen ausgehend, werden in der Zielsprache die Entsprechungen festgestellt, wobei ggf. *Spezialregeln* anzuwenden sind (vgl.u. 4.2). An Hand der *Wortstellungsregeln* wird dann aus den äquivalenten Angaben die Übersetzung in die endgültige Form gebracht. Der allgemeine Übersetzungsvorgang ist folgender:

Ausgangssprache		**Zielsprache**
↓		↓
Relationsbasis	↔	Relationsbasis
↓		↓
Baupläne ↓ Relationen ↓ abhängige Wörter/ Flexion	↔ ↔ ↔	Baupläne ↓ Relationen ↓ abhängige Wörter/ Flexion
		+ Regeln
↑		↓
Ausgangssatz		***Übersetzter Satz***

Abb. 9 Übersetzungsvorgang

4.1 Relationsbasen-Entsprechungen

Im folgenden wird nun an Hand von Beispielen aus den Sprachen Deutsch und Englisch gezeigt, wie es sich im einzelnen mit dem Aufbau der Entsprechungen verhält. Die Fragewörter (Konstruktionselemente) sind noch zu spezifizieren. (Problemfälle sind z.T. dem Verbmobil-Grundlagenpapier entnommen [DFKI et al., 1991]).

a)Zunächst soll gezeigt werden, daß einem Verb der Ausgangssprache (gefallen) mit bestimmten Relationen und Konkretisierungen ein Verb der Zielsprache (to like) mit bestimmten Relationen und den zugehörigen Konkretisierungen entsprechen kann, daß jedoch andere Relationen äquivalent sind, z.B.:

Ausgangssprache		**Zielsprache**
↓		↓
Relationsbasis	↔	Relationsbasis
↓		↓
Bauplan: gefallen: *gefällt* ↓ Relation: was(Nom)? → *dasAuto* Relation: wem? → *dem Jungen* Relation: ... usw.	↔ ↔ ↔ ↔	Bauplan: to like: *likes* ↓ Relation: what(Akk)? → *the car* Relation: who? → *the boy* Relation: ... usw.
Dem Jungen gefällt das Auto. ↑ ***Ausgangssatz***		***Übersetzter Satz*** ↓ The boy likes the car.

Abb. 10 Übersetzungsbeispiel: gefallen - to like

b) Bei diesem Beispiel werden zusätzlich noch Wortstellungsregeln einbezogen. Falls Englisch die Zielsprache ist, gelten die Regeln (Englisch), falls es umgekehrt ist, die Regeln (Deutsch). [Zur Unterscheidung und Terminologie der im folgenden aufgeführten Konjugationsformen s. Rolland, 1994: 58ff.]. BV = Basisverb; HV-F = Hilfsverb-Futur; HV-P: Hilfsverb-Passiv; sE = subjektbezogenes Element; prädE = prädikatbezogenes Element.

Ausgangssprache: Deutsch		**Zielsprache: Englisch**
↓		↓
Relationsbasis	↔	Relationsbasis
↓		↓
Bauplan: diskutiert werden:	↔	Bauplan: to be discussed
Form:		Form:
werden (HV-F) (sE)	↔	will (HV-F) (sE)
diskutiert (BV) (prädE)	↔	discussed (BV) (prädE)
werden (HV-P) (prädE)	↔	be (HV-P) (prädE)
↓		↓
Relationen:		Relationen:
wann? → *dann*	↔	when? → *then*
welche Sachen? → *Entwicklungen*	↔	what things? → *developments*
wie beschaffene Entwicklungen? → *neue*	↔	what developments? → *recent*
wie? → *kurz*	↔	how? → *briefly*
Zwischenergebnis		**Zwischenergebnis**
Dann wann?		Then when?
werden (HV-F) (sE)		will (HV-F) (sE)
neue wie besch. Entwicklungen?		recent what developments?
Entwicklungen welche Sachen?		developments what things?
kurz wie?		briefly how?
diskutiert (BV) (prädE)		discussed (BV) (prädE)
werden. (HV-P) (prädE)		be. (HV-P) (prädE)
Regeln (Deutsch)		**Regeln (Englisch)**
1. Im Aussagesatz HV-F (sE) an die 2. Satzgliedstelle		1. Subjekt vor Prädikat
2. Prädikatbezogene Elemente ans Ende des Satzes, und zwar ungetrennt (durch ein anderes Wort) und in der Reihenfolge: BV(prädE), HV-P(prädE)		2. Reihenfolge der Prädikatelemente: HV-F (sE), HV-P(prädE), BV(prädE)
		3. Adverb vor BV(prädE)
Dann werden neue Entwicklungen kurz diskutiert werden.		***Übersetzter Satz***
↑		↓
Ausgangssatz		Then recent developments will be briefly discussed.

Abb. 11 Übersetzungsbeispiel: diskutiert werden

c) Bei den Äquivalenten kann es auch so sein, daß einem Wort der Ausgangssprache (Deutsch) mit einer bestimmten Relation und einer bestimmten Wortgruppe, also Konkretisierungen, in

der Zielsprache (Englisch) jeweils eigene Ausgangswörter entsprechen, oder, umgekehrt betrachtet, daß verschiedenen Wörtern der Ausgangssprache (Englisch) mit einer speziellen Relation und einer zugehörigen Wortgruppe das gleiche Wort mit der gleichen Relation und Äquivalenten im Wortgut in der Zielsprache (Deutsch) entsprechen, z.B.:

	Deutsch	Englisch
Verb:	*Einhalt gebieten*	*control*
Relation:	welcher Sache?	what?
Konkretisierungen	*Gefühlen, ...*	*emotions, ...*
.		
Verb:	*Einhalt gebieten*	*stop*
Relation:	welcher Sache?	what?
Konkretisierungen	*Tätigkeiten, ...*	*activities, ...*
.		

Es sind dies (a, b, c) nur einige Beispiele fiir die vielfältigen verschiedenartigen Möglichkeiten der Entsprechungen innerhalb der Relationsbasen mehrerer Sprachen.

4.2 Wörter und Spezialregeln

Im folgenden wird an Hand von 14 Beispielgruppen demonstriert, wie man sich weiterhin die Parallelisierung bzw. die Regelungen in Problemfällen vorzustellen hat. Es sei nachdrücklich betont, daß hierbei Vollständigkeit nicht angestrebt ist und ggf. die einzelnen Regeln noch zu erweitern sind.

1) Vokabelentsprechungen
Es ist durchaus möglich, daß einem Wort der Ausgangssprache ein einzelnes Wort der Zielsprache entspricht - es können aber auch mehrere sein, z.B.:

Einzelwort/ Einzelwort:	Tisch/ table
Einzelwort/ Mehrwortausdruck:	Seegang/ motion of the sea
	Zufällig/ by chance
	Kostenlos/ free of charge
	Klingeln/ to ring the bell
	Berücksichtigen/ to take into account notfalls/ in case of need
	Schwarzbrot/ brown bread
	Geschwister/ brothers and sisters
Mehrwortausdruck! Mehrwortausdruck:	zu diesem Zweck/ to this end
	nach Belieben/ at will
	auf den Zehenspitzen/ on tiptoe
	unter anderem! among other things vertraut werden mit! to become familiar with mit dem Bus fahren! to go by bus
unbestimmter Artikel:	ein Blumenstrauß/ a bunch of flowers
ohne/ mit Artikel (Kopulaverb):	Arzt sein/ to be a doctor
Wort (Adjektiv: Stellung vor dem Substantiv/ Teilsatz: nachgestellt):	die semantikorientierte Vorgehensweise/ the approach based on semantics

2) Idiome sind zu übernehmen:
der Zweck heiligt die Mittel/ the end justifies the means
über Berg und Tal/ over hedge and ditch
stille Wasser sind tief/ still waters run deep
Himmel und Erde in Bewegung setzen/ to leave no stone unturned

3) Homonyme Adjektive sind als abhängig vom zugehörigen Substantiv anzusetzen:

ein schwieriger Fall	a hard case
eine schwierige Aufgabe	a difficult task
ein großer Raum	a large room
ein großer Stuhl	a big chair
eine große Hoffnung	a great hope
ein großer Junge	a tall boy

4) Verschiedene Wörter im Deutschen, im Englischen Homonyme, transitiv bzw. intransitiv:

wachsen; anbauen	to grow (intr.); to grow (trans.)
Das Getreide wächst.	The corn grows.
Der Bauer baut Getreide an.	The farmer grows corn.

5) Frage:
Wenn nicht schon ein Hilfsverb im Satz vorhanden ist, wird die Frage mit *to do* umschrieben, es sei denn, daß das Fragewort selbst Subjekt oder Teil des Subjekts ist:

When *can* we see your new house? - Where *do* you live now? - *Which* road leads to the castle?

6) Negation

a. Gebräuchlich nur als Subjekte und in Antworten, die aus einem Wort bestehen:
keine Hunde no dogs ("Verlaufssubstantiv" [vg1.Rolland, 1994, 84ff.])
keiner meiner Freunde none of my friends (substantivisch + Bezugssubstantiv)
keiner wußte ... nobody (no one) knew (Substantiv)
Wer hat meinen Füller gebraucht? Keiner. Wbo has been using my fountain pen? Nobody.

b. In allen andern Fällen wird ein mit *not* verneintes Verb (Vollverb im Präsens und Imperfekt mit to do umschrieben) mit *a, one* bzw. *any. anywhere, ...* verwendet:
I haven't got *a* German dictionary. - I haven't got *one.* - He *don't* read *any* books.
The plane *didn't* stop *anywhere.*

c. Adverbial gebraucht steht: no vor dem Komparativ: Ich konnte nicht weiter gehen. - I could walk *no* further.
none the + Komparativ: Wir werden kein bißchen früher zurück sein. - We will get back *none the sooner.*
none too + "Adverb": Die Brücke ist nicht allzu sicher.- The bridge is *none too safe.*

7) Akkusativ mit Infinitiv (AC!)
a) Ua. folgende Verben im Englischen stehen mit ACI *ohne to:*
to see (sehen); to hear (hören); to watch (beobachten, sehen); to notice (bemerken); to observe (beobachten); to feel (spüren); to make (lassen, veranlassen; zwingen); to have (lassen, veranlassen; sagen, daß ... soll); to let (lassen, zulassen); to bid (heißen, befehlen).

Beispiel:
Jedes Geräusch ließ sie hochfahren. - Every noise made her jump.

b) U. a. folgende Verben im Englischen stehen mit ACI *mit to:*
- to advice (raten); to allow (erlauben, lassen); to ask (bitten); to beg (dringend bitten); to cause lassen, veranlassen, der Grund sein, daß ...); to expect (erwarten = verlangen vonjmd., daß ...); to get Gdn. dazu bringen); to induce Gdn. dazu bewegen); to lead (veranlassen; es fUhrt dazu, daß ...); to order (auffordern, befehlen, den Auftrag geben); to perrnit (erlauben, lassen); to require (verlangen); to tell (sagen, daß ... soll); to warn nachdrücklich darauf aufmerksam machen, dringend raten).
- I want (ich wünsche, möchte, will haben, daß ...); I wish (ich wünsche, möchte, will haben, daß ...); I should like (ich möchte, daß ...); I should prefer (es wäre mir lieber, wenn ...); I hate (ich kann es gar nicht leiden, wenn ...); I should hate (ich will nicht haben, daß ...).

- In der Schriftsprache: to believe, to consider, to declare, to find, to hold, to judge, to know, to recognize, to suppose, to take (= to suppose), to think, to understand. (Die Umgangssprache verwendet stattdessen that-Sätze.)

Beispiel: Ich weiß, daß er ein reicher Mann ist. - I know him to be a rich man.

Extrakt: Bauplan (Deutsch)		Bauplan (Englisch)
wissen: weiß		to know: know
wer? *ich*		who? *I*
was?		what?
a) *daß*	->	-
↓		
b) Präd (finit)	->	Präd-Infinitiv
↓		↓
ein Mann sein: ... *ein Mann ist*		to be a man: *to be a man*
↓		↓
c) wer? -> *er* -> Subjekt		whom? -> *him* -> Akkusativobjekt
d) Mann		man
welcher? ein *reicher* Mann		what? a *rich* man

Ergebnis:

Ich	I
weiß,	know
daß	
er	him
ein reicher Mann ist.	to be a rich man.

8) Man:

(meistgebrauchte Übersetzungsmöglichkeit)

Beispiel 1: Man versichert jdm. - (someone) ("be") assured

man versichert:		("be") assured
wer?	man	-
wem?	mir	who? I (am)
	dir	you (are)
	ihm usw.	he (is) usw.

Beispiel 2: Man glaubt, daß er ehrlich sei. - He is thought to be honest.

(man) *glaubt*:	who? (= Subjekt des deutschen daß-Satzes) *is thought*
wer? *man*	-
was?	what?
a) *daß*	-
↓	↓
b) ehrlich sein: *ehrlich sei*	to be honest: *to be honest*
↓	↓
c) wer?-> *er* -> Subjekt	who? -> *he* -> Subjekt

9) Verschiedene Verben, aber z.T. homonyme Formen.

Beispiel:

Vollverb:	schließen	to close	wurde geschlossen	*was closed*
Kopulaverb:	geschlossen sein	to be closed	war geschlossen	*was closed*

a) Die Tür wurde geschlossen. — The door was closed.

Extrakt: Bauplan (Deutsch-D)	Bauplan (Englisch-E)
geschlossen werden: wurde geschlossen	to be closed (Aktiv: to close)
was(Nom)?	what(Nom)?

b) Die Tür war geschlossen. — The door was closed.

Extrakt: Bauplan (Deutsch)	Bauplan (Englisch)

geschlossen sein — to be closed (Aktiv: to be closed)
was(Nom)? what(Nom)?

Die Übersetzung D-E stellt kein Problem dar. Bei der Übersetzung E-D benötigt man eine Regel, in der der Zusammenhang, d.h. bestimmte Relationen berücksichtigt sind.
Beispiel:
They came to the museum, but the door was closed *already.*
Sie kamen zum Museum, aber die Tür war *schon* geschlossen.
They came to the museum, but the door was *still* closed.
Sie kamen zum Museum, aber die Tür war *noch* geschlossen.

Wenn im Bauplan die Relation: *in welcher Weise fortgeschritten?* angegeben ist mit z.B. *schon* ... oder:
wenn im Bauplan die Relation: *in welcher Weise andauernd* angegeben ist mit z.B.: *noch, ...,*
oder: ...,
lautet das englische Verb *to be closed* (dem im Deutschen ja: *geschlossen sein* entspricht). In allen anderen Fällen lautet das englische Verb: to close (Entsprechung: *schließen).*
Beispiel:
When the crush became too great, the door *was closed.*
Als der Ansturm zu groß wurde, *wurde* die Tür *geschlossen.*

10) Tempusunterschiede: Perfekt – Past Tense
Im Deutschen wird häufig das Perfekt gebraucht, wenn im Englischen das Past Tense steht. Dieses Problem läßt sich per Regel lösen.
Beispiel:
Ich habe ihn gestern gesehen. — I saw him yesterday.

Die *allgemeine* Regel, die aber konkret in einer Relation und ihren Konkretisierungen greifbar ist, lautet: Enthält ein Satz eine Zeitbestimmung der Vergangenheit, die dort abgeschlossen ist und zur Gegenwart keine Verbindung hat, dann steht im Englischen das Past Tense.
Konkrete Regel:
Wenn die Zeitangaben lauten:
wann, während welchen Zeitraums?: *gestern, heute, vorgestern, in der letzten Woche, wahrend der Sommer pause, zur Bauzeit, ...*
wann, zu welchem Zeitpunkt?: *am* 18.4.1995, ..,
dann entsprechen sich deutsches Perfekt und englisches Past Tense.

11) Transitives und reflexives Verb im Deutschen, Homonyme im Englischen.
Beispiel 1:

Peter bewegte das Auto.		Peter moved the car.	
Extrakt: Bauplan (Deutsch)		Bauplan (Englisch)	
bewegen: bewegte		to move: moved	
wer?	Peter, ... das	who?	Peter
was(Akk)?	Auto, ...	what(Akk)?	the car

Beispiel 2:

Das Auto bewegte sich.		The car moved. Bauplan (Englisch)	
Extrakt: Bauplan (Deutsch)		to move: moved	
sich bewegen: bewegte sich			
was(Nom)?	das Auto	what(Nom)?	the car

12) Wortstellung
Beispiel:

Ein Vorschlag, den zu machen er mutig wagte.		A proposal, which he dared courageously to make.	
Extrakt: Bauplan (Deutsch)		Bauplan (Englisch)	
Vorschlag		proposal	
welcher?	den ->Prädikat	which?	which -> Prädikat

wagen: wagte			dare: dared		
wer?	er		who?	he	
was?	(Substantiv)		what?	(noun)	
	->	Präd: zu-Infinitiv		->	predicate: to-infinitive
		machen: zu machen			to make: to make
wie?	mutig		how?	courageously	

Die Entsprechungen sind angegeben. Im übrigen gelten für das Englische u.a. als Wortstellungsregeln: Subjekt vor Prädikat, bei komplexen Verbformen: Adverb nach dem Hilfs- bzw. Modalverb (im weiteren Sinne), so daß man erhält:
Aproposal, wbich *he dared courageous/y to make.*
Bei Englisch als Ausgangssprache würde bei einer Übersetzung die übliche Wortstellung im Deutschen sein: Ein Vorschlag, *den er mutig zu machen wagte.*

13) Ellipsen

Durch Regeln ist hier eine korrekte Übersetzung herbeizuführen.

a) Beispiel:

John hat das Buch gelesen und Martha auch. John has read the book and Martha has too.

John hat das Buch gelesen und seine Freunde auch. John has read the book and bis friends have too.

John las das Buch und Martha auch. John read the book and Martha did too.

Regel: D - E:
Bei mehreren, durch *und, oder* ... verknüpften Subjekten mit nur *einem* Prädikat wird im Englischen im verknüpften Satz nicht das volle Prädikat wiederholt, sondern nur das Hilfsverb der Prädikatform, und zwar in der zum Subjekt passenden Numerusform. Gibt es kein Hilfsverb, dann wird die entsprechende Form von to do eingesetzt.

Regel: E - D:

Bei mehreren, durch *and, or* ... verknüpften Subjekten mit nur *einem* Prädikat und dem Hilfsverb im verknüpften Satz wird im Deutschen im verknüpften Satz das Hilfsverb eliminiert.

b) Beispiel:

Ich habe John gekannt, aber sie (hat) ihn nicht (gekannt). I have known John, but she has not. I know John, but she doesn't.
Ich kenne John, aber sie (kennt) ihn nicht.

Regel: D - E:
Bei mehreren, durch *aber* ... verknüpften Sätzen mit zwei gleichen Wörtern als Prädikat oder nur einem Prädikat wird im *but-Satz* nicht das volle Prädikat wiederholt, sondern nur das Hilfsverb. Gibt es kein Hilfsverb, dann wird die entsprechende Form von to do eingesetzt. Ein im verknüpften Satz wiederholtes Objekt in Form eines Pronomens wird eliminiert.

Regel: E - D:

Bei mehreren, durch *but* ... verknüpften Sätzen mit nur einem vollen Prädikat wird im *aber-Satz* das volle Prädikat wiederholt, und zwar in der vom Subjekt geforderten Form, wobei auch die Form von to do entsprechend umgesetzt wird. Ein ggf. im Hauptsatz angeführtes Objekt wird im Deutschen im Nebensatz zusätzlich als Pronominalobjekt eingefügt. Falls der "aber-Satz" jedoch eine Ellipse darstellen soll, entfällt hier das Prädikat.

c) Beispiel:
Er sah den König und dankte ihm. He saw and thanked the king.
(König = Basissubstantiv, ihm = Pronomen)
Regel: D - E:
Bei mehreren, durch *und, oder* ... verknüpften Prädikaten mit nur einem Subjekt und einem Objekt (Basissubstantiv), bedingt vom l. Prädikat, und einem darauf bezogenen Objekt (pronomen im gleichen Genus und Numerus), bedingt vom 2. Prädikat, wird (I) nach der Grundübersetzung (2) das Pronomen durch das Objekt (Basissubstantiv) im verbbedingten Kasus des 2. Prädikats ersetzt und (3) das Objekt (Basissubstantiv) beim l. Prädikat entfällt.
Die Schritte im einzelnen sind:

(1): He saw the king and thanked him. (2):
He saw the king and thanked the king. (3):
He saw and thanked the king.

Regel: E - D:
Bei mehreren, durch *and, or* ... verknüpften Prädikaten mit nur einem Subjekt und einem Objekt (Basissubstantiv) beim 2. Prädikat, wird (1) nach der Grundübersetzung (2) das Objekt (Basissubstantiv) durch das zugehörige Objekt (pronomen im gleichen Genus und Numerus) ersetzt und (3) das Objekt (Basissubstantiv) wird im verbbedingten Kasus zum 1. Prädikats gestellt.
Die Schritte im einzelnen sind:

(1): Er sah und dankte dem König.
(2): Er sah und dankte ihm (dem König).
(3): *Er sah den König und dankte ihm.*

14) Verlaufsform: *ing* (Vorgang von vorübergehender Dauer), Beispiele D-E:

die Formen mit -ing sind nur im Aktiv sowie im Präsens und Past Tense Passiv möglich, aber nicht bei Verben wie z.B.:

to hear, to notice, to see, so smell, so taste
to dislike, to hate, to like, to love, to mind, to prefer
to desire, to want, to wish
to agree, to believe, to doubt, to fell (= be the opinion that), to know, to remember, to seem, to suppose, to think, to understand
to be, to belong, to contain, to exist, to have (= possess), to possess, to own

Bei der Relation: wann: zu welchem Zeitpunkt? mit der Konkretisierung: "gerade" wird die ing-Form verwendet, z.B.: Er liest *gerade* den Brief. - He is reading the letter.
Ein Präsens mit Zeitraumangaben im futurischen Sinn, wie: *tomorrow, next week, in July, at the week-end,...* (wenn also etwas: beabsichtigt, geplant, vereinbart ist) steht in der *-ing* Form, z.B.:
Ich fahre morgen in Ferien. - I am leaving for my holiday tomorrow.
Ein Futur (wenn also etwas: beabsichtigt, geplant, vereinbart ist) steht in der *-ing* -Form, z.B.: Er wird uns alles über das Treffen berichten. - He will be telling us all about the meeting.
Ich werde weitere Vorträge über dieses Gebiet halten. - I will be giving further lectures in tbis area.

Bei: "jemand *wollte, aber",* d.h. er hat es *aber* nicht getan, steht: *going to,* z.B.:
Ich wollte in München studieren, aber ich habe meine Absicht geändert. - I was going to study in Munich, but I changed my mind.
Bei Verstärkung der Aussage durch z.B.: "bestimmt, sicher", z.B.:
In London zu leben kostet *bestimmt* viel. - Living in London is going to cost quite a lot.
Bei: *to be sure, to expext,* ... (vermutetes Eintreten einer Sache), z.B.:
Ich erwarte, daß du länger als 3 Jahre zur Beendigung deiner Studien brauchst. - I expect you are going to take longer than 3 years to finish your studies.

Beim Präsens bzw. Präteritum mit *schon* im Deutschen bzw. Zeitangaben, wie: *seit wann?; wie lange schon? ...* (*the whole evening, for the past few weeks, for fifteen minutes, for two hours,for two weeks,* ...) steht das Perfect Continuous bzw. das Past Perfect Continuous, z.B.: Ich *lese schon* den ganzen Abend. - I *have been reading* the whole evening.
Ich *las schon* zwei Stunden. - I *had been reading* for two hours.

5 Realisierbarkeit einer automatischen maschinellen Übersetzung

Ausgehend von der Erkenntnis, daß die Welt in der Sprache in spezifischer Weise anverwandelt ist, daß dieses Geistige, die "Bedeutung", den Wortinhalt ausmacht und daß alle Inhalte als geistige Wirklichkeiten existieren, im Wort greifbar und erforschbar sind – davon ausgehend, ist in den Eingangskapiteln in den Grundlagen aufgezeigt, daß und wie die deutsche Sprache aufgebaut ist. Die Prinzipien gelten generell.

Da jedes Wort seine potentielle Wortumgebung bedingt und damit die Sprache als Beziehungsgefüge vorliegt, wenn der Mensch auf sie zugreift und spricht, läßt sich dieses Beziehungsgefüge explizit machen, indem man die jeweilige Wortumgebung feststellt. Das kann systematisch geschehen, indem man das Wortgut gemäß seiner 3 Inhaltkomponenten, dem speziellen Inhalt sowie dem 1. und 2. generellen Zug ordnet. Auf diese Weise erhält man die W ort klassen, die Flexionsklassen und als zentrale Komponente die Baupläne, in denen der Zugriff auf die Wortklassen und die Flexionsklassen erfolgt. In den Bauplänen aller Wörter aller Wortarten liegt die Sprache als potentielles Relationsgefüge vor. Jede sprachliche Äußerung des Menschen ist ein Extrakt aus diesen Vorgaben.

Indem man nun dieses potentielle Relationsgefüge explizit macht und mit Kennungen versieht, erhält man die Grundlagen für den Aufbau eines natürlichsprachlichen Computersystems. Will man Sprachen übersetzen, dann gilt es, zunächst das individuelle potentielle Relationsgefüge der einzelnen Sprache festzustellen und mit für den Computer zugreifbaren Kennungen zu versehen, so daß man die sogenannten Relationsbasen erhält. Diese müssen nun parallel gesetzt werden; d.h. dem einzelnen Wort (A) der Ausgangssprache mit einer speziellen Relation und speziellen abhängigen Wörtern mit ihren Flexionsausprägungen entspricht in der Zielsprache ebenfalls ein Wort (Z) mit der zugehörigen Relation und abhängigen Wörtern mit ihren Flexionsausprägungen.

Dieses Beziehungsmiteinander, also das potentielle Relationsgefüge mehrerer Sprachen, ist von Übersetzungsexperten *einmal* festzustellen und gilt dann für alle zu bildenden Sätze und Aussagen dieser Sprachen. Wortstellungsregeln und für sprachspezifische Sonderfälle Spezialregeln müssen diese Vorgaben vervollständigen. Hierfür sind eine Reihe von Beispielen angeführt, die das Prinzip verdeutlichen und zeigen, wie man im einzelnen vorzugehen hat. Ist das für mehrere Sprachen geschehen, dann benötigt man nicht mehr viele Sprachpaare (Deutsch – Englisch, Deutsch - Französisch, Deutsch - Italienisch usw.), sondern es ist lediglich notwendig, daß die einzelnen Sprachen relational vernetzt sind. Angenommen, es existieren die Verknüpfungen der Relationsbasen in der Reihenfolge: Deutsch, Englisch, Französisch, Italienisch, Spanisch, Deutsch, dann kann man den Ausgangssatz z.B. in Deutsch eingeben, er wird relational transformiert ins Englische, von dort ins Französische, von dort ins Italienische und dann mit Hilfe von Wortstellungsregeln u.a.. zur italienischen Übersetzung, die man haben möchte.

Bei allen bisherigen Versuchen, die natürliche Sprache maschinell zu übersetzen, hat man maximal für einen kleinen Anwendungsbereich Erfolge erzielt, bzw. es liegen Rohübersetzungen vor. Jede Generalisierung ist gescheitert. Hier liegt nun ein neuer Ansatz vor. Es ist unbestreitbar, daß viel investiert werden muß, um dieses Ziel zu erreichen; denn es sind ggf. eine ganze Reihe von Spezialfällen pro Sprache zu berücksichtigen; aber es ist ebenso unbestreitbar, daß mit dieser Vorgehensweise, wie sie hier geschildert ist, eine tragfähige generelle Lösung zur Überwindung der Sprachbarrieren vorliegt.

Literatur

Bergenholtz H.; Schaeder, B. 1977: Die Wortarten des Deutschen. Versuch einer syntaktisch orientierten Klassifikation. Stuttgart: Klett

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) 1995: Mobiles Dolmetschgerät. Verb mobil. CeBIT Hannover, 13. März 1995

Copeland, C.; Durand, J.; Krauwer, S.; Maegaard, B. (eds) 1991: The EUROTRA Linguistic Specifications. Luxembourg: Office for Official Publications of the European Communities, 2 vols = Studies in Machine Translation and Natural Language Processing 1/ 2

Danzin, A.; Studiengruppe für Strategiefragen 1992: Wege zu einer europäischen Sprachinfrastruktur. Be richt an die Kommission der Europäischen Gemeinschaften (GD XIII) = 5210/92 DE

Der Spiegel 1995: Verärgerte Kunden, Heft 22, S. 97 DFKI, IAI; IBM Deutschland; Siemens AG; Siemens-Nixdorf AG; Univ. Düsseldorf; Univ. Hamburg;
Univ. Karlsruhe; Univ. Stuttgart 1991: Verbmobil. Mobiles Dolmetschgerät. Studie. München

European Commission (EC), Directorate Generale XIII 1995: MT Summit V, July 10-13, 1995. Procee dings. Luxembourg

Fibricz, K 1989: Machine Translation. Active Systems. East-European Countries. In: Batori, I.S.; Lenders, W.; Putschke, W. (Hrsg.): Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin, New York: de Gruyter, 645-652 = Handbücher zur Sprach- und Kommunikationswissenschaft 4

Frankfurter Allgemeine Zeitung (FAZ) 1995: Fachleute tragen ihren Dolmetscher bald in der Jackentasche, 11./12.11.1995, S.16

Hewes, G. W. 21975: Language origins. A bibliography. The Hague: Mouton

Humboldt, W. von 1820: Über das vergleichende Sprachstudium = Werke IV, Leitzmann, A. (Hrsg.) Berlin:Behr's 1905

Reuse, B. 1994: Förderung der Anwendung der Künstlichen Intelligenz durch den Bundesminister für For schung und Technologie. In: KI, Sonderheft zur 18. Deutschen Jahrestagung für Künstliche Intelligenz, Juli 1994, S. 48-52 Rolland, M. Th. 1969: Zur Inhaltbestimmung der Sprachverben. Diss. Bonn

Rolland, M. Th. 1994: Sprachverarbeitung durch Logotechnik. Sprachtheorie, Methodik, Anwendungen. Bonn: Dümmler

Rolland, M. Th. 1995a: Ein semantikorientierter Ansatz im Bereich der Sprachverarbeitung. In: Hitzenberger, L. (Hrsg.): Angewandte Computerlinguistik. Vorträge im Rahmen der Jahrestagung 1995 der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.Y., Regensburg, 30./31.3.1995. Hildesheim: Olms 1995, S. 195-209 = Sprache und Computer 15

Rolland, M. Th. 1995b: Sprachverarbeitung auf der Basis der semantikorientierten Grammatik. In: LDV Forum 12 (1995) 1,9-27

Saussure, F. de 1967: Grundfragen der allgemeinen Sprachwissenschaft. Bally, Ch.; Sechehaye, A. (Hrsg.). Berlin: de Gruyter

SCS 1990: Scientific Control Systems GmbH: Maschinelle Übersetzung. Grundlagen, Stand, Perspektiven. Hamburg: SCS-Informationstechnik GmbH = SCS-Studie

Slocum, J. 1989: Machine Translation. A Survey of Active Systems. In: Batori, I.S.; Lenders, W.; Putschke, W. (Hrsg.): Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin, New York: de Gruyter, 629-645 = Handbücher zur Sprach- und Kommunikationswissenschaft 4

Weinrich, H. 1976: Sprache in Texten. Stuttgart: Klett

Welsgerber, L. 1962: Grundzüge der inhaltbezogenen Grammatik. Düsseldorf: Schwann = Von den Kräftender deutschen Sprache I (41971)

Weisgerber, L. 1964: Das Menschheitsgesetz der Sprache als Grundlage der Sprachwissenschaft. Heidelberg: Quelle & Meyer

Information Retrieval und Computerlinguistik

SIFT - Selecting Information From Text

Ein Projekt an den Universitäten von Heidelberg, Limerick und Amsterdam sowie von Lotus Development.

Dauer: 24 Monate

Aufwand: 123 Personen-Monate

Projektbeteiligte

- University of Limerick (Koordinierung)
- Lotus Development, Ireland
- University of Amsterdam
- Universität Heidelberg

Kontakt
Dr Richard Sutcliffe
Dept. of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland
Telephone:
+353 61 333644 Ext. 5006
+353 61 330876 Fax

Ziele

Das SIFT Projekt erarbeitet den Prototypen eines intelligenten Online-Hilfesystems für Software Manuals. Zwei Konzepte bilden die Grundlage:

- Das Vektormodell für Information Retrieval
- Verteilte Muster, mit denen Textbedeutung eingefangen werden soll.

In seinem Endausbau soll der Prototyp eine Anfrage in natürlicher Sprache entgegennehmen, mit der ein Nutzer nach softwarerelevanten Aspekten sucht. Ergebnis wird eine nach vermuteter Relevanz geordnete Liste von Verweisen auf Textstellen sein (Ranking), die zur Klärung der Frage beitragen könnten. Das Projekt soll darüber hinaus zeigen, daß verteilte Muster effektiv in praktischen Sprachverarbeitenden Systemen eingesetzt werden können und daß eine Integration mit bestehenden Techniken und Systemen fur lexikalische Datenbanken und für robustes lexikalisches Parsing erfolgreich ist. Dies wurde im Prinzip bereits in einem praktisch eingesetzten Retrievalsystem mit Erfolg realisiert, wo man das Vektormodell mit verteilten semantischen Repräsentationen kombinierte. Allerdings wurde dieses System weitgehend manuell aufgebaut, und SIFT soll die Technologie dafür bereitstellen, solche Systeme automatisch zu erzeugen.

Ansatz

Das SIFT -System wird aus zwei Hauptkomponenten bestehen:

- Die Dokumentenanalyse wird SGML-strukturierte Manuals auf der Ebene von Kapiteln, Unterkapiteln und einzelner Sätze mit verteilten Mustern in Verbindung bringen, mit denen die Bedeutung der Texteinheiten repräsentiert werden soll.
- Die interaktive Query-Komponente wird die Anfrage entgegennehmen und aus dem Anfrageergebnis ein Ranking für Textstellen erzeugen.

Das System basiert auf der Anwendung robuster lexikalischer Parsing-Techniken, der Zuordnung semantischer Rollen zu syntaktischen Konstituenten und der Extraktion verteilter Repräsentationen aus maschinenlesbaren Wörterbüchern.

Praktischer Nutzen und Ausblick

Auf textuelle Information gezielt zuzugreifen ist ein Problem in jeder Organisation. Genau für diese Aufgabe demonstriert SIFT eine innovative Lösung. Darüber hinaus können und sollen die eingesetzten Techniken in ei

nem weiten Spektrum verwandter Aufgaben der Textbearbeitung eingesetzt werden. So wird an die Integration in Stil-Checker, Zusammenfassungs-Generatoren und in Werkzeuge zur computerunterstützten Übersetzung gedacht. Aber es werden auch theoretische Einsichten in die Anwendbarkeit von Techniken zur automatischen Textanalyse erwartet, z.B. inwieweit verteilte Repräsentationen tatsächlich zur gleichzeitigen Abdeckung von Wort- und Satzbedeutung benutzt werden können, inwieweit ein großes, robustes und nicht spezialisiertes semantisches Lexikon automatisch aufgebaut werden kann, und ob ein robustes lexikalisches partielles Parsen möglich ist. Teil des Projektes ist es, die Überführung der SIFT - Technologie in eine kommerzielles Produkt zu untersuchen.
Quelle: http://itdsrvl.ul.ie/SIFT/sift-home-page.html

Venia "Computerlinguistik" für Dr. Karin Haenelt

Habilitationsverfahren an der Universität Heidelberg erfolgreich abgeschlossen

Karin Haenelt (GMD, Darmstadt), hat bei der Neuphilologischen Fakultät der Universität Heidelberg eine Habilitationsschrift mit dem Titel "Das KONTEXT-Modell. Verarbeitung natürlichsprachiger Texte auf der Basis eines Textmodells" eingereicht und am 8. Mai 1996 "mit Bravour das Habilitationskolloquiuum bestanden" (p. Hellwig).

Eine schriftliche Ausarbeitung ihres Vortrags *"Nachschlageverfahren im (elektronischen) Bedeutungswärterbuch* " soll unter dem Titel *"Looking-Up Procedures in an Electronic Meaning Dictionary: Considerations on the Role of a Meaning Dictionary in Textual Communication"* in Lexicographica erscheinen.

Eine Kurzfassung der Arbeit ist zugänglich als Arbeitsbericht der GMD: *"Das KONTEXT-Modell und die Konzeption der textmodellbasierten Verarbeitung natuerlich-sprachiger Texte.*" (Nr. 1009. GMD: Sankt Augustin, Juli 1996).

Eine allgemeine - nicht in allen Aspekten ganz aktuelle Beschreibung - des Kontext-Projektes und -Modells findet sich unter

lnt: http://www.darmstadt.gmd.de/KONTEXT/kontext.html

In diese Beschreibung sind verständlicherweise auch Informationen aus der Habilitationsschrift eingeflossen (z.B. „Das Kontext-Modell").

Karin Haenelt verfugt nun über die Venia "Computerlinguistik", und das LDV-Forum gratuliert herzlich.

Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. Hrsg. von Roland Hausser. Tübingen: Niemeyer, Reihe "Sprache und Information", Bd. 34, 1996, 181 S., 112 DM.

Der vorliegende Band mit dem Titel "Linguistische Verifikation" dokumentiert, wie dem Untertitel zu entnehmen ist, die Erste Morpholympics, eine in Form eines Wettkampfes durchgeführte Veranstaltung zur Bewertung von Systemen zur automatischen Wortformenerkennung des Deutschen, die im März 1994 vom Arbeitskreis "Parsing in Morphologie und Syntax" der Gesellschaft für Linguistische Datenverarbeitung durchgeführt wurde. Der Haupttitel des Buches erschließt sich dem Leser in der Einleitung des Herausgebers, in der dieser die Frage nach Methoden der Verifikation in der Linguistik im Vergleich zu den Naturwissenschaften stellt. Dem programmatischen Einleitungskapitel folgt zunächst der Ausschreibungstext, der im Vorfeld der Morpholympics im LDV-Forum 2 (1993) und über zahlreiche elektronische Listen im Internet bekannt gemacht wurde.

Sodann werden die Meßdaten präsentiert, die bei den Testläufen für die acht Systeme, die sich dem Wettstreit gestellt hatten, ermittelt wurden, sowie die Stellungnahme der Jury, der die Bewertung der Systeme oblag. Die daran anschließenden Kapitel 5 bis 13 umfassen die Darstellungen der acht teilnehmenden Systeme durch deren Autoren bzw. die präsentierenden Personen. Es handelt sich um das System MORPHY von W. Lezius, PC-Kimmo von A. Schiller, MORPH von G. Hanrieder, MORPHIX von W. Finkler und O. Lutzy, PLAIN von H. Visser und H.-D. Koch, LA-MORPH von G. Schüller und O. Lorenz, GERTWOL von K. Koskeniemmi und M. Haapalainen sowie MPRO von H.D. Maas. Den Abschluß der Dokumentation bildet ein Erfahrungsbericht des Koordinators der Veranstaltung.

Im einleitenden Kapitel argumentiert Hausser, daß der ursprünglich bestehende Mangel einer Verifikationsmethode in der Linguistik, wie sie in den Naturwissenschaften durch das Prinzip der Reproduktion oder in der mathematischen Logik durch die Axiomatisierung und die regelbasierte Ableitung von Theoremen bereitsteht, durch die Methoden, die die Computerlinguistik zur Verfugung stellt, behoben ist; denn Parser, die grammatische Regeln umsetzen, können als linguistisches Verifikations- bzw. Falsifikationsinstrument betrachtet werden. Aus diesem Blickwinkel wird die effiziente Implementierbarkeit zu einem Qualitätskriterium grammatischer Formalismen, das allerdings zahlreiche der gegenwärtig konkurrierenden Theorien bzw. Grammatikformalismen, deren Vielzahl Hausser als für die Linguistik ungut kritisiert, nicht erfüllen. Hausser wagt die Hypothese, daß mit der durch die Implementierung von Parsern gegebenen Möglichkeit des vergleichenden Testens konkurrierender Systeme langfristig auch ein Mittel bereitsteht, das die verschiedenen Theorien in konstruktiver Weise zu kanalisieren vermag. Die zentrale Rolle, die dem Wettkampf computerlinguistischer Systeme aus der Perspektive der linguistischen Verifikation zukommt, erfordert schließlich auch "eine einfache, klare und konsensfähige Definition der Kriterien für korrekte Analysen". Diese versucht der Herausgeber, der auch Koordinator der Ersten Morpholympics war, im letzten, mit "The Coordi-

nator's Final Report on the First Morpholympics" überschriebenen Kapitel zu entwickeln, indem er die Rahmenbedingungen und die Parameter, die für den Wettstreit der Systeme auf der Morpholympics zugrunde gelegt wurden, einer Kritik unterzieht und daraus Vorschläge zur Standardisierung ableitet, die bei künftigen Veranstaltungen diesen Typs klarere Bewertungsrichtlinien beim Vergleich verschiedener Systeme zur Verfugung stellen.

Der Fragenkatalog, der den Teilnehmern als Grundlage für die Darstellung der Systeme diente, ist Bestandteil des Ausschreibungstextes der Morpholympics gewesen. Dieser bildet in dem Dokumentationsband ein eigenständiges Kapitel. Auf die hier formulierten Kriterien, die das Ergebnis eines Morphologieworkshops mit internationaler Beteiligung sind, auf die im dritten Kapitel dargestellten Meßdaten, die bei den Testläufen der einzelnen Analysesysteme ermittelt wurden, sowie auf Stichproben der von den Systemen analysierten Daten stützt sich das Urteil der Jury, das als viertes Kapitel des vorliegenden ‚Buches publiziert ist. In diesem Kapitel machen die Mitglieder der Jury (W. Lenders, I. Bätori, G. Dogil, G. Görz, U. Seewald) deutlich, daß das Kriterium der linguistischen Fundiertheit für sie bei der Bewertung gegen über dem Zeitverhalten der Systeme im Vordergrund stand. Außer den drei ausgezeichneten Systemen (GERTWOL, MORPH und LA-MORPH) werden in diesem Beitrag auch die übrigen Analysesysteme gewürdigt. Einzelheiten der Systeme, die die Implementierung betreffen, sowie Angaben zu den Lexikoneinträgen und Beispielanalysen finden sich in den Folgekapiteln in den jeweiligen Einzeldarstellungen der Morpholympics-Teilnehmer selbst. Auch über den Aufbau der Systeme, die linguistischen Grundlagen und das jeweils angewendete Segmentierungsverfahren erhält der Leser hier einen Eindruck.

Vor dem Hintergrund der Erfahrungen mit der Ersten Morpholympics und der dort getroffenen Auswahl von Texten für die Testläufe sowie dem Fragenkatalog, auf den die Teilnehmer bei der Darstellung ihrer Systeme einzugehen hatten, entwickelt Hausser im letzten Kapitel Kriterien, die er zur Standardisierung von Testdateien und Bewertungskriterien vorsieht. Er stellt fest, daß die Auswahl von Testsätzen unter dem Gesichtspunkt eines möglichst natürlichen Analyselaufs der einzelnen Systeme, wie sie bei der Ersten Morpholympics praktiziert wurde, als problematisch einzustufen ist, da dieses Verfahren nicht gewährleistet, daß die Ergebnisse im Hinblick auf eine vergleichende Bewertung der einzelnen Systeme aussagekräftig sind.

Um dieses Dilemma zu beheben, stellt Hausser Kriterien zur Aufbereitung von Testdateien vor. Die Testdaten untergliedert er zunächst in Texte, die konzeptionell schriftlich sind, solche, die konzeptionell mündlich sind und transkribiert vorliegen, in Wörterlisten, die aus Textkorpora extrahiert werden und in solche, die gezielt nach bestimmten morphologischen und orthographischen Gesichtspunkten zusammengestellt werden. Hausser plädiert für die Verwendung von kommentierten Testdateien und nennt Angaben, die als Information im Kopf einer Datei in SGML-Format erscheinen sollten. Zu diesen Angaben zählen die Art der Darstellung von Umlauten, die Methode zur Ermittlung der Anzahl von Wortformen (denn hier treten in Abhängigkeit von der Zählmethode Schwankungen von bis zu 18% auf), die Gesamtzahl der Wortformen, die Zahl der wohlgeformten Wortformen, die Zahl der als nicht wohlgeformt eingestuften Wortformen sowie eine Auflistung dieser Wortformen. An die Einträge der manuell erstellten Listen knüpfen sich, so die Vorstellung Haussers, Fragen, die als Grundlage zur Bewertung der Qualität der morphologischen Analyse und zum Vergleich der Analyseergebnisse gedacht sind. Hier können beispielsweise die Behandlung der Valenzrahmen von Verben oder die Behandlung von Fugenelementen bei Komposita berücksichtigt werden.

Ein weiterer Fragenkomplex bezieht sich auf die Einzelheiten der Implementierung, so z.B. Fragen nach der Trennung von Regeln und Parser, dem Regelformat und den Fehlermeldungen. In diesem Teil sieht Hausser auch den Entwurf einer kleinen Beispielgrammatik einer nur wenig bekannten Sprache durch die Autoren eines morphologischen Analysesystems vor, um die Anpaßbarkeit der

Systeme an neue Daten zu überprüfen, eine Forderung, die allerdings im Rahmen einer Vorführung, wie Hausser dies vorschlägt, nur mit erheblichem zeitlichen Aufwand einzulösen sein dürfte.

Abschließend nennt Hausser einige Kriterien, die bei der Organisation und dem Verlauf künftiger Veranstaltungen Berücksichtigung finden sollten.

Neben dem zeitlich vorgeschalteten Einreichen der schriftlichen Präsentation der Systeme zählt hierzu einerseits die Forderung an alle teilnehmenden Systeme, bei jedem Analysedurchlauf automatisch Meßergebnisse zu erzeugen, um das fehleranfällige nachträgliche Berechnen von Hand zu vermeiden, andererseits aber auch die Sicherstellung gleicher Testvoraussetzungen für alle Systeme, so daß ausgeschlossen werden kann, daß das ermittelte Laufzeitverhalten Ergebnis eines zweiten Analysedurchlaufs ist, bei dem auf bereits in den Speicher geladene Daten zurückgegriffen wird.

Hausser erhofft sich von einem Verfahren, das die genannten Parameter, also Datenabdeckung, Geschwindigkeit, Qualität der linguistischen Analyse und Qualität der Implementierung, in der von ihm dargelegten Art berücksichtigt, zu einer zum Teil vielleicht komplizierteren, insgesamt aber objektiveren Bewertung von morphologischen Analysesystemen zu gelangen.

Uta Seewald, Universität Hannover

Linguistik in Internet. Das Buch zum Netz. Neuerscheinung im Herbst 96. Elisabeth Cölfen / Hermann Cölfen / Ulrich Schmitz (erscheint im Westdeutschen Verlag (Opladen)).

"Linguistik im Internet" ist ein praxisorientiertes Handbuch (mit CD) zur aktiven Nutzung des Internet für Linguisten und Geisteswissenschaftler. Es enthält einen sorgfältig kommentierten ausführlichen Reiseführer (mit screenshots) durch das deutschsprachige und internationale Angebot zu Linguistik und einigen Nachbargebieten und stellt ausgewählte Leckerbissen für Geisteswissenschaftler vor. Sinn und Möglichkeiten des Internet als neues Arbeitsmittel werden diskutiert.

Ein Lehrgang zeigt, wie man eigene Angebote ins Netz geben, eine Homepage und einen eigenen Server einrichten kann. Ein umfangreiches Internet-Adreßbuch, ausgewählte Literaturhinweise sowie ein Glossar runden den Band ab. Die mitgelieferte CD enthält sämtliche notwendige Zugangssoftware zum Internet, Hilfsprogramme zur eigenen Erstellung von WWW-Homepages sowie Webseiten mit Links zur Linguistik. Außerdem sind die Formulare der wichtigsten Internet-Suchwerkzeuge enthalten, so daß das gesamte Internet unmittelbar durchsucht werden kann.

Inhalt

Zu den Verfassern

Elisabeth Cölfen arbeitet als freie Journalistin für verschiedene Computerzeitschriften. Sie studiert Germanistik und ist als Web-Mistress

für LINSE (Linguistik -Server Essen: `http://www.uni-essen.de/f.b3/1inse/home.htm`) an der Universität GH Essen verantwortlich. Hermann Cölfen ist Doktorand und Mitarbeiter im Fach Germanistik/Linguistik an der Universität GH Essen. Ulrich Schmitz ist Professor für Germanistik/Linguistik und Sprachdidaktik an der Universität GH Essen. Das Buch entstand aus der gemeinsamen Erfahrung der drei Autoren mit dem Internet und dem Linguistik-Server. LINSE existiert seit Mitte 1995, und die Resonanz auf das Angebot hat die anfänglichen Erwartungen weit übertroffen. Die Autoren möchten den Leser über die bloße Information hinaus dazu ermutigen, sich durch eigene Beiträge einzubringen oder gar selbst Informationsanbieter zu werden.

Raum für Ihre Initiative

Folgende Bücher stehen gegenwärtig als Rezensionsexemplare zur Verfugung und können für eine kompetente Besprechung von der Redaktion angefordert werden:

- Porteie, Thomas: Ein phonetischakustisch motiviertes Inventar zur Sprachsynthese deutscher Äußerungen. 170 S. (incl. Anhang), Sprache und Information, Bd. 32, Tübingen: Niemeyer, 1996
- Volk, Martin: Einsatz einer Testsatzsammlung im Grammar Engineering. 178 S., Sprache und Information, Bd. 30, Tübingen: Niemeyer, 1995
- Eherer, Stefan: Eine Software Umgebung für die kooperative Erstellung von Hypertexten. 161 S., Sprache und Information, Bd. 29, Tübingen: Niemeyer, 1995
- Hötker, Wilfried; Petra Ludewig (Hr.): Lexikonimport, Lexikonexport. Studien zur Wiederverwertung lexikalischer Informationen. 242 S. (incl.. Anhang), Sprache und Information, Bd. 32, Tübingen: Niemeyer, 1996
  Sammelband mit Beiträgen von Heid, u.; Hoelter, M.; Hötker, W.; Kanngießer, S.; Ludewig, P.; Schnelle, H.; Teufel, S.; Wegmann, F.; Wilkens, R.
- Schiltz, Guillaume: Der Dialektometrische Atlas von Südwest-Baden (DASB). Konzepte eines dialektometrischen Informationssystems. Mit Teil 1 (Textband) und den Teilen 2-4 (Kartenbände). Studien zur Dialektologie in Südwestdeutschland, Bd. 5, Marburg: N.G. Elwert Verlag, 1996

11. EUROPÄISCHE TAGUNG DES ARBEITSKREISES "MASCHINELLE ÜBERSETZUNG" AM 11./12.12.1995 BEIDERSAP AG IN WALLDORF

Ursula Bernhard

Der Arbeitskreis "Maschinelle Übersetzung" ist eine auf europäischer Basis agierende informelle Gruppe, die Anwender, potentielle Anwender, Entwickler und Hersteller maschineller Übersetzungssysteme, Universitäten und Forschungseinrichtungen zusammenbringt, um mit dem Thema "Maschinelle Übersetzung" verbundene praxisorientierte Fragen *zu* diskutieren und Informationen auszutauschen. Das Sekretariat des Arbeitskreises wird von der GMD-Forschungszentrum Informationstechnik GmbH (Darmstadt und Sankt Augustin) wahrgenommen. Der *Ar*beitskreis trifft sich einmal pro Jahr. 1995 fand die Jahrestagung am 11.und 12. Dezember bei der SAP AG in Walldorf statt.

Die Zahl von 41 Teilnehmern und Teilnehmerinnen aus Wirtschaftsunternehmen, Behörden, Universitäten und Forschungseinrichtungen zeigte das große Interesse am Thema "Maschinelle Übersetzung".

Zu Beginn des ersten Tages stellte Stefan Blaschke, Leiter der Informationsentwicklung der SAP AG, seine Firma, ihre Entwicklung und Aktivitäten sowie die Organisation der Übersetzungsdienstleistungen vor. Die SAP AG hat 100 *fest* angestellte, 38 freiberufliche und 30 Übersetzer und Übersetzerinnen in Tochterfirmen. Deutsch, Englisch, Französisch und Japanisch sind die wichtigsten Sprachen. Bei Englisch zeichnet sich eine Entwicklung von der Zielsprache zur Quellsprache ab. Stefan Blaschke bedauerte, daß für Englisch als Quellsprache die maschinelle Übersetzung gegenwärtig noch keine im SAP-Kontext einsatzfähige Umgebung bieten könne. Am Hauptsitz in Walldorf verlagert sich der Schwerpunkt der Aktivitäten zum Translation Consulting, es wird nicht mehr alles im Hause erledigt, allerdings sollen Englischbezogene Übersetzungsdienstleistungen weiterhin in Walldorf angesiedelt sein.

Daniel Grasmick präsentierte die MTGruppe der SAP AG. Sie umfaßt mit Aushil*fen* und Externen 13 Mitarbeiter und Mitarbeiterinnen. 1982 unternahm SAP mit LOGOS den ersten Versuch, ein maschinelles Übersetzungssystem praktisch anzuwenden, allerdings ohne Erfolg. 1990 erfolgte dann ein zweiter Versuch mit MET AL für Deutsch-Englisch und Englisch-Deutsch, wobei das Sprachpaar Deutsch-Englisch in eine erfolgreiche Anwendung integriert werden konnte, wogegen Englisch-Deutsch den SAP-Anforderungen nicht genügte.1995 wurde erneut eine LOGOS-Anwendung für das Sprachpaar Englisch-Französisch begonnen.

Die METAL-Anwendung wird mit 7 Vollzeitkräften und 1 Praktikantin betrieben. Die MT -Gruppe ist nach dem Prinzip der Mischarbeitsplätze organisiert, die einzelnen Mitarbeiter und Mitarbeiterinnen spezialisieren sich auf SAP-Fachgebiete wie Finanzwesen, Personalwirtschaft usw. Die Gruppe versteht sich als Serviceteam für Fachübersetzer, Landesgesellschaften, Berater und Entwickler. Ziel ist es, die bestmögliche Übersetzung in kürzester Zeit zu liefern, was u. a. durch eine maxi

male Terminologie-Abdeckung und ein Ausschöpfen der linguistischen Verbesserungsmöglichkeiten erreicht wird. Es gibt unterschiedliche "Ausgabequalitäten", nämlich Basisversion für Fachübersetzer und Dokumentationsentwickler, Vorabversion für Landesgesellschaften und Endversion für verschiedene Dokumentationstypen, wobei der Grad der Nachbearbeitung der maschinellen Rohübersetzung variiert.

Die Arbeit im Jahre 1995 war durch sehr umfangreiche Dokumente mit äußerst kurzen Durchlaufzeiten gekennzeichnet. Im Durchschnitt wurden 500.000 Quellwörter pro Monat bearbeitet. Für 1996 zeichnen sich kleinere Dokumente, andere Texttypen und Fachgebiete sowie noch engere Termine ab. Außerdem wird das Sprachpaar Englisch-Deutsch an Bedeutung gewinnen. Die Integration in SAP-Mail, WWW, die Schaffung einer einheitlichen Oberfläche (Otelo) sowie die Bereitstellung zusätzlicher Dienstleistungen sind vorrangige Ziele.

Eric Brunelle berichtete über die LOGOS-Anwendung, die von einem Mitarbeiter und zwei Praktikanten (plus 1 Springer) betrieben wird. Es werden 6-8 Handbücher zu ungefähr je 100 Seiten pro Monat übersetzt. In zwei Monaten wurden so 1,7 Millionen Wörter übersetzt und in neun Monaten 23.000 Begriffe und 400 semantische Regeln kodiert. Die Überarbeitung der maschinellen Rohübersetzungen erfolgt extern.

Anschließend präsentierten Jennifer Brundage, Mary Wells, John Wells und Dirk Lüke kurz die Anwendungen, die am Nachmittag Gegenstand von Demos waren: Terminologie mit MLIF (Makro für Terminologie-Schnellerfassung), Terminologievergleich SAP-term/MET AL, MET AL-interne Dateien in Win Word, Hinweisübersetzung mit MET AL.

Peter Quartier von Lotus und Iain Urquhart von der EU stellten das EU-Projekt Otelo vor. Otelo ist ein Kürzel für Open Translation Environment for Localisation. Es wird von einem Konsortium bestehend aus SAP, Lotus, MET AL, LOGOS und CST (Zentrum fur Sprachtechnologie in Dänemark) durchgeführt. Es gibt Associate Partners wie z.B. Boehringer Ingelheim und Sharp, letztere für den asiatischen Raum. Ziel ist ein Otelo-Client, d.h. eine PC-Anwendung, die einen organisierten Zugriff auf MT -Ressourcen gewährleistet. Übersetzungsdienste sollen in einer Netzwerk-Umgebung abrufbar sein. Das Projekt hat ein Forum, die Otelo User Group. Interessenten sind zur Teilnahme eingeladen. Das Konzept ist offen, alle, die etwas anzubieten haben, können anbieten, wobei nicht nur an MT -Dienste gedacht ist.

Am Nachmittag fanden die Demos der am Vormittag präsentierten Anwendungen statt, wobei sich hier insbesondere der Aspekt der Anwenderselbsthilfe für die Fälle manifestierte, in denen der vom Hersteller gelieferte Standard zu einer effizienten Problemlösung nicht ausreicht. Im Kontext des Makros für Terminologie-Schnellerfassung wurde berichtet, daß 95% der Wörterbuch-Neueingaben Substantive sind. Der Terminologievergleich SAPterm und METAL gewährleistet eine Konsistenz zwischen der SAP-eigenen Terminologie-Datenbank und dem METAL-Wörterbuch. Die METAL-internen Dateien in Win Word sollen einen "Zoo" von Bildschirmen vermeiden, und die Hinweisübersetzungen mit MET AL gewährleisten einen 24-Stunden-Service für die Übersetzung von Fehlermeldungen usw.

Am zweiten Tag berichtete Dr. Klaus Heller vom Institut für deutsche Sprache in Mannheim über die deutsche Rechtschreibreform. Er stellte die Geschichte der Reform dar, für die im Sprachgebrauch der Linguisten die Bezeichnung Weiterentwicklung bevorzugt wird. 212 bisherige Regeln werden auf 112 reduziert, insgesamt werden nur 33 Wörter geändert. Seit 1974 gab es dazu 1. 073 Publikationen, u. a .:

. Klaus Heller: Reform der deutschen Rechtschreibung. Die Neuregelung auf einen Blick. Bertelsmann Lexikon Verlag. br. DM 5,-

. Deutsche Rechtschreibung. Vorschläge zu ihrer Neuregelung. Hrsg.: Internationaler Arbeitskreis für Orthographie. Narr, kt. DM 38,-

Die Reform wurde also keineswegs hinter verschlossenen Türen behandelt, wie in der leider auf unrichtigen Informationen basierenden Pressekampagne beanstandet worden war (die immer wieder zitierte Schreibweise "Filosofie" war nie geplant). Der Start der Reform an deutschen Schulen wird am 1. 8.1998 erfolgen. Es gibt eine Übergangszeit bis 2005. In diesem Zeitraum wird nur noch die neue Orthographie an den Schulen vermittelt, alte Schreibweisen werden zwar angestrichen, aber nicht als Fehler bewertet. Nicht mehr der Duden als Privatverlag wird die Orthographie in Zukunft festschreiben, sondern eine zwischenstaatliche Kommission wird die Entwicklung beobachten und eventuelle Änderungen beschließen.

Anschließend bewertete Prof. H. Zimmermann von Softex, Saarbrücken, die Auswirkungen der Reform auf die maschinelle Übersetzung. Er sieht keine großen Probleme, es gibt ganz wenige Fälle, die elektronisch nicht beherrschbar sind (z.B. im Besonderen = im besonderen Falle). Lediglich bei der Datenbanksuche könnten in Zukunft durch heterogene Texte Schwierigkeiten auftreten.

Da Ged Pearson, Intergraph, London, durch einen Unfall verhindert war, stellte Jochen Hummel von TRADOS, Stuttgart, das Konzept der Integration von Terminologiedatenbank, Translation Memory und maschineller Übersetzung in der TRADOS Translator's Workbench vor. Als maschinelles Übersetzungssystem kann sowohl Transcend als auch LOGOS integriert werden.

Reinhard Schäler, University College Dublin, berichtete über Software Localisation in Irland, er stellte die Aufgaben und Ziele der Software Localisation Interest Group (SLIG) und des Localisation Resources Centre vor. Irland ist das Weltzentrum für Software-Lokalisierung, 40-50% der in Europa verkauften PC-basierten Software kommt aus Irland. Die SLIG wurde1994 gegründet, sie vertritt und koordiniert die Interessen der Industrie, dient als Informationsforum und fördert die Zusammenarbeit zwischen Industrie und Forschung. Gegenwärtig gehören ihr Apple, Lotus, Microsoft, Oracle, Corel, Berlitz u. a. an. Das Localisation Resources Centre wurde 1995 am University College Dublin gegründet. Es wird von 15 Lokalisierungsunternehmen unterstützt. Es kümmert sich um Forschung und Entwicklung, Evaluation, Beratung und Ausbildung. Es ist u. a . geplant, eine Werkzeug-Bibliothek bereitzustellen.

John Hatley von LOGOS Computer Integrated Translation GmbH, Eschborn, berichtete über den EUROLANG Optimizer und LOGOS, das innerhalb des Optimizers als Übersetzungskomponente vom PC aus angestoßen werden kann.

Bernhard Masion von SN!, München, stellte das Projekt LINGO vor, das von Telekom und SN! durchgeführt wird. Ziel des Projekts ist die Bereitstellung eines elektronischen Marktplatzes, auf dem Angebot und Nachfrage im Bereich Telelearning/Commerce zusammengebracht werden. Ziel ist ein integriertes, multimediales, netzbasiertes Sprachendienstleistungssystem. Im Produktportfolio sind unter Übersetzen die Komponenten Übersetzen durch Menschen (Human), Übersetzen durch Maschinen (maschinell), Übersetzen durch Conferencing (Joint Translating), Dolmetschen und Terminologie-Wörterbücher vorgesehen. Telekom und SN! werden Anfang 1996 eine gemeinsame Tochterfirma gründen, die erstmals zur CeBit auftreten soll. Die LINGO-Zielgruppe sind Unternehmen, große Institutionen, noch nicht der private Bereich.

Innerhalb des allgemeinen Informationsaustauschs berichtete Nicole Klingenberg, Univ. Saarbrücken, über ein Projekt zur Modellierung des Übersetzungsprozesses an der Universität des Saarland es, Fachrichtung 8.6: Angewandte Sprachwissenschaft, Übersetzen und Dolmetschen.

Ursula Bernhard, GMD-Sankt Augustin, referierte über die Tagungen der Benutzergruppen von LOGOS und METAL. Die LOGOS USER GROUP für Europa konstituierte sich am 5. Oktober 1995 in Frankfurt.

Das Sekretariat übernahm Jose Garcia Martinez von der Union Fenosa in Madrid. Bei der METAL-Benutzertagung am 31.10.1995 am Flughafen Frankfurt legte die Sietec die Gründe für ihren Ausstieg aus der

Weiterentwicklung und Betreuung des 11E-TAL-Übersetzungssystems dar, und die GMS (Gesellschaft für Multilinguale Systeme) stellte ihr neues Konzept für das METAL-Nachfolgeprodukt vor. Die Unix-Version von METAL wird nicht mehr weiterentwickelt. Als erstes Nachfolgeprodukt wird zur Cebit eine PC- Version für Laien herauskommen, die nicht mehr METAL heißen wird. Eine PC-gestützte Profiversion für den Einzelübersetzer sowie eine High-End-Version als Unternehmenslösung werden folgen. Das Sekretariat der Benutzergruppe ging an das Amt für Auslandsfragen (AfA), München, über.

Prof. Klaus-Dirk Schmitz, Fachhochschule Köln, kündigte die TKE96 an.

Dr. Hans Billing, GMD-Darmstadt, berichtete von einem Schreiben der Firma Boehringer Ingelheim an den Bundesminister für Bildung, Wissenschaft, Forschung und Technologie. In diesem Schreiben wird darauf hingewiesen, daß durch den Ausstieg von Sietec aus der Weiterentwicklung des METAL-Systems die Gefahr besteht, daß die Arbeiten im Bereich der maschinellen Übersetzung in Deutschland stagnieren und dadurch der exportorientierten Wirtschaft im Kontext der neuen Informationstechnologien Nachteile erwachsen könnten, weshalb eine staatliche Förderung in diesem Bereich erwogen werden sollte. In seiner Antwort verwies der Bundesminister auf das Verbmobil-Projekt.

Jean-Marie Leick von der EU stellte das neue EU-Programm "Multilingual Information Society" (MLIS) vor und teilte mit, daß die EU beabsichtigt, die Arbeiten an SYSTRAN einzustellen.

Die Diskussion über die Überlebensstrategien für MÜ in Entwicklung und Anwendung mußte leider praktisch ohne Hersteller stattfinden, da nur LOGOS Vertreter entsandt hat. Es wird vorgeschlagen, daß sich der Arbeitskreis aktiver als bisher in die Diskussion ein bringt, z.B. durch eine "Messe für Hersteller".

Die nächste Tagung des Arbeitskreises "Maschinelle Übersetzung" wird voraussichtlich im Januar 1997 bei der GMD in Darmstadt stattfinden.

Innovative Information-Retrievalsysteme für die Praxis

Workshop für Anwender und Anbieter 13. und 14. November 1996
am Institut für integrierte Publikations- und Informationssysteme
(IPSI) der GMD in Darmstadt Veranstaltet von: Fachgruppe
Information Retrieval der Gesellschaft für Informatik (GI), GMD -
Forschungszentrum Informationstechnik GmbH (GMD-IPSI),
Fachhochschule Darmstadt, Fachbereich luD

Ankündigung

Information Retrieval ist ein Gebiet im Aufbruch. Denn auf dem Weg in die viel beschworene Informationsgesellschaft treten die Schwachstellen einer Philosophie, die das Informationsproblem mit dem Speichern schon für gelöst hält, immer deutlicher und störender in Erscheinung.

Wir befinden uns mitten in einer Entwicklung, die immer mehr und vielfältigere Formen von Wissen und Daten elektronisch verfügbar hält, sie in den Bereichen von Wirtschaft, Industrie und Unterhaltung/Konsum immer näher an den Ort ihrer unmittelbaren Verwendung bringt und sich dabei an ein heterogenes Spektrum von Anwendern richtet. Neuartige Anwendungen für Informationssysteme werden realisiert, und immer ehrgeiziger werden Planungen und Visionen für die Zukunft.

Vergleicht man den Wandel bei den Anforderungen mit den Fortschritten, die gängige Retrievalsysteme auf der konzeptionellen Seite anzubieten haben, so ergibt sich keine allzu glänzende Bilanz. "Die Retrievalpraxis ignoriert schlichtweg die Ergebnisse von 30 Jahren Forschung auf dem Gebiet des Information Retrieval" lautet die Standardklage der Forschungsszene. Auf der anderen Seite fühlen sich Anwender allein gelassen, wenn es darum geht, sich ganz konkret für Innovation und ein Mehr an Retrievalqualität und -komfort zu entscheiden. "Wo sind denn die Systeme, die wir tatsächlich mit besseren Ergebnissen einsetzen könnten?" fragen Anwender zurück!

Ziel und Anspruch

Mit dem Workshop *Innovative Information - Retrievalsysteme für die Praxis* will die Fachgruppe *Information Retrieval* der Gesellschaft für Informatik eine Diskussion in Gang bringen und einen praktischen Beitrag dazu leisten, daß Anwender sich ein eigenes Urteil darüber bilden können, welche Möglichkeiten der Markt für kommerzielle Retrievalsysteme gegenwärtig bietet und wohin der Trend hinsichtlich neuer Funktionen und Konzeptionen fuhren wird. Gleichzeitig wird es auch Ziel sein, daß Anwender formulieren, wo "sie der Schuh drückt", was also aus Ihrer Sicht die vorrangig zu lösenden Probleme für die IR-Entwickler sein sollten.

Form und Organisation
Der Workshop gibt den Anbietern von IR-Systemen in Form von Vorträgen und Vorführungen Gelegenheit, ihre Systeme mit ihren speziellen Pluspunkten vorzustellen. Alle Systeme sind in einem Ausstellungsbereich präsent, wo sich die Teilnehmer gezielt und vertieft informieren können. An Poster-Ständen mit Demovorführungen werden auch Prototypen zu sehen sein, die über die Funktionalität marktgängiger IR-Systeme hinausgehen. Vorträge, insbesondere von IR-Anwendungen, werden die Anbietersicht ergänzen und insgesamt zu einer differenzierten Sicht beitragen. Plenumsdiskussionen sollen das Thema der Innovation im Information Retrieval aus Entwickler- und Anwendersicht aufarbeiten.

Wer sollte sich angesprochen fühlen?
Wenn Sie entweder

- an der Entwicklung oder Anpassung von IR-Systemen beteiligt sind,
- oder als Anwender Erfahrungen mit innovativen Retrievalsystemen haben oder Sie eine fundierte Meinung dazu haben, was "innovativ" in Bezug auf IR-Systeme bedeutet bzw. Bedeuten sollte,
- oder in der Situation stehen, ein für Ihre Anforderungen und Zwecke besonders geeignetes IR-System auswählen zu müssen, dann würde Ihre Teilnahme zum Erfolg

der Veranstaltung beitragen und sicher würden Sie auch von ihr profitieren.

Was sollten Sie tun?
Teilnahme
Sofern Sie Interesse an einer Teilnahme haben, so lassen Sie sich zweckmäßigerweise auf die Interessentenliste setzen:

per e-mail an:thiel@darmstadt.gmd.de
per gelbe Post an: Dr. Ulrich Thiel, GMD - Forschungszentrum Informationstechnik, Institut für integrierte Informations- und Publikationssysteme (IPSI) Dolivostraße 15, D-64293 Darmstadt

Sie erhalten dann rechtzeitig Programm und Anmeldungsunterlagen zugeschickt. Ansonsten können Sie sich über den Stand der Vorbereitungen unter folgender URL `http://www.cui.darmstadt.gmd.de/mind/IR-WS.html` auf dem laufenden halten.

System-Darstellung
Sind Sie als Anbieter der Meinung, daß Ihr System substanziell mehr bietet als die gängigen Standardfeatures? Weil es zum Beispiel speziell auf die deutsche Sprache zugeschnitten ist, weil es mit multimedialen Dokumenten im Netz intelligent umgeht, weil es in effektiver Weise Dokumente ranken kann, oder, oder Dann arbeiten Sie das aus Ihrer Sicht Innovative in einem Exposee (2 bis 3 Seiten) heraus und benennen Sie Personen, die Ihr System kompetent und anwendernah vorstellen können. Setzen Sie sich mit Ihrer Bewerbung in Verbindung mit Prof. Dr. Gerhard Knorz (FR Darmstadt) oder Dr. Ulrich Thiel (GMD).

Damit die Leistungsfähigkeit verschiedener Systeme und deren besondere Stärken für die Teilnehmer durchschaubarer und vergleichbarer wird, stellen die Veranstalter einen deutschsprachigen Textkorpus mit 100000 Dokumenten (Agenturmeldungen) und 50 Anfragen zur Verfugung, an dem die eingeladenen Systeme ihre Art des Retrieval demonstrieren sollen.

Anwendungsorientierte Prototypen
aus der Forschung bewerben sich analog zu Anbietern kommerzieller Systeme.

Anwenderberichte
Ihre Erfahrungen mit innovativen IR-Systemen in der Praxis erwarten wir von Personen, die differenziert über Informations- und Dokumentationspraxis berichten können. Einreichungen in der Form eines Exposees im Umfang von 2 bis 3 Seiten an Prof. Dr. Gerhard Knorz (FR Darmstadt) oder Dr. Ulrich Thiel (GMD).

Termine

16.9.96	Headline für Einreichungen (Systemdarstellung/Anwenderberichte)
4.10.96	Entscheidung des Programmkomitees, Übergabe von Testdokumenten und Anfragen an Systembetreiber

Kontakt

Dr. Ulrich Thiel, GMD - Forschungszentrum Informationstechnik, Institut für integrierte Informations- und Publikationssysteme (IPSI) Dolivostraße 15, D-64293 Darmstadt, Tel. +496151 869-855, Fax: -818, e-mail: thiel@darmstadt.gmd.de
Prof. Dr. Gerhard Knorz, FR Darmstadt, FB Information und Dokumentation, Raardtring 100, D-64295 Darmstadt, Tel. +49 6151 16-8499, Fax: -8980, e-mail: knorz@fh-darmstadt.de

Programmkomitee
Vorstand der GI-Fachgruppe Information Retrieval

6th European Workshop on Natural Language Generation

CALL For PAPERS
6th European Workshop on Natural Language
Generation March 24 - 26, 1997 Gerhard-Mercator
University, Duisburg, Germany

The workshop aims to bring together researchers interested in Natural Language Generation from such different perspectives as Linguistics, Artificial Intelligence, Psychology, Cognitive Science, and Engineering. The meeting continues the tradition of a series of workshops held biannually in Europe (Royaumont, 1987; Edinburgh, 1989; Judenstein, 1991; Pisa, 1993; and Leiden, 1995) but it is open to researchers from all over the world.

Program Committee:
Stephan Busemann, Saarbrücken
Alison Cawsey, Edinburgh
Robert Dale, Sydney
Wolf gang Roeppner, Duisburg
(chair) Richard Kittredge, Montreal
Stephan Mehl, Duisburg
Koenraad de Smedt, Bergen Michael
Zock, Paris

Papers, posters and demonstrations are invited on original and substantial work related to the automatic generation of natural language, including computational linguistics research, artificial intelligence methods, computer models of human language processing, empirical research, and the development and evaluation of applied systems. Contributions on all aspects of natural language generation are welcome, but the special theme of this workshop will be 'System Architectures for Text Generation'.

This topic comprises a variety of more specific questions, e.g. planning and/or schemata, pragmatic impact on content selection and form determination, serial or incremental processing, macro-planning and micro-planning.

To encourage a workshop atmosphere, while allowing a relatively large number of people to participate, selected papers will be given large time slots including ample discussion time; other papers will be grouped for shorter presentations and mutual interaction, and there will be sessions for posters and computer demonstrations.

Submissions:
Researchers wishing to present a PAPER are requested to submit three copies of an original unpublished article (10 pages). To allow for anonymous reviewing the name(s) and complete adress(es) of the author(s) have to be provided on a separate sheet. We would appreciate that you additionally send an electronic version of the paper (email or diskette).
Researchers wishing to present a POSTER are invited to submit three copies of a reduced version of their poster on 4 normal pages that together form an A2 size sheet. Use a normal character size. As with papers any personal information about the author(s) should appear on a separate sheet.

Researchers wishing to demonstrate a computer PROGRAM are invited to send three copies of a short description of their program together with so me examples of input and output and hardware requirements. Please include the name(s) and complete address(es) of the author(s) in the description.

All contributions must be sent BEFORE NOVEMBER 1, 1996 to the Program Chairman at the following address:

Prof. Dr. Wolf gang Hoeppner Gerhard-
Mercator University, Duisburg FB3;
Computational Linguistics
D-47048 Duisburg, Germany
Tel.: +49203 379-2006/2008
email: hoeppner@unidui.uni-duisburg.de

Authors will be notified about acceptance or rejection by January 17,

1997.

Local arrangements:
Local arrangements are handled by: Wolfgang Hoeppner and Stephan Mehl (University of Duisburg).

The meeting will be held from the morning of Monday, March 24, 1997 through afternoon on Wednesday 26, in 'Die Wolfsburg' situated in the municipal forest of Duisburg. This conference site hosts congresses and workshops from all scientific areas and is equipped with excellent presentation facilities and modern guest rooms.

The cost of the workshop to each participant is currently estimated at about DM 500 including accommodation and meals, but the participants' fee may turn out to be lower depending on funding. The workshop will also be open to a limited number of participants not contributing a paper, poster or demo. A call for participation including more information and a registration form will be sent out later as soon as the program has been put together.

Please direct any inquiries to the address above.

Workshop on Practical Applications of Information Filtering

To be held in conjunction with

First International Conference on Practical Aspects of Knowledge Management

(PAKM)
Basel, Switzerland,
October 30-31,
1996

Information filtering is an aspect of knowledge management which has been the focus of concerted research in recent times. This has arisen because of the increasing volumes of electronically stored information being made available.

Unlike information retrieval or data mining, both of which address problems associated with static document databases, filtering applies to transiently occurring information on a computer network. The basic aim of information filtering is to route through to a user those source documents deemed relevant to his/her needs, possibly ranking them by estimated relevance; documents deemed not to be relevant are filtered out. Estimation of relevance is carried out by comparing a user profile - embodying knowledge of a user's ongoing interests - with incoming documents in an information stream. It may be expected that, as a user's interests change or evolve, the corresponding user profile is adjusted accordingly.

Example user scenarios in which information filtering would be an appropriate tool might include the following:

A journalist in a newsroom may be following developments of a certain story - or particular aspects of a story - over a prolonged time period. As such, he/she may wish to have relevant newswire articles filtered through. Also, as certain aspects of the story begin to assume importance to the journalist, the nature of articles received might be expected to change.

A financial institution trades stock internationally. Beside the normal world-wide monitoring of stock exchanges, it also needs to have knowledge of world events which might affect stock prices: weather conditions. in a particular location; earthquakes; military coups; government changes or collapses; interest rate changes. Toward this end, the institution may employ people to monitor the API and FT newswires, pertinent USENET Newsgroups, etc. It is clearly desirable that such individuals be able to set up profiles for each stock item(s) being tracked, that this act as a filter against incoming information, and that the filter may undergo adaptation as new factors arise or existing factors increase or decrease in significance.

The manager of the Information Systems division of a company needs to keep up to date regarding the relative benefits of a competing range of software and hardware products. To do this, he/she wishes to monitor computer mailing lists and USENET News to locate articles containing meaningful comparisons. As with the other cases, the important issue is that only relevant articles are presented, with irrelevant contributions screened out, i.e., filtering takes place on the incoming information. Also, as the manager's needs change (e.g. a hardware purchase is made or a software product is eliminated from further consideration), the nature of information being routed through should automatically adapt.

Papers

The purpose of this workshop is to examine currently available practical applications of information filtering, to assess the impact of the technology, to evaluate its successes and failures and to appraise its future utility as a practical application of knowledge management. Papers are invited on any aspect of information filtering, but emphasis will be placed on real-world systems and approaches. It is thus desirable that the paper be linked to some specific user scenario, such as one of those listed above. A non-exhaustive list of topics is included below:

- Applications of Filtering
- Profile / Document Representation
- Profile Adaptation
- User Interfaces
- Filter System Architectures
- Profile/Document Comparison
- Profile Optimisation
- User Modelling
- Multimedial Hypermedia Filtering
- Evaluation Techniques

Demonstrations

Software demonstrations related to the workshop topics are also encouraged. These may or may not be associated with a paper being presented. Conference organisers will provide a room where such demonstrations can be given during lunch breaks and at other times. Lunch, exhibitions and demonstrations will take place in the same or adjacent rooms.

Participation

Beside being open to people presenting papers and demonstrating systems, the work-shop will be open to practitioners interested in concretely applying information filtering strategies. Workshop participants presenting a paper will, however, quality for a reduced conference fee. Refer to the main conference's general information (URL below) for participation details.

Organisers

Alan Smeaton
School of Computer Applications
Dublin City University
Dublin 9
Ireland
Alan.
Smeaton@CompApp.DCU.IE

Humphrey S0rensen Computer
Science Department University
College
Cork
Ireland
sorensen@odyssey.ucc.ie

Mitteilungen aus der GLDV

WWW-Home-Page der GLDV

Die GLDV hat seit Mai eine WWW-Home-Page:

http://cll.ikp.uni-bonn.de/GLDV /

Es bestehen jeweils *links* zu Seiten, auf denen die Gesellschaft vorgestellt wird, Veranstaltungen, Arbeitskreise und Publikationen der Gesellschaft zu finden sind und das LDV-Forum sowie die Planung der künftigen Hefte des LDV-Forums vorgestellt werden. Unter "Aktuelles" besteht derzeit ein *link* zum Programm der GLDV-Herbstschule in Magdeburg.

MORPHOLYMPICS

Die Dokumentation zur ersten MORPHOLYMPICS ist inzwischen im Max Niemeyer Verlag unter dem Titel "Linguistische Verifikation" erschienen. Die vollständige Angabe lautet:

Hausser, Roland (Hrg.): Linguistische Verifikation. Tübingen: Niemeyer, 1996.

Die zweite MORPHOL YMPICS wird entgegen den ursprünglichen Planungen erst 1997 stattfinden. Die Entscheidung über die Haupttestsprache wird im Laufe des Jahres fallen.

Herbstschule 1996 in Magdeburg

Vom 23. bis 27. September 1996 wird in Magdeburg die diesjährige Herbstschule stattfinden. Wie den Ankündigungen (e-mail, Plakate, WWW-Seiten) bereits zu entnehmen war, lautet das Thema der Herbstschule "Herausforderungen an die Computerlinguistik - Multilingualität, Multimedialität, Multidisziplinarität". Detailliertere Informationen können den WWW-Seiten entnommen werden:

http://www-ai.cs.uni-magdeburg.de/herbstschule96.html

Für studentische Teilnehmer besteht die Möglichkeit der Unterbringung in der Jugendherberge zu einem Preis von 28 DM (inklusive Frühstück) pro Nacht.

Schnupperangebot

Im Zusammenhang mit der GLDV-Herbstschule besteht ein Schnupperangebot. Für Teilnehmer an der GLDV-Herbstschule, die der GLDV beitreten möchten, gilt die Zahlung der Teilnahmegebühr für die Herbstschule gleichzeitig als Mitgliedsbeitrag für 1996.

Mitgliederversammlung

Die nächste Mitgliederversammlung wird im Rahmen der KONVENS am 8. Oktober 1996 in Bielefeld stattfinden. Die Einladung mit der Tagesordnung wird satzungsgemäß vorab an alle Mitglieder verschickt werden. Als Tagesordnungspunkte werden unter anderem ein Kassenbericht, der Haushaltsplan 1997, die Neuwahl der Kassenprüfer, eine Vorschau auf die GLDV-Jahrestagung 1997 sowie die im Frühjahr 1997 stattfindenden Wahlen behandelt werden.

Jahrestagung 1997

Die Jahrestagung 1997 der GLD V wird vom 17. bis 19. März 1997 in Leipzig, unmittelbar vor der Leipziger Buchmesse, stattfinden. Die Veranstalter der Buchmesse haben gegenüber Herrn Heyer, dem örtlichen Organisator der Jahrestagung, bereits Interesse an einer gemeinsamen Veranstaltung mit der GLDV bekundet.

Preis für eine hervorragende studentische Arbeit
Zur Jahrestagung soll ein mit *1.000* DM dotierter Preis für eine hervorragende studentische Arbeit auf dem Gebiet der LDV/CL ausgeschrieben werden. Die eingereichten studentischen Arbeiten, für die Diplom-, Magister- und Studienarbeiten in Frage kommen, werden von einer Jury begutachtet.