In: Reinhart Herzog (Hrsg., 1981): Computer in der Übersetzungswissenschaft.
Frankfurt am Main, Bern: Lang, S. 74-84

Harald H. Zimmermann

LEXIKON UND PARSING

Bei der Beschäftigung mit automatischer syntaktischer Analyse einer Sprache oder maschineller Übersetzung steht das Problem eines geeigneten Parsers, also die Frage der Umsetzung eines Grammatikmodells in einen Algorithmus zur Bearbeitung sprachlicher Informationen, im Vordergrund. Die Lexikonkomponente wurde bisher vorwiegend unter dem gleichen Gesichtspunkt betrachtet: Hauptprobleme waren dabei geeignete Darstellungs-, Speicherungs- und Suchmethoden, die Fragestellung war also formal-ökonomischer Natur. So notwendig auch eine formale Beschreibung sprachlicher Gegebenheiten gerade in der Linguistischen Datenverarbeitung ist, viel wichtiger ist der Adäquatheitsgrad der Beschreibung: Was dabei die allgemeinen Anforderungen angeht, die jede Grammatik erfüllen soll, so werden im folgenden einige theoretische Ansprüche der generativ-transformationellen Grammatik (TG; bezogen auf die Lexikonkomponente) diskutiert werden, in deren Rahmen Adäquatheitsfragen einen breiten Raum einnehmen. Diese auf die sprachliche Kompetenz eines idealen Sprechers/Hörers bezogenen Forderungen sind mit den Problemen eines realen Sprachmodells, hier besonders des Teilbereichs "maschinelles Lexikon", konfrontiert.

Die erste Frage, der wir uns zuwenden wollen, ist die nach dem Aufbau des Lexikons. Dazu gibt es eine These, die Chomsky so formuliert: "Die lexikalischen Eintragungen umfassen ... die Gesamtmenge von Irregularitäten einer Sprache." [1] Hier wird also eine Zweiteilung des Grammatiksystems in Regelsystem (Regularitäten enthaltend) und Lexikon (Irregularitäten enthaltend) vorgenommen. Ein wichtiger Begriff ist in diesem Zusammenhang die 'Redundanz', das heißt die Forderung, durch Verwendung vereinfachender Regeln die phonologische (daneben auch die syntaktisch-semantische) Komponente des Lexikons quantitativ einzuschränken. In Beispielen ausgedrückt: Statt der Wortformen VATER, VATERS, VAETER, VAETERN soll nur ein Eintrag VATER mit entsprechenden, in Regeln verwendbaren Merkmalen zur Flexion vorgenommen werden. Ähnliches gilt für Komposita wie SCHNEIDERMEISTER oder TISCHLERMEISTER oder für Wortableitungen wie TAENZER (zu TANZ oder TANZEN): Hier ist mittels entsprechender Regeln eine lexikalische Klassifizierung anhand der beteiligten freien oder gebundenen Morpheme anzustreben.

Der Maßstab, mit dem die Lexikoneintragung dabei gemessen wird, ist ökonomischer Natur. Daraus lässt sich folgern, dass sich die Verwendung von Redundanzregeln ebenfalls diesem Grundsatz der Vereinfachung des Systems stellen muss. Obgleich man nämlich prinzipiell zugestehen kann, dass redundante Eigenschaften durch Regeln erfasst werden sollten, lassen sich Fälle denken, in denen eine Redundanzregel ökonomisch nicht effektiv ist, da sie ihrerseits den Analyse- oder Syntheseprozess verkompliziert.

Wir nehmen als Beispiel die so genannten starken Verben im Deutschen und betrachten dabei die morphologische Komponente. Je nach Art der Abwandlung eines Stammvokals, eventuell auch der Art des Konsonantenwechsels, lassen sich diese Verben formal gruppieren: Einigen dieser Gruppen kann man eine Anzahl Verben zuordnen: Wie BINDEN werden etwa noch DINGEN, DRINGEN, GELINGEN, KLINGEN, SCHLINGEN, SCHWINGEN, SCHWINDEN, SINGEN

... flektiert; für manche 'Gruppen' gibt es nur ein einziges zugehöriges Simplex-Verb, z.B. FALLEN, FANGEN, ESSEN.[2] An einem Beispiel konkretisiert heißt dies, dass etwa bei ZIEHEN als morphologischer Lexikoneintrag nur noch Z- verbleiben dürfte, als einziges allen Wortformen (z.B. ZIEH, ZOG) gemeinsames Graphem, während IEH oder OG in einer Liste des Regelteils aufzuführen wären.[3] Die konsequente Verkürzung des Lexikons zieht also eine teilweise idiosynkratische Erweiterung des Regelteils nach sich. Ganz allgemein scheint zu gelten, dass die Redundanzregel zwar den Ablauf des Verstehensprozesses, aber nicht notwendig das Ergebnis beeinflusst, sondern nur eine Verlagerung des 'Wissens' von der Regel ins Lexikon oder umgekehrt nach sich zieht. Begreift man Lexikon und Regelsystem mehr als eine Einheit denn als zwei scharf zu trennende Komponenten des Sprachsystems, so fällt es nicht schwer, dem Redundanzproblem eher eine ökonomische als eine primär-linguistische Bedeutung zuzuschreiben.

In engem Zusammenhang mit der Frage der Vereinfachung des Lexikons steht das Problem der Wortbildung. Für die TG ist die Lösung dieses Problems von entscheidender Bedeutung, da im Grammatiksystem Regeln zur Komposition und Ableitung vorgesehen sein müssen, die zugleich den Anspruch zu erfüllen haben, dass nur 'grammatische' Wörter gebildet werden. Um sich dem Dilemma zu entziehen, dass die Anwendung von Wortbildungsregeln teilweise sehr starken Restriktionen unterliegt, die kaum oder nur sehr schwer zu formalisieren sind, werden in der TG im allgemeinen drei Möglichkeiten unterschieden:

a) vorkommend (im Lexikon zu vermerken)
b) möglich, aber nicht realisiert (im Regelteil zugelassen und entsprechend gekennzeichnet) und
c) nicht möglich (keine entsprechenden Regeln).

Vor allem bei der Komposition aber (man denke gerade an die im Deutschen möglichen Augenblickskomposita wie BAHR-PAPIER, KRIMBESUCH, PARTEITAGSDEBATTE) kann eine Prädiktabilität nach dem heutigen Stand der Linguistik nicht immer eindeutig festgelegt werden. Gewissen Wortbildungen beispielsweise, etwa Komposita wie ESELSBRUECKE oder BUERGERMEISTER, also vor allem solchen mit metaphorischem Charakter, lässt sich heute keine oder nur eine höchst komplizierte, kaum mehr formalisierbare Tiefenstruktur zuordnen, die auf den Kompositionselementen aufbaut. Vor allem lässt sich die semantische Komponente des Kompositums nicht oder nicht mehr aus den semantischen Merkmalen der Kompositionsteile erschließen. Derartige Wörter sind also als idiosynkratische Einheiten zu betrachten und als solche ins Lexikon aufzunehmen. Es ist jedoch - dies darf als allgemeine Forderung der TG angesehen werden - stets zu prüfen, ob sich formalisierbare Strukturen phonologischer, syntaktischer und semantischer Art erkennen lassen, die eine Aufnahme der komplexeren Fügung in das Lexikon überflüssig erscheinen lassen.

Bei der Umsetzung dieser Forderungen in ein reales Lexikon treten notwendig Begrenzungen auf. Dabei lassen sich zwei Gesichtspunkte unterscheiden: Einmal ist das Lexikon ein Teil der sprachlichen Kompetenz; es beschreibt also die Verwendungsmöglichkeiten, die ein Sprecher/Hörer hat. Zugleich gibt es einen bestimmten Zustand des Sprachsystems wieder, der vielleicht am Tag X um n Uhr gegolten hat. So betrachtet, ist das Lexikon unbeweglich. Doch die Welt und die menschlichen Vorstellungen, die das Lexikon in gewisser Weise widerspiegelt, sind veränderlich. Von daher ist an das Lexikon also die Forderung zu stellen, dass es sich mit den Vorstellungen ändert, dass es modifizierbar, dynamisch ist.

Unter diesen beiden Gesichtspunkten, der notwendigen Begrenztheit der Lexikoneintragungen und der daraus resultierenden Forderung nach einer Dynamisierung des Lexikon- und Regelsystems, sind die folgenden Ausführungen zu sehen:

Eine ideale deskriptive Adäquatheit ist für das maschinelle Lexikon nicht zu erreichen, wohl aber ein gewisser Grad an Adäquatheit, der sich mit der Kompetenz eines realen Sprechers/Hörers vergleichen ließe. Das computerorientierte Lexikon wird in der Regel dadurch aufgebaut, dass die augenblickssprachliche Kompetenz eines oder mehrerer Menschen, möglicherweise auch die in andere Wörterbücher eingegangene Kompetenz, im Bereich der zu erstellenden Informationen (also nicht unbedingt das ganze enzyklopädische Wissen des Informanten betreffend) in computerzugängliche Form gebracht wird. Das Lexikon kann erweitert werden, wenn sich die Kompetenz eines Bearbeiters entsprechend erweitert hat. Bestenfalls stellt das maschinelle Lexikon also die Vereinigung der bekannten Lexikoneintragungen (und eventuell des Wissens) der menschlichen Bearbeiter dar.

In diesem Zusammenhang stellt sich die Frage nach dem Umfang des Lexikons. Man denke an die numerisch kaum abschätzbare Vielfalt der Eigennamen (Familiennamen, Ortsnamen,...), an die fast ebenso zahllosen Fachtermini und an die erhebliche Zahl seltener, veralteter, das heißt ungebräuchlicher Wörter. Welcher kompetente Sprecher des Deutschen könnte etwa von sich behaupten, alle der folgenden im Wörterbuch von Wahrig[4] verzeichneten Lemmata zu kennen: ABHEBERN (mit Heber entfernen), ABKETTELN, ABKNAUPELN, ABMARKTEN (abfeilschen), ABRIFFELN als Beispiele für Verbalkomposita; FITTING, FITZ, FITZBUENDE, FITZE, FIXISMUS, FLABBE als Beispiele für Substantive - wenigen Spalten des Lexikons entnommen - mögen genügen, um die Unmöglichkeit deutlich zu machen, ein umfassendes Gesamtlexikon für eine Sprache zu erstellen. Selbst wenn dies gelänge - etwa auf der Basis eines Morphemwörterbuchs - so blieben doch eine auch im Deutschen nicht zu unterschätzende Zahl von Fremdwörtern und solche Neologismen übrig, die sich aus bisher nicht verwendeten, sprachlich möglichen Graphemkombinationen zusammensetzen.

Darüber hinaus ist der Umfang des maschinellen Lexikons mitbestimmt von Überlegungen ökonomischer Natur. Etwa kann man davon ausgehen, dass der Herstellungsaufwand direkt proportional ist zu der Anzahl der aufzunehmenden Lexikoneinträge (sieht man einmal von den häufigsten - zumeist wegen ihrer Unregelmäßigkeiten aufwendiger zu kodierenden Wörtern einer Sprache ab). Diese an sich triviale Feststellung hat durchaus ihre praktischen Konsequenzen: Wenn ein an häufigeren Wörtern orientiertes Lexikon mit 10 000 Eintragungen 98 % aller fortlaufenden Wörter eines beliebigen Textes erfassen könnte und eine Steigerung auf 99,5 % nur durch die Aufnahme von weiteren 50 000 Eintragungen möglich wäre, ist zu fragen, ob sich der Aufwand an Arbeitszeit dafür noch lohnt oder ob sich nicht durch eine geeignete Auswahl der Eintragungen (etwa nach Häufigkeitsgesichtspunkten) und durch zusätzliche Prozeduren eine brauchbarere, praktikablere Lösung finden ließe.

Ähnliche Argumente gelten für das Wörterbuchdurchsuchen, also die Zuordnung von Textwortformen zu einem Lexikoneintrag. Je mehr Eintragungen in einem maschinellen Lexikon stehen, desto mehr Vergleiche (wenn auch nicht direkt proportional) sind durchzuführen. Die Zuordnungsgeschwindigkeit wird unter Umständen bedeutend verlangsamt, ohne dass damit ein entscheidender Zugewinn an erfolgreich zugeordneten Textwörtern eintritt.

Es ist also zunächst nach Methoden zu suchen, deren Anwendung eine sinnvolle Optimierung des Lexikonumfangs erlaubt. Wenn von der Dynamik des Sprachschatzes gesprochen worden ist, so trifft dies doch nicht für alle Elemente in gleichem Umfang zu. Statisch oder doch über Jahrzehnte kaum wandlungsfähig sind die Elemente einiger Wortklassen, die gelegentlich - da sie vorwiegend syntaktische Funktionen ausüben und keine oder geringe eigenständige Bedeutung besitzen - als Funktionswörter oder Leerwörter bezeichnet werden. Zu ihnen werden die Konjunktionen, die Post- und Präpositionen, die 'reinen' (Orts- und Zeit-) Adverbien, die Partikel und Interjektionen sowie die verschiedenartigsten Pronomina einschließlich der Artikel gerechnet. Die absolute Anzahl der Elemente dieser Wortklassen ist verhältnismäßig gering; es gibt im Deutschen vielleicht 250 - 300 Präpositionen und einige Dutzend Konjunktionen, ähnliches gilt für die übrigen genannten Wortklassen. Sieht man einmal von den mehrwortigen Ausdrücken ab, die in ähnlicher Funktion verwendet werden können (z.B. DEN GANZEN TAG, AUS DIESEM GRUND) - auch ihre Zahl ist im Grunde überschaubar - so sind diese Elemente in Funktion und Bedeutung als wenig flexibel, auf eine kurze Sprachepoche bezogen sogar als statisch anzusehen. Ihre vollständige Aufnahme in das Lexikon scheint aus diesen Gründen bereits sinnvoll zu sein; hinzu kommt, dass sie in der Regel zu den am häufigsten gebrauchten Wörtern einer Sprache gehören.

Zu den verbleibenden Grundwortklassen der Sprache wären also - ohne hier über Grenzfälle oder die Wortklasseneinteilung rechten zu wollen - die Elemente folgender syntaktischer Kategorien zu rechnen: Finite und infinite Verben, Substantive, Eigennamen, Adjektive und (Adjektiv-) Adverbien. Ähnlich wie bei den Adverbien geschehen, scheint es noch sinnvoll, eine Subkategorie der Verben, nämlich die der Hilfs- und Modalverben, auszunehmen, da sie ebenfalls eher eine statische als dynamische Teilmenge darstellt. Von einigen noch zu behandelnden Fällen wäre damit der erste invariable Teil eines optimierten Lexikons bereits umgrenzt.

Einen zweiten festen Bestandteil dieses Basislexikons bilden die Ausnahmelisten; es handelt sich hier um Elemente der übrigen Wortklassen, die bei der Anwendung von Redundanzregeln zu mehrdeutigen (sprachlich irrelevanten) oder falschen Ergebnissen führten. Dafür einige Beispiele, zunächst eingeschränkt auf die syntaktische Merkmalerkennung: So kann eine einfache Regel vorgesehen sein, die auf der Ableitungssilbe -UNG basiert und zu einer Kategorisierung 'Substantiv femininum' führen würde (ohne eine weitergehende morphologische Kontrolle). Eine solche Regel führte jedoch bei Wörtern wie JUNGEN (zugleich Flexionsform des Adjektivs JUNG), auch beim Auftreten der Partizipien ERRUNGEN, GEZWUNGEN, BESUNGEN,... zu falschen, in Fällen wie LUNGEN, ZUNGEN zu zwar ad hoc richtigen, aber unbefriedigenden Ergebnissen. Ähnliches gilt für Regeln, bei denen andere Suffixe zu ihnen entsprechenden Kategorisierungen führen; so müssten bei dem Suffix -HEIT etwa die Ausnahmen GESCHEIT und auch SCHEIT verzeichnet werden, bei -IG (Ergebnis: Adjektiv) wären Wörter wie ESSIG, REISIG, DANZIG, KOENIG, ZEISIG, HONIG, PFENNIG, STEIG oder FEIG in der Ausnahmeliste zu führen. Das Auflisten weniger Ausnahmen erlaubt etwa im syntaktischen Bereich die Informationserschließung einer großen Zahl von Wörtern anhand einfacher morphologischer Regeln, ohne das Lexikon übermäßig zu belasten. Sofern zu diesen funktionalen Informationen weitere Angaben, eventuell zur semantischen Subkategorisierung, erschlossen und verarbeitet werden sollen, muss die Ausnahmeliste um die entsprechenden Sonderfälle erweitert werden; dies träfe etwa auf ESELSBRUECKE zu. Der Umfang der Ausnahmelisten ist also abhängig von Art und Umfang der Redundanzregeln.

Prinzipiell ist das Basislexikon mit der Erstellung der Ausnahmelisten abgeschlossen. Dennoch wird man aus ökonomischen Gründen einen weiteren Teil ansetzen müssen, der der Entlastung des Regelteils dient: Da bei jedem im Lexikon nicht unmittelbar, das heißt über die Graphemfolge, verzeichneten Wort die gesamte Kategorisierung durch das Regelsystem geleistet und das Ergebnis spätestens nach einem Analyseprozess wieder 'vergessen' wird, muss bei einem erneuten Auftreten des Wortes diese Arbeit vom Computer aufs neue durchgeführt werden. Treten derartige Wörter häufiger auf, so schlägt dies in einer Erhöhung des Rechenaufwands möglicherweise deutlich zu Buche. Eine Aufnahme solcher hochfrequenter Wörter (Morpheme, Stämme, Wortformen) - unabhängig von ihrer systematischen Einordnungsmöglichkeit - scheint daher sinnvoll zu sein.

Aus der gegebenen Begrenztheit der Lexikoneintragungen folgt notwendig die Frage nach der Behandlung von Textwortformen, die nicht mit Hilfe der vorhandenen Lexikoneinträge - seien es nun Wortformen, Stämme oder Kernmorpheme - erkannt und/oder mittels Redundanzregeln klassifiziert werden können. Durchaus sinnvoll - vor allem bei einer automatischen Sprachübersetzung - scheint der Aufbau eines interaktiven Kommunikationssystems Mensch-Maschine zu sein: In allen Fällen, in denen mit Hilfe des maschinellen Lexikons keine Informationszuordnung möglich ist, wendet sich der Computer an seinen menschlichen Partner und erhält von diesem die entsprechenden Angaben; der Mensch bildet dabei gleichsam ein peripheres Ersatzlexikon für den Computer. So praktikabel diese Methode auch erscheint, sie ist aus der Sicht des Linguisten noch unbefriedigend. Die zweite Möglichkeit, die sich anbietet, ist die der Integration eines unbekannten Wortes in das Klassifikationsverfahren einer maschinellen Analyse. Diese Vorgehensweise baut im Grunde auf der Vorstellung auf, dass der Kontext eines Wortes zu seiner Klassifizierung beitragen kann. Wenn wir beispielsweise den Satz lesen

HERR DAKAPOPULOS HAT SEINEN ONKEL BESUCHT.

so nehmen wir an, dass DAKAPOPULOS ein Name ist. Wir verwenden dabei etwa die Information, dass HERR häufig als Anredeform vor Namen steht oder dass das Verb BESUCHEN ein belebtes Subjekt erfordert. In dem Satz

ER BEWEGTE SICH DEGAGIERT.

kann man - ohne die semantische Struktur des Satzes genau zu erkennen - das letzte Wort, falls es nicht bekannt ist, zumindest als adverbial gebrauchtes Partizip bestimmen. Den beiden Beispielen ist gemein, dass aufgrund der Wortstellung und des bekannten syntaktischen oder semantischen Kontexts - eventuell auch aufgrund besonderer morphologischer Merkmale des Wortes selbst - eine gewisse Klassifikationsstufe eines unbekannten Wortes erreicht wird. Mit Hilfe der maschinellen Kontextanalyse könnte es daher auch möglich sein, die Kluft zwischen einem endlichen maschinellen Lexikon und den - wenn man so will - unendlichen Wortbildungsmöglichkeiten zu überbrücken.

Bei der Klassifizierung der unbekannten Wörter kann man sich ohne größere Problematik der Strategien und Regeln bedienen, wie sie zur Auflösung natürlicher Mehrdeutigkeiten entwickelt oder noch zu entwickeln sind. Das Problem der natürlichen Mehrdeutigkeit ist schon mehrfach behandelt [5]; hier nur einige Beispiele für natürlich mehrdeutige Wortformen: BILLIGEN (syntaktisch mehrdeutig: DIE BILLIGEN AEPFEL, WIR BILLIGEN ES), SCHLOSS (semantisch/syntaktisch mehrdeutig: DAS SCHLOSS DES KOENIGS, DAS SCHLOSS AN DER

TUER, SCHLOSS ER DAS TOR AUF?). Auch hier findet sich die bei der maschinellen Analyse auswertbare Information im Textzusammenhang.

Die unbekannten Wörter können a priori allen Merkmalklassen, deren Elemente nicht vollständig im Lexikon verzeichnet sind oder die sich nicht ausschließlich durch entsprechende Regeln erfassen lassen, angehören. Der Regelfall wird jedoch sein, dass sie in der sprachlichen Realisierung nur einige spezifische Merkmale aufweisen (Damit soll nicht ausgeschlossen werden, dass sie auch 'natürlich' mehrdeutig sein können). Dennoch sind sie zunächst mit allen noch möglichen Merkmalen auszustatten; dann ist durch die Kontextanalyse - nach ähnlichen Verfahren wie bei der Auflösung natürlicher Mehrdeutigkeiten - eine Reduktion dieser Merkmale auf die im Kontext noch möglichen Angaben durchzuführen.

Veranschaulicht sei dies am Beispiel der Wortklassenmerkmale: Für diesen Bereich wurden bereits entsprechende praktische Tests durchgeführt.

Klammert man einmal die Eigennamen aus (deren Erkennung besondere Strategien erfordert, da sehr viele Wörter einer Sprache zugleich Eigennamen sein können), bleiben noch die vier Hauptklassen Substantiv, Adjektiv, Adverb und Verb, wobei sich das Verb syntax-funktional noch einmal aufgliedern lässt nach finitem Verb, Infinitiv und Partizip II unflektiert. Wollte man alle unbekannten Wörter mit diesen Mehrdeutigkeiten ausstatten, so wären sie in dieser Hinsicht a priori sechsdeutig.

Der eigentlichen Analyse lässt sich aber noch ein quasimorphologischer Regelteil vorschalten, in dem aufgrund graphematischer Merkmale (hier die Endgrapheme) erste Wortklassenreduktionen durchgeführt werden. Setzt man etwa voraus, dass alle (auch präfigierte) starken Verben im Deutschen über das Lexikon erfasst werden, kann für die Kategorie 'Partizip II unflektiert' gelten, dass das letzte Graphem ein T ist. Trifft dies zu, wird zugleich die Möglichkeit des flektierten Adjektivs ausgeschlossen; endet ein Wort nicht auf N, so kann es kein Infinitiv mehr sein, endet es auf einen Vokal (außer E), werden Adjektiv, Verb, Partizip ausgeschlossen usw. Auf diese Weise lässt sich die künstliche Mehrdeutigkeit in vielen Fällen weitgehend reduzieren, so dass die endgültige Vereindeutigung anhand des Kontexts durch den Parser erleichtert wird.

Nicht immer lässt jedoch der engere Kontext (im allgemeinen ist dies der Satzzusammenhang) sichere Schlüsse auf die reale Funktion eines unbekannten Wortes zu, zumal wenn mehrere natürlich oder künstlich mehrdeutige Wörter zu klassifizieren sind. Die in der maschinellen Analyse noch auswertbare Information findet sich im übersatzmäßigen Textzusammenhang. Man kann von der Annahme ausgehen, dass ein unbekanntes Wort - vorausgesetzt, es kommt irgendwo in einem Text wiederholt vor - einen anderen Kontext im engeren Sinn aufweist, der eine Lösung entweder bestätigt, gegebenenfalls aber auch falsifiziert. Bestätigung bedeutet zugleich Verstärkung des vorherigen Ergebnisses, eine abweichende Lösung schwächt es eventuell in seiner Aussagekraft ab. Ein für den gesamten Analyseprozess erstelltes kurzzeitiges Zwischenlexikon könnte zur Rückkopplung und Verknüpfung der einzelnen Reduktionsergebnisse dienen.

Es ist denkbar, dieses Kurzzeitlexikon auch für andere Zwecke zu verwenden: etwa zur Kopplung neugewonnener Informationen mit dem Basislexikon; das heißt einige Wörter werden aufgrund ihrer Häufigkeit nicht nur automatisch klassifiziert, sondern anschließend in das Basislexikon überführt. Schließlich kann das Zwischenlexikon im Zusammenhang mit der Auflösung natürlicher syntaktischer oder semantischer Mehrdeutigkeiten (als übersatzmäßiges Korrektiv)

ebenso verwendet werden wie bei der Klassifikation von Eigennamen, sofern mehr als ein entsprechender Textbeleg auftritt.

Testergebnisse anhand eines Textmaterials von 282 Sätzen mit insgesamt 3178 laufenden Wortformen scheinen die Brauchbarkeit eines derartigen Verfahrens, zumindest für den Bereich der syntaktischen Klassifizierung unbekannter Wortformen nach Wortklassen, zu bestätigen. Ohne hier auf Einzelheiten eingehen zu wollen, seien zum Abschluss einige Ergebnisse angeführt:

Die Quote richtiger Lösungen liegt je nach Mehrdeutigkeitstyp zwischen 73 und 91%; bei den vergleichbaren Werten der Auflösung natürlicher Mehrdeutigkeiten, die 1969 in Saarbrücken ermittelt wurden, lag sie zwischen 74 und 93%; die Analysequote bei der Auflösung natürlicher Mehrdeutigkeiten war absolut gesehen nur 1,3% besser als die Quote bei der Reduktion künstlicher Homographen. Von 26 Wortformen, die für den Test vorgegeben waren (die Wahl der übrigen Wörter war den Testpersonen freigestellt) konnten 7 anhand des Kontexts eindeutig klassifiziert werden, obwohl dazu mindestens 8 richtige Auflösungen erforderlich waren und keine andere Auflösung auftreten durfte; 9 weitere konnten in der Mehrdeutigkeit zumindestens (richtig) reduziert werden. Eine Verbesserung des Modells könnte für diese Bereiche die Fehlerquote noch weiter senken.

Voraussetzung für eine (im Modell nicht in letzter Konsequenz angestrebte) Verbesserung der Analyseresultate sind eine den theoretischen Anforderungen an das Basislexikon stärker entsprechende Lexikonstruktur, ein feineres morphologisches Analysesystem und nicht zuletzt eine exaktere, eventuell den speziellen Bedingungen der dann noch nicht erfassbaren Wörter angepasste Analyse. Da das gegenwärtig funktionsfähige Saarbrücker Analysemodell diese Voraussetzungen nicht in hinreichendem Maße aufweist, waren fehlerfreie Ergebnisse nicht zu erwarten. Die erreichten Werte ermutigen aber dazu, den bisher eingeschlagenen Weg fortzusetzen.

Anmerkungen

1 Chomsky, N., *Aspekte der Syntax-Theorie* (Frankfurt, 1969), 181

2 Vgl. Dietrich, R., *Eine formale Beschreibung der starken und unregelmäßigen Verben der deutschen Gegerwartssprache*. Linguistische Arbeiten des Germanistischen Instituts der Universität des Saarlandes, 9 (Saarbrücken, 1970), Liste p. 35ff.

3 Billmeier, G., "*Simulation verbalen Verhaltens*". In: Vorabdruck der G.I.-Fachtagung: Information Retrieval Systeme, (Stuttgart, 1970), 104.

4 Wahrig, G., *Deutsches Wörterbuch*, (Gütersloh, 1968).

5 Z.B. Agricola, E., *Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und des Englischen* (Berlin, 1968); Weber, H.J., "*Bestimmung der Wortklassen*". In: Eggers, H. et al., *Elektronische Syntaxanalyse der deutschen Gegenwartssprache*. Ein Bericht (Tübingen, 1969).