Wolfgang Klein / Rainer Rath

AUTOMATISCHE LEMMATISIERUNG DEUTSCHER FLEXIONSFORMEN[1]

1 SKIZZE DES PROJEKTS "AUTOMATISCHE LEMMATISIERUNG DEUTSCHER FLEXIONSFORMEN" (abgekürzt : AL)[2]

1.1 Begriffe der AL[3]

(a) Ein Text bestehe aus den Sätzen:
DIE MASCHINE ARBEITET FEHLERLOS.
FEHLERLOSE MASCHINEN ARBEITEN.
FEHLERLOSE MASCHINEN HABEN GEARBEITET.
DIE MASCHINEN HAETTEN FEHLERLOS GEARBEITET.

(b) Ein Programm bringe den Text in die Form:

ARBEITEN	Häufigkeit:	1	Fundort:	Seite x,	Zeile y
ARBEITET	"	1	"	"	"
DIE	"	2	"	"	"
FEHLERLOS	"	2	"	"	"
FEHLERLOSE	"	2	"	"	"
GEARBEITET	"	2	"	"	"
HABEN	"	1	"	"	"
HAETTEN	"	1	"	"	"
MASCHINE	"	1	"	"	"
MASCHINEN	"	3	"	"	"

(c) Ein zweites Programm bringe den Text in die Form:

ARBEITEN	Verb	Häufigkeit 4	davon: 3. Pers. Sing. Präs. Ind. Aktiv usw. 1 mal	Fundort
DIE	...	...	...	...
FEHLERLOS	...	...	...	...
MASCHINE	...	...	...	...

(d) Die Operation (b) wird gewöhnlich Indizierung, das Ergebnis Wortformen-Index genannt. Die Operation (c) nennen wir Automatische Lemmati

sierung (AL), das Ergebnis lemmatisiertes Wörterbuch. Die AL hat also die Aufgabe, eine gegebene Folge von Flexionsformen ihrem jeweiligen Lemma zuzuordnen. Eine ausgezeichnete Flexionsform des Lemmas ist als "Lemmaname" definiert. Die Menge aller Lemmanamen bildet zusammen mit bestimmten grammatischen Informationen das lemmatisierte Wörterbuch.

(e) Wortform nennen wir eine von Leerzeichen (Zwischenraum) begrenzte Graphemfolge. (Statt des Leerzeichens können bestimmte Satzzeichen stehen, z.B. Komma, Punkt).

(f) Unter einer Flexionsform verstehen wir ein Paar $(G_i, , I_i)$, wobei G_i eine von Leerzeichen begrenzte Graphemfolge ist, die eine oder mehrere Wortformen umfassen kann. I_i ist ein Tripel (S_i, R_i, P_i), wobei S. ein Komplex semantischer, R_i und P_i Komplexe syntaktischer Merkmale sind. R_i enthält syntagmatische Merkmale wie Kasus, Numerus, P_i paradigmatische Merkmale wie Wortklasse, Rektion.

(9) Sind $F_1, F_2, \ldots, F_n$ alle Flexionsformen, für die gilt:

$$F_i = (G_i, (S, R_i, P))$$
$$L = \{G_1, G_2, \ldots, G_n\}$$
$$M = (\{R_1, R_2, \ldots, R_n\}, S, P)$$

dann bezeichnen wir das Paar $\mathfrak{L} = (L,M)$ als Lemma. Vereinfacht gesagt: Ein Lemma besteht aus der Menge der Flexionsformen, bei denen gilt:

(h) $S_1 = S_2 = S_3 = \ldots S_i$ und $P_1 = P_2 = P_3 = \ldots P_i$.

Als Lemmanamen setzen wir bei Substantiven den Nom, sing., bei Verben den Infinitiv, bei Adjektiven die unflektierte Positivform (Adverbialform) fest.

1.2 Index, lemmatisiertes Wörterbuch, Anwendungen

Das Ziel der Arbeiten ist mit der Operation 1.1 (c) beschrieben. Verbal ausgedrückt: Ein beliebiger Text der deutschen Gegenwartssprache soll automatisch lemmatisiert werden, das heißt die einzelnen Flexionsformen sollen ihrem jeweiligen Lemma, repräsentiert durch den Lemmanamen, maschinell zugeordnet werden. Die Liste der Lemmanamen soll alphabetisiert, und jedem Lemmanamen sollen Informationen über die Häufigkeit seiner einzelnen Flexionsformen, über seine grammatischen Eigenschaften und über seinen Fundort beigegeben werden. Die

Menge aller Lemmanamen und aller Informationen nennen wir das "lemmatisierte Wörterbuch des Textes X". In letzter Zeit sind eine Reihe automatisch hergestellter Wortformen-Indices erschienen.[4] Die Mängel derartiger Indices sind bekannt. Sie bestehen im wesentlichen darin, daß (a) verschiedene Flexionsformen eines Lemmas an verschiedenen Stellen des Indexes stehen, sofern ihre Graphemfolgen voneinander abweichen ($G_i \neq G_j$); daß (b) semantisch und syntaktisch homographe Wortformen[5] nicht oder nur unzulänglich getrennt werden können[6]; daß (c) Flexionsformen, die aus mehreren Wortformen bestehen (AM SCHOENSTEN, IST GESEHEN WORDEN usw.), nicht als solche erkannt und somit dem entsprechenden Lemmanamen (z.B. AM SCHOENSTEN zu SCHOEN und IST GESEHEN WORDEN zu SEHEN) auch nicht zugeordnet werden können; daß (d) außer Frequenzangaben zu den einzelnen Wortformen keine weiteren Informationen ermittelt und ausgegeben werden können.

Diese von den Indexherstellern bewußt in Kauf genommenen Mängel beruhen vor allem darauf, daß die zu lemmatisierenden Texte linguistisch nicht analysiert werden. Das Verfahren der Indexherstellung besteht - vereinfacht gesagt - darin, einen gegebenen Text in seine Wortformen zu segmentieren, (was relativ einfach ist, da die Wortformen von Leer- oder Satzzeichen begrenzt sind) und diese alphabetisch zu sortieren. Bei aller Unvollkommenheit bieten diese Indices auch "Vorteile gegenüber der herkömmlichen Wörterbuchkonzeption". Sie "bestehen ... nicht allein in der Schnelligkeit und praktischen Fehlerlosigkeit der Bearbeitung .... sondern auch in der bislang kaum erreichbaren Vollständigkeit der Belege".

Dieser Sachverhalt - einerseits eine linguistisch unbefriedigende Verarbeitung des Textes mit der Konsequenz einer unzulänglichen Textausgabe, andererseits eine überaus schnelle, fehlerfreie und umfassende Textbehandlung - war für die Saarbrücker Arbeitsgruppe der Anlaß, ein Verfahren zu entwickeln, nach dem Texte schnell, umfassend, fehlerfrei und linguistisch analysiert zu einem Wörterbuch der oben beschriebenen Art verarbeitet werden können. Ein solches Wörterbuch erfüllt zunächst den gleichen Zweck wie ein Index. Nur ist es wesentlich leichter zu benutzen, weil die oben aufgeführten Nachteile eines Indexes weitgehend wegfallen.

Es ist darüberhinaus vielseitiger verwendbar, weil außer den Frequenzangaben grammatische Informationen zu den einzelnen Lemmata angegeben sind. Das lemmatisierte Wörterbuch ist mithin zugleich ein grammatisch spezifizierter Stellennachweis. Irgendwelche sprachlichen oder literarischen Phänomene, die sich bis aufs "Wort" reduzieren lassen, können über das lemmatisierte Wörterbuch schnell ermittelt und weiter bearbeitet werden.

Ein solches Lemmatisierungsverfahren dient nicht allein zum Anlegen von literarischen Stellennachweisen und damit als technische Interpretationshilfe. Es können auch durchschnittliche Gebrauchstexte unter bestimmten - am Wortschatz orientierten - Gesichtspunkten untersucht und verglichen werden. Um ein praktisches Beispiel zu nennen: Mit Hilfe von lemmatisierten Wortverzeichnissen (mit Frequenzangaben und grammatischen Informationen) könnten Fragen des Sprachgebrauchs behandelt werden: Welche Wörter, welche Wortklassen werden in einem Text bevorzugt, welche kommen überhaupt nicht vor? Welche Formen eines Paradigmas werden gemieden, welche nicht? Das lemmatisierte Wortverzeichnis könnte das Material zur weiteren, nicht-maschinellen Verarbeitung geordnet bereitstellen.

Eine weitere Anwendung der AL: Man könnte auf der formalen Grundlage des Algorithmus zur AL, sofern sie sich an einer größeren Textmenge als kompetent erwiesen hat, eine rein formale Formen- und Paradigmenlehre entwickeln.

1.3 Lemmatisierungsverfahren

1.3.1 Wir gehen davon aus, daß sich eine Flexionsform (G_i, I_i) in der Regel so zerlegen läßt, daß die dabei entstehenden Teilgraphemfolgen jeweils die Informationen S_i, R_i und P_i repräsentieren. Die Information S_i wird von der Teilgraphemfolge G_{Si}, die Information R_i von G_{Ri} repräsentiert. P_i, die paradigmatische Information, wird entweder

(a) ebenfalls durch G_{Ri}, oder

(b) durch G_{Pi}

repräsentiert. Im Falle (a) ist die Menge der G_{Ri} repräsentierenden Teilgraphemfolgen paradigmentypisch. So gilt etwa für die folgende

Menge: {-e, -er, -em, -en, -es, -ere, -erer, -erem, -eren, -eres, -ste, -ster, -stes, -sten, -stem}, daß sie syntagmatische Informationen Tür Adjektive, das heißt aber syntagmatische und paradigmatische Informationen repräsentiert.[12] Im Falle (b) enthält die Flexionsform eine Graphemfolge, die neben eventuellen semantischen vornehmlich paradigmatische Informationen repräsentiert. Z.B. repräsentiert -ISIEREN die Information "Verb", -KEIT "Substantiv", -BAR "Adjektiv".

Es wird bei einer AL vor allem darauf ankommen, diejenigen Teilgraphemfolgen aus den Flexionsformen herauszufiltern, die die Information P_i repräsentieren. Denn mit deren Hilfe können bestimmte Lemmatisierungsprozesse in Gang gebracht werden. Diese Informationen werden in verschiedener Weise gewonnen, verwertet, interpretiert:

(a) durch Benutzung eines "Wörterbuchs zur Lemmatisierung",
(b) durch Wortzerlegungs- und Vergleichsprogramme und
(c) durch Programme für die syntaktische Analyse des Textes.

1.3.2 Eine Lemmatisierung ohne ein Wörterbuch wäre nur für eine kleinere Teilmenge der Flexionsformen (und dann auch noch mit bestimmten Einschränkungen) möglich, und zwar nur für diejenigen, die eine paradigmatisch eindeutig interpretierbare Graphemfolge enthalten. So sind beispielsweise die Graphemfolgen -LICHSTE, -LICHERE paradigmatisch eindeutig interpretierbar: Sie repräsentieren (u.a.) die paradigmatische Information "Adjektiv". Der Lemmaname dieser Adjektive endet auf -LICH, die einzelnen Flexionsformen können ihm zugeordnet werden. Die meisten für die Lemmatisierung zu untersuchenden Graphemfolgen aber sind mehrdeutig, das heißt sie können nicht mechanisch auf ein Lemma bezogen werden.[13]

Die Mehrdeutigkeit der paradigmatisch interessanten Graphemfolgen macht eine Kontrolle der Zerlegungsprozeduren erforderlich. Diese erfolgt durch ein Wörterbuch.

Das Wörterbuch soll das Vokabular des zu lemmatisierenden Textes enthalten. Da beliebige deutsche Gegenwartstexte lemmatisiert werden sollen, läuft diese Forderung praktisch darauf hinaus, daß das Wörterbuch das vollständige Vokabular der deutschen Gegenwartssprache in einer maschinengerechten Form enthält. Diese Forderung ist,

streng genommen, nicht zu erfüllen. Man denke nur an das umfangreiche fachsprachliche Vokabular, an die fast beliebig bildbaren und täglich neu gebildeten Augenblickskomposita, an die vielen, oft kaum eingedeutschten Fremdwörter. Es geht deswegen zunächst darum, ein Wörterbuch aufzubauen, das einerseits den gebräuchlichen Wortschatz der Gegenwartssprache enthält, also sehr umfangreich ist, das andererseits aber in einer bestimmten zeitlichen Frist fertiggestellt werden kann.

Wir haben uns entschlossen, dem "Wörterbuch zur Lemmatisierung" Gerhard Wahrigs *Deutsches Wörterbuch* [14] zugrunde zu legen. [15] Dafür waren folgende Gründe maßgebend: Es handelt sich bei diesem Wörterbuch um ein relativ neues Werk, das den modernen Wortschatz voll berücksichtigt. Das Wörterbuch ist ziemlich umfangreich, es enthält 90.000 verschiedene Stichwörter. Der Hauptgrund aber ist der, daß jedem Stichwort umfangreiche semantische und syntaktische Angaben beigegeben sind, die in das "Wörterbuch zur Lemmatisierung" eingebracht werden können.

Aus der Unvollständigkeit des Wörterbuchs ergeben sich zwei Folgerungen: (a) Das Wörterbuch muß ständig erweitert werden. Alle Wörter, die in dem zu lemmatisierenden Text vorkommen, nicht aber im Wörterbuch stehen, sollen in einem automatischen oder halbautomatischen Verfahren in der gewünschten Form ins Wörterbuch aufgenommen werden, (b) Es muß eine Prozedur vorgesehen werden, die es erlaubt, nicht im Wörterbuch verzeichnete Wörter unaufwendig, das heißt möglichst auch automatisch zu lemmatisieren. Das Wörterbuch wird nach explizit definierten Konventionen, sogenannten Kodierungsvorschriften, angelegt.

Ein Wörterbucheintrag besteht aus einem Wort- und aus einem Informationseintrag. Der Worteintrag ist bei den Verben der Infinitiv minus -EN (bzw. minus -N; vgl. z.B. HANDELN), also für GLAUBEN z.B. GLAUB,[16] bei den Adjektiven die unflektierte Positivform (SCHOEN) und bei den Substantiven der Nominativ Singular (HAUS). Der Informationseintrag enthält paradigmatische, syntagmatische und einige Subkategorisierungsangaben.

1.3.3 Die Wortzerlegungs- und Vergleichsprogramme ordnen mit Hilfe des

Wörterbuchs die Flexionsformen des Textes ihrem jeweiligen Lemma zu. Dabei sollen die syntagmatischen Merkmale der Flexionsformen (bei den Substantiven z.B. Kasus und Numerus) ermittelt und für eine spätere Ausgabe aufbewahrt werden. Die Zuordnung einer Flexionsform zu ihrem Lemma geschieht so, daß zunächst jede Wortform des Textes mit dem Worteintrag des Wörterbuchs verglichen wird.

Stimmen Textwortform und Worteintrag des Wörterbuchs überein,[18] tritt also eine Wortform auf, die als Lemmaname definiert ist, kann sogleich und ohne weitere Analyse der Wortform eine Lemmatisierung erfolgen.[19] Ist der Vergleich zwischen der Textwortform und dem Worteintrag des Wörterbuchs negativ (kein Worteintrag des Wörterbuchs stimmt mit der Textwortform überein) beginnt der Prozeß der Zerlegung der Wortform. Es muß geprüft werden, in welche Teilgraphemfolgen G_{Si}, G_{Ri} und G_{pi} die Textwortform zerlegbar ist. Die dabei entstehende Teilgraphemfolge G_{Si} - sie entspricht dem Worteintrag des Wörterbuchs - wird mit den Worteinträgen des Wörterbuchs verglichen. Zerlegungs- und Vergleichsprozeß sind erst dann abgeschlossen, wenn ein positives Vergleichsergebnis erzielt worden ist. Die Zerlegung in Teilgraphemfolgen ist für die einzelnen Wortklassen und innerhalb dieser für die einzelnen Paradigmenklassen verschieden. Die Zerlegungsprogramme müssen von vornherein alle paradigmatisch und syntagmatisch relevanten Grapheme und Graphemkombinationen für alle Wortklassen enthalten. Es muß weiter berücksichtigt werden, daß es innerhalb der einzelnen Wortklassen eine ganze Reihe verschiedener Paradigmenklassen gibt, und daß diejenigen Grapheme und Graphemkombinationen, die die Information P. repräsentieren, fast alle, wie schon angedeutet, mehrdeutig sind. Ein wesentliches Problem ist, eine hierarchische Ordnung der Zerlegungsprogramme und ihren kombinatorischen Einsatz festzulegen.

Eine weitere generelle Schwierigkeit sei hier nur kurz angedeutet: Die im Deutschen überaus häufige Umlautbildung macht den Zerlegungsprozeß innerhalb der AL komplizierter. Bei einer automatischen Behandlung eines nicht präparierten Textes - wie dies bei der AL der Fall ist - kann bei einem Auftreten der Graphemfolgen AE, OE, UE nicht sofort automatisch entschieden werden, ob sie das Ergebnis pa-

radigmatisch bedingter Veränderungen (z.B. STARK, STAERKER) oder unveränderliches Merkmal aller Flexionsformen eines Lemmas (z.B. SCHOEN, SCHOENSTE usw.) sind. Daher müssen die Zerlegungsprogramme die generelle Möglichkeit vorsehen, jede in einer Wortform auftretende Graphemfolge AE, OE, UE zu A, O, U zu reduzieren, um eventuell ein positives Vergleichsergebnis zu erzielen. Die Zerlegungsprogramme dienen also - zusammengefaßt - dem Zweck, die Wortformen des Textes in eine in bezug auf das Wörterbuch vergleichsfähige Form zu bringen. Das setzt voraus, daß den Zerlegungsprogrammen von vornherein alle paradigmatisch und syntagmatisch bedingten Veränderungsmöglichkeiten für alle Elemente der drei hier zur Diskussion stehenden Wortklassen Verb, Adjektiv, Substantiv bekannt sein müssen.

4 In einer Reihe von Fällen ist eine AL nur im Zusammenhang mit einer syntaktischen Analyse des Satzes möglich. Im wesentlichen handelt es sich dabei um die folgenden Probleme:

(a) Beseitigung von Wortklassenmehrdeutigkeiten;
(b) Erkennung von Flexionsformen, die aus mehreren Wortformen bestehen;
(c) Lemmatisierung von Flexionsformen, deren, Lemmaname nicht im Wörterbuch steht.

(a) Angenommen, die Flexionsform BILLIGE tritt in dem zu lemmatisierenden Text auf. Sie ist mehrdeutig, denn sie kann zum Adjektivlemma BILLIG oder dem Verblemma BILLIGEN gehören. Eine richtige Lemmatisierung setzt in einem solchen Fall die Kenntnis der Wortklasse der zu lemmatisierenden Flexionsform voraus. Diese kann nicht dem Wörterbuch entnommen werden, eben weil es sich um eine mehrdeutige Flexionsform handelt. Das Wörterbuch vermittelt indes die Information, welchen Wortklassen die in Rede stehende Flexionsform angehören kann; in Abhängigkeit von dieser Information kann nun durch eine syntaktische Analyse der Umgebung der mehrdeutigen Flexionsform die im Text realisierte Wortklasse ermittelt werden.[21]

(b) Die Flexionsformen, die aus mehreren Wortformen bestehen, sind zu unterteilen in solche, die nur kontinuierlich stehen können, und solche, die kontinuierlich oder diskontinuierlich angeordnet sein können. Im ersten Falle ist nur eine engere Umgebungsanalyse,

im zweiten in der Regel eine syntaktische Analyse des gesamten Satzes erforderlich.

Ein Beispiel für eine kontinuierliche mehrteilige Flexionsform ist der Superlativ der Adjektive mit AM: AM SCHOENSTEN; beide Wortformen zusammen bilden eine Flexionsform; sie werden als eine Einheit dem Lemma SCHOEN zugeordnet.[22]

Beispiele für diskontinuierliche Flexionsformen, die aber auch kontinuierlich stehen können, sind: HAT ... GESEHEN, TRIFFT ... EIN,[23] WURDE ... BEZAHLT; auch hier bilden die einzelnen Wortformen zusammen eine Flexionsform: HAT ... GESEHEN wird dem Lemma SEHEN, TRIFFT ... EIN dem Lemma EINTREFFEN, WURDE ... BEZAHLT dem Lemma BEZAHLEN zugeordnet.[24]

(c) Bei der Lemmatisierung von Flexionsformen, deren Lemmaname nicht im Wörterbuch steht, sind zwei Vorgänge zu unterscheiden: Einmal sollen diese Flexionsformen - soweit wie möglich ebenfalls automatisch - lemmatisiert werden, zum anderen soll das Wörterbuch zur Lemmatisierung automatisch in der vorgeschriebenen Form komplettiert werden.

Zunächst wird geprüft, ob die "unbekannte Flexionsform" eine paradigmatisch eindeutig interpretierbare Graphemfolge enthält (z.B. -KEIT, -LICHSTE). Das bedeutet, daß dem Zerlegungsprogramm eine Liste aller paradigmatisch eindeutig interpretierbaren Graphemfolgen zur Verfügung stehen muß, und zwar mit allen syntagmatisch bedingten Veränderungsmöglichkeiten. So müssen z.B. für adjektivische Ableitungssilben jeweils 16 verschiedene Graphemfolgen angesetzt werden; für -LICH beispielsweise: -LICH, -LICHE, -LICHES, -LICHER, -LICHEN, -LICHEM (Positiv); -LICHERE, -LICHERES, -LICHERER, -LICHEREN, -LICHEREM (Komparativ); -LICHSTE, -LICHSTES, -LICHSTER, -LICHSTEN, -LICHSTEM (Superlativ).

Enthält die "unbekannte Flexionsform" eine solche Graphemfolge, kann in den meisten Fällen eine Lemmatisierung erfolgen.[25] Ist dies nicht der Fall, wird die "unbekannte Flexionsform" zu einem künstlichen Homographen erklärt (Verb/Adjektiv/Substantiv; evtl. noch andere Kombinationen). Eine Umgebungsanalyse kann nun - genau so wie im Fall der "echten" Homographen, deren Mehrdeutigkeitsangaben dem Wörterbuch

entnommen werden - bestimmte paradigmatische und syntagmatische Informationen über sie ermitteln. Mit diesen Informationen ist eine AL in einer ganzen Reihe von Fällen möglich. Summarisch sei das folgende Beispiel behandelt. Ein Satz heiße: MODERNSTE FORMEN SIND ERWUENSCHT. Im Wörterbuch zur Lemmatisierung mögen die Lemmanamen FORM und ERWUENSCHT sowie sämtliche Formen von SEIN, also auch SIND, stehen.[27] Im ersten Vergleich zwischen den Textwortformen und den Worteinträgen des Wörterbuchs werden SIND und ERWUENSCHT erkannt und automatisch lemmatisiert. Ober ihre für die weitere Analyse sehr wichtigen grammatischen Informationen kann verfügt werden. Für die Textwortformen MODERNSTE und FORMEN ist der erste Vergleich negativ. Auf beide werden nun die Zerlegungs- und die weiteren Vergleichsprogramme angewandt. Dabei kann FORMEN dem Lemma FORM zugeordnet werden, dessen grammatische Informationen nun ebenfalls verfügbar werden. Für MODERNSTE ist eine Lemmatisierung zunächst nicht möglich, es fehlt ja der Wörterbucheintrag MODERN. MODERNSTE wird nunmehr zum Homographen erklärt, das entsprechende Lösungsprogramm tritt in Kraft. Da über das Wörterbuch bekannt ist, daß FORMEN ein Substantiv femininum Plural ist, da ferner die Position von MODERNSTE ermittelt werden kann (vor einem Substantiv stehend, am Anfang eines Aussagesatzes) kann MODERNSTE nunmehr als Adjektiv und aufgrund der Endgrapheme -STE als Superlativ klassifiziert werden. Das Lemma MODERN kann erschlossen und ins Wörterbuch aufgenommen werden.

Nicht in allen Fällen ist eine automatische Lemmatisierung unbekannter Flexionsformen möglich. Es werden oft Mehrdeutigkeiten bestehen bleiben.[28] In diesen Fällen wird es aber möglich sein, mehrere Lemmatisierungsvorschläge zur Disposition zu stellen. Sie werden einem menschlichen Bearbeiter zur Entscheidung übergeben; diese Entscheidung wird in geeigneter Form der Maschine übermittelt, die das Wort ins Wörterbuch zur Lemmatisierung aufnimmt und es auch als lemmatisierte Form in das lemmatisierte Wörterbuch des entsprechenden Textes stellt.

2 ALLGEMEINE PROBLEME DER KODIERUNG

Im folgenden werden einige Fragen aus dem Bereich der Kodierung be-

sprochen, die nicht wortklassenspezifisch sind. Eine scharfe Abgrenzung von wortklassenspezifischen und nicht wortklassenspezifischen Problemen ist allerdings nicht möglich, da die Problemstellungen unter theoretischem Aspekt die gleichen, ihre Lösungen aber aus praktischen Erwägungen verschieden sein können und umgekehrt. Darauf wird jeweils im einzelnen eingegangen.

2.1 Wortaufnahme und Worteintrag

2.1.1 Prinzipien der Wortauswahl

Die AL soll auf beliebige deutsche Texte anwendbar sein. Daher müßte das Wörterbuch im Prinzip sämtliche deutschen Wörter umfassen. Man schätzt den Wortschatz des Deutschen auf etwa 300.000 bis 500.000 Wörter; die einschlägigen Angaben schwanken ganz erheblich, aber 300.000 kann als akzeptables Minimum gelten. Wahrigs *Deutsches Wörterbuch*, von dem wir ja ausgehen, umfaßt weniger als ein Drittel davon; dennoch enthält es eine Reihe von Einträgen, die einem durchschnittlichen Sprecher des Deutschen und selbst einem Germanisten unbekannt sein dürften, z.B. Verben wie FRETTIEREN, ALFANZEN, RAETTERN, RAJOLEN, KAAKEN, RINKELN, Substantive wie ICHOR, GOEPEL, BALGE, RASTRAL, RAIZE, GIEMEN, RIESTER, KOSAETE, Adjektive wie ADENOID, RUECKENSCHLAECHTIG, RASS, RITTIG, um willkürlich einige anzuführen. Nun schwankt aber, wie man vielleicht schon an diesen Wörtern sieht, der passive Wortschatz außerordentlich stark von Sprecher zu Sprecher; wollte man alle Wörter ausschließen, die ein oder mehrere Kodierer nicht kennen (etwa nach der Regel: "Es werden nur Wörter aufgenommen, die mindestens einer der im Raum Anwesenden kennt"), dann hieße das ja bloß, die Kompetenz in Sachen Lexikon der - zudem noch vielfach wechselnden - Kodierer an die Stelle der Wahrigs und seiner Mitarbeiter, also immerhin geschulter und erfahrener Lexikographen, zu setzen. Das ist allem Anschein nach nicht sehr sinnvoll, und da es auch sonst kein praktikables Kriterium für Aufnahme oder Nichtaufnahme zu geben scheint, haben wir uns entschlossen, alle im "Wahrig" stehenden Wörter aufzunehmen, so daß das Verfahren im Prinzip auch in der Lage ist, Texte zu bearbeiten, die die lexikalischen Kenntnisse eines normalen Sprechers übersteigen.

[29]

Anders steht es mit der Abgrenzung nach oben. Es gibt eine ganze Anzahl von Wörtern - wobei man allerdings anfügen müßte, daß wir auch über kein Kriterium für "geläufig" verfügen - die sich nicht im "Wahrig" finden; es handelt sich dabei im wesentlichen um

- Fachtermini: DIAGONALVERFAHREN, WURZELGLEICHUNG, WAHRHEITSWERT, PRAEDIKATENLOGIK, AUSSAGENVARIABLE usw.
- Namen, die an der Grenze zu Appellativa stehen: DUDEN ist belegt, BROCKHAUS nicht bzw. nur in anderer Bedeutung,
- geographische Bezeichnungen: DEUTSCH, FRANZOESISCH, RUSSISCH sind vorhanden, nicht aber z.B. SARDISCH, BOTTNISCH, MEDISCH (aber MEDER und PERSISCH),
- aktuelle Fremdwörter: KONZERTIERT, ESTABLISHMENT,
- Ableitungen, die nicht systematisch aufgeführt sind: ABBRUCHREIF, aber nicht BAUREIF-, BASSGEIGE, BASSGITARRE, BASSGEIGER, aber nicht BASSGITARRIST; etwas ABONNIEREN, aber nicht ABONIERT SEIN auf etwas,
- zufällige Lücken: z.B. wird ABKEHR definiert durch ABWENDUNG, dieses steht jedoch nicht im "Wahrig".

Wie man sieht, handelt es sich durchweg um etwas abgelegene Wörter, die aber doch mindestens so bekannt sind wie die weiter oben angeführten belegten.

Um die Anwendbarkeit auf beliebige Texte zu gewährleisten, kann jeder Kodierer zusätzlich Wörter nach seiner Kenntnis aufnehmen. Eine Beschränkung der Wortaufnahme erfolgt also lediglich nach unten.

Es ist durchaus denkbar, daß zu einem späteren Zeitpunkt eine systematische Erweiterung mit Hilfe anderer Wörterbücher vorgenommen wird; aber das ist eher eine arbeitstechnische Frage. Aus denselben arbeitstechnischen Gründen werden zunächst auch nur die Hauptwortklassen Adjektiv, Substantiv, Verb verschlüsselt, da sie - jedenfalls unter morphologischen Aspekten - am wichtigsten sind; die übrigen Wortklassen sind für eine AL weit weniger interessant und zudem nicht so umfangreich; außerdem könnten sie eventuell aus dem in Saarbrükken vorhandenen syntaktischen Wörterbuch [30] übernommen werden, in dem sie so gut wie vollständig verzeichnet sind.

Weiterhin wurden von der Kodierung vorerst die Abkürzungen ausge-

nommen; sie sind im "Wahrig" nicht sehr systematisch aufgenommen und sollen daher später aus einem speziellen Abkürzungslexikon nachgetragen werden; zudem mUssen hier einige Besonderheiten wie z.B. der sogenannte AbkUrzungspunkt beachtet werden.[31]

Ein besonderes Problem schließlich stellen die verschiedenen idiomatischen Wendungen dar, die aus mehreren Wörtern bestehen, aber eine semantische Einheit bilden. Daher müssen unserem Lemmakonzept zufolge Ausdrücke wie **IN ZUSAMMENHANG STEHEN, ZUR ENTSCHEIDUNG BRINGEN, HUNGER HABEN, IM BEGRIFF SEIN,** auch **BLINDER PASSAGIER, AUSWAERTIGES AMT, KREUZ UND QUER, GUT UND GERN** unter ein Lemma gefaßt werden; deswegen wurde auch bei der Definition des Lemmas betont, daß eine Flexionsform durchaus aus mehreren Wortformen bestehen kann (s.o. 1.1 (f)). Die Erfassung derartiger "mehrwortiger Ausdrücke" ist allerdings sehr schwierig; das wird ersichtlich, wenn man die völlig parallelen Konstruktionen **ZUR ENTSCHEIDUNG BRINGEN : ZUR WERKSTATT BRINGEN** oder **KARL UND FRITZ : KRETHI UND PLETHI** oder **AUSWAERTIGES AMT : ZUSTAENDIGES AMT** vergleicht. Syntaktisch können diese Konstruktionen nicht differenziert werden. Die Wörter, die in derartigen idiomatischen Wendungen auftauchen, müssen daher im Wörterbuch besonders markiert werden; dazu müssen zunächst alle möglichen Wendungen zusammengestellt und systematisch analysiert werden; eine entsprechende Untersuchung ist derzeit im Gange; Teilergebnisse liegen auch schon vor;[32] bevor sie jedoch abgeschlossen ist, ist eine Kodierung der idiomatischen Wendungen, die eine Unterscheidung von parallelen syntaktischen Konstruktionen und damit eine korrekte Lemmatisierung gewährleistet, nicht möglich.

2.1.2 Der Eintrag

Jedes aufzunehmende Wort wird in einer orthographischen Normalform in den Spalten 2-33 (Substantiv), 2-29 (Verb) und 2-29 (Adjektiv) einer Lochkarte verzeichnet.

Was als "Normalform" angesehen wird, hängt durchaus von praktischen Gesichtspunkten ab; die Normalform braucht keineswegs mit dem späteren Lemmanamen übereinzustimmen; in der Tat ist sie beim Verb in keinem Fall mit dem (voraussichtlichen) Lemmanamen identisch. Normalform ist

bei den Verben:	der Infinitiv minus Endung:[33] GEH, LACH, AUSZIEH, WANDER, DESINTEGRIER, SAUF;
bei den Substantiven:	der Nominativ Singular: MANN, POSTBOTE, ABALIENATION; pluralia tantum müssen natürlich im Plural aufgenommen werden: LEUTE, MASERN;
bei den Adjektiven:	die unflektierte Positivform: STARK, SCHOEN, HELIOTROP.

Bei Adjektiven und Verben müssen außer dieser Normalform auch die möglichen graphematischen Varianten eines Stammes (Ablaut, Heteroklisie) als Sonderformen verzeichnet werden: GING, GANG (zu GEH), BEST, BESS (zu GUT) usw. Diese Sonderformen werden wie eigene Einträge behandelt, das heißt auf einer eigenen Karte in die oben angegebenen Spalten eingetragen; außerdem aber wird bei ihnen ein "Worteintrag II" besetzt, in den die Normalform zu der betreffenden Sonderform eingetragen wird (also etwa GEH bei GANG oder GUT bei BESS); auf diese Weise ist es möglich, die verschiedenen graphematischen Varianten eines Stammes wieder aufzufinden. Für den Worteintrag II ist bei den Verben Spalte 44 bis Spalte 71 vorgesehen, bei den Adjektiven Spalte 44 bis 59. Bei den Substantiven ist kein Worteintrag II erforderlich, da die einzigen hier auftretenden graphematischen Variationen - der Umlaut - über eine eigene Spalte erfaßt werden.

Neben diesen grammatisch bedingten Sonderformen bzw. Umlautformen gibt es gelegentlich graphematische Variationen eines Wortes, die grammatisch irrelevant sind und auf regionale Unterschiede oder - bei Fremdwörtern - unterschiedliche Eindeutschung bzw. unterschiedliche Transkription zurückzuführen sind. Wir führen hier nur einige Beispiele derartiger Nebenformen an:

(1) 1. Verben

ABKNAPPEN	-	ABKNAPSEN
ABLOHNEN	-	ABLOEHNEN
PANSCHEN	-	PANTSCHEN
PLANSCHEN	-	PLANTSCHEN
PIEPSEN	-	PIEPEN
POCHIEREN	-	POSCHIEREN

PHOTOGRAPHIEN		FOTOGRAFIEREN
2. Substantive		
BAUCHKNEIPEN	-	BAUCHKNEIFEN
BEISEL	-	BEISL
DRIFT	-	TRIFT
ALP	-	ALB
BANKEROTT	-	BANKROTT
BANDONEON	-	BANDONION
BASAR	-	BAZAR
MOSLEM	-	MUSLIM
3. Adjektive		
FLATTERIG	-	FLATTRIG
MICKERIG	-	MICKRIG
ENGELSGLEICH	-	ENGELGLEICH
EXTRAVERTIERT	-	EXTROVERTIERT

Bei den Substantiven erhält jedes derartige Wort einen Hinweis (in Spalte 78), daß dazu eine Nebenform existiert; das erlaubt es, später sämtliche im Wörterbuch vorhandenen Nebenformen herauszusondern. Bei Adjektiven und Verben wird eine der verschiedenen Formen - in der Regel diejenige, die zuerst im Wörterbuch auftritt - als Normalform gesetzt, die andere (oder die anderen, falls es mehr als zwei Möglichkeiten gibt) wird bei gleichen grammatischen Informationen als Sonderform behandelt; sie enthält also im Worteintrag II die andere, als Normalfall angesetzte Variante; zusätzlich wird in Spalte 41 (Adjektiv) bzw. über den Stammnummerncode (beim Verb) registriert, daß es sich um Nebenformen handelt. Im übrigen ist es, wie schon aus den wenigen Beispielen deutlich wird, vielfach eine Anschauungsfrage, ob man zwei gegebene Formen als Varianten oder als eigenständige Wörter ansehen will. Feste Prinzipien dafür lassen sich kaum aufstellen; es ist auch nicht nötig, da das Problem ziemlich marginal ist.

2.1.3 Sonderinformationen im Worteintrag

Die orthographische Wiedergabe des Wortes im Worteintrag wird durch eine Reihe zusätzlicher Informationen ergänzt, die mit Hilfe einiger Sonderzeichen kodiert werden. Die entsprechenden Konventionen sollen im folgenden

systematisch dargestellt und diskutiert werden.

2.1.3.1 Virgel ("/")

Es ist zwar augenblicklich noch nicht vorgesehen, zusammengesetzte Wörter unter einem Lemma, nämlich dem des "Grundwortes", zusammenzufassen; aber diese Möglichkeit sollte doch von Anfang an eingeplant werden. Daher werden die Kompositionselemente zusammengesetzter Wörter durch eine Virgel (Schrägstrich, slash) "/" abgetrennt:

HELL/ROT, ZINNOBER/ROT, BLAU/ROT - Grundwort ROT;
VOR/GEH, HINTER/GEH, ZU/GEH - Grundwort GEH;
BRIEF/TRAEGER, WASSER/TRAEGER, EISEN/TRAEGER {?), HOSEN/TRAEGER (?) - Grundwort TRAEGER.

Das letzte Beispiel wie auch die Vagheit des Ausdrucks "Grundwort" machen schon deutlich, daß bei dieser Abtrennung mit erheblichen Komplikationen zu rechnen ist. Sinn hat sie ja nur, wenn sie nach einem einheitlichen Prinzip vorgenommen wird und wenn die späteren "Grundwortlemmata" linguistisch vernünftige Zusammenfassungen darstellen. Es muß also ein Prinzip der Abtrennung formuliert werden, das zugleich praktikabel, das heißt möglichst unabhängig von der Intuition und der Intelligenz des Kodierers verwendbar, und linguistisch sinnvoll ist. Man könnte etwa immer dann eine Virgel setzen, wenn die rechts davon stehende Graphemfolge als selbständige Einheit auftreten kann; dieses Prinzip ist leicht anwendbar, es erlaubt eine rasche, spontane Entscheidung des Kodierers, führt aber, wie man leicht sieht, oft zu unsinnigen Abtrennungen; man denke nur an Fälle wie DEKAN/EI, SCHAUDER/HAFT, FURCHT/BAR usw. Diese extreme Lösung ist also nicht gangbar; die rechts stehende Graphemfolge und die entsprechende isolierte Einheit müssen sich zumindest ungefähr semantisch entsprechen, im Idealfall synonym sein; dieser Umstand muß bei der Formulierung eines Prinzips der Abtrennung berücksichtigt werden; dabei ergibt sich aber eine Reihe von Komplikationen, denn es ist vielfach nur sehr schwer zu entscheiden, ob ein zusammengesetztes Wort noch als Zusammensetzung zweier Kompositionselemente interpretiert werden kann oder ob die Kompositionselemente ihre ursprüngliche Bedeutung bereits so weit verloren haben, daß es nicht mehr sinnvoll erscheint, das Kompositum unter dem Grundwort aufzuführen. Im folgenden sollen die wichtigsten Schwierigkeiten an Beispielen kurz aufgezeigt werden; die Beispiele entstammen einer

Probekodierung der Substantive, bei denen ja die Möglichkeit, zusammengesetzte Wörter zu bilden, am stärksten entwickelt ist.

(2) (i) Sollen die Komposita DAUNENBETT, STERBEBETT, EHEBETT, WOCHENBETT, NAGELBETT, FLUSSBETT alle unter BETT subsumiert werden, und welche Kriterien kann man für die eine oder andere Entscheidung anführen? Es sei darauf hingewiesen, daß natürlich in vielen Fällen irgendwelche ad hoc motivierte Entscheidungen möglich sind. Die Schwierigkeit liegt aber darin, bei diesen stufenweisen Fortentwicklungen vom Grundwort ein praktikables, von jedem Kodierer leicht anwendbares Prinzip zu finden,

(ii) Das Wort SCHRIFT hat, wie sehr viele Wörter, eine ganze Anzahl von Bedeutungen, u.a. die folgenden: 1. DIE SCHRIFT IST TROCKEN. 2. DIE SCHRIFT IST AUSGEPRAEGT. 3. ... IST EINE ART KYRILLISCH MIT GRIECHISCHEN ELEMENTEN. 4. ... ENTHAELT ZAHLREICHE HINWEISE FUER DIE VERNICHTUNG VON KAKERLAKEN UND KELLERASSELN. Das Kompositum DEBATTIERSCHRIFT fällt nur unter 4.; es müßte also mit der Abtrennung durch "/" zugleich ein Hinweis gegeben werden, daß das Kompositum nur unter SCHRIFT in dieser Bedeutung paßt.[34] Das aber setzt voraus, daß die Grunwörter nach ihren verschiedenen möglichen Bedeutungen klassifiziert werden, eine Arbeit, die innerhalb der AL unmöglich geleistet werden kann.[35]

(iii) Eine Komplikation von Fall (ii) liegt vor, wenn auch ein Kompositum selbst mehrdeutig ist. Das Wort ABSCHLAG hat z.B. mindestens die beiden folgenden Bedeutungen: 1. ABSCHLAG VOM TOR. 2. VORAUSZAHLUNG AUF DEN ARBEITSLOHN. Beide sind (im Hinblick auf die weiter unten in Abschnitt 2.2 beschriebene Subkategorisierung) "abstrakt", aber allem Anschein nach läßt sich nur ABSCHLAG in der ersten Bedeutung unter eine der Bedeutungen von SCHLAG einordnen,

(iv) Ein verwandter, durch die Art der Bedeutungsunterschiede aber doch verschiedener Fall liegt vor bei Komposita wie HEIZSCHLANGE, PAPIERSCHLANGE usw., bei denen das Grundwort SCHLANGE normalerweise ein Merkmal /+belebt/ aufweist, die Komposita hingegen nicht. Entsprechende Fälle lassen sich auch für andere Merkmale belegen, etwa ABSCHLEPPDIENST /-abstrakt/ gegenüber DIENST /+abstrakt/.

(v) Gelegentlich gibt es zu einer Serie von Komposita kein Grundwort

mehr, wie etwa bei BUNDESTAG, LANDTAG, REICHSTAG, STAEDTETAG, KIRCHENTAG. (In diesem speziellen Fall ist ein Teil der Komposita mehrdeutig: alle können Institutionen bezeichnen, die ersten drei zudem Örtlichkeiten, die letzten drei - möglicherweise eintägige - Ereignisse; im übrigen sind alle mehrdeutig im Hinblick auf das Merkmal /+abstrakt/).

(vi) DAVIDSHARFE ist offenbar ein Kompositum von HARFE, das heißt es ist eine spezielle Art von Harfe gemeint; das ganze Wort bezeichnet eine Schneckenart mit einem an eine Harfe erinnernden Gehäuse. Im wesentlichen in dieselbe Rubrik fallen Komposita wie SEENELKE, SEESTERN, SEEWALZE, SEEKUH, MEERSCHWEINCHEN, WALFISCH, SEEJUNGFRAU, WASSERJUNGFER.

Die verschiedenen Problemklassen ließen sich um einige, die der Beispiele um zahlreiche ergänzen. Es sei noch einmal erwähnt, daß es natürlich nicht darauf ankommt, diese Fälle zu lösen, sondern für den Kodierer anwendbare Prinzipien zu formulieren. Das ist uns nicht gelungen. Wir haben uns daher zu einem Kompromiß entschlossen, der grobe Zuordnungsfehler weitgehend vermeidet, zwar zuviel Zuordnungen macht, aber so das gesamte Material für eine spätere eingehende Untersuchung der Wortbildung erfaßt und bereitstellt. Man kann das Prinzip der Abtrennung demnach etwa so formulieren:

(3) Eine Virgel wird gesetzt, wenn

(i) die rechts davon stehende Graphemfolge als selbständiges Wort vorkommt und

(ii) die rechts davon stehende Graphemfolge kein Ableitungsmorphem ist und

(iii) die rechts davon stehende Graphemfolge zur gleichen Wortklasse wie das ganze Wort gehört und

(iv) der übrige Teil des Wortes ein Morphem, Lexem oder eine Kombination davon ist.

Alle vier Bedingungen müssen erfüllt sein. Bedingung (ii) verhindert krasse Fehl Zuordnungen wie FURCHTBAR, SCHAUDERHAFT, DEKANEI; Bedingung (iii) verhindert, daß z.B. DAUMESDICK unter DICK auftaucht oder VERLANG unter LANG (diese Fälle sind nicht allzu häufig); Bedingung (iv) verhindert, daß unsinnige Abtrennungen wie BAL/LAST, VERD/RUSS oder VER/KUEN/

DUNG vorgenommen werden. Gelegentlich haben verschiedene der angeführten Bedingungen die gleiche Wirkung, aber das ist ja kein Schade.

Es ist ersichtlich, daß nach dieser Regel zu viele Komposita unter ein Grundwort aufgenommen werden. Dies scheint uns aber, solange es an einem zugleich praktikablen und linguistisch befriedigenden Kriterium mangelt, zweckmäßiger als umgekehrt möglichst wenig Komposita unter Grundformen zu verzeichnen (bzw. verzeichnen zu können, denn es handelt sich hier ja nur um die Möglichkeit); dadurch werden nämlich die Aussichten, die Wortbildung mit all ihren Idiosynkrasien, von denen einige oben erwähnt wurden, systematisch zu untersuchen, von vornherein stark beschnitten.

Die Virgel stellt nur eine der Informationen für den Aufbau eines Grundwortlemmas dar; über die grammatischen Informationen können weitere Restriktionen formuliert werden. Beispielsweise ist es plausibel anzunehmen, daß - bei den Substantiven - Grundwort und ganzes Wort im Genus übereinstimmen müssen; mit dieser Annahme kann z.B. ausgeschlossen werden, daß KOM/FORT unter FORT, GE/WICHT unter WICHT, GE/BISS unter BISS auftauchen. Ähnliche Restriktionen lassen sich auch - mit anderen grammatischen Informationen - für die anderen Wortklassen angeben; so kann man z.B. verlangen, daß Verben nur unter einem Grundwort auftauchen, wenn sie in die gleiche Paradigmenklasse gehören; so kann vermieden werden, daß etwa REITEN und BE/REITEN zusammengefaßt werden; allerdings werden dann z.B. HABEN und (SICH) GE/HABEN getrennt. Man kann sehr unterschiedlicher Meinung sein, ob das wünschenswert ist oder nicht; jedenfalls lassen sich die entsprechenden Bedingungen formulieren.

Daß dies im Einzelfall nicht immer ganz einfach ist, ist auch der Grund dafür, daß die Restriktionen nicht direkt in die Virgelkonvention (3) aufgenommen werden. Die genauen Bedingungen für den Aufbau eines sinnvollen Grundwortlemmas können daher erst nach eingehender Auswertung des Materials, das durch die weite Virgelkonvention, wie sie oben formuliert wurde, in erforderlicher Breite bereitgestellt wird, angegeben werden. Schließlich ist zu bemerken, daß die grammatischen Angaben ja auf jeden Fall kodiert werden müssen; sie auch noch einmal beim Worteintrag zu berücksichtigen, wäre also Doppelarbeit.

2.1.3.2 Bindestrich ("-")

Eine morphologische Zerlegung der Worteifiträge ist im Rahmen der AL nur

soweit von unmittelbarer Bedeutung, als die Flexionsmorpheme betroffen sind. Andererseits scheint es aber auch wünschenswert, gegebenenfalls sämtliche zu einem Lexem gehörigen bzw. von ihm abgeleiteten Wörter zu einem Lemma, einem "Lexemlemma" vereinigen zu können, alle Ableitungen mit einem bestimmten Ableitungsmorphem zusammenstellen zu können usw.; derartige Zusammenfassungen gehen zwar über die aktuellen Ziele der AL beträchtlich hinaus; doch sollten unseren Vorstellungen zufolge bereits bei der Anlage des Wörterbuchs möglichst viele Vorkehrungen getroffen werden, die in einem späteren Stadium der Arbeiten eine Untersuchung der deutschen Lexik nach verschiedenen morphologischen Aspekten erlauben. Voraussetzung dazu ist eine morphologische Zerlegung des Worteintrags bei der Kodierung. Es muß also in den Worteintrag ein Hinweis auf Morphemgrenzen aufgenommen werden; wir benutzen dazu den Bindestrich (Querstrich) "-", z.B. MISS-VER-STAEND-NIS, VER-GE-WALT-IGT, UN-WIDER-RUF-LICH. Eine Virgel impliziert stets einen Bindestrich, so daß bei Komposita nicht zusätzlich "-" gesetzt werden muß.

Die morphologische Zerlegung, die wir im Rahmen der AL vornehmen, muß aus mindestens zwei Gründen ein Provisorium bleiben. Erstens gibt es kaum für unsere Zwecke brauchbare Untersuchungen auf dem Gebiet der deutschen Morphologie (der historischen ausgenommen, aber die ist für uns nur in zweiter Hinsicht von Belang). Zweitens besteht schätzungsweise die Hälfte der Einträge des "Wahrig" aus teils überhaupt nicht, teils geringfügig ans morphologische System des Deutschen angeglichenen "Fremdwörtern". Diese Wörter, darunter sehr viele sehr gängige, können selbstverständlich nicht ausgeschlossen werden. Eine konsequente und systematische Zergliederung der Worteinträge müßte daher die morphologischen Regularitäten nicht nur des Deutschen, sondern auch der verschiedenen Quellen seiner Lexik mit berücksichtigen; die wichtigsten darunter sind Griechisch und Latein. Aber selbst damit ist es noch nicht getan, wenn man bedenkt, daß zahlreiche Wörter Mischformen aus verschiedenen Sprachen darstellen (AUTOMOBIL). Es ist klar, daß diese Arbeit nicht im Rahmen der AL geleistet werden kann; Ziel der morphologischen Zerlegung kann es daher nur sein, ohne Anspruch auf eine durchgehende Systematik und Vollständigkeit das Material für eingehende Untersuchungen der Morphologie des deutschen Wortschatzes aufzubereiten.

Bei der Abtrennung der Morpheme ist primär darauf zu achten, daß sich auf diese Weise sinnvolle Zusammenstellungen bilden lassen, z.B. Zusammenstellungen aller Wörter, die vom Lexem KOMM abgeleitet sind bzw. es in einem ihrer Kompositionselemente enthalten oder Zusammenstellungen aller Adjektive, die auf -LICH enden usw. Diese vage Orientierungshilfe wirft natürlich in der Praxis zahlreiche Schwierigkeiten auf, *von* denen im folgenden einige an Beispielen erläutert werden. Da die Probleme bei Wörtern deutscher Provenienz etwas anders sind als bei andern, trennen wir die Beispiele:

(4) Bei deutschen Wörtern ist vor allem nicht klar, bis zu welchem Ausmaß Wortbildungselemente abzutrennen sind:

(i) DACH/DECK-ER, SCHNEID-ER, MAL-ER, HAENDL-ER sind sicher zu zerlegen (vgl. DECKEN, SCHNEIDEN, MALEN, HANDELN); ist ER auch in NACHT/FALT-ER, SCHWYZ-ER, SCHWITZ-ER, SECHS-ER abzutrennen? Wie steht es schließlich mit dem gleich konstruierten MUELLER, bei dem Abtrennung des ER augenscheinlich unsinnig wäre?

(ii) Man betrachte die Reihe GE-BAEU-DE, GE-MEIN-DE, GE-TREI-DE, (die) KUN-DE. Bei GE-BAEU-DE ist die Abtrennung sicherlich sinnvoll, bei GETREIDE und KUNDE wohl kaum (es sei daran erinnert, daß es hier nicht um historische Morphologie geht.)

(iii) Ähnliches gilt für eine ganze Reihe von Wortbildungsmorphemen; man vergleiche etwa GEWINST, DIENST, GESPINST, KUNST, BRUNST. Welche nicbtnistorischen Gründe kann man anführen, ST bei GE-WINST, nicht aber z.B. bei OBERST abzutrennen, da weder bei dem einen noch bei dem andern die (unterschiedliche) funktionale Bedeutung des ST noch empfunden wird? Und soll man schließlich bei dem Verb GEWINNEN eine Abtrennung GE-WIN-N-EN vornehmen, weil sonst die Abtrennung von ST bei GEWINST sinnlos bleibt? Nur dann nämlich können sie aufeinander bezogen werden.

(iv) Soll man 0-MA, O-PA, MA-MA, PA-PA zerlegen? Und weshalb nicht?

(5) Bei Fremdwörtern liegt die Schwierigkeit vor allem darin, daß es kein Kriterium dafür gibt, wie "tief" die Zerlegung zu gehen hat:

(i) SITUATION: SITU-ATION, SITUA-TION, SITU-A-T-I-O-N u.a.

(ii) DELEATUR: DELE-ATUR, DELE-A-TUR, DELE-A-T-U-R, DEL-E-A-T-U-R?

(iii) Soll man DIFFUS, DIFFUSION, DIFFUNDIEREN so zerlegen, daß sie

aufeinander bezogen werden können? Dann müßte man konsequenterweise DIFFU-S, DIFFU-SIOM, DIFFU-NDIEREN (oder DIFFU-ND-IEREN) ansetzen. Ganz ähnlich liegt der Fall etwa bei CEMBAL-0, CEMBAL-IST, VIOLIN-E, VIOLIN-IST, PIAN-0, PIAN-IST usw. usw.

Die Liste problematischer Fälle ließe sich endlos fortsetzen; in der Tat sind einfache Fälle, die keinerlei Schwierigkeiten aufwerfen, schon fast Ausnahmen. Die angeführten Beispiele machen aber wohl schon hinlänglich klar, weshalb wir vorerst lediglich die oben genannten, relativ bescheidenen Ziele verfolgen.

2.1.3.3 Asterisk {"*"}

Die bei vielen deutschen Komposita auftretenden Fugenzeichen wie z.B. EN in SCHWANENGESANG, S in ENGELSGEDULD werden durch einen Asterisk markiert. Als Fugenzeichen gelten die dem Asterisk folgenden Grapheme bis zum nächsten Querstrich oder bis zur nächsten Virgel: SCHWAN*EN/GE-SANG, ENGEL*S/GE-DULD. Bei Fremdwörtern werden in der Regel keine Fugenzeichen berücksichtigt.

2.1.3.4 Dollarzeichen ("$")

Im "Wahrig" finden sich einige wenige Wörter mit diakritischen Zeichen, etwa ABBÉ'. Diese Fälle sind jedoch so selten belegt, daß es zu aufwendig wäre, für alle möglichen Diakritika spezielle Konventionen vorzusehen. Deshalb werden entsprechende Wörter lediglich durch Setzen eines Dollar-Zeichens in der letzten Spalte des Worteintrags markiert. Selbst dieser Hinweis wäre im Grunde nur bei den überaus seltenen Fällen notwendig, in denen das Fehlen des diakritischen Zeichens eine Mehrdeutigkeit zur Folge hätte: ROSE - ROSÉ".

2.1.3.5 Plus bzw. doppeltes Plus ("+") und ("++")

Eine vergleichsweise wichtige orthographische Komplikation hingegen ist der ß-ss-Wechsel in vielen deutschen Wörtern. Da bei den üblichen Fernschreiber-, bzw. Kartenlochertastaturen kein "ß" vorgesehen ist, entstehen bei der linguistischen Datenverarbeitung künstliche Homographen wie z.B. MASSE - MAßE. Ein unmarkiertes SS im Worteintrag kann daher dreierlei bedeuten:

1. ss 2. ß 3. ss wechselt mit ß bei Flexion.

Um diese Fälle auseinanderzuhalten, werden zwei diakritische Zeichen ver-

wendet: "+" und "++". Sie werden nach folgender Konvention gesetzt:

(6) (i) keine Markierung: ss; Beispiel: MASSE, HAUSSE.

(ii) + nach ss: ß; Beispiel: MASS+, FL0SS+, GR0SS+.

(iii) ++ nach ss: Wechsel von ß und ss im Paradigma; Beispiele: FASS++, FRESS++, KRASS++. Dabei gilt folgende Regel: SS++ ist in normaler Orthographie als ß realisiert vor 0, t, n, r, /; in allen anderen Fällen steht in normaler Orthographie ss (Kriterien wie "silbenschließend" o.ä. sind naturgemäß für unsere Zwecke nicht brauchbar).

2.1.3.6 Kurze Zusammenfassung der Sonderinformationen beim Wortei ntrag

Zeichen	Name	Bedeutung
/	Virgel	rechts folgt "Grundwort"
-	Bindestrich	Morphemgrenze
*	Asterisk	Fugenzeichen
$	Dollar	diakritisches Zeichen im Wort
+	Plus	ss ist als ß zu interpretieren
++	doppeltes Plus	ss und ß wechseln im Paradigma

2.2 Subkategorisierung

Unter Subkategorisierung verstehen wir im folgenden die Charakterisierung einzelner Einträge nach bestimmten syntaktisch-semantischen Merkmalen wie etwa /+abstrakt/, /-belebt/ usw. Diese Merkmale können sich direkt auf den betreffenden Eintrag beziehen (wir reden dann von inhärenten Merkmalen), aber auch auf die Kookkurenz eines Wortes mit einem andern, dem diese Merkmale zu- oder abgesprochen werden (dann handelt es sich um Selektionsbeschränkungen der Art /+___/+abstrakt// usw. Strikte Subkategorisierung figuriert bei uns unter "Rektion"; sie wird jedoch der Systematik wegen in diesem Abschnitt mitdiskutiert.

2.2.1 Motivation

Eine derartige Subkategorisierung ist für eine AL aus zweierlei Gründen von Belang.

Erstens kann sie dazu beitragen, sinnvolle Lemmata zu konzipieren. Définit ionsgemäß setzt die Zugehörigkeit mehrerer Flexionsformen zu einem

Lemma Bedeutungsgleichheit voraus. Wir verfügen nicht über eine semantische Beschreibungssprache, die es bei zwei gegebenen Wortformen unabhängig von der Beurteilung durch einen kompetenten Sprecher und unabhängig von formalen Übereinstimmungen (gleiche Graphemfolge) zu entscheiden erlaubte, ob sie zum gleichen Lemma gehören. Eine derartige Beschreibungssprache zu entwickeln, ginge über das Projekt der AL weit hinaus. Wir können daher nur auf die Übereinstimmung in der Graphemfolge (und natürlich der paradigmatischen Merkmale) rekurrieren; das ist natürlich in vielen Fällen unbefriedigend. So bezeichnen, um ein weiter oben in anderem Zusammenhang schon einmal angeführtes Beispiel wieder aufzunehmen, einige Komposita des Wortes SCHLANGE wie etwa HEIZSCHLANGE, PAPIERSCHLANGE keine belebten Wesen. Die Annahme eines Merkmals /+belebt/ könnte diesem Umstand Rechnung tragen; dabei ist es zunächst ganz gleichgültig, ob man für diese Komposita ein eigenes Lemma ansetzt oder ob man sie mit Komposita wie KLAPPERSCHLANGE, GIFTSCHLANGE usw. unter ein Grundformenlemma SCHLANGE zusammenfaßt und innerhalb dieses Lemmas als besondere Bildung markiert. Jedenfalls wären derartige Merkmale ein geeignetes Mittel, die oft unzulängliche graphematische Kongruenz durch zumindest partiellen Nachweis der Synonymie - genauer: Nachweis, daß eine bestimmte Heteronymie nicht vorliegt - zu ergänzen. Freilich bleibt auch innerhalb der durch gleiche Merkmalkombinationen gebildeten Subklassen ein erheblicher Spielraum; aber die Wahrscheinlichkeit der Heteronymie zweier Einträge bei gleicher Graphemfolge, gleicher Subkategorisierung und natürlich gleichen paradigmatischen Angaben ist doch relativ gering; sie wird umso geringer, je zahlreicher und je angemessener die Merkmale sind. Dies gilt im übrigen unabhängig davon, wie weit man den Begriff der "Synonymie" faßt; ganz gleich, wie weit man die Bedeutungsunterscheidungen treiben will, geht es hier darum, gemäß einer bestimmten, wie immer festgelegten Synonymiekonzeption festzustellen, und zwar im Prinzip automatisch festzustellen, ob zwei Wörter synonym sind.

Aus Gründen, die weiter unten deutlich werden, ist es jedoch nicht unproblematisch, bei der Bildung von Lemmata die Merkmale zu Hilfe zu nehmen. Von weitaus größerer Bedeutung sind die Merkmale aus einem anderen Grund: sie können zur Auflösung syntaktischer Mehrdeutigkeiten verwendet werden, wie sie im Deutschen vor allem im Zusammenhang mit der sogenannten

"freien Wortstellung" auftreten. Das soll im folgenden erläutert werden.[40]

In einem einfachen Satz wie

(7) DIE MUTTER SIEHT DAS KIND

können sowohl DIE MUTTER wie auch DAS KIND jeweils Subjekt oder Objekt sein. Es gibt bei der automatischen syntaktischen Analyse keinerlei Möglichkeit, diese Mehrdeutigkeit zu entscheiden. Zweifellos neigt ein durchschnittlicher Sprecher dazu, DIE MUTTER als Subjekt anzusehen, das heißt keine Umordnung vorzunehmen. Chomsky stellt dazu ganz allgemein die recht plausible These auf, derartige Umordnungen seien nur erlaubt "bis zur Grenze der Doppeldeutigkeit, das heißt, bis zu jenem Punkt, an dem eine Struktur entsteht, die durch die grammatischen Regeln auch unabhängig auf anderem Weg hätte erzeugt werden können." Der oben angeführte Satz wäre demnach überhaupt nicht mehrdeutig; doch ist dies möglicherweise eine Frage der Performanz, denn daß die Kompetenz im Prinzip Mehrdeutigkeiten toleriert, steht außer Frage; es ist aber möglich, daß sie einfach vermieden werden, ähnlich wie im Prinzip nicht 100 Genitivattribute nebeneinanderstehen, obwohl das gleichfalls durchaus möglich wäre. Für die syntaktische Analyse spielt es allerdings keine Rolle, aus welchen Gründen eine bestimmte Konstruktion nicht vorkommt.

Man kann den von Chomsky geäußerten Gedanken provisorisch als Prinzip für die Analyse etwa so formulieren:

(8) Von zwei möglichen Strukturen bei Mehrdeutigkeit wird stets die einfachste, d.h. jene, die mit den wenigsten Regeln auskommt, gewählt.

In dieser Form läßt sich das Prinzip sicher nicht halten, ganz abgesehen davon, daß nicht ganz einfach zu bestimmen ist, was man als **einfachste** Struktur anzusehen hat. Wahrscheinlich müssen die Regeln in irgendeiner Form gewichtet werden; wir können diesen Punkt hier aber nicht weiter verfolgen.

Prinzip (8) ist, entsprechend verfeinert, nur dann anzuwenden, wenn es sonst keinerlei Kriterium gibt, die Mehrdeutigkeit aufzulösen, wie es etwa in Satz (7) der Fall ist. Etwas anders ist die Lage nun in Sätzen **wie**

(9) DAS HAUS SIEHT DAS KIND

die auf keinen Fall mehrdeutig sind: DAS HAUS muß Objekt und DAS KIND Subjekt sein, denn SEHEN enthält eine Selektionsbeschränkung /- **/-belebt/**

.../, das heißt es darf nicht mit unbelebtem Subjekt stehen. Man kann dann in die syntaktische Analyse die folgende sehr einfache Teilstrategie einbauen:

(10) (i) Weist die erste nominale Gruppe ein Merkmal /+belebt/ auf? Wenn ja, ist sie Subjekt.(Wir akzeptieren hier also das in (8) formulierte Prinzip.)

(ii) Wenn nein: Ist die zweite nominale Gruppe belebt? Wenn ja, ist sie Subjekt,

(iii) Wenn nein: der Satz ist abweichend. Nach Prinzip (8) ist dann die erste nominale Gruppe Subjekt.

Der Gedanke, nach dem das Merkmal /+belebt/ zur Auflösung dieser Mehrdeutigkeit eingesetzt werden kann, ist also sehr einfach:

(11) Sind beide nominale Gruppen gleich spezifiziert, dann ist die erste (eventuell metaphorisches) Subjekt. Sind sie ungleich spezifiziert, dann ist die positiv spezifizierte Subjekt.

Ähnliche Prinzipien lassen sich auch für andere Merkmale leicht formulieren.

Es hat also den Anschein, als ließen sich derartige Selektionsmerkmale leicht in einen Analysealgorithmus integrieren. Nun gibt es jedoch zahlreiche Fälle, in denen die Wohlgeformtheit des Satzes, die ja Voraussetzung für irgendwelche Analysevorschriften sein muß, von Merkmalen abhängt, die sich mit den bisherigen Methoden nicht oder nur sehr umständlich erfassen lassen. Es gibt z.B. keinen Zweifel, wie der Satz

(12) DAS GRAS FRISST DIE KUH

zu analysieren ist. Dem kann aber im Parser nicht dadurch Rechnung getragen werden, daß man das lexikalische Formativ FRESS mit einem Merkmal /- /-belebt/.../ versieht; das Objekt kann ja gleichfalls belebt sein. Man- denke an Fälle wie z.B.:

(13) DEN SPERLING FRASS DIE KATZE

Dieser Satz kann jedoch aufgrund morphologischer Kriterien entscheidbar sein: DEN SPERLING ist nicht mehrdeutig, sodaß es ohnehin nicht nötig ist, auf Selektionsbeschränkungen zu rekurrieren. Anders steht es mit dem vielleicht etwas ungewöhnlichen, aber sicher doch nicht ungrammatischen Satz

(14) DAS KANINCHEN FRASS DIE SCHLANGE

FRESSEN hat hier offenbar eine zusätzliche Spezifikation, die sich auf eine Relation zwischen Subjekt und Objekt bezieht. Doch liegt der entscheidende Unterschied zwischen "freßbar von" und "belebt" nicht darin, daß ersteres eine Relation, letzteres ein (einstelliges) Prädikat ist; das ist vielmehr nur eine zusätzliche Komplikation. Dasselbe Problem tritt auch im folgenden Satz auf:

(15) DIE SCHLANGE FRISST DIE VORUEBERLEGUNG

Der Satz ist deshalb abweichend, weil - ganz unabhängig vom Subjekt - FRESSEN und VORUEBERLEGUNG nie kompatibel sind. Objekt zu FRESSEN muß banalerweise stets etwas Freßbares sein. Man müßte also ein Verb wie /-___/ -freßbar// beim Verb FRESSEN annehmen. Mit anderen Worten: auch die Verben müßten nach inhärenten Merkmalen subkategorisiert werden, denn ein Nomen mit dem Merkmal /+freßbar/ versehen, heißt ja nichts anderes, als es nach einem inhärenten Merkmal eines Verbs kennzeichnen; statt /+freßbar/ könnte man auch ein kontextuelles "Merkmal" /+ /FRESS/___/ beim Nomen ansetzen; dies gilt auch für sehr viele andere Verben wie TRINKEN, HINRICHTEN, VERSPOTTEN, SINGEN, überhaupt für die meisten Verben. Würde das System der Selektionsbeschränkungen schon dadurch ungeheuer kompliziert, so wird es durch die Einführung von Relationen, wie sie in manchen Fällen ja erforderlich sind, vollends unbrauchbar für die maschinelle syntaktische Analyse; denn es ist wohl kaum möglich, etwa bei jedem Substantiv anzugeben, was es alles fressen kann, sofern es überhaupt etwas fressen kann, bzw. von wem es gefressen werden kann, wenn überhaupt, usw.

Damit entfällt auch die Möglichkeit, für die beiden gegebenen Sätze

(16) (i) DIE SCHLANGE FRISST DAS KANINCHEN

(ii) DAS KANINCHEN FRISST DIE SCHLANGE

die richtige Struktur aufzufinden - obgleich in beiden Fällen nur jeweils eine Analyse korrekt ist. Dies gilt auch für komplexere Fälle, die über eine derartige stellungsbedingte Mehrdeutigkeit hinausgehen. So ist etwa der Satz

(17) DIE MAURER ASSEN IHRE BROTE UND BAUTEN

u.a. nicht nur im Hinblick auf Subjekt und Objekt mehrdeutig - diese Mehrdeutigkeit kann mit Hilfe des Merkmals Abelebt/ gelöst werden - sondern das Wort BAUTEN kann darin Verb und Akkusativobjekt sein. Diese Mehrdeutigkeit ist nicht lösbar, obwohl klar ist, wie sie hier interpretiert

werden muß.

Daß diese Fälle nicht lösbar sind, besagt keineswegs, daß die Verwendung von Merkmalen zur Auflösung syntaktischer Mehrdeutigkeiten insgesamt zu verwerfen sei. Wie wir gesehen haben, gibt es immerhin eine Reihe von Fällen, die auf diese Weise gelöst werden können.

Sehr störend wirkt sich allerdings auch in diesen Fällen aus, daß gegen Selektionsbeschränkungen relativ leicht verstoßen werden kann; wir sind bereits im Zusammenhang mit Prinzip (8) kurz darauf eingegangen.

Die Mehrdeutigkeit in dem Satz

(9) DAS HAUS SIEHT DAS KIND

ist lösbar, weil Verben des Sehens im allgemeinen als /-/-belebt/...___/ spezifiziert sind. (Es soll in diesem Zusammenhang keine Rolle spielen, ob gerade dieses Merkmal zutrifft oder ob man vielleicht besser /-/+abstrakt/...___/ ansetzen würde). Demnach ist der Satz

(18) AUS DEN FENSTERHOEHLEN STARRT DAS GRAUEN

abweichend. Man versteht ihn analog zu Sätzen wie

(19) AUS DEN FENSTERHOEHLEN STARREN KINDER

Einem normalen Sprecher bereitet es keinerlei Schwierigkeiten, Satz (18) richtig zu interpretieren. Er ist auch, obwohl er abweicht, anscheinend durch einen Parser vollständig, das heißt mit Einschluß der Abweichung und ihrer Kennzeichnung als solcher, analysierbar. Dabei ergeben sich aufgrund der Abweichung jeweils mehrere Möglichkeiten, die einander teilweise ausschließen. Die nominale Gruppe könnte ihrer Form nach Subjekt oder Akkusativobjekt sein. Gegen die Bestimmung als Akkusativobjekt spricht zweierlei:

(20) (i) STARREN weist u.a. ein Merkmal /-___NP/ auf, das heißt es ist als intransitives Verb subkategorisiert.
(ii) Im Satz sind keine weiteren Elemente vorhanden, die als Subjekt analysierbar wären. Mithin wäre gegen die fundamentale Regel verstoßen, daß ein Satz ein Subjekt haben muß.

Gegen die Bestimmung als Subjekt spricht lediglich ein Grund:

(21) (i) Das Verb erlaubt seiner Subkategorisierung nach kein nichtbelebtes Subjekt, während das Nomen GRAUEN als nicht belebt spezifiziert ist.

Da es offenkundig keinerlei Zweifel gibt, wie der Satz zu analysieren

ist, muß im Parser eine Möglichkeit vorgesehen sein, zwischen den beiden Regelverstößen (20) und (21) zu entscheiden. Es liegt sehr nahe, die Regeln hierarchisch zu ordnen und in Konfliktsituationen stets zugunsten der höherstehenden Regeln zu entscheiden. Diesem Verfahren liegt also die Annahme zugrunde, daß ein Verstoß gegen eine niedrigerstehende Regel eher zu erwarten ist als gegen eine höhere. Es ist beispielsweise eher anzunehmen, daß gegen eine Selektionsbeschränkung verstoßen wird als daß der Satz kein Subjekt enthält.

Es ist praktisch natürlich sehr schwierig und ohne Rekurs auf ein bestimmtes Modell auch gar nicht möglich, die Regeln hierarchisch zu ordnen. Aber ganz abgesehen davon ergeben sich auch bei dem Prinzip selbst gewisse Komplikationen. Der Satz

(22) IMMER ZUERST DIE EIGENEN FEHLER SEHEN!

besitzt eine eindeutige Struktur, analog etwa dem Satz

(23) MAN SOLL IMMER ZUERST DIE EIGENEN FEHLER SEHEN!

Wir haben also eine ähnliche Konfliktsituation wie im obigen Fall, mit dem Unterschied jedoch, daß sich hier nur zwei Regeln eindeutig gegenüberstehen:

(24) Die Regel, daß ein Satz normalerweise ein Subjekt zu enthalten hat, spräche dafür, FEHLER als Subjekt zu interpretieren.

(25) Das Merkmal /-/-belebt/...___/ bei SEHEN spräche gegen die Lösung als Subjekt.

Eine einfache Hierarchisierung, wie sie oben angedeutet wurde, erbrächte also eine falsche Lösung.

Die Schwierigkeit liegt darin, daß sich SEHEN nicht eindeutig als transitiv oder intransitiv spezifizieren läßt (bzw. daß es zwei Verben SEHEN - ein transitives und ein intransitives - gibt). Man vergleiche etwa die beiden Sätze

(26) (i) KARL SIEHT DEN SCHALLDAEMPFER
(ii) KARL SIEHT OHNE BRILLE BESSER

In Satz (22) ist SEHEN transitiv; da dies aber aus dem Satz nicht festzustellen ist und Intransitivität auch ohne Verstoß gegen die Grammatikalität möglich wäre, führt die einfache Hierarchisierung zur falschen Lösung. Sie genügt also nicht, allein schon deshalb nicht, weil ja unter Umständen mehrere Regelverstöße unterschiedlicher Schwere einander

gegenüberstehen können. Die Regeln müssen also in irgendeiner Weise zusätzlich gewichtet und kombiniert werden. Chomsky macht z.B. darauf aufmerksam,[42] daß das Merkmal /+menschlich/ in manchen Fällen von größerer Relevanz für die syntaktische Struktur ist als das hierarchisch höherstehende Merkmal /+abstrakt/. Das läßt darauf schließen, daß die automatische syntaktische Analyse in bestimmten Fällen zumindest auf eine bestimmte Kombinatorik zurückgreifen muß, die die Priorität determiniert. Beispielsweise können die Regeln (oder Merkmale) 1,2,3,... höherstehen als ...6, 7,..., die Kombination 6+7 aber höher *zu* bewerten sein als die Kombination 2+3 usw.

Dem ersten Vorschlag, einfach von einer Hierarchie der Regeln auszugehen, lag, wie schon angedeutet wurde, der Gedanke zugrunde, daß von zwei Möglichkeiten, einem Satz eine Struktur zuzuschreiben, stets die zu wählen ist, die am wenigsten abweicht. Dieser Gedanke erscheint dadurch gerechtfertigt, daß bei Sätzen, die nicht der Grammatik gemäß interpretierbar sind, sondern nur in Analogie zu wohl geformten, stets das nächstliegende Analogon gesucht wird; man versucht gleichsam, ein Maximum *an* syntaktischer Struktur zu finden.

Das oben angeführte Beispiel zeigt, daß diese Annahme, so plausibel sie scheint, nicht immer zutrifft, daß also unter zwei möglichen Strukturen auch die stärker abweichende vom Sprecher intendiert und vom Hörer verstanden werden kann. Denkbar wäre es natürlich auch, daß der Grad der Abweichung selbst von den gewichteten, kombinatorischen Regelverstößen abhängt. Das spielt jedoch für die maschinelle syntaktische Analyse keine Rolle.

Es sei noch erwähnt, daß eine derartige, durch Kombinatorik und Gewichtung erweiterte Hierarchie dann durch Prinzipien wie (8) ergänzt werden kann, die noch unterhalb der untersten Regelverstöße stehen - wie es den Anschein hat - und also erst eingesetzt werden, wenn alle anderen Möglichkeiten, dem Satz eine syntaktische Struktur gemäß den Regeln der Grammatik zuzuordnen, versagen.

2.2.2 Ein aufgegebenes Merkmalsystem

Der Wert, den die Subkategorisierung für die AL hat, hängt naturgemäß in erster Linie davon ab, wie die Merkmale angesetzt werden. Es wäre daher erforderlich, ein möglichst systematisches und umfassendes System, das

gleichzeitig für den Kodierer leicht anwendbar ist, zu entwickeln. Diese Entwicklungsarbeit ginge über den Rahmen unseres Projektes jedoch weit hinaus; wir mußten uns daher mit einer weniger auktionierten Lösung begnügen; nach sehr langen Vorüberlegungen haben wir uns entschieden, die • fünf von Chomsky in den "Aspects" vorgeschlagenen Merkmale zu verwenden, und zwar zur nominalen wie zur verbalen Subkategorisierung - bei letzterer in Form von Selektionsbeschränkungen. Sie wurden um drei Merkmale ergänzt, die sich nach unseren Erfahrungen als für die syntaktische Analyse besonders brauchbar erweisen konnten. Insgesamt wurden die folgenden acht Merkmale kodiert:

(27) (i) /+Klassenname/
(ii) /+abzahlbar/
(iii) /+abstrakt/
(iv) /+belebt/
(v) /+menschlich/
(vi) /+Zeitangabe/
(vii) /+Ortsangabe/
(viii) /+Namensanschluß möglich/

Die Merkmale "Ortsangabe" und "Zeitangabe" sollten dazu dienen, akkusativische Adverbialkonstruktionen von Akkusativobjekten zu unterscheiden. Man vergleiche etwa J. Seilers bekanntes Beispiel

(28) (i) KARL ISST DEN GANZEN KAESE
(ii) KARL ISST DEN GANZEM TAG

oder das Pseudoobjekt in dem Satz

(29) KARL FAELLT DIE TREPPE HERUNTER

usw.

Das Merkmal "Namensanschluß möglich" bezieht sich auf Einträge wie HERR, FRAU, DOKTOR, die einen spezifischen syntaktischen Anschluß von Namen erlauben (Namen sind durch /-Klassenname/ markiert):

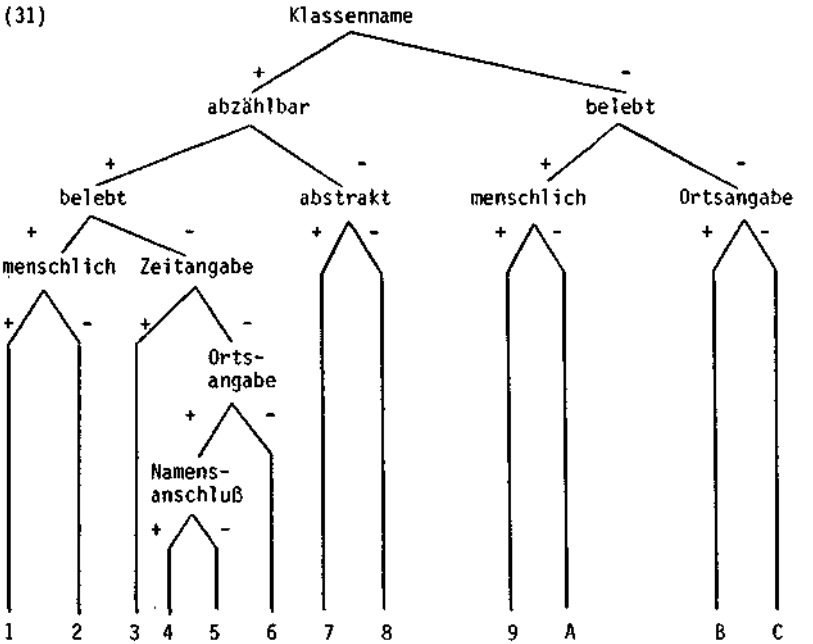
(30) (i) DORT KOMMT PROFESSOR SCHMALZLEDER
(ii)*D0RT STEHT DER STUHL THEODOR

(Es versteht sich, daß innerhalb der Namen noch einmal Restriktionen bestehen; vgl. etwa *FRAEULEIN EMIL; das spricht aber nicht dagegen, zumindest diese hier zu erfassen).

Der einfachste Weg, diese Merkmale zu kodieren, bestünde darin, acht Spal-

ten in der Lochkarte dafür vorzusehen und - z.B. - positive Spezifikation durch "1", negative durch "2" o.ä. zu markieren; da jedoch das Vorhandensein bestimmter Merkmale vielfach das Vorhandensein oder Fehlen anderer bedingt, wäre diese Methode - zumal bei acht Merkmalen - sehr aufwendig und höchst redundant. Deshalb wurden die acht Merkmale zu einem binär verzweigenden Graphen zusammengefaßt, wie man sie aus der Theorie der Codes kennt (man denke etwa an die bekannte entsprechende Darstellung des Morsealphabetes). Jeder seiner Endpunkte bezeichnet eine durch einen Merkmalkomplex gekennzeichnete Subkategorie; insgesamt ergeben sich so noch zwölf derartiger Kategorien; aus technischen Gründen sind sie von 1 bis 9 und dann von A bis C durchgezählt.

(31)



Beispiele:

1: RUSSE, ARZT[43]	2	PFERD, AMOEBE, ENGEL	3: ABEND, MITTWOCH
4: SCHIFF	5	SAAL	6: DUEBEL, PRIMZAHL[44]
7: MITLEID, HASS	8	WASSER, MEHL	9: WOLFGANG
A: FIDO	B	ASSUR, CHINA, NIL	C: ASPIRIN

Diese weitgehend, wenn auch nicht ganz redundanzfreie Form der Darstellung konnte allerdings nur bei den Substantiven selbst, nicht bei den entsprechenden Selektionsbeschränkungen der Verben angewendet werden, da die Kontextrestriktionen meist nicht für Merkmalkombinationen, sondern für einzelne Merkmale gelten; so erfaßt die Selektionsbeschränkung /+ /+abzählbar/...___/ (lies: Subjekt muß abzählbar sein) sämtliche Klassen von 1 bis 6, weil in ihnen allen das Merkmal /+abzählbar/ vorkommt. Daher wurde beim Verb die andere, oben erwähnte Form der Darstellung verwendet. Mit diesem, wie an den Beispielen schon zu ersehen ist, stellenweise etwas zweifelhaften System wurden in einem umfangreichen Text etwa 1000 Substantive und eine größere Anzahl von Verben subkategorisiert [45]. Die Ergebnisse dieses Tests waren wenig befriedigend. Außer weniger wichtigen, weil leicht reparablen Faktoren (Vgl. Anm. 44) waren dafür im wesentlichen zwei Gründe maßgeblich:

(32) (i) Die Merkmale sind vollkommen unzulänglich definiert; eigentlich sind sie überhaupt nicht definiert, sondern sie werden nach der individuellen Interpretation der jeweiligen Bezeichnung durch den Kodierer verstanden und gesetzt; zu einem gewissen Ausmaß läßt sich diese Interpretation allerdings durch gruppeninterne Absprachen normieren.

(ii) Welches Merkmal einem Wort zukommt, ist stärker satzabhängig als zu vermuten war. Ähnlich wie ja die Bedeutung eines Wortes in Abhängigkeit vom Zusammenhang, in dem es steht, sehr stark schwanken kann, ist auch die Merkmalspezifikation äußerst variabel.

Diesen beiden theoretischen Problemen korrespondieren als praktische Schwierigkeiten:

(33) (i) Die Kodierung ist heterogen; sie schwankt erheblich in Abhängigkeit vom Kodierer, ja selbst beim selben Kodierer zu verschiedenen Zeiten.

(ii) In der Regel ist eine bestimmte lexikalische Interpretation eines Wortes vorherrschend; an andere erinnert man sich beim Kodieren eben noch; es ist aber unmöglich, sich spontan alle möglichen Verwendungen eines Wortes zu vergegenwärtigen; die Ansichten, ob eine bestimmte Verwendung eines Wortes noch "nor-

mal" ist, gehen zudem weit auseinander (Standardeinwand: "Aber man kann doch auch sagen...").

Diese Schwierigkeiten erwiesen sich als so gravierend, daß unproblematische Fälle schon bald die Ausnahmen waren. Allerdings ist es mit den einzelnen Merkmalen etwas unterschiedlich bestellt; so sind die Merkmale /+ belebt/, /+menschlich/ relativ einfach zu kodieren (relativ); genau das Gegenteil gilt z.B. für /+ abstrakt/; ungefähr in der Mitte hält sich etwa /+ Klassenname/, für das wir im folgenden stellvertretend einige Typen problematischer Belege anführen:

(34) (i) chemische Bezeichnungen: GOLD, ASPIRIN, TRINITROTOLUOL, POLYVINYLCHLORID, STYROPOR, DDT, ...

(ii) Abkürzungen: AA, DGBM, DGB, ...

(iii) Titel: RHEINSTAHLDIREKTOR, DOKTOR, BUNDESPOSTMINISTER (Titel stehen, wie Namen, oft ohne Artikel).

(iv) Krankheiten: ANGINA, TBC, FEMUROHERNIE, MENINGOENZEPHALITIS, ...

(v) Getränke: CHERRY, HIGHBALL, MARTELL, MARTINI, ...

(vi) Speisen: BAUERNOMELETT, ROQUEFORT, ...

(vii) Feste: MARTINI, LICHTMESS, HIMMELFAHRT (muß z.B. von der faktischen Himmelfahrt unterschieden werden), ...

(viii) Gestirne: VENUS, ORION, MOND (als Erdtrabant, im Gegensatz zum allgemeinen astronomischen Begriff MOND), ...

(ix) Tänze: RUMBA, CHAÇONNE, WALZER, ...

(x) Autos: PEUGEOT, MERCEDES, HERON, ...

(xi) Schiffsnamen: (die!) BISMARCK, ...

(Einige dieser Bezeichnungen wird man nicht von Anfang an ins Wörterbuch aufnehmen wollen; es ist aber damit zu rechnen, daß sie in bestimmten Texten auftauchen und daher früher oder später doch ins Wörterbuch zu übernehmen sind).

Es scheint prinzipiell zwei Wege zu geben, diesen Schwierigkeiten beizukommen; eine "maximale" Lösung sähe vor, ein umfassendes, stichhaltiges und zugleich praktikables, das heißt auf die Bedürfnisse des Kodierers zurechtgeschnittenes System syntaktisch-semantischer Merkmale zu entwikkeln; die "minimale" Lösung bestünde darin, die Subkategorisierung drastisch zu reduzieren und auf einige Merkmale zu beschränken, die bei der Auflösung von Mehrdeutigkeiten relevant werden, weitergehende Ambitionen

aber aufzugeben; auch dieses System müßte natürlich leicht in der Praxis anwendbar sein.

Von diesen beiden Möglichkeiten ist zweifellos nur die erste linguistisch befriedigend; wir haben uns trotzdem für die zweite entscheiden müssen, weil die andere im Rahmen der AL nicht zu realisieren ist. Nach sorgfältiger Prüfung und Abwägung der verschiedenen in Frage kommenden Merkmale haben wir uns für fünf entschieden, die für die Auflösung von Mehrdeutigkeiten erfahrungsgemäß besonders wichtig sind und die sich so formulieren lassen, daß der Kodierer spontan, ohne lange reflektieren zu müssen, seine Entscheidung treffen kann. Diese Merkmale, die sowohl beim Verb wie bei den Substantiven verwendet werden - mit dem Unterschied allerdings, daß beim Verb zusätzlich noch die neutrale Spezifikation "+" vorgesehen werden muß - werden im folgenden kurz erläutert.

Weiterhin wurde nunmehr auch eine ganz vorläufige Subkategorisierung für die Adjektive entwickelt; sie ist allerdings für die syntaktische Analyse von geringerer Bedeutung; auch sie wird anschließend vorgestellt.

Zum Technischen ist zu bemerken, daß sich bei nunmehr nur noch fünf Merkmalen für Substantiv und Verb die Nachteile der Kreuzklassifikation relativ stärker geltend machen, so daß die Darstellung durch einen binären Graphen zugunsten der einfachen Matrizendarstellung aufgegeben wurde; Verb, Substantiv (und auch Adjektiv) werden nunmehr in dieser Hinsicht gleich behandelt.

2.2.3 Subkategorisierungsmerkmale für Verben und Substantive

1. /+abstrakt/

Wie schon bemerkt, ist bei diesem Merkmal oft sehr schwer zu entscheiden, ob es einem Eintrag zugesprochen werden soll oder nicht; außerdem können fast alle Konkreta abstrakt verwendet werden; das Gegenteil gilt allerdings nicht oder zumindest nicht allgemein; jedenfalls bildet dieser Umstand eine zusätzliche Erschwernis für den Kodierer. Da auf das Merkmal aber nicht gut verzichtet werden kann, haben wir uns entschlossen, eine Arbeitsdefinition aufzustellen, die die "eigentlichen" Abstrakta von den abstrakt gebrauchten Konkreta einigermaßen zu unterscheiden erlaubt. Wir sehen als konkret - d.i. /-abstrakt/ - alle jene Wörter an, deren Designat mit einer Waage gewogen und mit einem Meßstab gemessen werden könnte

(prinzipiell natürlich; die Sonne ist konkret, obwohl es in der Praxis schwer sein dürfte, sie auf eine Waage zu praktizieren).

Mit dieser Nominaldefinition sind selbstverständlich nicht alle Schwierigkeiten aus der Welt geschafft, aber sie garantiert eine ziemlich einheitliche und spontane Interpretation aller Kodierer sowohl beim Substantiv wie beim Verb; damit sind auch sinnvolle Selektionsbeschränkungen gewährleistet. Eine ganz andere Frage ist natürlich, ob damit auch getroffen ist, was viele Linguisten meinen, wenn sie von "abstrakt" und "konkret" reden; sie ist für uns ziemlich uninteressant.

2. /+belebt/
Dieses Merkmal ist relativ unproblematisch, vor allem, wenn man berücksichtigt, daß damit "belebtes Wesen", nicht etwa "Teil eines belebten Wesens" oder dergleichen gemeint ist (ARM, LEBER usw. gelten also als nichtbelebt) und wenn man einige Grenzfälle generell entscheidet; so gelten z.B. Pflanzen allgemein als nichtbelebt.

3. /+menschlich/
Auch hier ist "menschliches Wesen" gemeint, nicht Teil eines Menschen, für einen Menschen typische Aktivität usw.; ansonsten ist auch dieses Merkmal unproblematisch; einige Schwierigkeiten gibt es allenfalls insofern, als die Klasse /-menschlich/ sowohl Tiere wie auch z.B. Gespenster, Engel, Götter umfaßt; zwischen dem Gott Horus und dem biblischen Gott wird, unabhängig von theologischen Querelen zu dieser Frage, kein Unterschied gemacht. Auch Wiedergänger, Werwölfe, Nixen, Heinzelmännchen, Trolle und die anderen Elementargeister gelten als /-menschlich/.

4. /+617_____/ (lies: + 617 voraus).
Ober das Merkmal /^abzahlbar/ ist vielfach schwer zu befinden. Ähnlich wie bei der Nominaldefinition von "konkret" wird daher normativ festgelegt, daß statt dessen darüber entschieden wird, ob vor dem betreffenden Eintrag die Angabe 617 stehen kann; man kann z.B. sagen "617 Stühle", nicht aber "617 Mitleide" o.ä. Eine Zusatzkonvention muß allerdings für verkürzte Maßangaben eingeführt werden; BIER ist z.B. nach normalem Verstand nicht abzählbar; aber es ist möglich, statt "617 GLAS BIER" verkürzt "617 BIER" zu sagen; es wird daher gefordert, daß das auf 617 folgende Wort im Plural steht, wenn es positiv spezifiziert werden soll.

Daraus geht hervor, daß Wörter, die keinen Plural bilden, stets negativ spezifiziert werden. Dies gilt aber nicht umgekehrt; bestimmte, normalerweise als nicht abzählbar gewertete Wörter können nämlich einen sogenannten Artenplural bilden; so kann das Wort SALZE verwendet werden, wenn man damit mehrere Salzarten meint. In diesem Fall wird /-617/ spezifiziert, obwohl es eventuell denkbar ist, daß nach dem Sprachgefühl bestimmter Kodierer auch die Form "617 SALZE" möglich ist. Diese Unklarheit ist in der Praxis jedoch nicht so gravierend, wie sie hier scheinen mag. Soweit es sich beurteilen läßt, scheint jedenfalls die "Auslegung" /+617__/ für /+abzahlbar/ eine relativ einheitliche Kodierung zu gewährleisten.

5. /^kollektiv/

Eine Anzahl von Wörtern wie z.B. REGIERUNG, WERKSTATT, FIRMA wird im Hinblick auf das Merkmal /+belebt/ systematisch mehrdeutig gebraucht; sie können nämlich wie belebte Kollektiva verwendet werden. Man vergleiche etwa Sätze wie

(35) (i) DER MECHANIKER HAT MEINEN WAGEN REPARIERT
(ii) DIE WERKSTATT HAT MEINEN WAGEN REPARIERT
(iii) MEIN WAGEN STEHT IN DER WERKSTATT

Die entsprechenden Wörter können also mit Verben stehen, die ein belebtes Subjekt verlangen, obwohl sie nach üblichem Verständnis keine belebten Wesen sind. Sie werden daher als /-belebt/ /+kollektiv/ spezifiziert.

Wir sind uns, um allen entsprechenden Einwänden zuvorzukommen, völlig darüber im klaren, daß dieses System von fünf Merkmalen nicht der Weisheit letzter Schluß ist. Zahlreiche weitere Merkmale und Versuche, Arbeitsdefinitionen zu finden, wurden ausgiebig diskutiert und an Beispielsätzen überprüft. Der gegenwärtige Vorschlag soll auch nicht mehr sein als ein Versuch, mit beschränkten Mitteln möglichst viele Informationen zur Auflösung von Mehrdeutigkeiten in der syntaktischen Analyse zu erfassen. Er stellt gleichsam ein praktikables Maximum innerhalb des Projektes AL dar. Nichts spricht dagegen, ihn auszubauen.

2.2.4 Subkategorisierungsmerkmale für Adjektive

Der Subkategorisierung von Adjektiven hat man bisher kaum große Aufmerksamkeit gezollt. Sie spielt auch im Rahmen unseres Projektes nur eine ganz marginale Rolle, da sie zur Auflösung syntaktischer Mehrdeutigkeiten allem

Anschein nach wenig beiträgt; allerdings ist diese Möglichkeit bis jetzt auch kaum diskutiert, geschweige denn eingehend untersucht worden. Eine solche Untersuchung ist im Rahmen der AL natürlich nicht möglich; wir haben uns aber doch - ohne sonderlich weitgehende Ambitionen - entschlossen, einige Merkmalklassen zu bilden, um durch Vergleich mit den Substantivmerkmalen später einige mehr oder minder experimentelle Kookkurrenz-Untersuchungen anstellen zu können - also um z.B. festzustellen, welche Adjektive zusammen mit als /+menschlich/ spezifizierten Substantiven stehen.

Die Klassifizierung folgt im wesentlichen den drei Merkmalen, ob das Adjektiv (1) eine menschliche Eigenschaft bezeichnet oder nicht, (2) sich auf eine von Menschen ausgeübte Tätigkeit oder deren Resultate bezieht oder nicht und (3), ob es sich um ein Farbadjektiv handelt oder nicht. Die Farbadjektive bilden eine eigene Klasse. Bei den beiden erstgenannten Merkmalen wird gelegentlich auch ein - besonders gängiger - übertragener Gebrauch verzeichnet; insgesamt werden so die folgenden Subklassen gebildet (met. = metaphorisch):

(36)	menschliche Eigenschaft	menschliche Tätigkeit	Beispiel
0	-	-	PHYSIKALISCH
1	+ (met. auch -)	-	FREUNDLICH
2	±	-	DICK, GROSS
3	+	+	ABGEFEIMT
4	- (met. auch +)	-	SAUER, KLAR, SPITZ
5	-	+	ABGEDROSCHEN, ABENDLAENDISCH
6	Farbadjektive		

Das Vorläufige und Ungenaue dieses Vorgehens fällt auf; ob es sich später in geeigneter Weise präzisieren, ausbauen und vervollständigen läßt, muß sich erweisen.

2.2.5 Strikte Subkategorisierung (Rektion) des Verbs

Angaben über die syntaktischen Einheiten, mit denen eine Verbform zusammen in einem Satz stehen kann oder muß, spielen in der traditionellen Grammatik unter Bezeichnungen wie "Valenz", "Wertigkeit", "Rektion", "government" eine große Rolle. Die Behandlung dieses Problemkomplexes, wie sie Chomsky 1965 unter dem Rubrum "Strikte Subkategorisierung" vorgeschlagen hat, ist nicht sehr befriedigend, vor allem nicht für Sprachen mit relativ stark ausgebildeten Kasussystemen, in denen sich die Fakten, die es zu beschreiben gilt, nicht auf einige einfache kategoriale Kookkurrenzen reduzieren lassen. Daher wurden in den letzten Jahren schon einige Alternativentwürfe im Rahmen der Transformationsgrammatik vorgelegt, etwa von Ch. Fillmore, J. Robinson, J. Anderson u.a. Inwieweit diese Ansätze für unsere Zwecke nutzbar zu machen sind, ist noch unklar. Da die syntaktische Analyse stets an der Oberflächenstruktur des Satzes ansetzen muß, können bei der Kodierung der "Rektion", wie wir, um den Zusammenhang mit vergleichbaren Erscheinungen bei Substantiv und Adjektiv ("Adjektivrektion, Substantivrektion") etwas zu betonen, sagen, nur die syntaktischen Verhältnisse, wie sie im aktuellen Satz qua Folge von Graphemen vorliegen, nicht aber irgendwelche tieferliegende Strukturen unmittelbar berücksichtigt werden. Jedes Verb wird daher seiner Rektion nach durch eine Folge von Symbolen gekennzeichnet, die für eine Folge von syntaktischen Kategorien stehen und fragmentarisch die Oberflächenstruktur eines Satzes repräsentieren. Die syntaktischen Kategorien und die entsprechenden Symbole sind

(37)

N	nominale Gruppe[46] im Nominativ
G	nominale Gruppe im Genitiv
D	nominale Gruppe im Dativ
A	nominale Gruppe im Akkusativ
K_i	nominale Gruppe im Dativ oder Akkusativ
P	präpositionale Gruppe
ADV	adverbiale Gruppe
N'	Gleichsetzungsnominativ
A'	Gleichsetzungsakkusativ
S	Satz[47]
V	Verb

Das Verb BEDUERFEN läßt sich dann in seiner Rektion etwa durch die Sequenz N V G, (DER PAPST BEDARF GOTTES BEISTANDS), das Verb TAUFEN durch die Sequenz N V A A' (DER PAPST TAUFTE IHN WLADIMIR) gekennzeichnet; TAUFEN kann allerdings, wie sehr viele Verben, auch andere Rektionen haben, etwa N V A (DER PAPST TAUFTE SEINEN ENKEL); durch die Angabe der verschiedenen möglichen Rektionen eines Verbs bei der Kodierung ist es möglich, verschiedene Bedeutungen bzw. Verwendungsweisen dieses Verbs zu unterscheiden. Man vergleiche etwa die folgenden vier Verwendungsweisen des Verbs STIMMEN:

(38) (i) DIE RECHNUNG STIMMT. N V

(ii) DIE RECHNUNG STIMMT MICH TRAURIG. N V A ADV

(iii) DER PAPST STIMMT DAS KLAVIER. N V A

(iv) DER PAPST STIMMT FUER DIE PILLE. N V P

Die Rektionsangaben können also unter Umständen zur Auflösung semantischer Mehrdeutigkeiten und damit zur Feingliederung von Lemmata beitragen; beispielsweise muß STIMMEN in irgendeinem zu analysierenden Satz in irgendeinem zu lemmatisierenden Text in der Bedeutung (i) gemeint sein, wenn die Analyse zeigt, daß dieser Satz nur eine nominale Gruppe enthält. Dieses Prinzip gilt allerdings nur, wenn der Satz syntaktisch eindeutig ist.

In dem Satz

(39) TRAURIG STIMMT PAPST PAUL DAS KLAVIER.

kann STIMMEN in Bedeutung (ii) "Erzeugung einer Gemütsbewegung" und in Bedeutung (iii) "Bewirken bestimmter akustischer Eigenschaften bei gleichzeitiger Depression" verwendet sein. Dasselbe gilt, um ein anderes Beispiel wieder aufzugreifen, für den Satz

(40) DER PAPST TAUFT SEINEN ENKEL WLADIMIR.

in dem über die semantische Mehrdeutigkeit von TAUFEN (mit oder ohne Namensgebung) nicht befunden werden kann, weil die syntaktische Mehrdeutigkeit nicht entscheidbar ist.[48]

In der Regel ändert sich die Rektion nicht, wenn die Anordnung der Elemente im Satz vertauscht wird: statt N V A kann durchweg auch A V N stehen (einfache Inversion). Bei der Auswertung der Rektionsangaben in der syntaktischen Analyse werden selbstverständlich diese möglichen Umstellungen gemäß den bekannten Wortstellungsregeln des Deutschen berücksichtigt, ebenso wie die wenigen Fälle, in denen sich bei Umstellung -

zumindest an der Oberfläche, mit der wir es hier zu tun haben - die Rektion ändert.

Man vergleiche etwa die Sätze

(41) (i) MICH GELUESTET NACH SCHOENEN FRAUEN.
(ii) NACH SCHOENEN FRAUEN GELUESTET MICH,
(iii) ES GELUESTET MICH NACH SCHOENEN FRAUEN.
(iv) NACH SCHOENEN FRAUEN GELUESTET ES MICH.

Die Angabe N V D A bei einem Verb, z.B. bei GEBEN, ist also eine Abkürzung für N V 0 A, N V A D, D V N A, D V A N, A V D N, A V N O sowie in Nebensätzen, bei denen das Verb ans Ende rUckt, N D A V, N A D V, D N A V, D A N V, A N D V und A D N V anzusehen, wobei natürlich eventuelle Restriktionen wie die in (41) beschriebene in Rechnung zu stellen und entsprechend zu formulieren sind. Da all diese Formen aber aus der einen Angabe N V A D abgeleitet werden, genügt bei der Kodierung diese eine repräsentative Sequenz.

2.3 Zur Homographie

Die Ergebnisse der AL hängen unter anderem wesentlich von der Beseitigung morphologischer und syntaktischer Mehrdeutigkeiten ab. Eine dieser Mehrdeutigkeiten ist die Homographie von Flexionsformen.[49]

Bei ihrer Behandlung sind zwei verschiedene Vorgänge zu unterscheiden:

(i) Das Erkennen nomographer Flexionsformen

(ii) Die Beseitigung nomographer Mehrdeutigkeiten.
Dies bedeutet praktisch die "richtige" Lemmatisierung einer homographen Flexionsform. So muß beispielsweise die Flexionsform LIEBE in dem Satz: MAN SPRICHT VON LIEBE dem Substantiv LIEBE und nicht dem Verb LIEBEN zugeordnet werden.

Die Beseitigung von Homographen ist hier nur insoweit von Belang, als durch ihr Auftreten die Lemmatisierungsergebnisse beeinflußt werden können. Die Auflösung homographer Mehrdeutigkeiten basiert auf syntaktischen Umgebungsanalysen, das heißt der syntaktische Kontext liefert Anhaltspunkte zur Auflösung homographer Flexionsformen. Dieses Verfahren wird schon längere Zeit bei der Saarbrücker maschinellen Syntaxanalyse angewandt.[50]

Von Interesse ist im obigen Zusammenhang das Erkennen homographer Fle-

xionsformen. Bisher wird zur Syntaxanalyse ein Flexionsformenbuch benutzt; es enthält bei jeder Flexionsform die vom Kodierer eingetragene Information, ob es sich um eine homographe Flexionsform handelt und wenn ja, die Homographenklassenangabe (z.B. FLOH = Ho-Klasse 4, Verb/Substantiv).[51]

Das Wörterbuch zur Lemmatisierung aber ist als lemmatisiertes Wörterbuch konzipiert. Deswegen sind hier von vornherein andere Verfahrensweisen zur Feststellung homographer Flexionsformen geboten. Prinzipiell bieten sich folgende Möglichkeiten an:

(iii) In einem mehrstufigen Verfahren läßt sich die Homographenangabe automatisch erzeugen:
   (a) Von jedem Lemma werden automatisch sämtliche Flexionsformen erzeugt und gespeichert.
   (b) Jede Flexionsform eines Lemmas wird mit sämtlichen Flexionsformen aller anderen Lemmata verglichen. Bei übereinstimmender Graphemfolge wird Homographie notiert.
   (c) Alle Homographen erhalten aus dem Informationsteil ihrer jeweiligen Wörterbucheinträge die genaue Homographenspezifikation. (Z.B. erhält SAGEN die Angabe: Substantiv, ...Dativ Plural / Verb, ... Infinitiv, 1. und 3. Person Indikativ und Konjunktiv Präsens).
   (d) Zusammenfassung sämtlicher gleichspezifizierter Homographen zu Homographenklassen.
   (e) Geeignete Integration der genau spezifizierten homographen Flexionsformen in das Wörterbuch.

(iv) Wie (iii) (a) und (b), jedoch mit dem Zusatz, daß jedes Lemma, das eine oder mehrere homographe Flexionsformen enthält, durch ein (unspezifiziertes) Homographenkennzeichen markiert wird.

(v) Die Homographeninformation wird nicht automatisch erzeugt; vielmehr wird im Rahmen der Zerlegungs- und Vergleichsprogramme als obligatorischer Schritt die Homographenfeststellung durchlaufen. Das bedeutet aber, daß jede Flexionsform eines zu lemmatisierenden Textes, ob homograph oder nicht, obligatorisch - durch den Vergleich mit den Flexionsformen der im Wörterbuch benachbarten Lemmata - der Homographenprüfung unterworfen wird.

Die Verfahren (iii) und (iv) sind nur dann sinnvoll, wenn das Wörter-

buch zur Lemmatisierung "stabil" ist, das heißt unverändert bleibt. Das Herausnehmen und Hinzufügen bestimmter Wörterbucheinträge müßte unterbleiben, da sonst die einmal festgesetzten Homographeninformationen beeinträchtigt werden könnten. Weiterhin ist noch nicht abzusehen, ob der mit (iii) und (iv) verbundene Rechen- und Speicheraufwand in einem vertretbaren Verhältnis zu den zu erwartenden Ergebnissen steht. Allerdings hätte (iii) den Vorzug, daß die spezifizierten Homographeninformationen direkt (für eine weitere Analyse) verfügbar wären. Die durch (iv) erzeugte Information würde darüber entscheiden, ob (v) durchgeführt wird oder nicht.

Der in (v) dargelegte Vorschlag hätte den Vorzug, daß der enorme Erzeugungs-, Speicherungs- und Klassifizierungsprozeß wegfallen würde. Dafür müßte dann aber jede einzelne Flexionsform im Rahmen der AL die Homographenprüfung durchlaufen'. Diese Prüfung ist relativ unaufwendig, weil nur in der engeren (graphematischen) Umgebung der jeweiligen Wörterbucheinträge mit Homographie zu rechnen ist. Nur ein kleiner, verhältnismäßig genau festlegbarer Bereich in der Umgebung der zu untersuchenden Flexionsform muß abgesucht werden.

Eine weitere Lösung wäre denkbar:

(vi) Auf der Grundlage eines größeren Textmaterials wird die Homographie nach (v) festgestellt; sodann werden die häufigsten homographen Flexionsformen aus dem Wörterbuch herausgenommen und mit einem relativ kleinen Hochfrequenzwörterbuch vereinigt, das vor dem Wörterbuch zur Lemmatisierung benutzt wird. Auf diese Weise kann der Prozeß zur Feststellung von Homographien erheblich beschleunigt werden.

Wie ersichtlich, ist (iv) eine Variante von (iii), während bei (v) der Prozeß von (iii) in die Analyseprogramme verlagert wird unter gleichzeitiger starker Begrenzung des Such- und Vergleichsbereichs. Gemeinsam ist den hier gemachten Vorschlägen, daß die Homographeninformation nicht von den Kodierern im Rahmen der Kodierungsvorschriften auf den Kodierungsblättern eingetragen werden muß; dies ist ein großer zeitlicher und sachlicher Gewinn: Denn die Kodierung würde, sollte der Kodierer sämtliche Flexionsformen eines jeden Lemmas auf Homographie hin prüfen, zeitlich außerordentlich belastet. Darüberhinaus würde eine auf diese Art zustandegekommene Homographeninformation sicher unvollständig sein, weil sie nur auf

der aktiven Sprachkompetenz eines einzelnen Kodierers beruhen würde: Wer denkt schon, wenn er das Adjektiv SCHLICHT bearbeitet, an die Verbform SCHLICHT (von SCHLEICHEN!)!

Es wird mit einiger Sicherheit der Vorschlag (vi) realisiert werden. Welche Modifikationen sich ergeben, müssen größere Experimente im Zusammenhang mit dem fertigen Wörterbuch zur Lemmatisierung ergeben. Zunächst wird dieses ohne Homographenkennzeichnung angelegt.

3 TECHNISCHE BEMERKUNGEN ZUR ERSTELLUNG DES MASCHINELLEN WÖRTERBUCHS

Die hier nicht näher beschriebenen Kodierblätter werden, wenn sie entsprechend ausgefüllt sind, auf Lochkarten übertragen. Für Adjektive und Substantive genügt eine Karte, für Verben müssen je zwei angelegt werden. Damit beide zusammengeordnet werden können, werden Worteintrag und Bedeutungsnummer aus der ersten in die zweite Verbkarte dupliziert. Bei allen drei Wortklassen sind darüberhinaus gegebenenfalls Folgekarten abzulochen.

Die Lochkarten bilden den Ausgangspunkt aller weiteren Arbeiten, auch eventueller Korrekturen und Ergänzungen; sie werden nach Abschluß der Locharbeiten zunächst mit Hilfe eines ersten Maschinenprogramms LEMWOBU auf Alphabetfolge sowie auf formale Fehler(d.h. daraufhin, ob sie womöglich irgendwelche im Kodierungsschlüssel überhaupt nicht vorgesehenen Lochungen enthalten)überprüft. Weit wichtiger sind natürlich die Korrekturen von "echten" Kodierungsfehlern, etwa falsche Flexionsangaben, falsche Angaben über Steigerungsfähigkeit usw. Dazu wurden bzw. werden in einem Programm AUSWLEM verschiedene Auswahlmöglichkeiten vorgesehen, so daß die Einträge nach kleinsten - auch kombinierten - Merkmalen (Rektion, Subkategorisierung usw.) zusammengestellt werden können. Darüber hinaus werden Generierungsprogramme (ADJGEN...) erstellt, die sämtliche oder einige typische aus den Angaben zum Worteintrag ableitbare Formen, z.B. einige Steigerungsformen, Negationspräfigierungen, Flexionsformen erzeugen; wenn etwa bei dem Wort MANN als Genitivendungskode versehentlich eine 1 statt einer 2 angegeben wäre, so würde die falsche Form MANNEN statt MANNES erzeugt (vgl. 3.2.1.); diese falsche Form ist leicht erkenn- und korrigierbar.

Ein Sortierprogramm, das alle graphematischen Einträge nach den verschiedensten Gesichtspunkten (Präfixe, Suffixe, Fugen, Morpheme, ...) al-

phabetisch ordnet, soll in der Korrekturphase die Vereinheitlichung der Abtrennungen erleichtern.

Nach Abschluß dieser Korrekturphase, die erfahrungsgemäß außerordentlich viel Aufwand erfordert, wird das eigentliche maschinelle Lexikon nach bestimmten Konventionen (auf Maqnetband/-platte) erstellt, besser gesagt: die maschinellen Lexika werden erstellt, denn es wird sich voraussichtlich um mehrere inhaltlich verschiedene Lexika handeln. Neben dem "eigentlichen" Lexikon, das sämtliche Einträge in alphabetischer Reihenfolge samt Informationen enthält, ist z.B. für die Verben mit abtrennbaren Verbzusätzen (VOR-GEHEN, AB-SCHLACHTEN, AUS-BALDOWERN) ein Speziallexikon erforderlich, das während der Analyse angesprochen werden kann; wenn etwa zunächst nur die beiden abgetrennten Einheiten ... GING ... VOR ... für sich analysiert wurden und es sich ergibt, daß sie nach Stellungskriterien eventuell zusammengehören (Man vergleiche etwa die beiden Sätze SIE GING FUENF MINUTEN VOR - SIE GING FUENF MINUTEN VOR ACHT!), muß dieses Lexikon zur Verifizierung herangezogen werden. Weiterhin wird aus ökonomischen Gründen (Ersparung von Rechenzeit) ein Häufigkeitswörterbuch angelegt, das die 60 bis 120 häufigsten Wörter des Deutschen umfaßt, die ja weit über die Hälfte der Wortformen eines laufenden Textes ausmachen. Sämtliche Lexika werden selbstverständlich so angelegt, daß sie jederzeit mit mehr oder weniger großem Aufwand erweitert werden können.

Anmerkungen

1 Der hier gekürzt wiedergegebene Bericht, der unter Mitwirkung von M. Bartoli, R. Dietrich, A. Rothkegel, H.J. Weber und H. Zimmermann erstellt wurde, ist im Juni 1971 als Arbeitsbericht Nr. 10 im Rahmen der Linguistischen **Arbeiten des Germanistischen Instituts und des Instituts** für **Angewandte Mathematik der Universität des Saarlandes in hektographierter Form erstveröffentlicht worden. Die Kürzungen erstrecken sich** vor **allem auf den zweiten Teil der Original Veröffentlichung, in dem detaillierte Kodierungsvorschriften für die einzelnen Wortartklassen abgebildet und kommentiert werden.**

2 Vgl. dazu auch: Rainer Rath, "Vorschläge zur automatischen Lemmatisierunq (AI) deutscher Adjektive". In: *Linguistische Berichte,* 12 (1971), 53-59. Ders., "Probleme der automatischen Lemmatisierung". In: *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung,* 5 (1971).

3 Die folgenden Definitionen erheben nicht den Anspruch besonderer Origi-

nalität. Sie versuchen, bestimmte, längst bekannte Tatsachen formal festzulegen. Sie beziehen sich nur auf geschriebene Sprache.

4 Vgl. z.B. Helmut Schanze/Hans Schwerte (eds.), *Indices zur neueren deutschen Literatur*, bisher erschienen Index 1-4, 1968-1970. Roy Wisbey, *A Complete Word-Index to the Speculum Ecclesiae* (1968). Siehe aber: Wolfgang Klein/Harald Zimmermann, *Index zu Georg Trakls Werken* (Frankfurt,1971) (= *Indices zur neueren deutschen Literartur,* Bd. 5). Es handelt sich hierbei um ein lemmatisiertes Wörterbuch. Die Lemmatisierungen und die Zuordnung der grammatischen Informationen sind teilweise automatisch, teilweise von Hand vorgenommen. Vgl. dazu das Vorwort des Bandes, in dem die Herausgeber über ihr Verfahren und dessen Schwierigkeiten berichten.

5 Unter syntaktischen Homographen verstehen wir solche Wortformen, "die verschiedenen Wortklassen zugewiesen werden" können, z.B. ALTER (Substantiv) und ALTER (Adjektiv). Vgl. dazu Hans Eggers et al., *Elektronische Syntaxanalyse der deutschen Gegenwartssprache* (Tübingen 1969) p. 62 ff. Semantische Homographen bestimmen wir als Wortformen, die - innerhalb einer Wortklasse - verschiedene lexikalisierte Bedeutungen haben können. Nach unseren Definitionen wäre ein syntaktischer Homograph (mit Zweideutigkeit) als eine Flexionsform darzustellen, bei der $G_i = G_j$, $R_i =/ R_j$, $P_i =/ P_j$ ist. Bei einem semantischen Homographen gelten die folgenden beiden Fälle: (a) $G_i = G_j$, $P_i = P_j$, $S_i =/ S_j$; z.B. DAS SCHLOSS (an der Tür, auf dem Berg); (b) $G_i = G_j$, $P_i =/ P_j$, $S_i =/ S_j$; z.B. DER TOR, DAS TOR.

6 *Index zur deutschen Literatur* Nr. 1, a.a.O., D. VII; vgl. auch H. Schanze, "Computerunterstützte Literaturwissenschaft". In: *Muttersprache* 9/10 (1969), 315-321. - In den Indices Nr. 2 und 3 "werden die Einzelbelege jeweils nach ihrer Zugehörigkeit zu einer bestimmten Wortart aufgeschlüsselt." Die Mehrzahl der Homographen lasse sich durch "Klassifikation des Wortmaterials nach Wortarten trennen." (Index Nr. 2, S. V f.) Wie diese Klassifikation erfolgt ist - ob automatisch, halbautomatisch oder manuell durch Bearbeiter - wird nicht mitgeteilt.

7 *Index zur deutschen Literatur* Nr. 1, a.a.O., p. VII.

8 ebd. S. VIII "...ein präzises, jederzeit greifbares und nicht allzu kostspieliges Hilfsmittel für die literarische und sprachliche Analyse..."

9 Allerdings werden auch nach dem hier vorgeschlagenen Verfahren semantische Homographen nicht oder nur in Ausnahmefällen getrennt werden können. Dies liegt an der ungenügenden semantischen Klassifizierungsmöglichkeit des Wortschatzes. Vgl. dazu weiter unten.

10 Ähnlich H. Schanze, *Computerunterstützte Literaturwissenschaft,* a.a.O., p. 316: "In der Erkenntnis, daß literarische Phänomene immerzu den 'Wörtern' anhängen, können 'Wortindices' sinnvoll vom Literaturwissenschaftler gebraucht werden."

11 Syntagmatische Probleme, die sich bei bestimmten Zusammensetzungen ergeben, sind hier außer Acht gelassen: In der Flexionsform BAHNHOFSVORSTEHERS beispielsweise steht nach der entsprechenden Zerlegung innerhalb der Teilgraphemfolge G_{Si} (BAHNHOFSVORSTEHER) noch ein Graphem

(das sogenannte Fugen-s), das eine syntagmatische Information repräsentiert: Das Verhältnis der beiden Kompositionselemente zueinander. Dies bleibt hier unberücksichtigt.

12 Oft ist es so, daß die spezielle Kombination G_{Si}. und G_{Ri} eine paradigmatische Information liefert. Die Kombination z.B.: LIEB und ST repräsentiert die paradigmatische Information "Verb", während die Kombination LIEB und EM die paradigmatische Information "Adjektiv" repräsentiert.

13 Die Endqraphemfolqe -TE ist beispielsweise ein solcher Fall. Auf -TE enden (u.a.) WORTE, HOERTE, SCHOENSTE. Ohne Kenntnis der Wortklassen der zu lemmatisierenden Flexionsformen ist eine Lemmatisierung hier nicht möglich: Dem Substantiv muß zur Lemmatisierung ein -E abgestrichen werden, beim Verb wird das -TE durch -EN ersetzt, während beim Adjektiv schließlich die Graphemfolge -STE abgestrichen wird. Eine Lemmatisierung kann in diesen Fällen - und dies sind die meisten - nur in Kenntnis der paradigmatischen Information der Wortklasse erfolgen.

14 Gütersloh, 1968.

15 Das vorhandene, auf Magnetband gespeicherte syntaktische Wörterbuch für die Saarbrücker maschinelle syntaktische Analyse (Vgl. *Elektronische Syntaxanalyse*, a.a.O., p. 40 f. u. p. 55 ff.) wird nicht verwendet: Es handelt sich um ein Wortformenbuch, wir wollen künftig jedoch mit einem Grundformenbuch arbeiten. Außerdem ist es zu klein, und die grammatischen Angaben reichen für eine Lemmatisierung nicht aus.

16 Die sogenannten "starken Verben" erhalten zusätzliche Kennzeichnungen.

17 Eine Flexionsform kann, wie definiert, mehrere Wortformen umfassen. Der erste Schritt der AL befaßt sich nur mit Wortformen, und zwar deswegen, weil der Text zunächst automatisch nur in solche zerlegt werden kann. Ob nun die Wortform Teil einer Flexionsform oder selbst Flexionsform ist, wird mit Hilfe des Informationsteils des Wörterbuchs durch das syntaktische Analyseprogramm ermittelt.

18 Bei den Substantiven und Adjektiven tritt dieser Fall häufig auf: In vielen Substantivparadigmata haben die verschiedenen Kasus nur ein oder zwei Formen.

19 Es sei denn, es liegt ein Hinweis auf Homographie vor.

20 Zur Illustration ein weiteres Beispiel: Das Endgraphem -E tritt als paradigmatisch zu interpretierendes Graphem bei allen drei hier interessierenden Wortklassen auf: DIEBE, KLEINE, GEBE.
Aber auch innerhalb der einzelnen Wortklassen ist es als paradigmatisch interessantes Endgraphem mehrdeutig: DIEBE: Pluralendung; TASCHE: Teil des Lemmanamens; MUEDE: flektierte und unflektierte (Lemmaname) Form; KLEINE: nur flektierte Form.

21 Zu den Prinzipien der Beseitigung von Wortklassenmehrdeutigkeiten vgl. *Elektronische Syntaxanalyse*, a.a.O., p. 62 ff.

22 Sogenannte feste oder idiomatische Wendungen wie HOCH UND HEILIG, MIT LEIB UND SEELE werden ebenfalls als mehrteilige Flexionsformen behandelt.

23 Die trennbaren Verben können in Hauptsatzstellung zwar kontinuierlich stehen, jedoch muß bei der Lemmatisierung darauf geachtet werden, daß der Verbzusatz, der dem Verb folgt, vor das Verb gestellt wird, damit beispielsweise TRIFFT ... EIN dem Lemma EINTREFFEN zugeordnet werden kann.

24 Ein mögliches Analyseverfahren ist dargestellt in: *Elektronische Syntaxanalyse,* a.a.O., pp. 110-115. Vgl. auch: "Die automatische Behandlung diskontinuierlicher Konstituenten im Deutschen". In: *Muttersprache* 9/10 (1969).

25 Allein aufgrund der Kenntnis der Ableitungssilben (mit den syntagmatisch bedingten Veränderungsmöglichkeiten) können z.B. die folgenden Fälle nicht unterschieden und deswegen nicht lemmatisiert werden: BLUMENSAMEN (Substantiv), SELTSAMEN (Adjektiv); LUSTIGEN (Adjektiv), SCHWEIGEN (Substantiv).

26 Es bleibt natürlich zunächst offen, ob es sich bei dem künstlichen Homographen zufälligerweise um einen echten Homographen oder um eine eindeutige Wortform handelt.

27 Zur Kodierung von SEIN existieren eigene Kodierungsvorschriften.

28 Gewisse Schwierigkeiten ergeben sich dadurch, daß die Endgraphemfolge -STE bei Adjektiven nicht eindeutig ist. Vgl. z.B. FESTE, WUESTE; zu erwägen wäre hier, die wenigen Adjektive auf -ST im Positiv besonders zu kennzeichnen.

29 Nach einer privaten Mitteilung von Gerhard Wahrig sind es 90.000. Ein derzeit von ihm vorbereitetes Wörterbuch soll etwa 270.000 Einträge umfassen.

30 Vgl. dazu oben Anm. 15

31 Vgl. dazu *Elektronische Satzanalyse,* a.a.O., 31.

32 Siehe dazu: A. Rothkegel, "Funktionsverbgefüge als Gegenstand maschineller Sprachanalysen". In: *Beiträge zur Linguistik und Informationsverarbeitung* 17 (1969), 7-26.

33 Außer SEIN und TUN.

34 Auch für die anderen Bedeutungen gibt es natürlich Komposita, z.B. MAYASCHRIFT, KURSIVSCHRIFT usw.

35 Vgl. jedoch weiter unten die entsprechenden Ansätze im Rahmen der Subkategorisierung in Abschnitt 2.2

36 Vgl. dazu insbesondere: K. Bünting, *Morphologische Strukturen deutscher Wörter* (Hamburg,1970), der auch die ältere Literartur verzeichnet und diskutiert; Büntings Problemstellung ist allerdings eine andere als die, die wir verfolgen.

37 Vgl. dazu oben Abschnitt 1.1 (Lemmadefinition)

38 Eine derartige Beschreibungssprache der Semantik des Deutschen wird z.B. von Brockhaus, v. Stechow, "On Formal Semantics: A New Approach". In: *Linguistische Berichte* 11, 7-36, skizziert.

39 Ober Stellung und Bedeutung der syntaktischen Analyse im Rahmen der AL

vgl. 1.3.4.

40 Vgl. dazu: W. Klein, *Parsing*, (Frankfurt,1971), Kap. 4.4, wo diese Problematik ausführlich diskutiert wird; einige Formulierungen und Beispiele werden von dort übernommen.

41 Chomsky, "Aspects; deutsche Übersetzung". In: *Aspekte der Syntax-Theorie* (Frankfurt,1969), p. 163.

42 *Aspekte*, a.a.O., p. 190.

43 Bei 1 ist immer Namensanschluß möglich; ähnlich sind z.B. A, B, C immer abzählbar; für diese und andere Fälle sind daher in geeigneter Weise Redundanzkonventionen zu formulieren.

44 Hier müßte noch einmal /+ abstrakt/ aufgenommen werden, um diese unsinnig weite Klasse sinngemäß aufzugliedern.

45 Eine Subkategorisierung der Adjektive war vorerst nicht vorgesehen.

46 Diese Gruppe - und das gilt für die folgenden entsprechend - kann selbstverständlich aus mehreren Einzelgruppen bestehen: DER VATER - DER VATER MEINER FREUNDIN - DER UNLAENGST VERSTORBENE VATER MEINER AUS FREUDE UEBER SEINEN JAEHEN ABGANG AUS DIESEM JAMMERTAL EINEM PLOETZLICHEN HERZSCHLAG ERLEGENEN, ABER OHNEHIN NICHT ALLZU HUEBSCHEN FREUNDIN UND GEFÄHRTIN UNTERSCHIEDLICH ANGENEHMER STUNDEN.

47 Bei S wird noch einmal unterschieden zwischen DASS-Sätzen, OB-Sätzen, ACI-Sätzen und einfachen Infinitivsätzen mit ZU.

48 Über die Bedeutung der Rektion zur Auflösung syntaktischer Mehrdeutigkeiten vgl. *Elektronische Syntaxanalyse*, a.a.O., pp. 133-145.

49 Vgl. zum Homographenbegriff: *Elektronische Syntaxanalyse*, a.a.O., p. 62 ff. Für die AL empfiehlt sich eine gewisse Korrektur des Homographenbegriffs: Homographen sind Flexionsformen, die mehreren Lemmata zugeordnet werden können. Dies stellt eine gewisse Erweiterung gegenüber unserer früheren Auffassung dar, denn nunmehr können auch gewisse Mehrdeutigkeiten innerhalb einer Wortklasse berücksichtigt werden. Allerdings ist dies nur dann möglich, wenn eine Mehrdeutigkeit auch formal - im Rahmen unserer Kodierungsvorschriften - zu Buche schlägt, d.h. wenn für eine Graphemfolge aufgrund unterschiedlicher Informationseinträge mehrere Lemmata angesetzt werden. So kann beispielsweise DRUCK, von $DRUCK_2$ (Substantiv) unterschieden werden durch die Merkmale (+abs) (=$DRUCK_1$: "Vorgang des Drückens und Druckens") und (-abs) (=$Druck_2$: "Ergebnis des Druckens").

50 Vgl. *Elektronische Satzanalyse*, a.a.O., pp. 76-89 (insbesondere die Tabellen der Lösungsergebnisse D. 86 ff). Die Einteilung der Homographenklassen wird sich durch die Erweiterung des Homographenbegriffs ändern.

51 *Elektronische Syntaxanalyse*, a.a.O., p. 65 ff.