Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX

Stefan Langer

10. November 2009

Vorwort

Vorliegender Text ist die durchgesehene Fassung meiner Dissertation im Fach Computerlinguistik. Die Dissertation entstand 1993-1995 am Centrum für Informations- und Sprachverarbeitung (CIS) der Universität München im Rahmen des interdisziplinären Graduiertenkollegs SIL (Sprache, Information, Logik). Thematisch war sie in den Rahmen des CISLEX-Projekts eingebettet.

Einerseits hat sie von daher aus den anregenden interdisziplinären Diskussionen des Autors mit den Stipendiaten des SIL-Kollegs Gewinn gezogen, die mehrheitlich aus ganz anderen Fachgebieten entstammen, und deren Sicht der Dinge aus einem etwas anderen, meist formal besser geschulten Blickwinkel einige allzu verschwommene Gedankengänge schließlich in klarere Fahrwasser brachte. Andererseits hat sie in hohem Maße auch von den zahlreichen wichtigen Anregungen und Ratschlägen profitiert, die mir von den Mitarbeitern der Lexikongruppe im Rahmen von Seminaren und in Diskussionen auf den Fluren und Treppen des Instituts gegeben wurden.

Herzlich danken möchte ich meinem Doktorvater Prof. Dr. Guenthner, dem Initiator und Motor des CISLEX-Projektes. Ihm verdanke ich den Anstoß zum Thema dieser Arbeit sowie immer wieder neue Anregungen und hilfreiche Ratschläge während ihrer Entstehung.

Danken möchte ich auch den bereits oben erwähnten Kollegiaten des SIL-Kollegs und den Mitarbeitern des CIS — insbesondere den Mitgliedern der Lexikongruppe — sowie all jenen, die mir beim Korrekturlesen behilflich waren.

Herzlich Dank auch an Prof. Altmann und Prof. Zaefferer vom Institut für Deutsche Philologie der Universität München, die mir in ihren Oberseminaren einige Sitzungen zur Vorstellung von Teilen meiner Arbeit zur Verfügung stellten. Sie und die Teilnehmer der Oberseminare gaben mir einige wichtige Anregungen.

Diese Arbeit wurde im Rahmen des SIL-Graduiertenkollegs mit Mitteln der DFG durch ein Doktorandenstipendium gefördert. Ohne diese Finanzierung hätte sie kaum in so kurzer Zeit — wenn überhaupt — entstehen können.

München, im April 1998

Stefan Langer

Typographische Konventionen

Objektsprache wird wie üblich durch Kursivschreibung markiert. Einzelne Großbuchstaben in Belegen sind als Platzhalter für Nominalphrasen zu verstehen:

Fahrzeug

Karl fährt mit dem X

Semantische Klassen im Sinne der CISLEX-Kodierung werden durch Schreibung in Kapitälchen ausgezeichnet. Das Gleichzeichen vor dem Bezeichner einer taxonomischen Klasse ist der Kode für das Hyperonym der Lexeme in der Klasse:

Fahrzeuge

=Fahrzeuge

Merkmale der semantischen Klassen werden in Kapitälchenschreibung in spitze Klammern gesetzt:

¡taxonomische Klasse¿

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Zielsetzung

Die vorliegende Arbeit beschreibt die Hintergründe und die Vorgehensweise zur Erreichung eines konkreten lexikographischen Ziels: Für das am CIS (Centrum für Informations- und Sprachverarbeitung der Universität München) entstehende umfassende elektronische Wörterbuch CISLEX mußte eine Form der semantischen Kodierung der Einträge entwickelt und ein Teilausschnitt des Lexikons entsprechend kodiert werden.

Der kodierte Wortschatzausschnitt war das Teillexikon der 'einfachen' - d. h. nicht-präfigierten und nicht-komplexen — Nomina. Die theoretische Vorüberlegungen zu dieser Kodierung, die Form ihrer Durchführung und die Ergebnisse zweier Tests zur Anwendbarkeit der Kodierung sollen im Rahmen dieser Arbeit beschrieben werden.

Hervorzuheben ist, daß im Rahmen des Lexikonprojekts eine semantische Kodierung für das gesamte Lexikon angestrebt wird, und damit eine semantische Kodierung für den gesamten deutschen Wortschatz. Im Bereich der 'einfachen Nomina' mußten von daher Kodierungskriterien und Formalismen entwickelt werden, die auf alle thematischen Bereiche des Wortschatzes anwendbar sind, und nicht nur auf einen Teilbereich, wie dies etwa bei zahlreichen domänengebundenen Maschinenlexika, etwa für die maschinelle Übersetzung oder das Information Retrieval der Fall ist, deren Anwendungsdomäne von vornherein feststeht. Selbstverständlich sollte dabei sein, daß die dargestellten Kriterien und Vorgehensweisen zur Kodierung eines Ausschnitts aus dem Nominallexikon auch auf die anderen — also präfigierten und komplexen — Nomina Anwendung finden können. Zudem sollte die Kodierung eine Grundlage für die semantische Kodierung der Argumentstruktur von Verben, Adjektiven und relationalen Nomina zu Verfügung stellen.

Ergänzend soll in dieser Arbeit gezeigt werden, daß die konzipierte Kodierung für eine breite Palette computerlinguistischer Aufgabenbereiche sinnvoll anwendbar ist. Als exemplarische Anwendungsdomänen wurden Teilaufgaben der maschinellen Übersetzung und Textgenerierung einerseits und einige Teilaspekte des Information Retrievals andererseits gewählt. Sie stehen im einen Fall für die Eignung der semantischen Kodierung zur Beschreibung von distributionellen Regulariäten und im anderen Fall für deren Verwendbarkeit für eine thematische Zuordnung der Lemmata.

1.2 Das CIS-Lexikon

Das CIS-Lexikon ist ein umfassendes elektronisches Wörterbuch des Deutschen. [1]. Es wird ständig durch Abgleich mit aktuellen Texten (Tageszeitungen, Fachzeitschriften) auf dem neuesten Stand gehalten.

Das CISLEX ist eines der wenigen elektronischen Wörterbücher mit Vollständigkeitsanspruch. Weitgehende Vollständigkeit heißt:

- Alle einfachen und präfigierten Lexeme des Deutschen sind im Wörterbuch erfaßt. Dieses Teillexikon der 'einfachen Formen plus' umfaßt alle Lexeme, die nicht aus zwei Lexemen bestehen, so daß die morphologischen Eigenschaften der Gesamtform aus denen der Teile eindeutig zu ermitteln sind. Das Lexikon der nicht aus mehreren Lexemen bestehenden Formen ist wiederum aufgeteilt nach 'einfachen Formen' (EF) und 'präfigierten Formen' (zusammen: EF+).

- Zudem umfaßt das Lexikon sämtliche komplexen Formen, die in Wörterbüchern des Deutschen und in den bisher ausgewerteten Texten angetroffen wurden. Diese konstituieren das Teillexikon der 'komplexen Formen'. Den größten Teilbestand macht hierbei die Sammlung von Nominalkomposita aus. Beim Bestand an Komposita ist allerdings keine Vollständigkeit zu erreichen, da Komposition im Deutschen ein hochproduktives Wortbildungsmuster ist, und Komposita ständig neu gebildet werden. Vollständigkeit heißt für dieses Teillexikon aber zumindest, daß alle lexikalisierten Komposita enthalten sind.

Die morphologische Information für die einfachen und präfigierten Formen ist bereits kodiert (zur Kodierung und zu den morphologischen Kriterien vgl. Maier 1995). Das CISLEX der nicht-präfigierten Formen (EF) umfaßt ca. 37 000 Nomina, 12 000 Verben, und 10 000 Adjektive.

1.3 Kodierter Wortschatzausschnitt

Im Rahmen der Arbeit wird die Kodierung der 'einfachen Nomina' beschrieben. Der Begriff des 'einfachen Nomens' ist folgendermaßen definiert:

- Ein 'einfaches Nomen' ist ein Nomen *w*, von dem es keine sinnvolle Zerlegung *w1w2* gibt, so daß

- *w1* ein Präfix oder ein anderes Wort ist und

- *w* dieselben morphologischen Eigenschaften wie *w2* besitzt.

Unter morphologischen Eigenschaften ist dabei die Bildung der Kasus- und Numerusformen zu verstehen[2].

[1] Zum CISLEX vgl. Guenthner/Maier (1994). Auch die Dissertation von Petra Maier (1995) setzt sich mit Aspekten der Kodierung des CISLEX auseinander — allerdings in erster Linie der morphologischen Beschreibung der Lemmata — und beschreibt die Verwendung elektronischer Wörterbücher für die automatische Lemmatisierung.

[2] Ein Problem in diesem Zusammenhang sind die Fugenformen der Nomina. Während die Kasus- und Numerusformen eines komplexen Lexems wie *Bahnhof* aus denen des Simpliziums *Hof* abgeleitet werden können, ist dies für die Fugenform *Bahnhofs-* nicht möglich, denn *Hof* tritt als Erstglied in Komposita ohne ein Fugen-*s* auf; vgl. *Hofeingang* versus *Bahnhofseingang*.

. Der kodierte Wortschatzausschnitt umfaßt also sowohl suffigierte Nomina als auch von einer präfigierten Basis derivierte Nomina, wenn die unpräfigierte Basis keine nominale Ableitung mit denselben morphologischen Eigenschaften hat. So ist das Lexem *Veränderung* nicht im kodierten Wortschatzauschnitt enthalten (wegen *Änderung*, das dieselben morphologischen Eigenschaften besitzt), wohl aber das Lexem *Vergebung* (es existiert zwar das Verb *geben*, nicht aber ein Nomen **Gebung*).

Die semantische Kodierung der 'einfachen Nomina' ist Thema dieser Arbeit. Explizit ausgeschlossen vom zu kodierenden Wortschatz waren bisher neben den präfigierten und komplexen Nomina zunächst solche mit adjektivischer Deklination. Deren Kodierung wird die semantische Beschreibung der einfachen Nomina abschließen. In einem nächsten Kodierungsschritt sollen die präfigierten Nomina (ca. 15 000) mit einem semantischen Kode versehen werden, gefolgt von den lexikalisierten Komposita.

Komposita stellen ein Problem für die semantische Beschreibung in großen Lexika dar, da es sich hier um eine offene Klasse von Wörtern handelt. Das Lexikon der (morphologisch) komplexen Formen umfaßt derzeit ca. 1 000 000 morphologisch segmentierte Nominalkomposita. Dieser Bestand wird zudem ständig erweitert. Der größere Teil dieser Nominalkomposita wird damit nicht einzeln kodiert werden können. Es sind also Algorithmen erforderlich, die eine automatische Kodierung dieses Teilwortschatzes ermöglichen. Hierzu muß die Kodierung der einfachen Nomina eine Grundlage bieten. Wenn ein Kompositum nicht bereits vollständig lexikalisiert ist, hängt die Gesamtbedeutung von den Bedeutungen der Teile und von der Relation zwischen diesen Teilbedeutungen ab. Die Untersuchungen zu Selektionspräferenzen von Köpfen von Nominalkomposita in Kapitel 4 können neben ihrer Aussagekraft für die Selektionsrelevanz der kodierten semantischen Klassen als eine Grundlage zur automatischen semantischen Beschreibung von Nominalkomposita angesehen werden.

1.4 Theorie und Praxis der Lexikographie

Für den Zusammenhang zwischen Anforderungen von Seiten möglicher Anwendungen, theoretischen Überlegungen zur Kodierung und der Kodierungspraxis gilt für die elektronischen Lexikographie, was Calzolari (1994) so zusammenfaßt:

> "If we aim at a reusable multifunctional lexicon, obviously our choices, for each parameter (such as size, depth of information, levels of linguistic description, explicitness, etc.), should be guided by the most demanding, the widest, and the deepest requirements. However, these should be corrected by a careful analysis of a) the possibility of actually achieving these most general requirements, in accordance with the state-of-the-art (in linguistic theory, software technology, etc.), and b) the cost-effectiveness of such ambitious prospects as opposed to more modest but more realistic and economically viable strategies." (Calzolari 1994: 269).

Diese Aussage findet auch auf die Kodierung für das CISLEX Anwendung, welches als multifunktionelles Lexikon konzipiert ist. Für die Konzeption der semantischen Kodierung sind dabei besonders intensive Abwägungen erforderlich, da für diesen Bereich die Kodierung elektronischer Wörterbücher besonders wenig standardisiert ist, während für die morphologische Kodierung die Zahl der Möglichkeiten relativ eingeschränkt ist (vgl. Calzolari 1994: 270).

Auf die Anforderungen an die semantische Kodierung und den Stand der Theorie der lexikalischen Semantik in der theoretischen Linguistik und der Computerlinguistik wird in Kapitel 2 dieser Arbeit ausführlich eingegangen. Dabei wird festgestellt werden, daß es die

Theorie der Wortsemantik keineswegs erlaubt, festzulegen, was eine vollständige semantische Kodierung für ein Lexikon an Beschreibungselementen zu enthalten hat. Die Erörterung aktueller wortsemantischer Paradigmen gibt allerdings, wie sich besonders auch im Vergleich mit den theoretischen Hintergründen anderer elektronischer Lexika — z. B. Miller u. a. (1993) oder Gross (1992, 1994) — zeigen wird, einige Anregungen zur Modifizierung gängiger computerlinguistischer Kodierungspraxis.

Bei der konkreten Ausführung der Kodierung der semantischen Klassen für die einfachen Nomina des CISLEX wurde aufgrund des eingeschränkten Zeitraums für die gestellte Aufgabe auf zwei Aspekte besonders geachtet:

- Modularität: Das Ergebnis sollte eine in sich abgeschlossene Teileinheit der semantischen Kodierung bilden, die als Informationsmodul für eine Vielzahl von Anwendungen eingesetzt werden kann.

- Erweiterbarkeit: Die kodierte Information sollte für die verschiedensten Anwendungen eines elektronischen Wörterbuches sukzessiv erweiterbar sein.

Das Ergebnis der Kodierungsarbeit war primär die Einteilung des Wortschatzes der 'einfachen Nomina' in semantische Klassen. Diese Einteilung eignet sich zur Beschreibung einer Reihe von semantische Regularitäten im Bereich von Selektionspräferenzen einerseits und zur Erstellung einer thematischen Gliederung des deutschen Wortschatzes andererseits — und damit auch zur Implementierung grundlegender Funktionen für die nachfolgend angesprochenen Anwendungsbereiche der maschinellen Übersetzung und des Information Retrieval. Zusätzlich bietet die vorgenommene Einteilung eine solide Grundlage für eine detailliertere semantische Beschreibung.

1.5 Mögliche Anwendungen eines elektronischen Wörterbuchs

Im folgenden soll unter dem Gesichtspunkt des Stellenwerts der semantischen Beschreibung eine kurze Übersicht über die möglichen Anwendungen eines elektronischen Wörterbuches gegeben werden. Eine ausführliche Erörterung der genauen Erfordernisse an die semantische Kodierung für die entsprechenden Anwendungen findet sich in Kapitel 2.

1.5.1 Rechtschreibkorrektur, Grammatik-Prüfung und Thesauri

Beinahe jedes Textverarbeitungsprogramm enthält mittlerweile eine Subroutine zur Rechtschreibkorrektur. Oft jedoch sind die vorhandenen Lexika relativ klein oder enthalten nicht alle morphologischen Formen oder Rechtschreibvarianten jedes Lexems.

Damit ist die automatische oder halbautomatische Rechtschreibkorrektur sicher eine der naheliegendsten Anwendungen eines umfassenden elektronischen Wörterbuchs. Allerdings ist hier die rein morphologische Kodierung sämtlicher Lemmata weitgehend ausreichend, da für die Rechtschreibprüfung nur eine Unterscheidung korrekter und unkorrekter Wortformen aufgrund der Möglichkeit der Zuordnung zur graphemischen Gestalt der Lexemformen im Lexikon nötig ist. Zwar können bei der Rechtschreibkorrektur auch semantische Aspekte ins Spiel kommen und zusätzliche Kriterien zur Beurteilung der Zulässigkeit eines komplexen Lexems oder einer Wortfolge bieten, wie zum Beispiel bei der Korrektheitsbeurteilung bestimmter Wortwiederholungen oder der Überprüfung nicht im Lexikon enthaltener Nomi-

nalkomposita, doch ist der Aufwand zur Implementierung entsprechender Routinen hier im Vergleich zur Verbesserung des Ergebnisses relativ hoch[3].

Unverzichtbar ist eine semantische Kodierung dagegen für bessere Varianten der automatischen Grammatik- und Stilüberprüfung, d. h. der Überprüfung bestimmter grammatischer und stilistischer Wohlgeformtheitsbedingungen. [4]. Moderne Textverarbeitungsysteme gehen zusehends mehr dazu über, solche Subroutinen zur Grammatik- und Stilüberprüfung einzusetzen, wobei über ihre Akzeptanz und Nützlichkeit allerdings bisher meines Wissens keine umfassenden Untersuchungen vorliegen.

Für solche Anwendungen ist die Berücksichtigung verbaler oder adjektivischer Selektionsrestriktionen nötig, zu deren Überprüfung zumindest elementare semantische Merkmale wie ¡belebt¿, ¡menschlich¿ oder ¡konkret¿ verfügbar sein müssen.

Den sogenannten 'Thesauri' in Textverarbeitungprogrammen liegt von vornherein eine semantische Kodierung zugrunde. Diese elektronischen Synonymwörterbücher dienen dazu, auf Anfrage alternative Lexeme vorzuschlagen, indem das Interface zu einem solchen Thesaurus für jedes eingegebene Wort die (Teil-)synonyme zur Verfügung stellt. Sie helfen damit, Wortwiederholungen zu vermeiden und für einen gegebenen Ausdruck einen passenderen, bedeutungsähnlichen Ausdruck zu finden. Dabei wird der sprachliche Kontext in den mir bekannten Systemen nicht berücksichtigt. Eine solche Anwendung setzt ein nach den Sinnrelationen Synonymie und Hyponymie geordnetes Lexikon voraus.

1.5.2 Information Retrieval

Das Information Retrieval schließt die thematische Klassifizierung, Inhaltszuordnung, Strukturierung und Inhaltsbeschreibung eines Textes ein (vgl. Kuhlen 1989).

Die thematische Klassifizierung ist hierbei die grundlegendste Aufgabe. Sie umfaßt folgende Teilaufgaben:

- Zum Erkennen eines bestimmte Lexems ist zunächst eine eindeutige Lemmatisierung notwendig. Dies ist zunächst Aufgabe der morphologischen Komponente eines Lexikons. Allerdings schließt Lemmatisierung letztendlich auch die Disambiguierung semantisch mehrdeutiger Formen ein.

- Nächster Schritt ist die Gewichtung der Lexeme nach inhaltlicher Aussagekraft und Klassifikationsvermögen. In der einfachsten Form ist dies eine Aussortierung von thematisch irrelevanten Wörtern. Dies ermöglicht dann die Indexierung eines Textes.

- Die Einordnung eines Textes wird möglich durch einen thematischen Thesaurus, der die Lexeme enthält, welche für die ausgewählten Themengebiete relevant sind.

Die semantische Komponente eines elektronischen Wörterbuchs, das für das Information Retrieval geeignet sein sollte, beinhaltet also zunächst einen Thesaurus. Dabei muß eine Strukturierung des Lexikons, die für die Einordnung bestimmter Texte geeignet sein soll, auf das betroffene Themengebiet zugeschnitten sein. Ein allgemeinsprachlicher Thesaurus, wie er zunächst für das CISLEX anvisiert wird, erlaubt eine erste Klassifikation allgemeinsprachlicher Texte. Allerdings ist auch eine fach- bzw. fachbereichsspezifische thematische Unterteilung aus einem nach allgemeinsprachlichen Gesichtspunkten bereits klassifizierten Wortschatz wesentlich leichter zu erstellen als ohne diese Vorarbeiten, da hier zum einen

[3]Eine semantische Kodierung könnte auch bei der Auswahl sinnvoller Korrekturvorschläge innerhalb eines Rechtschreibkorrekturprogramms von Nutzen sein. Vgl. dazu St-Onge (1995), der sich mit Aspekten der Rechtschreibkorrektur auf Basis der semantischen Kodierung in WordNet beschäftigt.

[4]Zu Programmen zur Grammatik und Stilüberprüfung s. Ravin (1993), Eglowstein (1991).

bestimmte Klassen übernommen worden können, zum anderen die Klassen leicht zugänglich sind, innerhalb derer eine fachspezifische Klassifizierung vorgenommen werden soll.

1.5.3 Maschinelle Übersetzung und Textgenerierung

Zur (regelbasierten) maschinellen Übersetzung sind im Lexikon eine große Menge linguistischer Informationen zu kodieren, darunter auch verschiedene Typen von semantischer Information (vgl. Calzolari 1994: 279f). Eine für die maschinelle Übersetzung geeignete semantische Kodierung setzt in jedem Fall die Wahl einer Übersetzungsstrategie und die Definition der Übersetzungsalgorithmen voraus. Ein einsprachiges elektronisches Wörterbuch kann jedoch einen Teil der notwendigen Information bereitstellen und ist damit als ein Modul für die maschinelle Übersetzung einsetzbar. Eine der wichtigsten Voraussetzungen für eine solche Verwendung ist die Einteilung des Wortschatzes in Klassen von Lexemen, die Selektionsrestriktionen in bestimmten Kontexten gehorchen (auch die Beschreibungselemente, die üblicherweise als 'semantische Merkmale' bezeichnet werden, können als Klassenbezeichner in diesem Sinne angesehen werden). Solche Klassen erlauben sowohl teilweise Disambiguierung in der Quellsprache als auch eine korrekte Wortwahl in der Zielsprache.

1.6 Überblick

In den vergangenen Abschnitten wurden die wichtigsten Grundvoraussetzungen kurz skizziert. Der folgende Hauptteil der Arbeit gliedert sich in drei Kapitel. Das Kapitel 2 beschreibt die Hintergründe der vorgenommenen Kodierung unter Berücksichtigung der Wortsemantik in der theoretischen Linguistik, der Computerlinguistik und der Praxis der Lexikographie. In Kapitel 3 werden die semantischen Beschreibungselemente dargestellt und durch ausgewählte Kodierungsbeispiele veranschaulicht. Das Kapitel 4 ist zwei Tests zur Anwendbarkeit der vorgenommenen Klassifikation gewidmet. Eine Zusammenfassung wichtiger Ergebnisse und ein Überblick über die anstehenden Kodierungsschritte in Kapitel fünf runden die Arbeit ab.

Selektiver Forschungsüberblick der linguistischen Semantik, Lexikographie und der Computerlinguistik

Es wird zunächst ein Überblick über die Hauptparadigmen der lexikalischen Semantik in der theoretischen Linguistik gegeben. Berücksichtigt werden sowohl klassische Paradigmen zur semantischen Beschreibung von Lexemen wie die Komponentenanalyse und die Theorie der Sinnrelationen, daneben aber auch jüngere Ansätze wie die Prototypentheorie. Unter Berücksichtigung unserer Zielvorgaben wird diskutiert, welche Implikationen Aspekte dieser Paradigmen für die semantische Kodierung haben könnten. Im Zusammenhang mit wortsemantischen Theorien werden in diesem Kapitel auch Möglichkeiten zur korpusgestützten Ermittlung von Wortbedeutungen angesprochen.

Zudem wird das Unterfangen einer semantischen Kodierung für ein elektronisches Wörterbuch unter lexikographischer Sichtweise beleuchtet. Dazu werden zunächst Bedeutungsbeschreibungen in herkömmlichen und maschinenlesbaren Wörterbüchern sowie die zugrundeliegenden lexikographischen Prinzipien abgehandelt. Bereits bestehende Kodierungen für elektronische Wörterbücher und Wortdatenbanken werden auf ihre Funktionalität und Tragfähigkeit untersucht. Besondere Schwerpunkte liegen hier auf der englischsprachigen Wortdatenbank WordNet (vgl. Miller u.a. 1990, 1993) und der Bedeutungskodierung für das Wörterbuchsystem der Forschergruppe um G. Gross in Paris (vgl. Gross 1992, 1994). Dabei wird auch die Frage nach der Übertragbarkeit von Teilen der Konzeption der Semantikbeschreibung in diesen beiden Systemen auf ein elektronisches Wörterbuch des Deutschen gestellt.

Entwicklung eines eigenen Formalismus und Kodierungskriterien für das CIS-Lexikon sowie Darstellung der Strukturierung des Nominallexikons

Aufbauend auf der erfolgten Auswertung linguistischer und lexikographischer Theorien und der Untersuchung bereits bestehender Kodierungen werden die für das CISLEX entwickelten Formalismen zur semantischen Beschreibung vorgestellt.

Die Einbeziehung der Überlegungen des vorhergehenden Kapitels zu Klassentypen und zur Prototypentheorie führt für die semantische Gliederung des Nominalwortschatzes zu Kodierungselementen, die meines Wissens in ihrer Gesamtheit neu sind für ein solch umfangreiches elektronisches Wörterbuch. Die semantischen Klassen von Nomina werden nicht mehr als gleichartige Objekte betrachtet, sondern erhalten selbst semantische Metaeigenschaften, die sich auf die Aufstellungkriterien, den zugrundliegenden Kategorientyp und die Klassenstruktur beziehen.

Es wird eine Liste von semantische Klassen vorgestellt, die zur semantischen Grobklassifizierung des Lexikons dienen. Die Feinklassifikation einiger Teilbereiche des Nominallexikons (Menschenbezeichnungen, Fahrzeuge und Formen) wird ausführlich beschrieben. Für die einzelnen Klassen werden die Kriterien angegeben, die 1) zur Aufstellung der semantischen Klasse geführt haben und 2) die Einordnung eines Nomens in diese semantische Klasse bedingen.

Ferner werden die verwendeten semantischen Beschreibungselemente vorgestellt, die neben der Kodierung der semantischen Klassen eine Rolle spielten. Der Schwerpunkt liegt dabei auf der Erfassung von Sinnrelationen.

Untersuchung von Selektionsrestriktionen der Köpfe von Nominalkomposita und zur thematischen Klassifizierung von Zeitungstexten

Die Arbeit wird abgeschlossen mit der Darstellung zweier Untersuchungen, welche die Relevanz der semantischen Kodierung für die Beschreibung von Selektionspräferenzen und für die thematische Zuordnung von Dokumenten demonstrieren sollen.

Die Beschreibung des Selektionsverhaltens sprachlicher Ausdrücke auf Basis von Kollokationen in umfangreichen Korpora erlangt zusehends Bedeutung für die semantische Beschreibung von Lexemen. Im Deutschen bieten sich für eine Analyse semantischer Kollokationen zunächst die Nominalkomposita an, denn sie setzen kein kompliziertes Parsing voraus, wie es nötig wäre, um Verb-Komplement-Kollokationen zu ermitteln. Eine Analyse von Nominalkomposita fügt sich zudem in den Rahmen dieser Arbeit gut ein, indem die Kodierung der ' einfachen Nomina' verwendet wird, um weitere Aussagen über die Semantik von Lexemen derselben Wortart zu machen. Die Ergebnisse der Analyse zweigliedriger Nominalkomposita sollen die Konsistenz der vorgenommenen Klassifizierung und deren Relevanz für die maschinelle Sprachverarbeitung demonstrieren.

Der im zweiten Teil des Kapitels beschriebene Test zur thematischen Klassifizierung einer größeren Menge von Texten aus dem "Vermischten" der "Süddeutschen Zeitung" soll abschließend die Eignung der semantischen Klassen für das Information Retrieval demonstrieren. Hierzu wird aufgrund einer thematischen Auswahl der kodierten Klassen eine größere Zahl von Artikeln vier vorab festgelegten Themenbereichen automatisch zugeordnet. Diese Zuordnung wird dann mit einer manuellen Klassifikation verglichen.

Kapitel 2

Wortsemantik und Lexikographie

2.1 Lexikalische Semantik und Theorie des Lexikons

Eine computerlinguistische Arbeit zur Kodierung eines elektronischen Wörterbuchs könnte legitimerweise die Erörterung der lexikalische Semantik in der theoretischen Linguistik zum größten Teil auslassen und sich auf Traditionen in Anwendungen innerhalb der maschinellen Sprachverarbeitung berufen. Mir erscheint jedoch die Reflexion über die sprachwissenschaftliche Basis fruchtbringend zu sein, vor allem, da es darum geht, die Konzeption für die Semantik in einem elektronischen Wörterbuch vorzulegen, das nicht von vornherein auf eine spezielle Anwendung zugeschnitten ist. Ein solcher Überblick über die linguistischen Grundlagen computerlinguistischer Anwendungen ermöglicht zu vergleichen, welche Erkenntnisse aus dem Forschungsbereich der Wortsemantik in natürlichsprachliche Systeme integriert werden konnten, und welche bisher kaum berücksichtigt wurden, obwohl sie eine stärkere Berücksichtigung verdient hätten. Die Erörterung — auch nicht formalisierter — linguistischer Theorien zur Wortsemantik macht zudem deutlich, welche Arbeit auf diesem Feld auch von der Computerlinguistik noch zu leisten ist.

Ich meine, in der theoretischen Linguistik allgemein und speziell in der Wortsemantik zwei eng verwandte Trends feststellen zu können:

- Zunächst den Trend weg von Theorien, deren Grundlage die Auffassung ist, man könne ein gesamtes sprachliches Subsystem mit wenigen, präzisen Regeln und der ergänzenden Auflistung einer begrenzten Zahl von Ausnahmen beschreiben, und hin zu Ansätzen, die in höherem Maße eine Modellierung von idiosynkratischen oder nur wenige Beschreibungsobjekte umfassenden Phänomenen verlangen - d. h. es findet eine Verschiebung der Schwerpunkte hin zum Lexikon statt. Dies führt zu einem Anwachsen der Information, die herangezogen wird, um Modelle für das Sprachsystem und dessen Funktion zu entwerfen. Im Bereich der formalen Grammatik läßt sich dies an der zunehmenden Wichtigkeit von lexikonorientierten Formalismen wie etwa der HPSG (Pollard/ Sag 1987, 1994) erkennen. Zu diesem Trend gehört auch die Problematisierung und teilweise Auflösung einfacher terminologischer Begriffe in Bündel von Eigenschaften (man vergleiche etwa die Problematisierung des Valenzbegriffs in J. Jacobs 1994).

- Zudem einen teilweisen Trend weg von der Betrachtung einiger weniger sprachlicher

> Daten, aufgrund derer dann ein ganzes Theoriegebäude entwickelt wird, und hin zu einer (wieder) zunehmenden Berücksichtigung konkreter sprachlicher Fakten. Dies führt auch zu eine verstärkten Bedeutung von Korpora (vgl. hierzu vor allem die statistischen Ansätze, wie sie in Abschnitt 2.3.4. dargestellt werden).

In der Wortsemantik zeigen sich die oben genannten Trends besonders deutlich. Ansätze, die postulieren, daß die erschöpfende semantische Beschreibung eines Lexems, vielleicht sogar unabhängig vom Kontext, mit einem relativ kleinen, oder zumindest überschaubaren Instrumentarium zu erreichen ist, haben sich als unzureichend erwiesen und werden zusehends verdrängt von solchen, die von der Annahme ausgehen, daß es kaum Abschlußkriterien für die Darstellung der Bedeutung lexikalischer Einheiten gibt, und die in der Bedeutung eines Wortes ein "kleines Universum" sehen (Velardi/ Pazienza/ Fasolo 1991: 154), das wiederum ein nicht eindeutig abgegrenztes Subuniversum der Repräsentation des gesamten Lexikons ist. Einen nicht zu unterschätzenden Beitrag zu diesem Paradigmenwechsel hat auch die seit dem Ende der siebziger Jahre an Bedeutung gewinnende kognitive Linguistik geleistet, deren Erkenntnisse auch von anderen Strömungen in der theoretischen Linguistik und der Computerlinguistik rezipiert wurden. Auch solche Ansätze in der theoretischen Linguistik, die durchaus nicht der Forderung nach psychologischer Adäquatheit entsprechen wollen, kommen mittlerweile nicht darum herum, die Phänomene zu berücksichtigen, die sich mit früheren, scheinbar exakteren Ansätzen nicht vereinbaren lassen. So wurde selbst in der logischen Semantik — allerdings mit geringem Erfolg — versucht, prototypische Strukturen und Vagheit von Wortbedeutungen mit Hilfe der Fuzzy-Logic in die formale Beschreibung zu integrieren (vgl. Parikh 1994: 525f).

Der Transfer der genannten Entwicklungen im Bereich der Wortsemantik in die Computerlinguistik bringt allerdings notwendigerweise Schwierigkeiten mit sich, denn eine computerlinguistische Implementierung der Wortbedeutung ist auf Formalisierung angewiesen. Formalisieren lassen sich jedoch nur bestimmte Phänomene und Aspekte dessen, was in zahlreichen neueren Arbeiten zur Wortsemantik beschrieben wird. Andere Aspekte der Wortsemantik können zunächst noch kaum befriedigend systematisiert werden. Metapher und Metonymie beispielsweise sind Untersuchungsgegenstand zahlreicher Arbeiten der neueren Linguistik (z. B. Lakoff/ Johnson 1980), eine formalisierbare Theorie der zugrundeliegenden Mechanismen liegt hier allerdings trotz erster Ansätze in der Computerlinguistik (vgl. Fass 1991, Behrens 1993: 262f) nicht vor.

Neuere Ansätze haben also in der linguistischen Wortsemantik die herkömmlichen, formalisierbaren Beschreibungsinstrumentarien in Frage gestellt, ohne sie vollständig ersetzen zu können. Es bleibt damit zu fragen, ob nicht bestimmte ältere Ansätze, unter den neuen Prämissen umformuliert, in dieser revidierten Form immer noch verwendet werden können. Seit klar geworden ist, daß viele Ansätze zur Wortsemantik in ihrer klassischen Form keine vollständige Bedeutungsbeschreibung erlauben, wird in der Literatur das "Kind mit dem Bad" (Lüdi 1985: 99) ausgeschüttet: die — keineswegs essentiellen — Annahmen, die mit bestimmten Theorien assoziiert sind — so etwa ein Vollständigkeitsanspruch in der Komponentenanalyse — werden zum Anlaß genommen, diese Theorie als ganzes zu verwerfen, und durch sicher teilweise richtige, aber doch recht vage allgemeine Äußerungen zur Wortbedeutung zu ersetzen (vgl. Lüdi 1985: 90, 99). Dabei bleibt oft unberücksichtigt, daß eine Vereinfachung auch und gerade im Rahmen neuerer Ansätze notwendig bleibt, denn ist die Beschreibung eines Wortes so aufwendig und beinhaltet sie so viele Dimensionen, wie in diesen Theorien postuliert, muß notwendigerweise eine Reduktion des Beschriebenen auf das erfolgen, was im jeweils konkreten Fall als notwendig angesehen wird; eine solche Reduktion ist auch nicht zu verwerfen, solange nicht postuliert wird, daß diese Idealisierungen nun die ganze Wortsemantik seien.

Der Trend weg von der Auffassung, daß eine exakte und vollständige Beschreibung der Wortbedeutung möglich ist, wird unterstützt durch die Tendenz, anstatt wenige 'Eigenbelege' als Datenbasis zugrunde zu legen, größere Korpora zu betrachten und auszuwerten. Dieser Trend hat sicher teilweise praktische Ursachen. In den letzten Jahren wurden zusehends größere Korpora verfügbar, und diese lassen sich immer effektiver und schneller bearbeiten. Während noch vor einiger Zeit Belege mühsam von Hand aus Druckwerken herausgesucht werden mußten, können heute durch effiziente Suchalgorithmen große Mengen von relevanten Belegstellen aus in elektronischer Form vorliegenden Texten extrahiert werden, oder es kann zumindest eine Vorauswahl von potentiellen Belegstellen getroffen werden (vgl. Church/ Hanks 1990: 26-28). Dieser Trend kommt natürlich vor allem der Beschreibung des Verhaltens einzelner Wörter — und damit der Lexikographie — zugute, da Wortformen auch in großen Korpora relativ unproblematisch identifiziert werden können, während die Beschreibung von Mehrwortausdrücken, syntaktischen Konstruktionen oder gar textlinguistischen Regularitäten linguistisch aufbereitete Daten verlangt.

Für die gewachsene Bedeutung von realen Texten spielen aber sicher auch in der linguistischen Forschung der letzten Jahre selbst begründete Ursachen eine nicht zu vernachlässigende Rolle. Es hat sich gezeigt, daß die bloße Betrachtung einiger weniger Beispiele zu vorschnellen Generalisierungen führt, die dem Untersuchungobjekt Sprache nicht immer angemessen sind. Die weitgehende Beschränkung der Forschung auf die Beschreibung der Kompetenz hat dazu geführt, daß unliebsame Daten von der linguistischen Forschung einfach ausgeblendet und als für die Sprachbeschreibung uninteressant abgetan wurden. Diese Selbstbeschränkung ging einher mit der Konzentration der meisten Kräfte, auch in der Computerlinguistik, auf das Regelsystem der Syntax. [1]. Doch innerhalb der Syntax gibt es in den letzten Jahren ebenfalls einen Trend dahin, mehr Information im Lexikon unterzubringen, und damit für die Beschreibung lexikalischer Idiosynkrasien besser gerüstet zu sein (vgl. z. B. den HPSG-Formalismus wie beschrieben in Pollard/ Sag 1987, 1994). Nur die Beschränkung auf relativ wenige Satzmuster hat es auch ermöglicht, daß im Bereich der Satzsemantik die Montaguesemantik mit ihrem übergeneralisierten Kompositionalitätsanspruch [2] lange Zeit eine dominierende Stellung innehatte, ohne daß hinterfragt wurde, inwieweit sich dieser Formalismus in seiner ursprünglichen Form wirklich jemals dazu eignen könnte, die Bedeutung von unrestringiertem Text zu beschreiben, und auf welche Weise sich in diese Ansätze eine adäquate Wortsemantik integrieren lassen könnte. Doch in den letzten Jahren haben auch im Bereich der logischen Satzsemantik Formalismen an Bedeutung gewonnen, die prinzipiell in der Lage sind, in realen Texten auftretende Phänomene eher zu beschreiben. Zu den Entwicklungen aus diesem Bereich zählt die Berücksichtigung von anaphorischen Prozessen in der dynamischen Prädikatenlogik (DPL) und in der Diskursrepräsentationstheorie (Kamp/ Reyle 1993). In letzterem Formalismus ist es auch eher als in der klassischen Logik möglich, eine Beschreibung der Wortsemantik zu integrieren; [3] man vergleiche hierzu etwa den Versuch der Beschreibung deutscher Nominalkomposita im Formalismus der DRT durch

[1] Der hohe Stellenwert der Syntax in der Linguistik der letzten Jahrzehnte ist in dieser Hinsicht kein Zufall. Es ist vor allem in diesem Bereich möglich, mit Regelwerken, die sich auf wenige Daten stützen, die imponierendsten Ergebnisse zu erzielen, weil vieles, was sich hiermit nicht fassen läßt, in die Semantik und Pragmatik quasi "abgeschoben" werden kann. Die Renaissance der Lexikontheorien zeigt, daß sich hier etwas ändert.

[2] Zadrozny (1994) zeigt, daß jede Semantik als "kompositionell" im weitesten Sinne bezeichnet werden kann. Kompositionell soll im Zusammenhang mit dieser Aussage heißen, daß mit einer relativ kleinen Menge an Kompositionsregeln die Bedeutung eines Satzes aus der Bedeutung der Teile ermittelt werden soll. Es sollen nicht die Verdienste der logischen Semantik geschmälert werden; es wird nur darauf hingewiesen, daß eine solche Semantik in ihrer herkömmlichen Form mir prinzipiell nicht ausreichend erscheint, einige wesentliche Aspekte der Bedeutung von Texten zu erfassen.

[3] Insbesondere sei auf die unterschätzte Rolle des Lexikons in der Montaguegrammatik hingewiesen. Eine Darstellung der Wortsemantik in diesem Theorierahmen findet sich in Dowty (1979).

Meyer (1993).

Die Untersuchung großer Korpora macht effiziente Auswertungsalgorithmen für größere Datenmengen notwendig. Andererseits erfordert die Beschreibung von Phänomenen in Syntax und Semantik, die nicht durch ja/ nein Regeln beschrieben werden können, die Anwendung von Methoden zur Beschreibung von Unschärfephänomenen. Diese beiden Ursachen haben zu einer wachsenden Bedeutung statistischer Methoden in verschiedenen Domänen der Linguistik und vor allem in der Computerlinguistik geführt. Während Systeme zur Spracherkennung sich schon länger auf statistische Methoden stützen, sind solche Ansätze in der Syntaxanalyse, der Textgenerierung (z. B. Schabes/ Roth/ Osborne 1993) und der automatischen Übersetzung (vgl. z. B. Brown u. a. 1990) relativ neu und wurden, als genuin computerlinguistische Ansätze, in der theoretischen Linguistik bis vor kurzem kaum rezipiert. Doch sind heute die statistikbasierten oder zumindest durch Statistik ergänzten Ansätze mit regelbasierter Grundlage trotz eines noch vorhandenen Theoriedefizits gegenüber den seit vielen Jahren betriebenen, rein regelbasierten Ansätzen im Bereich natürlichsprachlicher Anwendungen durchaus schon teilweise konkurrenzfähig und lassen sich auch von der linguistischen Forschung nicht mehr ignorieren.

In der Wortsemantik hat die verstärkte Berücksichtigung von konkreten Texten und großen Korpora zu einer erhöhten Bedeutung kollokationeller Ansätze geführt [4]. Neu ist hier neben der Untersuchung umfangreicher Textmengen auch die Anwendung statistischer Methoden zu einer Annäherung an die Semantik von Wörtern. Diesen Ansätzen liegt die Annahme zugrunde, daß Aspekte der Lexembedeutung mit Hilfe statistischer Methoden aufgrund der Kombinatorik sprachlicher Ausdrücke determiniert werden können. In der Praxis der Lexikographie ist die Untersuchung der Bedeutung von Wörtern aufgrund von Belegsammlungen allerdings schon viel älter und wurde auch nie wesentlich in Frage gestellt. Wenn auch den lexikographischen Ansätzen noch die statistische Methodik fehlte, so beruhte sie doch zumindest auf der Sammlung konkreter Textstellen; die momentane Entwicklung läßt sich also auch als eine Wiederannäherung der Wortsemantiktheorie an die Praxis der Lexikographie verstehen.

Ich fasse kurz zusammen, was mir als die Hauptentwicklungen speziell für den Bereich der Wortsemantik in den letzten Jahren erscheint:

- Zunehmende Problematisierung der exakten Ansätze, die aus der Tradition des Strukturalismus und der logischen Semantik stammen, und — damit einhergehend — eine zunehmende Wichtigkeit von Ansätzen, die zur Beschreibung von Unschärfephänomenen geeignet sind.

- Hinwendung zu konkreten Texten sowohl in Form der detaillierten Analyse einzelner Textstellen als auch der statistischen Auswertung von Korpora im großen Stil.

Nachdem ich im folgenden zunächst auf die Grundsatzfrage des Beschreibungsgegenstandes der Wortsemantik eingehen werde, sollen im anschließenden Forschungsüberblick drei Fragenkomplexe besonders im Auge behalten werden:

- Welche formalisierbaren Ansätze aus der strukturalistischen Tradition der lexikalischen Semantik lassen sich so verwenden, daß sie mit den neueren Forschungsergebnissen der Wortsemantik aus der kognitiven Linguistik und der Psycholinguistik noch kompatibel sind?

[4] Als Beispiele für Untersuchungen aus diesem Bereich seien genannt: Agarval (1995), Basili/Pazienza/Velardi (1991), Church u.a. (1994), Granger (1977), Ribas (1994, 1995).

- Welche Aspekte aus einer kognitiv orientierten Wortsemantik könnten in einem computerlinguistischen Ansatz in Betracht gezogen werden, und auf welche Weise sollen sie berücksichtigt werden?
- Welchen Stellenwert kann man den statistischen Methoden zur Beschreibung von Wortsemantik zubilligen, für welche Aufgaben lassen sich die durch sie gewonnenen Daten verwenden, und inwiefern korrelieren sie mit den veränderten Ansätzen paradigmatischer Analysen?

Der folgende Überblick wird sich auf die Beschreibung der Nominalsemantik konzentrieren. Einerseits, weil diese in der Wortsemantikforschung schon immer einen prominenten Stellenplatz hatte, andererseits, weil die in Kapitel 3 beschriebene Kodierung für das CISLEX ausschließlich die Semantik der Nomina zum Gegenstand hat.

2.2 Beschreibungsgegenstand der Wortsemantik

Wenn von Wortbedeutung gesprochen wird, geht es in allen folgenden Ausführungen um die Beschreibung der Intension von Lexemen. Die Frage der Extension wird mich nicht weiter interessieren. Letztlich werden bei der Bedeutungsbeschreibung in den meisten Theorien nur Eigenschaften der Wörter oder der Konzepte beschrieben. Ob diese Beschreibung mit irgendwelchen tatsächliche Eigenschaften der Elemente der Extension zusammenhängt, mag dabei für unsere Zwecke gleichgültig sein, insofern nur die Beschreibung der Konzepte intersubjektiv nachprüfbar ist. Es liegt dieser Außerachtlassung der tatsächlichen Eigenschaften der Denotate eine abgeschwächte Form der Haltung zugrunde, die Dahlgren (1988) als "naiven Realismus" bezeichnet: "Realismus" deshalb, weil die Annahme zugrundeliegt, daß die durch einen Ausdruck bezeichneten Gegenstände natürliche Klassen bilden; naiv deshalb, weil Dahlgren lexikalische Konzepte als naive Theorien der Sprecher über die Welt ansieht (Dahlgren 1988: 35f). Sprecher verhalten sich also so, als hätten die Entitäten, für die sie eine gemeinsame Bezeichnung haben, auch tatsächlich gemeinsame Eigenschaften, und als bezeichneten sie damit real existierende Klassen. Im Gegensatz zu Dahlgren, die annimmt, daß es in der Welt tatsächlich stabile Klassen von Objekten gibt (Dahlgren 1988: 68), soll es mir jedoch als Grundannahme genügen, daß es sich aus der Sicht einer Sprachgemeinschaft so verhält; ob es der Fall ist, daß diese Kategorien tatsächlich objektiv existierende Klassen von Objekten bezeichnen, mag dahingestellt bleiben. Bezüglich des Gegensatzpaares Realismus — Nominalismus soll also kein Standpunkt bezogen werden.

Ebenfalls eine Grundhaltung der Indifferenz will ich gegenüber der Frage nach der psychologischen Realität einer Bedeutungsbeschreibung einnehmen. Eine Theorie der Wortsemantik sollte, zumindest hinsichtlich ihrer Verwertbarkeit für die Computerlinguistik, daran gemessen werden, ob sie das Verhalten von Ausdrücken annähernd richtig beschreibt, und — im Sinne einer effizienten Kodierung — ob sie nachvollziehbare Bedeutungbeschreibungen liefert, nicht jedoch daran, ob sie irgendwelchen ohnehin kaum bekannten mentalen Mechanismen entspricht[5].

2.2.1 Aspekte der Wortbedeutung: Abgrenzungprobleme

Jede Theorie der Wortsemantik sieht sich mit dem Problem der Definition des Umfangs ihres Beschreibungsgegenstandes konfrontiert. Vor der Beschreibung der Bedeutung einzelner

[5]Dies ist kein Einwand gegen die Berücksichtigung der psycholinguistischen Forschung auch in anderen sprachwissenschaftlichen Strömungen. Ergebnisse aus der Psycholinguistik sollten in der theoretischen Linguistik durchaus rezipiert werden. Nur ist hiervon die Frage nach der Verpflichtung zur direkten Modellierung dieser Regularitäten in der lexikographischen Sprachbeschreibung scharf abzugrenzen.

Lexeme in einem Lexikon gleich welcher Art muß der Gegenstand dieser Beschreibung zumindest umrissen werden. Außerdem sollte geklärt werden, welche Distinktionen bezüglich möglicher Bestandteile der Bedeutung einer lexikalischen Einheit zu treffen sind. In den folgenden Abschnitten werden zu diesem Zweck die Unterscheidungen Wortbedeutung versus Weltwissen, analytische versus kontingente Aussagen und Denotation versus Konnotation angesprochen und auf ihre Tauglichkeit zur lexikographischen Beschreibung von Bedeutung hin überprüft.

2.2.2 Wortbedeutung und Weltwissen

Eine grundlegende Frage im Zusammenhang mit der Bedeutungsbeschreibung von lexikalischen Einheiten ist, ob man von der Bedeutung eines Wortes im linguistischen Sinn sprechen kann und ob diese 'linguistische' Bedeutung von einer Repräsentation des Wissens abgrenzbar ist, das mit Begriffen im Zusammenhang steht. In der computerlinguistischen Modellierung läuft die Frage darauf hinaus, ob es zweckmäßig ist, die semantische Repräsentation im Lexikon von einer allgemeinen Wissensbasis strikt zu trennen. Während in den lange Zeit dominierenden strukturalistischen Ansätzen meist Position I vertreten wurde, hat in den letzten Jahren ein Umbruch zugunsten der nachfolgend als Position II geschilderten Ansicht stattgefunden (vgl. Cruse 1990).

Die Extrempositionen stellen sich folgendermaßen dar:

A Lexikalisches Wissen ist trennbar von enzyklopädischem Wissen.

Die Bedeutung eines Wortes im linguistischen Sinn ist abgrenzbar von dem mit einem Konzept assoziierten Weltwissen. Eine damit oft zusammenhängende Annahme ist, daß die Semantik eines Wortes, im Gegensatz zum enzyklopädischen Wissen, das ja beinahe beliebig erweiterbar ist, vollständig beschrieben werden kann.

Die Abgrenzung von linguistischer Bedeutung und Weltwissen ist mit Sicherheit nicht trivial. Doch bestehen zahlreiche Semantiker nach wie vor auf dieser Trennung (so etwa Strukturalisten der Tübinger Schule; vgl. die Beiträge von Geckeler in Lutzeier (Hrsg.) 1993). Das Beharren auf dieser Position ist verständlich, bringt doch ihre Aufgabe mit sich, daß eine linguistische Theorie der Wortbedeutung im strengen Sinn nicht mehr möglich ist und so die Außengrenzen des Fachs für diesen Bereich verwischt werden.

Eine sehr zurückhaltende Unterscheidung zwischen Lexikon und enzyklopädischem Wissen findet sich in folgendem Zitat von Cruse:

> "I interpret encyclopedic knowledge as everything that is represented in the conceptual network. This includes knowledge of, for instance, logical and conceptual properties ... Dictionary information, for me, is the information that is peculiar for words. Words and concepts exist in an intimate and mutually dependent relationship, but I would like to suggest that it is profitable to regard words as having properties, including semantic properties, that are not at the same time properties of their associated concepts" (Cruse 1990: 395f)

Er beharrt zwar auf der Abtrennung des enzyklopädischen Wissens, bezeichnet es allerdings als eine der wichtigsten Informationen in einem Lexikoneintrag, daß das mit einem Wort assoziierte Konzept genannt wird. Die speziell mit einem Wort selbst und nicht mit einem Konzept assoziierten semantischen Eigenschaften, die er in den folgenden Abschnitten herausarbeitet, zeigen zwar, daß es eher mit dem Wort und eher mit dessen Bedeutung assoziierte Bedeutungsaspekte gibt, es gelingt ihm aber nicht, eine scharfe Grenze zwischen beiden aufzuzeigen.

Keiner der Verfechter dieser Position war also bisher in der Lage, eine praktisch umsetzbare Grenze zu ziehen zwischen der Bedeutung im linguistischen Sinne und dem Weltwissen;

während sich bestimmte semantische Eigenschaften eines Lexem-Konzept-Ganzen dem einen oder anderen Bereich zuordnen lassen, finden sich zahlreiche andere Merkmale, deren Zugehörigkeit zum 'Kern' der Wortbedeutung keineswegs trivial ist.

II Lexikon = Enzyklopädie

Die Schwierigkeiten der Trennung von linguistischer Bedeutung und Weltwissen haben dazu geführt, daß in zahlreichen neueren Arbeiten zur linguistischen Semantik diese Unterscheidung ganz fallen gelassen wurde. So hat bereits Haimann (1980) gezeigt, daß keiner der Vorschläge zur Trennung der beiden Bereiche befriedigen kann.

Vor allem im Bereich der kognitiven Linguistik ist die Nicht-Unterscheidbarkeit von linguistischem und enzyklopädischem Wissen in bezug auf die Semantik eine zentrale Annahme. Diese Extremposition faßt einer der Hauptvertreter der kognitiven Linguistik so zusammen:

> "The distinction between semantics and pragmatics (or between linguistic and extralinguistic knowledge) is largely artifactual, and the only viable conception of linguistic semantics is one that avoids such false dichotomies and is consequently encyclopedic in nature." (Langacker 1987: 154)

In einen enzyklopädischen Ansatz lassen sich prinzipiell alle Methoden der Bedeutungsbeschreibung integrieren, ja er fordert im Grunde einen mehrdimensionalen Ansatz: eine völlig elaborierte Theorie der Wortsemantik wäre in diesem Sinne immer auch zugleich eine Beschreibung der Welt, wie sie von einer Sprechergemeinschaft gesehen wird.

In der computerlinguistischen Wortsemantik hat eine Umsetzung des Postulats der Nicht-Unterscheidbarkeit zur Folge, daß eine Abtrennung der Sprachbedeutungsrepräsentation von der Wissensrepräsentation im strengen Sinne nicht möglich ist. Die Bezeichnung 'lexikalische Wissenbasis' (vgl. Amsler 1994) ist in diesem Fall der angemessene Begriff für die Semantikkomponente eines elektronischen Wörterbuchs.

2.2.3 Denotation und Konnotation

Die Dichotomie Denotation versus Konnotation[6] beinhaltet zunächst die Unterscheidung zwischen dem Kern der Bedeutung eines Ausdrucks und den peripheren Bedeutungsbestandteilen. In dieser Verwendung ist sie eng verwandt mit der Unterscheidung zwischen lexikalischem und enzyklopädischem Wissen. Sie wird allerdings häufig auch verwendet, um das von allen Sprechern einer Sprachgemeinschaft geteilte Bedeutungswissen von regionalen, individuellen oder okkasionellen Bedeutungskomponenten abzugrenzen.

Unter zentralen Bedeutungsbestandteilen sind solche zu verstehen, die eine möglichst große Relevanz für die Beschreibung linguistischer Phänomene besitzen, wie unter anderem für die Modellierung von Selektionspräferenzen einer möglichst großen Zahl von Operatoren[7] und zur Erklärung anaphorischer Wiederaufnahme durch sinnverwandte Lexeme. Kriterien für eine scharfe Abgrenzung von Denotation und Konnotation im Sinne der Zentralität von Bedeutungsaspekten lassen sich allerdings kaum aufstellen. Anstatt einer direkten Umsetzung der Dichotomie in die Lexikographie läßt sich diese Unterscheidung auf die Praxis der Bedeutungsbeschreibung übertragen, indem einigen Prinzipien Folge geleistet wird:

- Zunächst sollten zentrale Aspekte der Bedeutung beschrieben werden.
- Regionale und andere Sonderbedeutungen müssen als solche markiert sein.

[6] Eine relativ neue Arbeit zum Thema ist Garza—Cuarón (1991).

[7] Unter 'Operatoren' und 'Operanden' werden im folgenden alle sprachlichen Ausdrücke verstanden, die im weitesten Sinne in einer Bestimmungrelation zueinander stehen. Eine Operator-Operand-Beziehung besteht in diesem Sinne zwischen dem Verb und seinen Komplementen, Adjektiven und dem modifizierten Nomen.

- Individuelle und okkasionelle Bedeutungen sollten nicht in das Wörterbuch aufgenommen werden.

2.2.4 Analytische versus kontingente Aussagen

In den vorigen beiden Abschnitten wurde deutlich, daß die Abgrenzung verschiedener Aspekte innerhalb von Wortbedeutungen Schwierigkeiten bereitet. Um die Problematik der Abgrenzung innerhalb der Bedeutungsbeschreibung für einzelne Wörter aus einem weiteren Blickwinkel zu beleuchten, möchte ich zuletzt noch auf das Problem der Abgrenzung von kontingenten und analytischen Aussagen — ein zunächst sprachphilosophisches Thema — eingehen.

In der philosophisch-logischen Tradition der Sprachbeschreibung hat die Unterscheidung zwischen analytischen und nicht-analytischen Aussagen über einen Begriff eine lange Tradition. Die Unterscheidung ist vor allem in die Modellierung der Wortbedeutung in der logischen Semantik eingegangen. Mit den 'Bedeutungspostulaten' (vgl. Dowty 1979) steht in der modelltheoretischen Semantik ein Formalismus zur Formulierung analytischer Wortinhalte zur Verfügung.

Eine analytische Aussage drückt eine dem Begriff immanente Eigenschaft aus, und somit einen Teil der lexikalischen Bedeutung eines Wortes. Immer wieder zitierte Beispiele für analytische Aussagen sind:

1 Ein Junggeselle ist unverheiratet

2 Ein Junggeselle ist männlich.

Als Beispiel für im Sinne dieser Unterscheidung nicht-analytische Folgerungen kann man beispielsweise die folgende Aussage heranziehen:

3 Vögel können fliegen

Im Gegensatz zu den ersten beiden Aussagen gilt für Satz (3) nicht, daß er notwendigerweise wahr ist. Er beinhaltet vielmehr eine generische Aussage, die nicht für alle Vögel gilt, denn es sind sowohl Sachverhalte (Flügellahmheit) als auch untergeordnete Begriffe (z. B. *Pinguin*) beschreibbar, für welche die Feststellung nicht wahr ist.

Zunächst scheint es, als ob sich für die kategorische Distinktion von analytischen versus nicht-analytischen Aussagen dieselben Probleme ergeben, wie bei der schon genannten Trennung von Weltwissen und lexikalischem Wissen, und daß es letztlich auch hier um eine Abstufung von Zentralität in der Bedeutung eines sprachlichen Ausdrucks geht, womit eine klare, operationalisierbare Abgrenzungsmethode auch in diesem Fall nicht formuliert werden kann. Im Rahmen der Sprachphilosophie hat sich Quine (1951) sehr ausführlich mit der Möglichkeit zur Unterscheidung von analytischen und nicht-analytischen Aussagen beschäftigt. Er kommt letzten Endes aufgrund von Schwierigkeiten bei der Abgrenzung dazu, diese Unterscheidung im strikten philosophisch-logischen Sinne zu verwerfen. Es ist allerdings unbestreitbar, daß es in nicht wenigen Fällen, wie etwa den Implikationen des schon genannten Begriffs *Junggeselle,* kaum denkbare Kontexte gibt, in denen die analytisch genannten Folgerungsbeziehungen aufgehoben werden könnten. Hier besteht offensichtlich die Möglichkeit, Begriffe in eine Ableitungsrelation zueinander zu setzen. Im Falle des Beispiels *Junggeselle* läßt sich sogar eine wechselseitige Folgerungsbeziehung formulieren. Wird die Bedeutung von *erwachsen* zu der von *unverheiratet* und *männlich* hinzugenommen, liegt eine brauchbare Definition des Begriffs *Junggeselle* vor.

Folgerungsbeziehungen dieser Art gelten für zahlreiche Lexeme des Wortschatzes. So lassen fast alle movierten weiblichen Formen analytische Aussagen nach dem folgenden Muster zu:

4 Eine Arbeiterin ist weiblich

Es ist von daher trotz der sprachphilosophischen Problematik möglich, sich die Unterscheidung von analytischen und kontigenten Aussagen für die semantischen Beschreibung eines Lexikons zunutze zu machen, indem analytische Bedeutungsbestandteile als Folgerungsbeziehung zwischen Lexemen modelliert werden.

Anders modelliert werden müssen Default-Eigenschaften, d. h. Folgerungsbeziehungen, die nur typischerweise gelten, wie etwa die genannte Eigenschaft von Vögeln, fliegen zu können. Im Rahmen der klassischen Logik sind solche Folgerungen nicht zu modellieren — die logische Beschreibung solcher Bedeutungsinhalte erfordert eine Logik mit modalen Operatoren oder eine Defaultlogik.

2.2.5 Modellierung von Wortbedeutung

Ich glaube, im Vorigen zumindest gezeigt zu haben, daß die geschilderten Unterscheidungen zwischen lexikalischer und enzyklopädischer Bedeutung und auch von Denotation und Konnotation — man mag zu ihrer Berechtigung sagen was man will — zumindest nicht direkt in die lexikographische Praxis umsetzbar sind. Ich ziehe daraus die Konsequenz, diese Unterscheidungen nicht in das Beschreibungsinventar aufzunehmen.

Eine Möglichkeit, die Beschreibung verschiedener Bestandteile der Wortbedeutung in gewissem Umfang zu systematisieren, ergab sich durch die Betrachtung der Unterscheidung zwischen analytischen und nicht-analytischen Aussagen. Dies zeigt, daß trotz möglicherweise nicht immer präzise anzugebender Kriterien innerhalb der Bedeutung lexikalischer Einheiten Typen von Bedeutungsaspekten aufgezeigt werden können. Es wird im folgenden unterschieden zwischen definitorischen, zentralen und Default-Aspekten. Hinzu kommen die peripheren Aspekten der Bedeutung eines Wortes.

So gibt es Komponenten in der Bedeutung mancher Lexeme, denen eine definitorische Kraft zukommt (so die Zuordnung des Adjektivs *unverheiratet* zu *Junggeselle*). Diese Merkmale sind im obigen Sinne analytische Bedeutungsbestandteile und zentral mit der richtigen Anwendung eines Ausdrucks verbunden. Diese definitorischen Bedeutungsbestandteile sind Aspekte der Wortbedeutung, die ohne Einschränkungen an untergeordnete Konzepte vererbt werden, und die sich in Beschreibungssprachen als nicht-aufhebbare Folgerungsbeziehungen modellieren lassen [8]. Nicht für alle Lexeme lassen sich solche definitorischen Bedeutungsbestandteile festlegen.

Ein zweiter Typ von zentralen Aspekten der Wortbedeutung sind die nichtanalytischen zentralen Bedeutungsaspekte. Die meisten Fälle der hyponymischen Unterordnung eines Begriffes unter einen anderen (z. B. *Vogel* zu *Lebewesen*)[9] gehören zu dieser Gruppe. Diese Komponenten sind nicht in derselben Weise essentiell für die richtige Anwendung eines

[8] Man vergleiche die semantischen Relatoren in Kapitel 3. Mit einigen von ihnen (z. B. JUV) werden nicht aufhebbare analytische Bedeutungsbestandteile modelliert.

[9] Eine solche Zuordnung als analytisch zu klassifizieren, ist meiner Ansicht nach nicht möglich: Wie Putnam (1975) überzeugend dargelegt hat, ist eine Zuordnung, wie sie in der Aussage *Eine Katze ist ein Lebewesen* vorgenommen wird, nicht im strikten Sinne analytisch wahr, da beispielsweise die Entdeckung der Tatsache, daß alle Katzen vom Mars ferngesteuerte Roboter sind — zugegebenermaßen ein etwas unwahrscheinliches Szenario — diese Aussage falsch machen würde, ohne daß die Kategorie 'Katze' und damit die extensionale Bedeutung des Ausdrucks *Katze* deshalb aufgegeben werden müßte. Allerdings sind solche hyponymischen Unterordnungen tatsächlich sehr stabil.

Ausdrucks wie die zuerst genannten, denn sie sind, wenn auch nicht extrem kontextabhängig, so doch zumindest unter bestimmten Umständen kontextuell aufhebbar.

Ein dritter Typ von relativ zentralen Bestandteilen der Wortbedeutung soll als Default-Komponenten (etwa die Zuordnung der von *flugfähig* denotierten Eigenschaft zu *Vogel*) bezeichnet werden. Diese Default-Aspekte sind nicht definitorisch und somit nicht essentiell für die richtige Anwendung eines Ausdrucks. Default-Bedeutungskomponenten sind aber immer dann deduzierbar, wenn der Kontext nichts gegenteiliges impliziert; sie werden an untergeordnete Begriffe nur bedingt vererbt. Sie können in eine Beschreibungssprache als Defaultregeln integriert werden[10].

Zu diesen Bedeutungsbestandteilen kommen peripherere Bedeutungsaspekte, die für die semantische Beschreibung in einem Lexikon allerdings nur dann relevant sind, wenn sie von einem großen Teil einer Sprechergemeinschaft geteilt werden. Der Beitrag solcher Konnotationen zur Lexembedeutung und auch deren Relevanz für die Textstruktur sollten jedoch nicht unterschätzt werden. Letzten Endes müssen auch diese Aspekte der Bedeutung in eine umfassende Wortsemantikkodierung integriert werden, denn auch sie sind relevant für lexikalische Selektion[11], und auch ihre Modellierung ist essentiell für Anwendungen in der Computerlinguistik. Es ist beispielsweise nicht möglich, die weitgehend bedeutungsgleichen Lexeme *Velo* und *Fahrrad* ohne Berücksichtigung des regionalen Sprachgebrauch zu verwenden; *Glotzophon, Flimmerkiste* und *Fernseher* sind unter Berücksichtigung der pejorativen Konnotation bei ersteren Lexemen zu verwenden und damit ebensowenig beliebig austauschbar. Vergleichbares gilt für *Piano* und *Klavier*.

Die genannten Distinktionen verschiedener Aspekte der Bedeutung sind nicht immer kategorisch aufzufassen. Sie sind als Eckpfeiler einer semantischen Beschreibung gedacht. Es gibt zwischen ihnen sicherlich Übergangsphänomene.

Zusammenfassend läßt sich folgern: Es ist möglich, einige Kriterien für Unterscheidungen zwischen verschiedenen Teilen der Wortbedeutung zu finden. Es läßt sich aber — zumindest im Normalfall — nicht festlegen, wann die semantische Kodierung eines Lexems vollständig ist. Die Bedeutungskodierung sollte von daher mit möglichst zentralen Aspekten der Bedeutung beginnen und problemlos erweiterbar sein.

Die genannten Typen von Bedeutungsaspekten sind mit jedem der folgenden paradigmatischen Paradigmen der Wortsemantik kompatibel, d. h. jede Beschreibungseinheit, sei dies nun ein Merkmal, eine semantische Relation zu einem anderen Lexem oder auch eine Prototypenbeschreibung, kann innerhalb der Bedeutungsstruktur einen unterschiedlichen Stellenwert haben.

Die im folgenden beschriebenen Hauptparadigmen der lexikalischen Semantik lassen sich somit nicht a priori einer der beiden vorhin genannten Extrempositionen bezüglich der Struktur von Lexembedeutungen zuordnen. Doch beruhen klassische Ansätze der zunächst vorgestellten Komponentenanalyse auf dem Grundsatz der vollständigen Analysierbarkeit einer Wortbedeutung , und setzen durch diesen Vollständigkeitsanspruch die Möglichkeit der Trennung des semantischen Wissens vom Weltwissen voraus. Dies ist allerdings keine Grundsatzbedingung jeglicher semantischer Dekomposition. Es wird deutlich werden, daß die Beschreibung durch Bedeutungsbestandteile an sich ein sehr flexibles und unter ganz verschiedenen theoretischen Grundvoraussetzungen anwendbares Verfahren ist. Wie bereits deutlich wurde, eignet sich die Komponentenanalyse zumindest zur Annäherung an die Bedeutungsbeschreibung semantisch komplexer Begriffe wie *Junggeselle*, die sich relativ einfach als Summe verschiedener Teilbedeutungen darstellen lassen.

[10] Ein Beispiel für die Modellierung von Default-Bedeutungsaspekten ist die Kodierung der bedingten Unterordnung (vgl. den Abschnitt zur Hierarchisierung in Kapitel 3).

[11] Vgl. zum Beispiel den Relator pejorativ, der in Kapitel 3 beschrieben wird.

Die Beschreibung von Sinnrelationen verhält sich gegenüber beiden Ansätzen neutral. Obwohl sie einer strukturalistischen Tradition entspringt, kann eine Beschreibung von Relationen zwischen Wortbedeutungen in jeder Art von wortsemantischer Theorie als grundlegend für die Bedeutungsbeschreibung angesehen werden.

Dagegen können die Theorien, die von einem Prototypenbegriff ausgehen, eher dem Ansatz zugerechnet werden, der fließende Übergänge in der Struktur von Lexembedeutungen annimmt. Dieses Konzept stammt ursprünglich aus der psychologischen Forschung und wurde in der Sprachwissenschaft zunächst hauptsächlich von der Strömung der kognitiven Linguistik rezipiert, zu deren Programm es ja von vornherein gehörte, die klassischen Paradigmen der strukturalistischen Sprachwissenschaft in Frage zu stellen. Bei der Beschreibung des Begriffs des Prototypen geht es primär um die menschliche Kategorisierung und erst sekundär um die Konsequenzen für die Linguistik; dennoch hat die Prototypentheorie inzwischen eine bedeutende Stellung in der Linguistik erlangt, und dies sogar über den Bereich der Wortsemantik hinausgehend. Eine Diskussion dieses Begriffes kann als exemplarisch für den Einfluß der Kognitionswissenschaften auf die wortsemantische Theoriebildung in den letzten beiden Jahrzehnten angesehen werden.

Als letzte Kategorie von Beschreibungsmethoden in der lexikalischen Semantik sollen Ansätze diskutiert werden, in denen versucht wird, die Semantik eines Wortes über dessen sprachliche Umgebung zu erfassen. Diese syntagmatischen Ansätze beschreiben das Verhalten von Worten und Wortformen in Kontexten und sind damit bezüglich des Umfangs der Bedeutungskodierung nicht festgelegt. Sie können auch als Mittel zur Determinierung der Information in paradigmatischer Hinsicht gesehen werden, da paradigmatische Ansätze ja ohne syntagmatische Information nicht auskommen könnten. Andererseits zeigen die Arbeiten über Kollokationen in großen Korpora, daß Versuche zur statistischen Erfassung von Bedeutung, die auf linguistische Vorabinformation verzichten, die vielschichtigen Regularitäten, denen die Kombinatorik von Lexemen gehorcht, auf problematische Weise vermischen.

2.3 Paradigmen lexikalischer Semantik

2.3.1 Komponentenanalyse

Die Komponentenanalyse ist ein im Rahmen der strukturalistischen Semantik (vgl. z. B. Hjelmslev 1959) entstandener Ansatz[12], der zeitweise das beherrschende Paradigma in der lexikalischen Semantik war, und der auch heute noch, freilich in modifizierter Form, große Bedeutung besitzt.

Die Theorie der Komponentenanalyse geht von der Grundannahme aus, die Bedeutung eines Lexems sei beschreibbar mit Hilfe semantischer Merkmale, in welche die Bedeutung zerlegt werden kann. Ausschließlich auf Dekomposition gestützte Theorien der Wortsemantik gehen davon aus, daß die Wortbedeutung durch die Angabe der einzelnen Merkmale vollständig beschreiben werden kann. Nach einer noch idealisierteren Ansicht sind diese kleinsten Einheiten zudem alle oder zumindest zu einem nicht geringen Teil universell, d. h. es gibt eine endliche Menge semantischer Primitiva, in die jeder Begriff in jeder beliebigen Sprache zerlegt werden kann. Ein solcher universalistischer Ansatz findet sich etwa in Katz (1972). Wären sie praktisch umsetzbar, würden diese Annahmen die Möglichkeit einer (wort-)semantischen Interlingua postulieren.[13]

[12] Zur Geschichte der Merkmalssemantik bis in die siebziger Jahre vgl. Lyons (1977: 317f.)

[13] Bisher wurde allerdings nur für wenige Wortschatzbereiche die Existenz von Universalien aufgezeigt. Dies gilt etwa für den Bereich der Farbbezeichungen in verschiedenen Sprachen, wobei die universellen Aspekte in diesem Fall in hohem Maße durch die Funktion des menschlichen Perzeptionsapparates bedingt sind (vgl.

Die Bedeutungskomponenten werden, entsprechend dem Vorgehen in der Phonologie, auch in der wortsemantischen Komponentenanalyse durch Minimalpaarbildung ermittelt. Hierzu werden in der bestimmte Wortfelder (s. u.) systematisch untersucht.

Es werden in Dekompositionstheorien häufig verschiedene Typen von Merkmalen angenommen. So unterscheidet der europäische Strukturalismus zwischen den auf ein bestimmtes Feld bezogenen 'Sememen' und den für große Bereiche des Wortschatzes wichtigen 'Klassemen' (wie ¡belebt¿/ ¡unbelebt¿, vgl. Lyons 1977: 326). In der Theorie von Katz/ Fodor (1963) werden ebenfalls zwei Typen von Bedeutungskomponenten unterschieden. Die 'Distinguishers' beschreiben den unsystematischen Rest innerhalb einer Wortbedeutung, der nach der Aufschlüsselung in die 'Markers' übrigbleibt, die deren systematischen Teil repräsentieren.

Die linguistische Literatur zu diesem Thema behandelt meist nur einige ausgewählte Beispiele. Standard sind die schon genannten Verwandtschaftsbezeichnungen sowie die Lexeme *bachelor/ Junggeselle* (nach Katz/ Fodor 1963) und *kill/ töten* (nach McCawley 1974), die in der Diskussion über Dekomposition immer wieder aufgegriffen werden. Bei diesen Lexemen ist die Bedeutungszerlegung noch relativ einfach vorzunehmen. Sobald man jedoch versucht, die Komponenten der Bedeutung beliebiger Lexeme anzugeben, stößt man auf unlösbare Probleme. Doch selbst in den bekannten und häufig zitierten Beispielen der Beschreibung durch Bedeutungsbestandteile läßt sich deren Problematik mühelos aufzeigen.

Ein klassisches, und gleichzeitig fragwürdiges Beispiel ist die Einteilung des Wortfeldes der Sitzgelegenheiten im Rahmen der Wortfeldtheorie durch Pottier (1964). Die Bedeutungen von *Stuhl* und *Sessel* werden dadurch unterschieden, daß letzterer das Merkmal ¡+ Armlehne¿ zugewiesen wird, das bei ersterem Wort negativ spezifiziert ist. Nun wird bei einigem Nachdenken schnell klar, daß dieses Merkmal keineswegs den definitorischen Unterschied zwischen beiden Begriffen ausmacht, da sowohl ein Stuhl mit als auch ein Sessel ohne Armlehnen vorstellbar ist. Das Merkmal ist also allenfalls ein Defaultmerkmal und hat also keine distinktive Bedeutung, Es scheint eher das typische Denotat beider Begriffe zu beschreiben.

Auch die gerne als Musterbeispiel für Dekomposition herangezogene Bedeutung von *Junggeselle* (zuerst in Katz/ Fodor 1963) wurde von Lakoff (1987: 69) im Gefolge von Fillmore (1982) zumindest teilweise in Frage gestellt, indem er auch hier Prototypikalitätseffekte aufzeigte, die eine Prädikation wie *Tarzan ist ein Junggeselle* zumindest fraglich machen. Hierdurch wird zumindest die Vollständigkeit der Dekomposition dieses Begriffs zweifelhaft[14].

All diese Einwände haben eine Reihe von Grundpositionen der Komponentenanalyse in erheblichen Maße in Frage gestellt. Als nicht haltbare Annahmen können sicher genannt werden:

- Der Universalitätsgedanke. Dieser postuliert die Universalität aller kleinsten Einheiten, in die eine Bedeutung zerlegbar ist. Selbst wenn diese Universalität gegeben wäre, könnten diese kleinsten Einheiten nicht mit den traditionellen Formen der Dekomposition von Bedeutungen ermittelt werden, da Verfahren nur für einzelne Sprachen vorliegen. Diese Annahme ist somit für die Praxis der Lexikographie zunächst irrele-

Berlin/ Kay 1969). Auch auf dem Gebiet der Verwandtschaftsrelationen ist eine universalistische Auffassung der Grundrelationen relativ unproblematisch (vgl. Lyons 1977: 319f, 333) und auch nützlich für den Vergleich von Verwandtschaftsbezeichnungen in verschiedenen Sprachen. Mit diesen Beispielen sind allerdings die Wortschatzbereiche, für die sich die Universalismushypothese bisher sinnvoll anwenden ließ, wohl weitgehend erschöpft. Allerdings gibt es nach wie vor Versuche zur Aufstellung eines Inventars semantischer Primitiva, so z. B. Goddard (Hrsg.) (1994).

[14]Der kritische Punkt ist hier hauptsächlich die Fragwürdigkeit der Dekompositionsanalyse. Natürlich läßt sich das Konzept zu *Junggeselle* durch eine etwas modifizierte Umschreibung in natürlicher Sprache weiterhin klar umreißen (vgl. Wierzbicka 1990: 348).

vant.

- Vollständige Zerlegbarkeit. Dies impliziert, daß die Bedeutung eines Wortes als mengentheoretische Funktion aus kleinsten Komponenten gesehen werden kann. Diese Hypothese setzt einen scharfen Bedeutungsbegriff voraus. Zudem wird der Tatsache keine Rechnung getragen, daß bei der Zerlegung von Begriffsbedeutungen meist ein großer unanalysierter Rest bleibt.

- Determiniertheit der Extension. Die Merkmale bilden eine intensionale Definition der Klasse von Objekten, die unter die Extension fallen. Diese angebliche Definitheit der Extension ignoriert die Tatsache unscharfer Kategoriengrenzen und interner Strukturierung von Kategorien.

- Operationalisierbarkeit der Ermittlung von Bedeutungskomponenten. Das übliche Vorgehen der Merkmalsermittlung durch Minimalpaarbildung ist problematisch. Insbesondere ist es in vielen Fällen schwierig, ähnliche Distinktionen zwischen verschiedenen Minimalpaaren auf ein und dasselbe Merkmal zurückzuführen. Merkmale können somit nicht vollständig durch Minimalpaarbildung ermittelt werden (vgl. Lüdi 1985: 98).

Die Dekompositionssemantik in ihrer klassischen Form ist also sicher zur Bedeutungsbeschreibung nur eingeschränkt nutzbar. Eine endgültige Menge an Merkmalen ist nicht einmal für eine Einzelsprache, geschweige denn für eine Interlingua aufstellbar. Bedeutungskomponenten sind kein strukturell so einfach zu fassendes Phänomen wie Phonemeigenschaften.

Das muß nicht heißen, daß jegliche Theorie, die von einer Zerlegbarkeit von Wortbedeutungen ausgeht, völlig untauglich zur Bedeutungsbeschreibung sein muß. Keine der genannten extremen Grundpositionen ist eine notwendige Eigenschaft jeder semantischen Dekompositionstheorie. Es ist durchaus möglich, das Phänomen unscharfer Kategoriengrenzen, auf das ich im nächsten Abschnitt eingehe, und auch die Tatsache der prinzipiellen Indeterminiertheit der Grenzen der eigentlich "linguistischen" Bedeutung in ein solches Paradigma zu integrieren. Eine Theorie, die von semantischen Merkmalen ausgeht, ist unter der Voraussetzung, daß sie nicht als Theorie der Gesamtbedeutung angesehen wird, hervorragend zur Formalisierung der Beschreibung bestimmter Bedeutungsbestandteile eines Lexems zu verwenden. Beinahe jede semantische Beschreibung kann als merkmalssemantische Beschreibung aufgefaßt werden, wenn man die Abtrennung einzelner Aspekte der Bedeutung als Darstellung der Bedeutungskomponenten auffaßt. Hat man einen so weiten Merkmalsbegriff, gibt es zu einer Beschreibung von Bedeutungskomponenten kaum eine Alternative. Das Problem ist also nicht zu entscheiden, ob Dekompositionssemantik oder nicht, sondern zu bestimmen, welchen Status innerhalb einer Bedeutungsbeschreibung man den semantischen Merkmalen zubilligen will. In den nächsten Abschnitten wird bei fast allen angesprochenen Paradigmen immer wieder deutlich werden, daß sie mit einer moderaten Dekompositionsannahme durchaus kompatibel sind und daß man wesentliche Bedeutungsaspekte von Lexemen innerhalb dieser Paradigmen durch merkmalssemantische Beschreibung effizient formalisieren kann. Dies heißt, Komponentenbeschreibung nicht mehr als theoretische Grundposition und damit als Theorie der lexikalischen Semantik aufzufassen, sondern als effizienten Kodierungsformalismus für Bedeutung.[15]

[15] Dies wird auch in der in Kapitel 3 geschilderten Kodierung für das CISLEX deutlich werden. Zumindest die Zuweisung eines Lexems zu solchen semantischen Klassen, die als taxonomische Klassen definiert sind, läßt sich auch als Zuweisung eines semantischen Merkmals verstehen, das eine Teilbedeutung des Lexems wiedergibt.

2.3.2 Semantische Felder und Sinnrelationen

Sowohl die Feldtheorie als auch die Theorie der Sinnrelationen entspringen wie schon die eben diskutierte Dekompositionstheorie der Tradition des Strukturalismus[16].

Die Theorie der Sinnrelationen beschäftigt sich mit der Strukturierung des Lexikons durch Sinnrelationen zwischen Lexemen oder Bedeutungen von Lexemen. Zur Beschreibung von Sinnrelationen wird aufgrund semantischer Kriterien ein bestimmter Teil des Vokabulars herausgegriffen, der dann aufgrund der semantischen Beziehungen der in ihm enthaltenen Lexeme strukturiert wird. Ein solcher Teil des Vokabulars, der Lexeme umfaßt, die sich einer gemeinsamen, unspezifischeren Bedeutungseinheit zuordnen lassen, ist ein semantisches Feld.

Aus der Feldtheorie interessiert uns der Grundgedanke, daß zahlreiche semantische Relationen zwischen Wörtern sinnvoll nur beschreibbar sind, wenn zuvor eine semantikbasierte Einteilung des Wortschatzes in Gruppen von Lexemen mit geteilten Bedeutungsbestandteilen erfolgt ist. Die Feldtheorie hat mit denselben Problemen zu kämpfen, wie die anderen strukturalistischen Ansätze zur Wortsemantik auch: die scharfe Abgrenzung von Beschreibungsentitäten, in diesem Fall dem Feld, ist in vielen Fällen unmöglich (vgl. Geckeler 1993: 18, wo zumindest die Problematik der Abgrenzung eines Feldes eingeräumt wird). Die begriffliche Ausformulierung der Feldtheorie — einem semantischen Feld liegt ein sogenanntes Archilexem als komplexe semantische Einheit zugrunde, für die in einer Sprache ein Bezeichner existieren kann oder nicht (vgl. Lüdi 1985: 81) — ändert insofern nichts an dieser Schwierigkeit, als diese 'Archilexem' genannte, gemeinsame Bedeutungskomponente der Lexeme eines Wortfeldes erst anhand des bereits konstituierten Feldes beschrieben werden kann, und durch die Feldtheorie keine ausreichenden Kriterien zur Verfügung gestellt werden, um festzustellen, wie man ein Wortfeld von einer Sammlung von beliebigen, intuitiv bedeutungsähnlichen Lexemen unterscheidet. Für die lexikographische Arbeit ist aber eben diese Frage entscheidend.

Stark vergröbernd und unter Weglassung der üblichen strukturalistischen Terminologie lassen sich aus der Feldtheorie zwei Schlüsse ziehen, die auch für Reihenfolge und Form der Kodierung eines elektronischen Wörterbuchs von Belang sind:

- Der Wortschatz einer Sprache läßt sich in Klasse von Lexemen einteilen, die bestimmte gemeinsame Bedeutungskomponenten besitzen. Häufig enthält eine solche Klasse ein Lexem, das als Überbegriff für die anderen Lexeme verwendet werden kann (in der Wortfeldtheorie Archilexem genannt).

- Die semantische Beschreibung von Lexemen sollte stets in bezug auf andere, bedeutungsverwandte Lexeme erfolgen, und nicht in Isolation.

Während es um die Feldtheorie im engeren Sinn in den letzten Jahren recht still geworden ist (trotz Lutzeier (Hrsg.) 1993)[17].

, bleibt die Theorie der Sinnrelationen der wichtigste Grundpfeiler der linguistischen Wortsemantik jeder Ausprägung. Zudem ist die Beschreibung von Sinnrelationen auch eine zentrale Aufgabe der lexikographischen Arbeit. Im Gegensatz zur Definition und Abgrenzung von semantischen Feldern gibt es für die Determinierung der verschiedenen Sinnrelationen relativ verläßliche sprachliche Tests.

Die wichtigsten Sinnrelationen seien im Anschluß genannt:

Synonymie: Die Synonymie ist eine symmetrische Relation zwischen Lexemen. Ausser zwischen vollständigen Synonymen ist die Synonymie keine transitive Relation. Diese

[16] Hier soll kein vollständiger Forschungsüberblick über die beide Theorien gegeben werden (dazu Lyons 1977: 230-317, Cruse 1986, Lüdi 1985).

[17] Vgl. Geckeler (1993: 11): "Die Wortfeldforschung ist nicht tot, ... aber als quicklebendig und kraftstrotzend, kurz als 'megain' wird man ihre derzeitliche Befindlichkeit auch nicht einschätzen können."

Sinnrelation ist notorisch schwierig zu definieren (vgl. Cruse 1986: 265ff). Das wichtigste Kriterium für Synonymie zwischen zwei Lexemen ist ihre Austauschbarkeit in Sätzen bei gleichbleibender Bedeutung. Bei Synonymenpaaren ist der Austausch der beiden Lexeme, im Gegensatz zur Ersetzung durch ein Hyperonym, auch in negierten Kontexten möglich. Definiert man, um den definitorischen Problemen aus dem Weg zu gehen, Synonymie als vollständige Bedeutungsgleichheit, d. h. vollständige Austauschbarkeit in allen Kontexten, wird sie als Sinnrelation irrelevant, da bis auf wenige Ausnahmen kaum zwei vollständig bedeutungsgleiche Lexeme existieren. Synonymie zwischen zwei Lexemen ist also häufig lesarten- und kontextabhängig. Trotz der Problematik ihrer Definition ist die Synonymierelation eine der grundlegendsten Relationen zur lexikographischen Sprachbeschreibung. (vgl. Sparck Jones 1986). Auch in der maschinellen Lexikographie ist die Synonymie grundlegend für den Aufbau einer semantischen Struktur des Wortschatzes. So baut die weiter unten genauer beschriebene Wortdatenbank WordNet auf der Zusammenstellung synonymer Lexeme auf (vgl. Miller u. a. 1990).

Die Synonymie ist abzugrenzen vom Begriff der Variante (Schreibvariante, phonologische Variante). Als Varianten werden im folgenden völlig bedeutungsgleiche Lexeme bezeichnet, die gleichzeitig eine systematische Formähnlichkeit aufweisen, wie z. B. *Telephon* und *Telefon*.

Antonymie/ Opposition: Die Relation des Bedeutungsgegensatzes ist untergliederbar in verschiedene Sinnrelationen, die z. T. nur für bestimmte Wortarten Relevanz haben: Antonyme (im engeren Sinne) (Adjektive, Nomina[18]), Komplementäre (Adjektive, Nomina), Reversiva (Verben), Konversen (Nomina, Verben, Adjektive). Während die anderen Sinnrelationen — außer der Synonymie — als Relationen zwischen Bedeutungen von Lexemen gesehen werden können, handelt es sich bei der Antonymie um eine Relation, bei der die Lexemform eine große Rolle spielt, was unschwer daran erkennbar ist, daß in Antonymenpaaren nie ein Lexem gegen ein anderes, bedeutungsgleiches, ausgetauscht werden kann (vgl. Miller u. a. 1993: 7), während dies bei allen anderen Sinnrelationen möglich ist.

Hyponymie/ Hyperonymie. Die Hyponymie ist die Unterordnungsrelation zwischen Lexembedeutungen, und damit die neben der Synonymie wichtigste Sinnrelation, um eine lexikographische Gliederung des Wortschatzes vorzunehmen (vgl. Cruse 1986: 88-92). Im Gegensatz zur Synonymie ist sie nicht symmetrisch, aber transitiv. Die konverse Relation zu Hyponymie ist die Hyperonymie. Ist das Lexem X Hyponym zu Y, ist folglich Y Hyperonym für X. Ist Z ebenfalls (unmittelbares) Hyponym zu Y, sind X und Z Kohyponyme.

Es existieren einige sprachlichen Tests für diese Relation. Der übliche Testrahmen für das Vorliegen einer Hyponymiebeziehung zwischen zwei Nomina *X* und *Y* sind *ein X ist ein Y* oder *ein X ist eine Art Y* (vgl. Lyons 292f). Aufgrund dieser Testrahmen wird die Hyponymiebeziehung in der englischischsprachigen Forschungsliteratur auch IS_A-Relation genannt. Hyponymie kann auch über eine Folgerungsbeziehung zwischen Satzbedeutungen definiert werden (vgl. Lyons 1977: 292-293). Wenn in einer logischen Repräsentation *X(e)* substituiert werden kann durch *Y(e)* — wie etwa die Aussage *ein Vogel fliegt* stets folgert aus *ein Spatz fliegt* — ohne daß der umgekehrte Schluß zulässig ist, liegt eine Unterordnungsbeziehung vor. Eine solche Substitution ist in negierten Kontexten nicht möglich (s. a. Cruse 1986: 88-92).

Die Hyponymierelation zwischen zwei Lexemen ist oftmals nur in einem bestimmten semantischen Bereich oder in einem bestimmten Kontext relevant oder gültig (so wird aus der Sicht eines Biologen — und damit in einem biologischen Text — die Hyponymiebeziehung *Hund - Haustiere* eine weniger große Rolle spielen, und dafür die Kategorie *Hundeartige*

[18] Antonymie im engeren Sinne spielt im Bereich der Nomina keine besonders große Rolle bzw. läßt sich hier meist auf bestimmte Merkmale (die dann wiederum Adjektivbedeutungen entsprechen) zurückführen.

Raubtiere, die in der Allgemeinsprache nicht relevant ist, berücksichtigt werden).

Hyponymie ist eine für diverse sprachliche Regularitäten wichtige Relation:

- Eine Reihe anderer semantischer Relationen wie Synonymie und Opposition lassen sich nur für Kohyponyme beschreiben.
- Selektionsrestriktionen operieren in vielen Fällen auf Mengen von kohyponymen Lexemen.
- Bei der anaphorischen Wiederaufnahme von bereits über einen sprachlichen Ausdruck in den Diskurs eingeführten Entitäten werden häufig Hyperonyme zu diesem Ausdruck verwendet.

Die Hyponymiebeziehung ist damit — neben der Synonymie — die Bedeutungsrelation, die für die lexikographische Strukturierung großer Teile des Wortschatzes am grundlegendsten ist.

Meronymie: Teil-Ganzes Relation, auch Partonymie genannt. Die Meronymierelation ist zentrale Komponente der Bedeutung einiger Lexeme (z. B. *Dach* als Teil eines Gebäudes), bei der Bedeutung anderer Lexeme spielt sie zwar eine Rolle in der Beschreibung der Semantik, ist aber nicht notwendiger Bedeutungsbestandteil (etwa *Blatt* als Teil eines *Schreibblocks*). Für wieder andere Lexembedeutungen spielt sie keine Rolle (z. B. bei Tierbezeichnungen). Tversky (1990: 342f) stellt fest, daß in bestimmten Bereichen des Wortschatzes die meronymische Struktur wichtiger wird als die Hyponymierelation (etwa im Bereich der Körperteile). Dies muß bei Versuchen zur Aufstellung einer Gliederung für den Gesamtwortschatz insofern berücksichtigt werden, als meronymisch strukturierte Wortschatzbereiche dabei gesondert behandelt werden sollten.

Es lassen sich für bestimmte Wortschatzbereiche meronymisch strukturierte Hierarchien mit bis zu vier Ebenen aufstellen (Cruse 1986: 157-180, er benutzt als Beispiel wiederum die Körperteile). Die Meronymierelation ist dabei im Gegensatz zur Hyponymierelation nicht prinzipiell transitiv.

Die Meronymierelation wird häufig gesehen als ein Überbegriff für unterschiedliche Relationen mit gewissen gemeinsamen Merkmalen (hierzu Winston/ Chaffin/ Hermann 1987). So ist etwa ein *Baum* Teil eines *Waldes* in einem anderen Sinn als der *Arm* Teil des *Körpers* ist.

Meronymie ist für die Analyse eines nicht geringen Prozentsatzes von Nominalkomposita mit einem nominalen Erstglied relevant. So in den Lexemen *Haustür*, *Hundebein, Tannenwald* etc., wo zwischen den Gliedern eine Teil-Ganzes-Beziehung besteht.

Keine semantische Beschreibung des Wortschatzes in einem Lexikon kann auf den Rückgriff auf die bisher genannten Sinnrelationen in der einen oder anderen Form verzichten. Sie nehmen v. a. unter folgenden Aspekten eine zentrale Stellung in der Bedeutungsbeschreibung ein:

- Eine erste Gliederung des Wortschatzes in Gruppen von bedeutungsähnlichen Wörtern. Hierzu ist neben der Synonymie vor allem die Hyponymierelation von Belang; wie man in späteren Abschnitten sehen wird, hat sie insbesondere für die Gliederung des Allgemeinwortschatzes eine herausragende Bedeutung.
- Das Hyperonym und teilweise auch die Hyponyme eines Lexems gehören unbedingt zu seinen zentralen Bedeutungskomponenten.
- Auch die anderen semantischen Relationen, insbesondere die Meronymierelation, sind bei einigen Lexemen wichtige Bestandteile der Bedeutung; ohne deren Berücksichtigung ist die Beschreibung der Kernbedeutung vieler Lexeme nicht möglich.

2.3.3 Herausforderung strukturalistischer Ansätze: Die Prototypentheorie

Ein elektronisches Wörterbuch muß sich sicher nicht der Forderung nach psycholinguistischer Adäquatheit stellen. Dies gilt umso mehr, als die Struktur des mentalen Lexikons trotz großer Fortschritte auf dem Gebiet der Psycholinguistik alles andere als geklärt ist. Dennoch lassen sich Ergebnisse aus diesem Forschungsfeld in einem Überblick über die Wortsemantik in der theoretischen Linguistik nicht einfach ignorieren, denn einige wesentliche Anregungen zur lexikalischen Semantik in den letzten beiden Jahrzehnten, die ihren Ursprung in der kognitiven Psychologie haben, sind in der Theorie der linguistischen Wortsemantik intensiv rezipiert worden und haben in der einen oder anderen Form auch Eingang in die computerlinguistische Bedeutungskodierung gefunden.

Eines der wichtigsten Paradigmen aus der kognitiven Semantik ist zweifelsfrei die Prototypentheorie[19]. Sie kommt ursprünglich aus der Kognitionspsychologie (Rosch 1975) und wurde in der Linguistik rezipiert als Gegentheorie zu den damals vorherrschenden Annahmen der strukturelle Wortsemantik [20], insbesondere der Komponentenanalyse, deren Vertreter glaubten, Wortbedeutung vollständig und eindeutig durch Bündel von Merkmalen beschreiben zu können[21]. Die Grundidee der Prototypentheorie ist, daß konzeptuelle Kategorien um einen Prototypen als bester Vertreter einer Klasse zentriert sind. Diese Prototypenzentriertheit von Kategorien erklärt deren vage Grenzen (man vergleiche z. B. die Abgrenzung in Kategorien wie Busch/ Baum) und bietet Erklärungsansätze für ihre interne Strukturierung. Prototypeneffekte korrelieren jedoch nicht stets mit unscharfen Kategoriengrenzen. Sie treten auch bei einigen klar abgegrenzten Begriffen, z. B. Vögeln (vgl. Wierzbicka 1990: 350f) und sogar Zahlen (Lehrer 1990: 368) auf , wo jeweils eines oder mehrere 'beste Beispiele', d. h. prototypische Unterkategorien, zu finden sind.

Das wichtigste Verdienst der Prototypensemantik ist sicher die teilweise Erklärung von bis dahin vernachlässigten Phänomenen der Wortbedeutung. Durch die Rezeption dieser Theorie in der lexikalischen Semantik wurde klar, daß Unschärfe und Vagheit von Wortbedeutungen mit Charakteristika der menschlichen Kategorisierungsfähigkeit korrelieren, namentlich mit der Unschärfe von Kategoriengrenzen und der Tatsache, daß viele Kategorien eine interne Typikalitätsstruktur besitzen. Die Berücksichtigung der Erkenntnisse der Prototypentheorie in der Theorie der linguistischen Semantik räumte endgültig mit dem Irrglauben auf, die Grenzen der meisten oder gar aller Wortbedeutungen seien klar zu ziehen und die Extension eines Begriffs eindeutig festzulegen. Es wurde klar, daß der Versuch, Wortbedeutung ausschließlich als Menge notwendiger und hinreichender Merkmale zu beschreiben, zum Scheitern verurteilt war.

[19] Vielleicht aufgrund des Suffixes "-typen" wird der Begriff des Stereotypen (geprägt von Hilary Putnam 1975) gerne in einem Atemzug mit dem Begriff des Prototypen genannt, gleichwohl es sich hier nicht um übereinstimmende Konzeptionen handelt. Putnam, der eine realistische Bedeutungtheorie vertritt, geht davon aus, daß die Sprecher "Stereotypen" von den durch einen Begriff bezeichneten Gegenständen im Kopf haben, diese aber letztendlich nicht die Bedeutung eines Lexems ausmachen, da jene (insbesondere für "natural kind terms") letzlich durch die Festlegung der Extension bestimmt ist. Die putnamschen Stereotypen haben jedoch zunächst keine Implikationen bezüglich des Typikalitätsgrades bestimmter untergeordneter Kategorien oder Referenten. Die Verwandtschaft zum Prototypenbegriff entsteht dadurch, daß die stereotypen Eigenschaften, die einem Konzept zugeordnet werden, meist die Eigenschaften eines typischen Subkonzeptes/ Referenten sind.

[20] In den Beiträgen in Tsohatzidis (1990) wird die Rezeption und die Nützlichkeit der Prototypentheorie in der lexikalischen Semantik unter verschiedenen Aspekten ausgelotet. Zu Kurzschlüssen in der Rezeption s. vor allem Wierzbicka (1990). Zur Anwendung in der lexikalischen Semantik s. a. Lehrer (1990), Cruse (1990).

[21] Cruse führt dieses Deneken auf die Anfänge der überlieferten Reflexion über Sprache zurück und nennt den Ansatz der hinreichenden und notwendigen Merkmalsbündel den 'aristotelischen Standpunkt' (Cruse 1990: 383).

Dieser Aspekt der Strukturierung von Kategorien hat auch Erklärungsfunktionen für die Schwierigkeiten der Hierarchisierung von Begriffen mittels der Hyponymierelation. Auch hier wirken sich die Prototypikalitätsphänomene aus, d. h. ein Begriff kann mehr oder weniger Unterbegriff eines anderen sein.

Die Rezeption der Prototypentheorie in der linguistische Semantik ist allerdings auch voll von Mißverständnissen. So ist die Prototypentheorie sicherlich keine Gesamttheorie der Bedeutung. Prototypikalitätseffekte spielen zwar zweifellos eine wesentliche Rolle für die Beschreibung von Kategorienstruktur und -abgrenzung, und damit für die Beschreibung der entsprechenden Wortbedeutungen, aber die gesamte Struktur kann der Prototypenbegriff bei zahlreichen Kategorien nicht erklären; damit kann über die Festlegung eines prototypischen Konzepts allein auch nicht die Bedeutung von Lexemen vollständig erklärt werden, die solche Kategorien denotieren. Zudem wurde die Prototypentheorie in der Linguistik etwas nachlässig rezipiert, indem sie direkt von der Kognitionspsychologie auf die lexikalische Semantik übertragen wurde. Sie ist aber zunächst nicht die Beschreibung der Bedeutungen von Begriffen, sondern eine Beschreibung eines Teilaspektes humaner Kategorisierung. Durch diesen Kurzschluß wurde kaum berücksichtigt, daß sich die kognitiven Aspekte der Kategorisierung mit anderen Phänomenen überlagern, so etwa mit der Zuschreibung der "richtigen" Definition von Begriffen zu bestimmten Sprechern mit hohem Prestige oder zu Sprechern, die in bestimmten Gebieten als Fachleute ausgewiesen sind.

Ein Beispiel für vorschnelle Verallgemeinerungen ist die Übertragung der Prototypentheorie auf biologisch bedingte Einteilungen. Die Tatsache, daß in der menschlichen Kognition bestimmte Vogelarten als "typischere" Vögel angesehen werden als andere, heißt nicht, daß das prototypische Konzept die Bedeutung des Wortes *Vogel* ausmacht und heißt auch nicht, daß Vögel nicht eine Kategorie wären, die durch ein von allen Sprechern geteiltes Wissen über die tatsächliche Extension bestimmt ist — also scharfe Grenzen hat[22]. Etwas anders ist dies bei dem Beispiel *Busch* vs. *Baum*, wo eine Expertenbestimmung nicht vorliegt: hier spielt für die Begriffsabgrenzung der jeweilige Prototyp sicher eine zentrale Rolle; doch auch hier ist es Reduktionismus, allein den Prototypen einer Kategorie als Bedeutung des bezeichnenden Ausdrucks anzunehmen. Nur in Abgrenzung gegen die verwandten Lexeme kann eine Bedeutungsbeschreibung erfolgen. Die Beschreibung von Bedeutung aufgrund eines Prototypen darf also nicht dazu führen, das Sprachsystem nicht mehr als ganzes zu betrachten, und glauben zu machen, die Bedeutung eines Lexems ließe sich unabhängig von der Bedeutung anderer Lexeme beschreiben.

Wierzbicka (1990: 354f) weist darauf hin, daß das Konzept, das dem Ausdruck *Spielzeug* zugrundeliegt, kaum mit denselben Methoden beschrieben werden kann, wie die meisten anderen der diskutierten Kategorien. Sie bezeichnet solche Sammelkonzepte, die hauptsächlich aufgrund der Funktion der Gegenstände konstituiert werden, für welche die untergeordneten Konzepte stehen, als kollektive Konzepte. Der Verbindung *Puppe - Spielzeug* liegt ihrer Meinung nach in einer Taxonomie ein gänzlich anderes Verhältnis zugrunde als der Relation *Spatz - Vogel*. In einem semantischen Eintrag zum Lemma *Ball* ist *Spielzeug* nicht das übergeordnete Konzept, da es auch *Volleybälle* etc. gibt, die nicht als *Spielzeug* zu bezeichnen sind. Ich glaube, daß sie hier einen kategorischen Unterschied macht, wo es sich nur um einen graduellen handelt: Taxonomische Strukturen sind stets mehr oder weniger aspektabhängig. Die Zugehörigkeit eines Konzeptes zu einer übergeordneten, funktional definierten Kategorie gehört dabei sich zu den am ehesten vom Kontext beeinflußbaren Relationen. Wierzbicka ignoriert hier also, daß eine Taxonomie ohnehin in vielen Bereichen nur unter einem be-

[22]Inzwischen ist in der Wortsemantik deutlich gezeigt worden, daß das prototypische Konzept in unterschiedliche Typen von Kategorien eine unterschiedliche Rolle spielt. Wierzbicka (1990) diskutiert ausführlich dessen Bedeutung für die Kategorie Vogel (ebd.: 350f, 361f), die Farbenbezeichnungen (ebd.: 358f) u. a.

stimmte Aspekt konstruiert werden kann; indem sie als Gegenbeispiel die Kategorie *Vögel* nennt, sucht sie nur eine extrem kontextunabhängige Kategorie aus. Unbestreitbar ist hingegen ihre Beobachtung, daß die genannten kollektiven Kategorien nicht in gleicher Weise um eine Prototypen strukturiert sind wie andere der diskutierten Kategorien.

Während grundsätzlich die Relevanz von Prototypikalitätseffekten für die Bedeutungsbeschreibung wohl von keinem Semantiker mehr ernsthaft bestritten werden kann, ist ein ernsthaftes Problem die Art und Weise, wie die beschriebenen Phänomene in die semantische Beschreibung eines Lexikons integriert werden können. Es ist dabei zu fragen, in welcher Form graduelle Kategorienzugehörigkeit und Zentralität einer Unterkategorie in bezug auf die übergeordnete Kategorie in der lexikalischen Beschreibung modelliert werden können.

Es gibt Versuche der Einbettung der Erkenntnisse aus der Prototypenforschung in die klassischen Semantiktheorien. Ein solcher Versuch ist die Modellierung von Typikalitätseffekten durch mehrwertige Logiken oder eine Logik mit beliebigen Werten zwischen Null und Eins, also der Fuzzy-Logic (vgl. Zadeh 1975). Die Fuzzy-Logik hat sich jedoch insofern als ein relativ problematischer Weg in diese Richtung herausgestellt, als hier wieder eine scheinbare Präzision in der Zuordnung der Werte ins Spiel kommt, die in der menschlichen Zuordnung zu übergeordneten Klassen nicht vorhanden ist. Insbesondere scheint die Zuordnung von Unschärfewerten zu Aussagen der Form *X ist ein Y* nicht weniger fraglich zu sein als die Zuordnung eines von zwei diskreten Werten. So weist Parikh (1994: 525) darauf hin, daß unter Sprechern bei der Aufgabe, Ausdrücke in eine übergeordnete Kategorie einzuordnen, keine größere Einigkeit über Werte in einer Fuzzy-Logic zu erzielen ist, als über die Zuordnung in einer normalen zweiwertigen Logik. Da auch keine objektiven Maße für Prototypikalität existieren, ist die Zuweisung von Werten kaum möglich. Dieser Versuch zur Integration der Prototypentheorie in die logische Semantik muß also, zumindest in seiner naiven Form, als gescheitert betrachtet werden.

Zudem ist es falsch, Prototypikalitätseffekte gleichzusetzen mit unscharfen Kategoriengrenzen, wie dies durch die Fuzzy-Logik suggeriert wird. Cruse (1990: 386f) weist darauf hin, daß eine rund um einen Prototypen strukturierte Kategorie wie die der Vögel trotzdem klare Grenzen aufweisen kann. So ist ein Pinguin zu hundert Prozent ein Vogel, wenn auch ein untypischer. Im Gegensatz zur Kategorienzugehörigkeit können Sprecher über Typikalität in vielen Fällen Einigkeit erzielen, weil Typikalität ein gradierbares Phänomen ist. Dies gilt nicht im selben Maße für Kategorienzugehörigkeit.

Für den Grad und die Zentralität der Kategorienzugehörigkeit stehen mit den sogenannten Heckenausdrücken ("Hedges", ein Begriff der durch Lakoff 1973 kreiert wurde) allerdings sprachliche Tests zur Verfügung. Solche Heckenausdrücke sind für das Deutsche adjektivische und Adverbialausdrücke wie *typisch*, *im strikten Sinne*. Die Akzeptabilitätswertung der beiden folgenden Aussagen läßt jeweils Schlüsse darüber zu, ob *Pinguin* und *Abt* zentrale Vertreter der übergeordneten Kategorien *Vogel* bzw. *Beruf* sind.

5 ? Ein Pinguin ist ein typischer Vogel

6 Abt ist kein Beruf im strikten Sinne

Die Akzeptabilität von Zuordnungssätzen mit solchen Heckenausdrücken kann herangezogen werden, um zu bestimmen, ob das Denotat eines sprachlichen Ausdrucks von den Sprechern als typischer, untypischer oder marginaler Vertreter der übergeordneten Kategorie aufgefaßt wird. Ist also die exakte Zuordnung von Werten einer unscharfen Logik problematisch, so ist es doch wenigsten möglich, Typizität und Marginalität bei der Zuordnung zu Oberbegriffen durch Tests zu bestimmen.

Zwar wurde die Prototypentheorie als Gegentheorie zur Dekompositionstheorie rezipiert, doch sind Prototypen durchaus nicht unvereinbar mit der Merkmalstheorie in einer gemäßig-

ten Form ("modified checklist theory" nennt dies Cruse 1990: 391f). Sieht man die Bestandteile einer Lexembedeutung nicht mehr als notwendige und hinreichende Merkmale an, ist der Weg frei, prototypische Merkmale anzunehmen (z. B. Vogel -¿ ¡flugfähig¿), die dann als Defaultmerkmale aufgefaßt werden, welche typische Eigenschaften der Denotate beschreiben, die durch das gegenteilige Merkmal (der Pinguin erhält das Merkmal ¡-flugunfähig¿) aufgehoben werden können.

Eine Integration der Prototypensemantik in die semantische Klassifikation des Wortschatzes für ein elektronisches Wörterbuch ließe sich auf verschiedene Weisen realisieren:

- Für jede taxonomische semantische Klasse wird ein Merkmal eingeführt, das festlegt, ob die entsprechende Kategorie eine prototypenzentrierte Struktur hat oder nicht.
- Für jede semantische Klasse wird festgelegt, ob sie unscharfe Ränder hat oder ob die Zuordnung von Lexemen zu ihr nach einer klaren ja/ nein-Entscheidung erfolgen kann.
- In semantischen Klassen mit prototypenzentrierten Struktur werden zum einen die Lexeme und Unterklassen markiert, deren Denotat in bezug auf die übergeordnete Kategorie prototypennah ist, und zum anderen jene, welche sich am weitesten von der Prototypikalität entfernen. Hierzu können als Tests Sätze mit Hekkenausdrücken herangezogen werden.

Die so kodierten Eigenschaften könnten dann verwendet werden, um auf einer weiteren Stufe der Semantikkodierung die Merkmalsvererbung zu regeln. Zentrale Subkategorien erben alle Merkmale des übergeordneten Knotens, während als peripher markierte nur die definitorischen Merkmale übernehmen.

Dies gilt auch für die Kontexte, die dazu dienen, distributionelle Klassen zu definieren und abzugrenzen. Während für die zentralen Mitglieder einer Klasse alle Kontexte gültig sind, können die peripheren Mitglieder nur in einer Auswahl von typischen Kontexten der Klasse auftreten.

2.3.4 Kontextuelle Theorien in der linguistischen Wortsemantik

Firths Aussage "You shall know a word by the company it keeps" (Firth 1957 : 194-196) formuliert im Gegensatz zu den paradigmatischen Ansätzen der oben genannten Dekompositionstheorie und der Prototypentheorie eine Betrachtungsweise der Wortsemantik, die syntagmatische Aspekte, d. h. Wortbedeutung gesehen als eine Funktion der Beziehung zu anderen Wörtern im sprachlichen Kontext, in den Vordergrund rückt. Syntagmatischer und paradigmatischer Ansatz schließen sich allerdings keineswegs aus. So baut etwa die Determinierung von Merkmalen in der Komponentenanalyse auf syntagmatischen Aspekten auf, indem in bestimmten Kontexten austauschbare Wörter auf die unterschiedlichen Bedeutungsbestandteile untersucht werden. Auch die Beschreibung von Sinnrelationen schließt insofern bereits eine Beschreibung von Kollokationen ein — wenn auch meist nicht in konkreten Texten — als Vorkommen von Lexemen in bestimmten ähnlichen Kontexten und (partielle) Substituierbarkeit bei gleichbleibender bzw. entgegengesetzter Bedeutung zur Feststellung von Synonymie und Hyponymie bzw. von Oppositionsbeziehungen verwendet werden. Die stärkere Akzentuierung syntagmatischer Beschreibung v. a. in der Korpuslinguistik kann also in gewisser Weise gesehen werden als die wieder verstärkte Berücksichtigung der Grundlagen einer paradigmatischen Beschreibung, die ja häufig — wie oben gezeigt wurde — unter übermäßiger Abstraktion von den Fakten leidet.

Kollokationelle Ansätze sind oberflächennäher als die bereits diskutierten Beschreibungsmodelle, da in ihnen Mengen von ähnlichen Wörtern oder Wortformen aufgrund konkreter

distributioneller Ähnlichkeiten konstituiert werden. Eine distributionelle Beschreibung ist letzten Endes die Beschreibung des Gebrauchs eines Wortes oder eines Mehrwortausdrucks[23]. Dabei müßte Kontext im weiteren Sinne natürlich auch das pragmatische Umfeld einer Äußerung berücksichtigt werden. Uns geht es jedoch in diesem Abschnitt nur um Theorien, die den Ko-Text, d. h. die sprachlichen Umgebung lexikalischer Einheiten, untersuchen[24].

Ich möchte nachfolgend bei kontextuellen Theorien der lexikalischen Semantik zunächst zwischen qualitativen und quantitativ-statistischen Ansätzen unterscheiden.

Qualitative Ansätze untersuchen die Distribution von Lexemen in Kontexten unter Berücksichtigung linguistischer Kriterien. Es kommt hier weniger auf die Häufigkeit bestimmter Distributionen an, als auf die Tatsache ihres Auftretens beziehungsweise Nicht-Auftretens. Sonderfälle, die sich auf Hyponymie, Polysemie, Bedeutungsübertragung oder Idiosynkrasien zurückführen lassen, werden manuell aussortiert. Dabei wird das Belegmaterial häufig durch Sprachbeispiele ergänzt, die mittels Introspektion gewonnen werden. Dieses Verfahren wird vor allem verwendet, um Datengrundlagen für die semantische Beschreibung in paradigmatischen Ansätzen zu gewinnen.

Quantitative Verfahren untersuchen das Verhalten von Lexemen in größeren Textmengen auf Basis statistischer Berechnungsmethoden. Die Ergebnisse rein quantitativer Verfahren sind für Anwendungen in der Computerlinguistik, die Bedeutungsaspekte mit einschließen, nur von bedingtem Interesse, solange sie blind für alle linguistischen Fragestellungen an Korpora herangehen. Sie werden erst dann relevant, wenn zumindest eine Möglichkeit der Rückführung von Wortformen auf das zugrundeliegende Lemma vorhanden ist, und gewinnen mit jeder für die statistische Auswertung zur Verfügung gestellten linguistischen Information an Wert für die Bedeutungsbeschreibung.

Die uns interessierende Klasse distributioneller Ansätze möchte ich im folgenden 'kollokationell' nennen. Darunter sind alle jene Versuche der syntagmatischen Beschreibung zu verstehen, die nicht durch Introspektion erzeugte, konkrete Texte bezüglich der tatsächlichen Distribution von Lexemen unter Zuhilfenahme statistischer Methoden untersuchen und dabei auf mehr oder weniger linguistisches Vorwissen und/ oder eine entsprechende Nachbereitung der Daten zurückgreifen.

Die paradigmatischen Veränderungen, die in der linguistischen Forschung auf allen Ebenen zu einem Trend hin zur Berücksichtigung konkreter Texte und vor allem zur Untersuchung großer Korpora geführt haben, wurden in der Einleitung zu diesem Kapitel bereits genannt. Hier seien noch die speziellen Bedingungen genannt, die große Korpora für die Wortsemantik interessant machen:

- Die manuelle Kodierung von Bedeutungen ist extrem aufwendig, und es ist alles andere als klar, welche der zugrundeliegenden Konzepte im Endeffekt wirklich tragfähig sind.

- Die Lemmatisierung, das Tagging und die Syntaxanalyse unrestringierter Texte hat in den letzten Jahren bereits zu einigen Ergebnisse geführt. Damit stehen zusehends größere Korpora zur Verfügung, die — zumindest gilt dies für das Englische — teilweise syntaktisch aufbereitet sind, bzw. durch robuste und effiziente Parsingmethoden aufbereitet werden können (vgl. Hindle 1990, Ribas 1994), was die Beachtung der Wortart und die Disambiguierung mehrdeutiger Wortformen bei der statistischen Analyse

[23] Velardi/ Pazienza/ Fasolo (1991: 156) unterscheiden zwischen "conceptual meaning", das in Theorien durch Merkmale ausgedrückt werde, und "collocative meaning", das den Gebrauch eines Wortes beschreibt.

[24] Dies sicher nicht, weil ich glaube, daß der situationelle Kontext irrelevant zur Bedeutungsbeschreibung sei. Ich blende ihn aber aus, weil das Hauptaugenmerk der Beschreibung der Wortbedeutung in geschriebenen Texten gelten soll. Hier wird das Kriterium des situationellen Kontextes weitgehend durch das der Textart ersetzt.

ermöglicht und sogar eine Berücksichtigung der grammatischen Funktion einzelner Phrasen erlaubt. [25]

- Es stehen inzwischen bewährte statistische Methoden zur Auswertung der gefundenen Daten zur Verfügung. Hier sind im Bereich der Kombinatorik zwischen Wortformen Maße wie der t-Test und die Transinformation (Mutual Information; für einen Überblick siehe Church/ Gale/ Hanks/ Hindle 1991) zu nennen. Es stehen aber auch komplexere Clusteringmethoden aufgrund von statistischer Nähe in allen Kontexten eines Lexems zur Verfügung (vgl. Pereira/ Tishby/ Lee 1993, Rooth 1994, Rieger 1984).

Die Kollokationsanalyse kann mit oder ohne Berücksichtigung linguistischer Kriterien erfolgen. Folgende Typen von Untersuchungen wurden durchgeführt:

- Rein quantitativ bestimmter Kontext, also ein Fenster von n Wörtern nach links und/ oder rechts (so etwa in Church/ Hanks 1990), wobei n sinnvollerweise nicht zu groß sein sollte. Die Ergebnisse solcher Untersuchungen sind sehr unscharf, insbesondere dann, wenn nicht einmal Satzgrenzen berücksichtigt werden.

- Syntaktisch bestimmter Kontext (z. B. Hindle 1990) um z. B. die Kollokationen zwischen Verben und ihren Komplementen zu ermitteln (dies setzt zumindest partielles oder lokales syntaktisches Parsing voraus).

Variiert werden v. a. Fenstergröße, Umfang und Art der verwendeten Korpora sowie statistische Schwellenwerte. Bei der Untersuchung von Kollokationen zwischen Einzelwörtern wird ein Schwellenwert für die Abweichung von einem Mittelwert festgelegt. Beim distributionellen Clustering spielen auch Faktoren wie Klassengröße und Zahl der Klassen eine Rolle.

In den Untersuchungen zum Englischen ähneln sich die Ergebnisse der Untersuchungen zur Kollokationsanalyse mit unterschiedlichen statistischen Maßen (vgl. Church/ Hanks 1990 mit dem Maß Transinformation, Church/ Gale/ Hanks/ Hindle 1991 zum t-Test). Die gefundenen Kollokationen lassen sich verschiedenen Arten von Phänomenen zurechnen, die hauptsächlich entlang der Linie der semantischen Kompositionalität variieren. Einige gefundene Paare entstammen voll opaken, andere gänzlich transparenten komplexen Ausdrücken; das Extrem auf der einen Seite sind Idiome, auf der anderen Seite stehen häufige Verb-Objekt-Paare (zu den verschiedenen erkannten Phänomenen s. Smadja 1993). Die Extremfälle lassen sich allerdings anhand des invarianten Abstands zwischen den Wortformen bei Idiomen ansatzweise durch statistische Mittel unterscheiden (Church/ Hanks 1990: 23). Problematischer sind die Übergänge zwischen beiden Extremen, wo sich die mehr oder weniger erweiterbaren Mehrwortausdrücke finden.

Werden Wortklassen aufgrund einer großen Menge von Kollokationen aufgestellt (distributionelles Clustering), haben die Resultate meist folgende Eigenschaften (vgl. Rooth 1994):

- Es treten stets einige Klassen auf, die semantisch relativ kohärent sind, d. h. deren Zusammensetzung aufgrund linguistischer Kriterien nachvollziehbar ist.

- Andere Klassen enthalten neben einem Kern an semantisch ähnlichen Lexemen zahlreiche Querschläger, d. h. semantisch keinerlei oder kaum Ähnlichkeiten aufweisende Lexeme.

[25] Es gibt für das Deutsche schon frühere Ansätze zur statistischen Wortbedeutungsbeschreibung, die allerdings weniger intensiv rezipiert wurden, als die jetzt aus dem angelsächsischen Raum kommenden Überlegungen. Vgl. z. B. die 'semantischen Räume' in Rieger (1984).

- Manche Klassen sind insgesamt völlig inkohärent, d. h. es sind keine linguistischen Kriterien für die statistische Affinität der in ihnen enthaltenen Lexeme oder Wortformen erkennbar.

Die falsch eingeordneten Lexeme und die inkohärenten Klassen erklären sich relativ einfach. Zum einen werden in Untersuchungen von Kollokationen in nicht-getaggten Korpora zunächst weder Homographie noch Polysemie berücksichtigt. Auf der anderen Seite werden wortsemantische Phänomene wie metonymische und metaphorische Bedeutungsverschiebung nicht in Betracht gezogen, da die Statistik auf bloßen Wortformen, oder im besten Falle auf Lemmata operiert, und keine ausreichenden Algorithmen zur Disambiguierung von Wortformen und zur Detektion von Bedeutungsverschiebung existieren.

Während ein Vorgehen ganz ohne linguistische Aufbereitung des Korpus für das Englische noch relativ brauchbare Resultate liefert, ergeben sich für das Deutsche noch zusätzliche Schwierigkeiten:

- Es gibt von den Wortklassen, die für die semantische Beschreibung interessant sind (Nomina, Verben, Adjektive), wesentlich mehr unterschiedliche Formen pro Lexem, was heißt, daß eine vorherige Lemmatisierung zur Erzielung brauchbarer Ergebnisse für das Deutsche eine wesentlich größere Rolle spielt als für das Englische. Für die Kategorie Verb kommt erschwerend hinzu, daß im Deutschen Präfixverben sehr häufig sind, und in einem Text zunächst die abgetrennte Präfixe richtig zugeordnet werden müßten, was nicht ohne weiteres möglich ist, da die meisten dieser Präfixe homograph mit Partikeln, Präpositionen oder Adverbien sind (vgl. Ott 1992; zum Problem der Lemmatisierung s. Maier 1995). Werden Präfixverben nicht berücksichtigt, verfälscht dies zusätzlich das Ergebnis für die zugrundeliegenden einfachen Verben.

- Die Wortstellung im Deutschen ist freier als im Englischen. Während dort syntaktische Funktionen von Nominalgruppen stark mit der Stellung im Satz korrelieren, hängen sie im Deutschen in erster Linie mit der Kasusmarkierung zusammen. Hinzu kommt, daß im Deutschen abhängige Konstituenten, in Wortabständen gemessen, weiter von der subkategorisierenden Kategorie entfernt stehen können als im Englischen, da die morphologische Markierung eine Zuordnung der grammatischen Funktion zu den Komplementen erlaubt. Somit wäre für das Deutsche eine syntaktische Korpusaufbereitung wesentlich wichtiger, um bei einer statistischen Kollokationsanalyse brauchbare Resultate zu erzielen. Dazu fehlt aber ein robuster Parser für unrestringierte deutsche Texte (vgl. Breidt 1993).[26]

Diese zusätzlichen Schwierigkeiten schlagen sich in deutlich in den Ergebnissen der bisher vorgenommenen Kollokationsuntersuchungen zum Deutschen nieder (vgl. Ott 1992, Breidt 1993). Damit ist die Verwendung von statistisch ermittelten distributionellen Klassen für die Wortsemantik, die schon im Englischen Probleme bereitet, für das Deutsche zunächst wenig offensichtlich.

Auch aus anderen Gründen kann schon die Schwierigkeit der Interpretation der Ergebnisse im Englischen nicht verwundern. Kollokationen sind nicht ausschließlich semantisch determiniert. Neben den genuin semantischen Regularitäten stammen die Phänomene, die sich in der Bildung von Kollokationsklassen niederschlagen, auch zu einem großen Teil aus

[26]Den bereits genannten Untersuchungen (z. B. Hindle 1990) für das Englische liegt eine syntaktische Analyse mit einem solchen Parser zugrunde. Wenngleich ein Parser für unrestringierten Text beim jetzigen Forschungsstand nie fehlerfrei arbeiten kann, ist die Fehlerrate doch so gering, daß die Zuordnung einer syntaktischen Funktion als verläßlich genug angesehen werden kann, um die Parsingfehler statistisch nicht ins Gewicht fallen zu lassen.

dem Bereich lexikalisierter Mehrwortausdrücke. Zudem lassen sich die Effekte der unterschiedlichen semantischen Phänomene nicht klar trennen.

Da alles andere als klar ist, was kollokationelle Analysen letztlich beschreiben, halte ich sie — auf dem jetzigen Stand der Forschung und speziell im Falle des Deutschen — nicht für geeignet, eine vollständige Klassifizierung des Wortschatzes ohne weitere Grundlagen durchzuführen.

In Verbindung mit einer vorherigen semantischen Grobklassifizierung des Wortschatzes lassen sich bessere Ergebnisse erzielen. So werden in einer Untersuchung zum Italienischen (Velardi/ Pazienza/ Fasolo 1991) Kollokationen — genauer gesagt Verb-Objekt-Paare — erfolgreich zur Bildung von Teilklassen manuell kodierter semantischer Klassen herangezogen. Ein solches Vorgehen wäre auch auf einer weiteren Stufe der Kodierung für das Deutsche zu erwägen.

Während eine statistische Klassifikation aufgrund von Korpora für die erste semantische Kodierung nicht in Frage kommt, gilt doch andererseits, daß eine manuelle semantische Beschreibung des Wortschatzes der Überprüfung anhand von Korpora standhalten muß. Die genannten statistischen Methoden eignen sich dazu, eine semantische Klassifikation des Wortschatzes an konkreten Texten oder anderen Korpora (wie z. B. Korpora von Nominalkomposita) zu überprüfen. Doch auch hierbei ist die Einschränkung zu machen, daß eine statistische Überprüfung nur für die Lexeme in Frage kommt, die mit ausreichender Häufigkeit in den verfügbaren Korpora auftreten - und damit nur für einen Bruchteil des zu kodierenden Wortschatzes.

Diese Einschätzung statistischer Methoden wurde auch dem Vorgehen zur Kodierung des CISLEX zugrundegelegt: Während die grundlegende Kodierung manuell — wenn auch unter Zuhilfenahme von sprachlichen Daten — durchgeführt wird, soll die erreichte semantische Beschreibung im Anschluß anhand von Daten aus Korpora — in unserem Falle Nominalkomposita und Zeitungstexten — mit Hilfe statistischer Methoden überprüft werden.

2.3.5 Faktoren der Wortbedeutung

Quasi als Zusammenfassung der Übersicht zu verschiedenen Aspekten der Wortbedeutung aus den vergangenen Abschnitten läßt sich das folgende Zitat von Dahlgren (1988: 17) lesen:

> "In summary, some words do not have criterial attributes in their meaning representations. Apparently the representation of word meaning varies across the lexicon. Some words may have criterial verbal features in their representations. Obvious examples come form mathematics. A *triangle* can be defined as a 'three-sided figure'. Others have features which correspond to naive theories, or stereotypes of the extensions, but are used with the intention of referring to kinds, that is, to classes of objects with some sort of stable essence independent of the human view of them (*water, lemon, tiger*). Dahlgren (1988b) argues that social terms such as *knife* and *programmer* belong to this group. Still others are descriptive. Human viewpoint determines the classes (*weed, junk, witch, game*). Other words are represented in terms of visual and other perceptual features which are not readily translated into verbal predicates."

Dieses Zitat faßt relativ bündig die verschiedenen Phänomene zusammen, welche die Beschreibung von Wortbedeutung vielschichtig und problematisch machen.

Das Versagen der Versuche in der theoretischen Linguistik, Wortbedeutung eindeutig zu beschreiben, liegt nicht in einem Defizit der Theoriebildung, sondern an der Tatsache, daß sich in der Konstitution von Wortbedeutung verschiedenste Mechanismen überlagern. Das

Studium der Wortbedeutung im linguistischen Sinne läßt sich nicht eindeutig und vollständig von einer enzyklopädischen Beschreibung der Welt abgrenzen. Auch wenn man versuchen möchte, sich im Rahmen der lexikalischen Semantik auf die zentralen Aspekte der Wortbedeutung zu beschränken, die noch eindeutigen Niederschlag in beschreibbaren sprachlichen Phänomenen finden und sich durch linguistische Akzeptabilitätstests eingrenzen lassen, sollte die lexikographische Beschreibung von Lexembedeutungen immer im Bewußtsein der zahlreichen relevanten Faktoren vorgenommen werden:

- Zum einen spielen Grundkonstanten humaner Kategorisierung für die Wortbedeutung eine große Rolle. Gegenstände werden vor allem aufgrund wahrgenommener äußerer Formähnlichkeiten und ihrer Funktion für den Menschen kategorisiert (hingegen gibt es zur Kategorisierung von Abstrakta bisher wesentlich weniger Untersuchungen, und über die Mechanismen ist hier wesentlich weniger bekannt). Aus dem Bereich der Kognitionsforschung kommt z. B. die oben ausführlich diskutierte Prototypentheorie.

- Zum zweiten dürfen soziale Faktoren nicht vernachlässigt werden: Sprecher gehen zumindest bei bestimmten Typen von Begriffen davon aus, daß ihnen real existierende Kategorien zugrundeliegen. Kennen sie selbst die Kriterien nicht, verlassen sie sich auf die Mehrheitsmeinung oder auf Autoritäten in der Sprechergemeinschaft. Dies wird z. B. deutlich im Bereich der Unterteilung der Bezeichnungen für Tiere, bei denen die Fachtaxonomien aus der Biologie partiell Allgemeingut geworden sind.

- Neben solchen Einflüssen aus anderen Systemen sind die diachronen Aspekte innerhalb des Sprachsystems selbst zu berücksichtigen. Sprachgeschichtlich bedingte semantische Einteilungen können sich, teilweise in modifizierter Form, oft lange halten. Als Beispiel sei die in den letzten Jahren durch die Zunahme der *-in* und *-Innen* Formen besonders ins Bewußtsein gerückte Referenz der Berufsbezeichnungen (und anderer Menschenbezeichnungen) auf alle einen Beruf ausübenden Menschen und gleichzeitig nur auf alle solchen Männer (früher der Normalfall) genannt. Die Versuche, diese inzwischen den Verhältnissen nicht mehr adäquate semantische Struktur zu ändern (was immer aus linguistischer Sicht davon zu halten ist), haben inzwischen allerdings nur dazu geführt, daß eine neutrale Referenz auf eine Berufsgruppe mit einem einfachen Ausdruck kaum mehr möglich ist.

- Als letzter Punkt seien die Bedeutungsübertragungen durch Metaphorik und Metonymie genannt, die synchron wie diachron wirksam ist. Sie sind wesentlicher Teil des Sprachsystems. Die Mechanismen, die einer solchen Bedeutungsübertragung zugrundeliegen, sind alles andere als geklärt. Somit können diese Phänomene auch nur partiell formal gefaßt werden.

Diese Einflüsse aus ganz unterschiedlichen Systemen erklären auch, warum mit allen Typen von eindimensionalen Ansätzen, seien diese nun eher regelbasiert ausgerichtet oder rein statistisch orientiert, keine allgemein brauchbaren Ergebnisse erzielt werden können. Ein Formalismus zur umfassenden Beschreibung der Wortsemantik muß vor allem eine Eigenschaft haben: Es darf ihm nicht das Prinzip der Abgeschlossenheit der Wortbedeutung zugrundeliegen, sondern er muß offen sein für verschiedenste Ergänzungen.

Den genannten Herausforderung hat sich die praktische Bedeutungsbeschreibung von Lexemen schon immer stellen müssen. In den folgenden Abschnitten wird von daher versucht werden, die bisher diskutierten Aspekte der Wortsemantik aus dem Blickwinkel der Praxis der Lexikographie zu beleuchten.

2.4 Semantische Beschreibung in der Praxis der Lexikographie

Es ist im folgenden zu trennen zwischen der traditionellen lexikographischen Praxis, deren Ziel die Erstellung eines gedruckten oder maschinenlesbaren Lexikons für einen menschlichen Benutzer ist, und der Erstellung eines Lexikons für computerlinguistische Anwendungen, das von vornherein für die Auswertung durch eine Maschine erstellt wurde (vgl. Calzolari 1994: 268). Bei letzteren unterscheide ich wiederum zwischen Maschinenlexika für eine bestimmte Anwendung und elektronischen Wörterbüchern, denen eine umfassendere Konzeption zugrundeliegt.

Die folgende Darstellung der verschiedenen Typen von Lexika soll dabei unterschiedliche Zielsetzungen verfolgen:

- Bei der Diskussion der semantischen Beschreibung in gedruckten und maschinenlesbaren Lexika kommt es mir vor allem darauf an, herauszuarbeiten, ob und wie die in ihnen enthaltene Information für die Kodierung eines elektronischen Wörterbuches genutzt werden könnte und welche semantischen Beschreibungselemente eines herkömmlichen Wörterbuchs sich für die Übertragung in die elektronische Lexikographie eignen könnten.

- Die Erörterung von Eigenschaften der semantischen Kodierung elektronischer Wörterbücher soll dem Abgleich mit der relativ kurzen Tradition elektronischer Lexikographie dienen.

- Bei der Diskussion der Anforderungen an Lexika für computerlinguistische Anwendungen soll anhand von zwei Anwendungsbeispielen, der maschinellen Übersetzung und dem Information Retrieval, exemplarisch ermittelt werden, welche Bedeutungsinformationen ein elektronisches Wörterbuch enthalten sollte, um eine Grundlage für verschiedene Aufgaben der maschinellen Sprachverarbeitung zu bieten

2.4.1 Gedruckte und maschinenlesbare Lexika

Obwohl gedruckte und maschinenlesbare[27] Lexika auf unterschiedlichen Medientypen dargeboten werden, sind Inhalt und Struktur der angebotenen Information im wesentlichen dieselben.

Die Beschreibung der Semantik in dieser Art von einsprachigen Lexika dient der Benutzung durch eine Person, die der beschriebenen Sprache weitgehend mächtig ist und die über ein allgemeines Hintergrundwissen verfügt. Ein solches Wörterbuch gibt also nur Hinweise auf die Bedeutung eines Lexems, während für eine maschinelle Anwendung die exakte Definition der beschriebenen Bedeutungskomponenten nötig ist. Auch maschinenlesbare Wörterbücher verhalten sich in dieser Beziehung wie ihre gedruckten Vorlagen und unterscheiden sich damit grundsätzlich von elektronischen Wörterbüchern, die für die elektronische Sprachverarbeitung geeignet sind.

Ein zusätzliches Problem für die Auswertung der Information gedruckter und maschinenlesbarer Lexika für die computerlinguistische Lexikographie ergibt sich durch die Form der Bedeutungsbeschreibung. Diese ist kaum je vollkommen systematisch. Die mangelnde Systematik äußert sich in zweierlei Weise: Einerseits ist der Semantikteil in den verschiedenen Einträgen nicht unbedingt gleich aufgebaut — andererseits werden innerhalb eines

[27] Es sind bereits zahlreiche maschinenlesbare deutsche Lexika auf CD-ROM erhältlich. Z. B. 'Duden. Deutsches Universalwörterbuch A-Z' (Dudenverlag), Herrmann Paul: 'Deutsches Wörterbuch' (Verlag Rheinbaben & Busch) u. a.

Wörterbucheintrags verschiedene Typen der semantischen Information nicht sauber unterschieden und von syntaktischer Information in Form von Satzmustern und der Darstellung von Gebrauchsbedingungen (z. B. in Form von Sprachbeispielen) und kombinatorischen Idiosynkrasien nicht getrennt.

2.4.2 Semasiologische Wörterbücher

Die Angabe der Semantik in herkömmlichen Wörterbüchern erfolgt durch zwei prinzipiell verschiedene Methoden, die allerdings in den Einträgen nicht immer voneinander getrennt sind:

- Angabe sinnverwandter Wörter: Es finden sich in fast jedem Wörterbucheintrag Angaben von (Teil-) Synonymen und/ oder Hyperonymen. Hyponyme, Meronyme sowie in Opposition stehende Lexeme werden nur unsystematisch angegeben. Die Angabe von sinnverwandten Wörtern erfolgt häufig nicht separat, sondern innerhalb von nichtformalisierten Umschreibungen der Bedeutung. Synonyme und Teilsynonyme werden manchmal speziell ausgezeichnet, sie stehen oft aber auch innerhalb einer Bedeutungsumschreibung. So in Wahrig (1986) unter dem Lemma *Nase*: "[...];Sy Riechorgan [...]", aber unter *Auge*: "[...] 1 Sehorgan des Menschen oder der Tiere [...]". Zusätzlich zur Auszeichnung "Sy" taucht im selben Wörterbuch zur Nennung eines Synonyms im Eintrag noch gelegentlich ein Gleichzeichen auf, so im Eintrag *Grashüpfer*: "= Heuschrecke".[28]

- Umschreibungen: die Bedeutung eines Lexems wird umschrieben, meist indem ein Teilsynonym oder eine Hyperonym durch Adjektive oder andere Komplemente näher spezifiziert wird. Als Beispiel kann der eben zitierte Teileintrag zu *Auge* dienen.

  Anders verhalten sich Wörterbücher, die von vornherein der semantischen Beschreibung des Wortschatzes dienen. Es gibt folgende wichtige Typen von Bedeutungswörterbüchern[29]

2.4.3 Synonym- und Antonymwörterbücher

Von diesen beiden Wörterbuchtypen spielen vor allem Synonymenwörterbücher eine Rolle in der Lexikographie. Sie liefern Synonyme und Teilsynonyme zu einem Lexem (z. B. Radszuweit/ Spalier 1982). Daneben existiert noch das relativ seltene Antonymenwörterbücher (vgl. Hausmann/ Reichmann/ Wiegand/ Zgusta 1990: Bd. 2: 1081-1083).

Beide Typen von Wörterbüchern enthalten Information, die auch in der maschinellen Sprachverarbeitung einsetzbar ist. So können aufgrund von solchen Wörterbüchern minimale Klassen von Lexemen mit geteilten Bedeutungskomponenten erstellt werden, die dann wiederum zu Klassen von Kohyponymen zusammengefaßt werden. Die semantische Kodierung großer Wortschatzbereiche wird durch Synonymwörterbücher erleichtert, indem man während der Konzeption eines Eintrags unmittelbaren Zugang zu den Lexemen erhält, die denselben oder einen ähnlichen semantischen Kode erfordern. Doch gilt für diese Art von Lexika, was oben schon in bezug auf andere gedruckte Wörterbücher gesagt wurde: Die Information innerhalb eines Eintrags ist oft nicht vollständig systematisch aufbereitet. Sie

[28] Relativ unproblematisch auswertbar sind die lateinischen Tier- und Pflanzenbezeichnungen im Linnéschen System, wie sie in einige Wörterbücher aufgenommen sind (z. B. Wahrig 1986). Im Wortschatzbereich der Tier- und Pflanzenwelt kann die Angabe der wissenschaftlichen lateinischen Bezeichnungen für ein elektronisches Wörterbuch, wenn sie für die verschiedenen Sprachen erfolgt, eine hervorragende Grundlage für die maschinelle Übersetzung dieser Ausdrücke bieten.

[29] Zu einer Übersicht verschiedener Wörterbuchtypen s. Hausmann/Reichmann/Wiegand/ Zgusta (1990): Band II.

können damit nur über eine vorherige manuelle Aufbereitung für die elektronische Lexikographie herangezogen werden.

2.4.4 Bildwörterbücher

Die in Bildwörterbüchern enthaltenen Informationen sind für den Transfer in die elektronische Lexikographie zunächst wenig geeignet. Sie können zwar einige Angaben zu pragmatischen Bereichen und zu Hyponymie- bzw. Meronymiebeziehungen liefern, diese Information wird aber teilweise recht unsystematisch gegeben und ist schwierig zu extrahieren.

2.4.5 Thesauri

Der Thesaurus[30] ist der wohl wichtigste Typ eines onomasiologischen Wörterbuchs. Thesauri sind nach Sachgebieten bzw. Themenbereichen geordnete Lexika und liefern damit auch eine (mehr oder minder intersubjektiv nachprüfbare) Einteilung oder Bereichszuordnung des in ihnen enthaltenen Wortschatzes. Die Einteilung erfolgt auf Grundlage einer mehrstufigen Begriffshierarchie. Für das Deutsche existieren als umfangreiche allgemeinsprachliche Thesauri nur Wehrle (1968) und Dornseiff (1970), beide sind zudem nur Neuauflagen von Werken früheren Datums. Die Originale entstanden 1954 (Wehrle) bzw. 1934. Ein Thesaurus neueren Datums existiert meines Wissens für das Deutsche nicht, d. h. alle seitdem neu hinzugekommenen Wortschatzbereiche sind in keinem Thesaurus für das Deutsche kodiert.

Im angelsächsischen Sprachraum erfreuen sich Thesauri einer wesentlich größeren Beliebtheit. Eines der verbreitetsten Werke dieser Art, das auch immer wieder neu aufgelegt wird, ist hier *Roget's Thesaurus of English Words and Phrases* (Kirkpatrick 1992).

Die Auswertung von Einträgen aus Thesauri für eine semantische Beschreibung des Wortschatzes in einem elektronischen Wörterbuch ist alles andere als evident. Die Zuordnung von Lexemen zu den einzelnen Kategorien ist keine Einteilung des Wortschatzes in semantische Klassen im Sinne von Kohyponymen, ebensowenig jedoch eine Einteilung nach pragmatischen Bereichen. In einem Eintrag finden sich durch verschiedenste Relationen miteinander verbundene Lexeme. Das folgende Beispiel aus Dornseiff (1970) macht dies deutlich. Unter dem Eintrag "Schlafen" (ebd.: Abschnitt 2.36., S. 141) finden sich folgende, durch das Separatorzeichen "#" getrennte Gruppen von Nomina:

> Dusel # Halbschlaf # Indolenz # Lethargie # Nickerchen # Rast # Rucks # Ruhe # Schlaf # Schlafsucht # Schlummer # Siesta # Traum # Rast # Nur ein Viertelstündchen # der Schlaf des Gerechten # Ätherrausch # Narkose

Vor allem innerhalb der ersten, alphabetisch geordneten Gruppe finden sich Lexeme, die verschiedene Zustände des Schlafens bezeichnen (*Schlaf*, *Nickerchen* u. a.), daneben solche, die ähnliche Zustände beschreiben (*Rast*, *Ruhe*), aber auch die Nomina *Schlafsucht* und *Indolenz*, die nur die Neigung zum Zustand des Schlafens bezeichnen. Hinzu kommen die zwei Idiome *Nur ein Viertelstündchen* und *der Schlaf des Gerechten*. In der zweiten Gruppe befinden sich zwei sinnverwandte Nomina, die schlafähnliche Zustände bezeichnen. An keiner Stelle wird explizit gemacht, in welchem Verhältnis die einzelnen Nomina zum Eintrag "Schlafen" stehen. Somit kann aufgrund eines solchen Eintrages keine Zuordnung der entsprechenden Nomina zu einer kohärenten Klasse, sondern nur eine (im Eintrag nicht näher spezifizierte) vage thematische Zuordnung zum Thema "Schlafen" vorgenommen werden.

[30] Die Aussagen hier gelten nur für allgemeinsprachliche Thesauri. Fachsprachliche Thesauri sind erst auf einer späteren Kodierungsstufe für eine elektronisches Wörterbuch relevant.

Der Nutzen der existierenden Thesauri für die elektronische Lexikographie beschränkt sich weitgehend auf Anregungen, die sie für eine erste Einteilung des Wortschatzes geben können, und auf erste Anhaltspunkte, welche Wörter welchen Klassen von Kohyponymen zugeordnet werden könnten. Die Vermischung verschiedener Einteilungskriterien in solchen Wörterbüchern macht jedoch die Übernahme ganzer Klassen nur in wenigen Einzelfällen möglich.

2.4.6 Auswertung der Information aus maschinenlesbaren herkömmlichen Lexika

Bei maschinenlesbaren Lexika handelt sich um Ausgaben von Wörterbüchern herkömmlicher Art in elektronischer Form, entweder in Form der elektronischen Drucksatzdateien oder als Edition auf CD-ROM bzw. anderen Datenträgern. Für den englischen Sprachraum, in dem eine größere Zahl von Lexika in maschinenlesbarer Form verfügbar ist, arbeiten eine Reihe von Forschern und Forschungsgruppen an der Aufbereitung der in maschinenenlesbaren Ressourcen enthaltenen Information für Maschinenlexika und elektronische Wörterbüchern[31].

Für die Art der Kodierung der semantischen Information in maschinenlesbaren Lexika gilt dasselbe wie für das gedruckte Pendant. Informationen aus solchen Lexika können zwar leichter zur Auswertung für computerlinguistische Anwendungen herangezogen werden als die Information aus den gedruckten Lexika, sie können jedoch ebenso schwierig systematisch in eine für die maschinelle Sprachverarbeitung verwendbare Form überführt werden. Hierzu ist für das Gros der Angaben ein aufwendiges Parsen der Wörterbucheinträge notwendig. Hinzu kommt die Schwierigkeit, daß die in den Einträgen enthaltene Information — insbesondere was die Semantik betrifft — überhaupt nur teilweise systematisierbar ist (vgl. Heyn 1992). Zudem ist für das Deutsche — abgesehen von den zahlreichen Lexika auf CD-ROM, bei denen die Information allerdings nicht als Textdatei vorliegt — kein maschinenenlesbares Wörterbuch frei verfügbar. Angesichts dieser Schwierigkeiten und der problematischen Resultate erschien es nicht sinnvoll, die semantische Kodierung des CISLEX auf der Auswertung eines maschinenlesbaren Wörterbuchs aufzubauen.

Zusammenfassend läßt sich zu der Auswertung der Information aus gedruckten und maschinenlesbaren Wörterbüchern folgendes sagen:

- Die systematische Übernahme der semantischen Angaben aus solchen Lexika ist nicht ohne beträchtlichen Aufwand möglich. Selbst wenn die Lexikoneinträge in elektronischer Form zur Verfügung stehen, ist es kaum möglich, mit Hilfe einfacher Algorithmen den Transfer in formalisierte Information zu erreichen.

- Dennoch ist die Konsultation dieser Lexika für die Kodierung eines elektronischen Wörterbuchs unerläßlich, will man nicht den gesamten deutschen Wortschatz aufgrund von Textbelegen neu beschreiben. Insbesondere ist die Konsultation von Lexika nötig, um die Bedeutung seltenerer Nomina und weniger oft auftretende Bedeutungen häufigerer Wortformen nicht zu übersehen. Der Transfer der Information in das elektronische Wörterbuch macht allerdings einige nicht automatisierbare Denkarbeit zur jeweiligen Umsetzung der nicht normierten Information in einen maschinenverwendbaren Formalismus nötig.

[31] Zur Verwendung maschinenlesbarer Lexika in der Computerlinguistik und für einen Überblick der Literatur zum Thema s. Boguraev (1994, insbesondere: 150-154).

2.4.7 Elektronische Wörterbücher und Wortdatenbanken

Im folgenden werden mit WordNet und der semantischen Kodierung durch die Forschungsgruppe um Gaston Gross zunächst zwei Versuche zur semantischen Beschreibung des Nominalwortschatzes des Englischen bzw. Französischen ausführlicher diskutiert, die auf die Konzeption und Kodierung der Nominalsemantik im CISLEX einen nicht unbedeutenden Einfluß hatten. Die beiden Systeme wurden ausgewählt, weil ihre Konzeption ähnlichen Prämissen unterlag wie die semantische Kodierung der einfachen Nomina des CISLEX. Die zwei wichtigsten geteilten Voraussetzungen waren dabei namentlich, daß erstens das Ziel angestrebt wurde, den gesamten Wortschatz einer Sprache weitgehend vollständig zu erfassen und zweitens die Kodierung anwendungsunabhängig vorgenommen wurde[32].

2.4.8 WordNet

WordNet (Miller u. a.1990, 1993) ist eine Wörtersammlung von englischen Nomina, Adjektiven, Verben und Adverbien. Das Ziel bei der Entwicklung von WordNet war die Kodierung von Synonymie, Hyponymie und anderen Sinnrelationen für den gesamten Wortschatz. Ergänzt wird diese semantische Beschreibung durch ein morphologisches Interface, das es ermöglicht, von beliebigen Wortformen ausgehend auf die Einträge zuzugreifen.

Die bisher letzte Version von WordNet wurde im Frühjahr 1995 herausgegeben. Aufgrund der freien Verfügbarkeit[33] der Datenbank und auch der Software zur Konsultation der Lexikondaten wurde WordNet bereits einige Male im Bereich der computerlinguistischen Forschung und zur Implementierung von natürlichsprachlichen Systemen eingesetzt, so etwa von Resnik (1993) und Ribas (1994, 1995) zur Beschreibung und zur automatischen Erlernung von Selektionsrestriktionen englischer Verben. WordNet ist meines Wissen die einzige weitgehend vollständige Wortdatenbank einer Sprache, in der systematisch alle wichtigeren Sinnrelationen zwischen sämtlichen Lemmata erfaßt sind. Von den Entwicklern der Datenbank selbst wird WordNet folgendermaßen charakterisiert:

> "WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets." (Miller u. a. 1990: 234)

Der Hinweis auf die Verarbeitung neuerer Erkenntnisse aus der Psycholinguistik erscheint allerdings angesichts der tatsächlich vorhandenen Information etwas übertrieben: WordNet berücksichtigt kaum die aktuellen Paradigmen aus der kognitiven Psychologie. Insbesondere die Prototypentheorie ist bisher in keinster Weise in die Kodierung eingearbeitet. Prinzipien der Merkmalsvererbung bei Nomina wie beschrieben in Miller (1990: 245ff) sind bereits aus der strukturalistischen Semantik bekannt und können in der Form, wie sie in WordNet implementiert sind, meiner Ansicht nach keine psychologische Adäquatheit für sich beanspruchen. Dies ist jedoch nur vor dem Hintergrund der oben zitierten Behauptung als Defizit anzusehen. WordNet erscheint mir als eine weitgehend solide Strukturierung des englischen Wortschatzes mit Hilfe von Sinnrelationen auf Grundlage von durchaus bewährten lexikographischen Kriterien.

[32] Letzteres soll nicht implizieren, daß bei der Kodierung der Semantik im CISLEX nicht Rücksicht genommen wird auf mögliche Anwendungen eine Lexikons, sondern nur, daß die Konzeption nicht primär auf eine Anwendung zugeschnitten ist.

[33] Das WordNet-System ist derzeit (Herbst 1995) per anonymem ftp auf den Servern ftp.ims.uni-stuttgart und ftp.princeton.edu frei für alle gängigen Rechnertypen erhältlich. Dort findet sich auch die als 'Miller (1993)' zitierte Dokumentation.

act, action, activity	natural object
animal, fauna	natural phenomenon
artifact	person, human being
attribute, property	plant, flora
body, corpus	possession
cognition, knowledge	process
communication	quantity, amount
event, happening	relation
feeling, emotion	shape
food	state, condition
group, collection	substance
location, place,	time

Tabelle 2.1: Grundklassen in WordNet

In der folgenden Erörterung der wichtigsten Eigenschaften des Systems soll nur die Einteilung des Nominalwortschatzes berücksichtigt werden. Die Einteilung der anderen Wortklassen wird beschrieben in Miller u. a. (1990 und 1993). Hauptstrukturierungskriterien für den gesamten Wortschatz in WordNet sind die oben angesprochenen Sinnrelationen, allen voran die Synonymie und die Hyponymie. Nach den Angaben in Miller u.a. (1993: 6) wurde eine erste Einteilung des Wortschatzes in Synonymencluster ("Synsets") durchgeführt, also in Gruppen von Lexemen, die mindestens eine Bedeutung teilen. Die meisten anderen kodierten Relationen, abgesehen von Oppositionsbeziehungen, setzen dann nicht die Lemmata in Beziehung zueinander, sondern operieren auf den Synonymenclustern. Die Kodierung von WordNet setzt also nicht bei Wortformen an, sondern bei der Klassifizierung von Wortbedeutungen, wobei die diskreten Bedeutungseinheiten durch die Synonymencluster definiert sind. Eine Ausnahme von dieser Regel macht die Sinnrelation der Opposition, die als Relation zwischen Lemmata kodiert ist (vgl. Miller u. a. 1993: 7) — diese Relation spielt allerdings im Bereich der Nomina eine untergeordnete Rolle.

Neben den genannten Sinnrelationen, die bereits in das System eingearbeitet sind, nennen Miller u. a. (1993: 19f, 22f) als anzustrebende Kodierungselemente noch Zeiger auf Attribute (etwa *klein* für *Kanarienvogel*, in bezug auf das Hyperonym *Vogel*) und Angaben der Funktion (etwa *schneiden* für *Messer*). Diese Relationen, die zwischen den bisher separaten Bedeutungkodierungen für die verschiedenen Wortarten vermitteln könnten, sind allerdings bisher nicht in WordNet implementiert.

Der Nominalwortschatz in WordNet umfaßt ca. 57 000 Nomina, deren Bedeutungen zu ca. 49 000 Synonymensets zusammengefaßt werden (vgl. Miller u. a. 1993: 10). Dieser Wortschatz ist aufgrund der Hyponymierelation zwischen den Synonymensets hierarchisch strukturiert. Für die Hierarchie der Nomina wurden 25 semantische Grundklassen ("unique beginners") aufgestellt, die zunächst als Wurzelknoten jeweils separater Hierarchien dienten. Allerdings wurden diese Grundklassen dann aufgrund 'erkannter natürlicher Gruppierungen' (nach Miller u. a. 1993: 17) wieder zusammengefaßt in Oberklassen wie "thing, entitity", so daß im Endeffekt eine Hierarchie mit wesentlich weniger Wurzelknoten entstanden ist. Die 25 angenommenen Grundklassen, die nach Angaben von Miller u. a. (1993: 16f) aufgrund von Adjektiv-Nomen Kombinationen aufgestellt wurden, sind in alphabetischer Ordnung die folgenden (nach Miller u. a. 1990: 248 und Miller u. a. 1993: 16f) (2.1).

Während im Bereich der Konkreta die Klassifikation in WordNet relativ unproblematisch nachzuvollziehen ist, sind im Bereich der Abstrakta die Probleme der Klassifizierung deutlich zu erkennen. Als Illustration seien hier die angegebenen Hyperonyme zum Lemma *chance*

herangezogen:

Sense 1

opportunity, chance – (a possibility due to a favorable combination of circumstances; “now is your chance“)

=> possibility, possibleness – (capability of existing or happening or being true)

=> being, beingness, existence – (the state or fact of existing: “a point of view gradually coming into being“; “laws in existence for centuries“)

=> state – (the way something is with respect to its main attributes; “the current state of knowledge“; “his state of health“; “in a weak financial state“)

Sense 2

luck, fortune, chance, hazard – (an unknown and unpredictable natural phenomenon that causes an event to result one way rather than another)

=> phenomenon – (any state or process known through the senses rather than by intuition or reasoning)

Sense 3

chance – (a risk involving danger; “you take a chance when you let her drive“)

=> risk, peril, danger – (a venture undertaken without regard to possible loss or injury; “he saw the rewards but not the risks of crime“)

=> venture – (any venturesome undertaking esp one with an uncertain outcome)

=> undertaking, project, task – (any piece of work)

=> work – (activity directed toward making or doing something; “she checked several points needing further work“)

=> activity – (any specific activity or pursuit; “they avoided all recreational activity“)

=> act, human action, human activity – (something that people do or cause to happen)

Sense 4

probability, chance – (a measure of how likely it is that some event will occur; “what is the chance of rain?“)

=> measure, measurement – (magnitude as determined by measurement or calculation)

=> magnitude – (relative size or extent)

=> property – (an attribute shared by objects)

=> attribute – (an abstraction belonging to or characteristic of an entity)

=> abstraction – (a concept formed by extracting common features from examples)

(Quelle: WordNet, Version 1.5)

Es wird besonders beim Vergleich der Bedeutungen 1, 2 und 3 deutlich, daß hier augenscheinlich relativ ähnliche Bedeutungen nicht in dieselbe Kohyponymklasse fallen und selbst bei einer Einbeziehung der sehr weit oben in der Hierarchie liegenden Hyperonyme keine gemeinsamen übergeordneten Begriffe besitzen. Dies geht einher mit der Fragwürdigkeit der genannten Hyperonyme, wenn man die üblichen Tests (X *ist ein* Y oder einen Austauschtest unter Beibehaltung der Wahrheitsbedingungen) für die Ausdrücke aus den aufgeführten hyponymischen Hierarchien heranzieht. Es handelt sich in diesen Beispielen jedenfalls um eine sehr eingeschränkte Form der Hyponymiebeziehung (vgl. die Quasihyponymie im Sinne von Lyons 1977: 299). Dieselbe Problematik bei der Aufteilung von Bedeutungen von Abstrakta zeigt sich allerdings in allen mir bekannten herkömmlichen Lexika; nur fällt sie hier weniger auf, da normalerweise keine explizite und formalisierte Begriffshierarchie vorhanden ist.

An diesem Beispiel zeigt sich, daß es keineswegs trivial ist, eine hyponymische Struktur für große Wortschatzbereiche aufzustellen. Dem sollte Rechnung getragen werden, indem Kodierungskriterien erarbeitet werden, die es ermöglichen, in problematischen Wortschatzbereichen einer Einteilung in irgendwelche undurchsichtige semantische Klassen vorzubeugen. Hierbei kommen zwei Lösungen in Betracht:

- Einerseits die striktere Beachtung des kontextuellen Verhalten der Lexeme, auch wenn dies zur Aufstellung von relativ kleinen Klassen führt.

- Andererseits der Verzicht auf eine ohnehin kaum für irgendwelche Anwendungen verwertbare hierarchische Strukturierung in den unklaren Fällen. Kohyponymklassen zu größeren Klassen zusammenzufassen ist immer dann sinnlos, wenn sich weder ein gemeinsames sprachliches Verhalten der verschiedenen Klassen, noch eine sinnvolle ontologische Klassifizierung ergibt.

Trotz der teilweise fragwürdigen Einordnungen — v. a. im Bereich nicht-konkreter Nomina — kann die Klassifikation in WordNet, die ja immerhin eine der umfangreichsten elektronisch verfügbaren, formalisierten Wortschatzklassifikationen auf Basis der Bedeutung darstellt, in mancher Hinsicht als Muster für ähnliche Vorhaben betrachtet werden. Es stellt sich von daher die Frage nach der Übertragbarkeit einiger Aspekte der Klassifikation auf das Deutsche.

Da die Unterschiede zwischen der semantischen Grobstruktur des Wortschatzes innerhalb des Bereichs der meisten Sprachen europäischer Herkunft aufgrund des gemeinsamen kulturellen Hintergrunds und ihrer engen Verwandtschaft nicht allzu groß sein dürften, ist anzunehmen, daß sich für eine Strukturierung des Deutschen ähnliche Klassen — natürlich abzüglich der individuellen Präferenzen der Kodierer — ergeben müßten. Wo Probleme in der Übereinstimmung dieser Grobklassen auftreten, lassen sie sich meist darauf zurückführen, daß die Kodierung dieser Wortschatzbereiche sprachübergreifend problematisch ist. Eine Übereinstimmung von Klassen einer Einteilung des deutschen Nominalwortschatzes, die auf der hyponymischen Struktur des Wortschatzes beruht, mit der zitierten Einteilung in WordNet ist ohne weiteres Nachdenken sofort deutlich für eine große Anzahl der Klassen. Als Beispiele seien die Kategorien genannt, die durch Lexeme wie *Tier*, *Pflanze*, *Artefakt*, *Mensch* bezeichnet werden — sie gehören in der semantischen Beschreibung seit jeher zum Standard. Bei einigen anderen Klassen, so etwa *Prozesse* vs. *Aktionen* und *Ereignisse*, erklärt sich die Unterscheidung nicht selbst, läßt sich aber bei einer Betrachtung der untergeordneten Klassen schnell nachvollziehen.

Ich glaube von daher, daß die Klassifizierung des Nominalwortschatzes in WordNet in vieler Hinsicht als Quelle für die Kodierung auch des deutschen Wortschatzes dienen kann und werde bei der Darstellung der semantischen Klassen im CISLEX später noch gelegentlich auf diese Einteilung Bezug nehmen. Dieser Abgleich der Klassen kann jedoch nur bis nur bis zu einer bestimmten Klassifikationstiefe vorangetrieben werden. Nicht grundsätzlich übernommen werden kann im Gegensatz zur Grobklassifizierung die Mikrostruktur des Wortschatzes, da eine direkte Zuordnung von Lexem zu Lexem aufgrund gleicher Bedeutung auch zwischen zwei so eng verwandten Sprachen wie dem Deutschen und dem Englischen nicht möglich ist, insbesondere, wenn man das konkrete sprachliche Verhalten der Lexeme bei der Klassifizierung berücksichtigen möchte. Ähnliches gilt für die meronymische, antonymische und synonymische Strukturierung des Nominalwortschatzes; speziell letztere ist nur in sehr geringem Maße von einer Sprache auf die andere übertragbar.

2.4.9 Forschergruppe um G. Gross

Im *Laboratoire de Linguistique Informatique* an der Universität Paris 13 wird seit einigen Jahren von einer Forschergruppe um Gaston Gross an einem Wörterbuch für das Französische gearbeitet, das eine morphologische, syntaktische und semantische Beschreibung der Lemmata beinhaltet. Das Kernstück der semantischen Beschreibung ist die Zuordnung der Lemmata zu syntakto-semantischen Selektionsklassen, die als 'classes d'objets' (Objektklassen) bezeichnet werden. Während die morphologische Kodierung des französischen Nominalwortschatzes weitgehend abgeschlossen ist, sind im Bereich der semantischen Beschreibung nur bestimmte Teile des Nominalwortschatzes (Menschenbezeichnungen, Verkehrsmittel) weitergehend semantisch kodiert. Eine Einteilung in Grobklassen liegt allerdings für den gesamten Nominalwortschatz bereits vor.

Gaston Gross beschreibt die Objektklassen in seinen Arbeiten Gross (1992, 1994) als inhaltliche Klassen, die nicht primär auf einer ontologischen Klassifikation der Denotate oder auf psychologisch fundierten Ähnlichkeiten zwischen den Lexemen zugeordneten Konzepten beruhen. Sie sind vielmehr Klassen von Lexemen, die aufgrund linguistischer —

genauer: syntakto-semantischer Regularitäten — aufgestellt werden. Die größte Rolle spielt dabei der strukturelle Kontext der Lexeme. Diejenigen Lexeme kommen in eine gemeinsame Objektklasse, die mit denselben Operatoren (Verben, aber auch Nomina und Adjektive) kombinierbar sind (die Bezeichnung 'Objektklasse' ist zu verstehen als 'Komplementklasse'). Es handelt sich also dem Ansatz nach um Selektionsklassen, nicht um taxonomische oder ontologische Klassen:

> "On est amené alors à postuler, pour rendre compte de l'emploi précis d'un opérateur donné, des informations supplémentaires concernant les arguments. Il s'agit de sous-catégoriser les traits en sous-classes sémantiques, que nous appelons classes d'objets, et qui seules permettent de discriminer le sens de l'opérateur avec la précision nécessaire ou à la reconnaissance ou à la génération de phrases correctes." (Gross 1994: 18)

Gross formuliert hier als Ziel seiner semantischen Klassifizierung von Nomina die Kodierung selektionsrelevanter Klassen. Dabei sind unter 'opérateur' in diesem Zitat in erster Linie Verben zu verstehen. Selektionsklassen spielen — wie Gross immer wieder betont (vgl. Gross 1994: 17f) — als Grundlage für die maschinelle Übersetzung eine wichtige Rolle, da vor allem solche Klassen hierfür relevant sind, die es erlauben, verschiedene Bedeutungen eines Verbs oder eines Adjektivs im Kontext zu disambiguieren und die Wahl des richtigen Operators zu steuern.

Eine semantische Klasse von Nomina ist in diesem Ansatz durch eine Menge von Operatoren definiert, die sich — bei festgelegter Argumentstelle — mit der Klasse kombinieren lassen. Diese Konzept läßt sich ohne weiteres auf das Deutsche übertragen. Eine Objektklasse beispielsweise, die man mit einem intuitiv einsichtigen Klassenbezeichner Kleidung nennen könnte, wäre jene, die sich definieren ließe als die Menge aller Nomina, die als Akkusativobjekt zu *anziehen/ ausziehen/ tragen* auftreten können (hierunter fiele also im Deutschen nicht ein Lexem wie *Hut*).

Die von Gross angestrebte Kodierung von selektionsrelevanten Klassen unter Zuhilfenahme ihres strukturellen Kontextes entspricht sicher — wie aus den bisherigen Ausführungen zu möglichen Anwendungen bereits deutlich wurde — zumindest einem der Ziele, die auch für die semantische Strukturierung des Nominalwortschatzes im CISLEX zu setzen sind. Allerdings ist die Übertragung des genannten theoretischen Anspruchs auf die konkrete Kodierungsarbeit an einem Lexikon in vielerlei Hinsicht problematisch. Vor einer Aufstellung selektionsrelevanter Klassen für eine Klassifikation des Nominalwortschatzes im Sinne von Gross wären nämlich folgende grundlegenden Fragen zu klären:

- Zunächst entsteht das Problem der Klassenauswahl. Welche der Selektionsklassen sind interessant für den Kodierer, d. h. welche Verben oder Adjektive werden ausgewählt, um zunächst die relevantesten Objektklassen zu definieren? Für das Deutschen existieren ca. 20 000 Verben (Lemmata im CISLEX, ohne komplexe Verben), von denen viele wiederum polysem sind[34].

- ; fast jedes dieser Verben hat eine bis zu einem gewissen Punkt idiosynkratische Argumentstruktur. Werden hochfrequente Verben ausgewählt, stößt man auf das Problem, daß gerade sie meist keine klar definierten Klassen abgrenzen, da Häufigkeit und Polysemie korreliert sind. Andere Kriterien für die Auswahl von Verben zur Aufstellung

[34]Zudem zieht Gaston Gross, wie man beispielsweise an dem in Gross (1994:20f) aufgeführten Beispiel sieht, nicht nur verbale Kontexte, sondern beinahe jeden Typ von Kontext zur Definition einer Objektklasse heran; für die Klasse ¡défauts humains¿ ('menschliche Fehler': Bezeichnungen für Menschen aufgrund von negativen Eigenschaften) nennt er dort unter anderen die Kriterien: Kann im Vokativ stehen, kommt im Kontext "Espèce de X" oder "X que tu es" vor.

von Klassen sind jedoch schwer anzugeben. Es wird also zunächst schwierig sein, solche Verben auszuwählen, die für die Kodierung von Selektionsklassen aufschlußreich sein können. Läßt man beliebige Kombinationen von Verben zu, die in bestimmten Lesarten Mengen von Nomina selegieren, deren Teilmenge nicht leer ist, kommt es zu einer kombinatorischen Explosion, die das Unternehmen kodierungstechnisch ad absurdum führt. Nicht wenige der so erlangten Klassen wären aber zumindest für einige wenige Operatoren wichtig zur Lesartenunterscheidung. Ein zusätzliches Problem ist der häufig übertragene Gebrauch von Verben, den vom nicht-figürlichen Gebrauch abzugrenzen notorisch schwierig ist. Eine Aufstellung von selektionsrelevanten Klassen von Nomina muß also doch mit der Aufstellung von (intuitiven) Nominalklassen beginnen, für die dann die richtigen Operatoren gesucht werden. Eine solche Aufstellung intuitiv relevanter Klassen wird ihre Motivation durch die taxonomische Struktur des Wortschatzes nicht verleugnen können.

- Wer entscheidet aufgrund welcher Kriterien, ob ein bestimmter Operator mit einem Nomen kombinierbar ist? Diese Entscheidung ist in vielen Fällen nicht objektiver, als eine Zuordnung aufgrund ontologischer Kriterien, da auch die Akzeptanzentscheidungen für Verb-Objekt-Kombinationen sicher in nicht wenigen Fällen auf Basis von ontologischen Überlegungen erfolgen. Das genannte Verfahren der Abgrenzung durch Operatoren ist also weniger objektiv als es zunächst scheinen mag. Ein Kodierer, der versucht, Zuordnungen der genannten Art vorzunehmen, wird bald feststellen, daß ihm die Kriterien, die er anwendet, teilweise selbst unklar sind und daß sein Gefühl für Richtigkeit und Falschheit bei längerer Betrachtung eines Beispiels unscharf wird. Bei entsprechend großem Wortschatz und mehreren Kodierern ist also sicher auch mit dieser theoretischen Prämisse keine vollständige Einheitlichkeit der Kodierung erreichbar[35].

- Eine Entscheidung über die Zuordnung aufgrund großer Textkorpora zu fällen, würde hier für einige Operator-Nomen-Paare sicher Klarheit bringen, aber die Benutzung von Texten löst weder das Problem der figürlichen Bedeutung noch das der Polysemie. Hinzu kommen früher genannten generellen Probleme des Zugangs zu Korpora und der Auswertung von Daten aus großen Textmengen. Der Glaube, man könne klare Kriterien finden, um bei der semantischen Zuordnung eindeutige ja/nein-Entscheidungen fällen zu können, läßt zudem völlig die Unschärfephänomene außer acht, die gerade im Bereich von Selektionsrestriktionen unübersehbar sind.

Wie werden nun die genannten Punkte in Gross' eigenen Kodierungsbeispielen behandelt? Bei einer Betrachtung seines Vorgehens in konkreten Fällen wird relativ schnell klar, daß die Aufstellung der Klassen in den meisten Fällen zunächst eher nach ontologischen Kriterien erfolgt, und erst in der Folge nach Operatoren gesucht wird, die diese Klasse dann eingrenzen. Es wird also zur konkreten Kodierung die Berücksichtigung der Hyponymiestruktur des Wortschatzes und ein distributionsorientiertes Vorgehen vermischt. Man vergleiche hierzu die folgenden Aussagen aus Mathieu-Colas (1993), einem Schüler von Gross, der die Kodierung der französischen Bindestrich-Nomina beschreibt. Die Vermischung der verschiedenen Kriterien, wie sie hier angedeutet wird, kann sicher auch für die semantische Kodierung anderer Teile des Nominalwortschatzes durch Gross selbst gelten:

[35] Gross (1994: 18f) verleitet die Betrachtung der semantischen Klassifizierung im SYSTRAN-Übersetzungsystem zu der Bedauernsäußerung "On imagine le désarroi du codeur!" ('Man stelle sich die Verwirrung des Kodierers vor') (ebd. 19). Allerdings ist jede semantische Klassifizierung zunächst nur ihrem Urheber voll durchsichtig und muß von einem neuen Kodierer stets in einem langen Lernprozeß nachvollzogen werden. Dies gilt sicher auch für die semantische Klassifikation von Gross.

> "Les classes dont nous parlons sont des catégories génériques subsumant des groupes de mots sémantiquement et syntaxiquement apparentés, et désignés comme telles par des hyperonymes" (Mathieu-Colas 1993: 162)
>
> "Chacune des classes d'objets ne doit sa validité, en dernière analyse, qu'aux propriétés linguistiques qui lui sont associés" (ebd. 163)
>
> "Il conviendrait idéalement de n'introduire de classes qu'en les accompagnant d'une définition rigoureuse, c'est-à-dire une description grammaticale. Nous avons préferé adopter provisoirement une attitude plus empirique ... " (ebd.: 168)

Hier werden als Kriterium zur Einteilung des Nominalwortschatzes nacheinander angeboten: Kontext, gemeinsames Hyperonym und schließlich "une attitude plus empirique", eine pragmatischere Haltung, in diesem Falle die Übernahme von — natürlich nicht ursprünglich distributionell sondern ontologisch motivierten — Klassifikationen aus Lexika.

Die Tatsache, daß es sich bei der Klassifizierung im Sinne von Gross nicht um primär distributionell entwickelte Klassen handelt, drückt sich schon in der Bezeichnung der Klassen aus (z. B. "moyens de transport", 'Verkehrsmittel', "fleurs", 'Blumen', vgl. Gross 1994). Bei der ersten Aufstellung scheint es nicht um eine Beschreibung von Selektionsklassen zu gehen, sondern um Klassen von Kohyponymen. Der Status von Selektionsklassen wird erst erreicht, wenn die zunächst erfolgte Klassifizierung durch die Festlegung der Operatoren insofern revidiert wird, als dann Selektionsrestriktionen zum einzigen für die Klassenzugehörigkeit relevanten Kriterium werden, und wenn größere Klassen aufgrund der Kombinatorik der in ihnen enthaltenen Lexeme weiter unterteilt werden. Allerdings gelingt es Gross (1992, 1994) meines Dafürhaltens nur in einigen Fällen, klar zu demonstrieren, daß er auf Basis einer vorausgegangenen hyponymischen Einteilung kohärente, relativ klar abgrenzbare Selektionsklassen aufstellen kann. Ein gut dokumentiertes Beispiel für eine fundierte Kodierung aufgrund von Operatoren sind vor allem für die Unterklassen der "moyens de transport", der Verkehrsmittel (1994: 23-26), die er in mehrere, sich teils überschneidende Unterklassen aufteilt, so daß das Resultat wie eine gründliche merkmalssemantische Beschreibung dieses Wortfelds wirkt. Er teilt den Wortschatzbereich[36] nach den folgenden Aspekten auf: öffentliche - nicht-öffentliche (aufgrund der Operatoren wie *verpassen*), zweirädrige - andere (*umkippen/ stürzen von*), Verkehrsmittel in Luft/ Land/ Wasser, schienengebunden - nicht-schienengebunden (*entgleisen*), motorgetrieben - nicht-motorgetrieben (keine Operatoren angegeben). Wie man an den genannten Distinktionen erkennt, handelt es sich hier zwar um selektionsrelevante, aber auch aufgrund ontologischer Kriterien beschreibbare Klassen. Die genannten Operatoren sind hier zunächst nur Motivation zur Aufstellung der Klassen, die dazugehörigen Lexeme lassen sich daraufhin ohne weiteres aufgrund nicht-distributioneller Überlegungen zuordnen.

Etwas problematisch vor dem Hintergrund der Ansprüche wirkt die vorgenommene Grobklassifizierung (Gross 1994: 20). Er unterscheidet diese Grobklassifizierung von der Einteilung in semantische Klassen, indem er ihre Bezeichner als syntakto-semantische Merkmale bezeichnet. Seiner Argumentation folgend sind sie allerdings als die umfassendsten Objektklassen seiner Kodierung aufzufassen. Er nennt acht Klassen:

> *humain, animal, végétal, inanimé concret, inanimé abstrait, locatif, temps, événement*

[36] Die Modellierung wird im Folgenden zum leichteren Verständnis und aufgrund der unproblematischen Reproduzierbarkeit der Einteilung nicht auf Französisch wiedergegeben, sondern gleich auf das Deutsche übertragen.

> ('menschlich, tierisch, pflanzlich, unbelebt konkret, unbelebt abstrakt, lokativ, Zeit, Ereignis')
>
> (Gross 1994: 20)

Nur für die Gruppe *végétal* gibt Gross einige distributionelle Kriterien für die Einteilung an, die allerdings relativ trivial erscheinen.[37] Interessieren würden hier vor allem die Abgrenzungen, die er zwischen Abstrakta, Lokativa und Konkreta zieht, da die Angabe der Kriterien für die Einteilung von Grenzfällen in diese Klassen bei jedem Typ von Klassifikationsmotivation — ontologisch wie distributionell — problematisch ist.

Aus den angegebenen Beispielen geht deutlich eine Parallelität der Einteilungsproblematik hervor: in den Bereichen, in denen die Einteilung aufgrund von Kohyponymie relativ einfach vornehmbar ist, scheint auch die Einteilung aufgrund sprachlichen Verhaltens relativ unproblematisch. Das Gegenteil gilt für viele Klassen allerdings ebenso, insbesondere im Bereich der Abstrakta. Wenn die Einteilung aufgrund ontologischer Überlegungen fehlschlägt, bieten auch die Operatoren hierfür keine Grundlage.

Eine weitere sehr aufschlußreiche Falluntersuchung auf der Basis der von Gross gemachten Annahmen findet sich in Le Pesant (1994). Er untersucht die nominalen Komplemente des Verbs *lire* ('lesen') und teilt sie aufgrund des Vorkommens als Komplemente anderer Verben (wie *imprimer* ('drucken') und *inscrire sur* ('beschriften')), in Unterklassen ein. Es zeigt sich hier meines Erachtens, daß eine solche Untersuchung in erster Linie dazu dient, herauszufinden, wie fein die Einteilung eines Wortfeldes sein muß, um für die Selektionspräferenzen auch nur in einem Bereich des Verbalwortschatzes aufzukommen. Zudem zeigt sich, daß bei den betroffenen Verben immer nur einige wenige sich zu Gruppen mit einheitlichem Selektionsverhalten verbinden lassen, und daß auf der anderen Seite nur einige wenige Nomina in eine weitgehend einheitliche Selektionsklasse fallen. Le Pesant selbst sagt:

> "La richesse de vocabulaire est parfois telle qu'on peut faire de fines subdivisions de classes d'objets grâce à la seule prise en compte des operateurs appropriés"
> (Le Pesant 1994: 43)

Die zitierten Kodierungsbeispiele können nicht davon überzeugen, daß ein Vorgehen im Sinne von Gross prinzipiell verschieden ist von herkömmlichen Formen der Klasseneinteilung aufgrund von begrifflicher Unterordnung, denn zahlreiche taxonomische Klassifikationen des Wortschatzes ziehen den sprachlichen Kontext als Entscheidungshilfe heran (vgl. z. B. Selektionsrestriktionen als Kriterien für die Erstellung der Grundkategorien für das oben genannte WordNet oder für die Klassifikation der Nomina wie dargestellt in Dahlgren 1988: 45). Ein Verdienst der Ausführungen von G. Gross zu den Objektklassen ist es aber sicherlich, darauf hingewiesen zu haben, daß das tatsächliche sprachliche Verhalten eines Lexems zu seiner semantischen Einteilung in höherem Maße als üblich herangezogen werden sollte. Es muß allerdings aus seinen und verwandten Arbeiten gefolgert werden, daß eine nur auf distributionellen Kriterien fußende Einteilung zu Problemen bei der praktischen Umsetzung führt, was es letztlich nötig macht — zumindest für eine große Zahl der aufgestellten Klassen — doch wieder auf zunächst nicht unmittelbar durch den sprachlichen Kontext bedingte Klassifikationen zurückzugreifen. Eine weitere wichtige Schlußfolgerung ergibt sich aus den vorgestellten Untersuchungen, die an seinen Begriff der Objektklassen orientiert sind: einige wenige semantische Klassen für Nomina sind nicht ausreichend, um eine größere Anzahl von Verbrahmen mit ihnen adäquat semantisch beschreiben zu können. Bei jeder größeren semantische Klasse von Nomina werden eine Reihe von Detailstudien hierfür notwendig sein.

[37] Gross teilt sie in eine von den Konkreta gesonderte Klasse ein, da sie mit Verben wie *wachsen*, *leben*, *sterben* zusammen auftreten (vgl. Gross 1994: 20).

Ein wichtiges Erfordernis erscheint mir darüber hinaus die Kodierung der Information über die Motivation für die Aufstellung einer semantischen Klasse, d. h. ob es ich um eine rein distributionelle Klasse handelt, oder ob taxonomische Einteilungen relevant waren.

Ein weiterer wichtiger Punkt in Gross' Ausführungen ist die Ablehnung einer Baumstruktur für die vorgenommene semantische Klassifizierung (vgl. Gross 1994: 19). Er argumentiert vor allem damit, daß bestimmte Lexeme gleichzeitig zu verschiedenen Klassen gehören können. Seine Aufstellung von Objektklassen verzichtet von daher auch darauf, einer solchen Klasse einen exakten Stellenwert in einer Hierarchie zuzumessen. Er nimmt jedoch jeweils eine Zuordnung seiner Objektklassen zu einer seiner acht Grundklassen vor und schließt eine nachträgliche feinere Hierarchisierung der Klassen nirgends aus. Die Ablehnung einer Baumstruktur findet sich auch bereits in der in Dahlgren (1988) beschriebenen Klassifizierung und ist auch in WordNet praktisch umgesetzt; bei beiden Kodierungen entsteht der Bedarf von Mehrfachzuordnungen von Unter- zu Oberklassen. Der Konsens für die Kodierung größerer Wortschatzausschnitte scheint also zu sein, daß eine adäquate Ordnung semantischer Klassen es zulassen muß, daß ein Knoten mehrere Mutterknoten hat.

2.4.10 Zusammenfassung

Die beiden diskutierten Konzeptionen zur semantischen Kodierung in WordNet und durch G. Gross wiesen einige Gemeinsamkeiten auf. Die Grundlage der semantischen Beschreibung ist in beiden Fällen die Gruppierung der Lexeme zu Klassen von bedeutungsähnlichen Ausdrücken, die in beiden Fällen in der Mehrzahl als Klassen einer hyponymischen Gliederung des Wortschatzes interpretierbar sind, wobei als Klassifizierungskriterium jeweils das kontextuelle Verhalten der zu klassifizierenden Lexeme eine Rolle spielt. Während in WordNet aus den minimalen Bedeutungsklassen — Synonymenclustern — von unten nach oben eine vollständige Taxonomie aufgebaut wird, verzichtet Gross zunächst auf die Definition einer vollständigen Struktur des Wortschatzes. Im Gegensatz zur Konzeption von WordNet stellt das Vorgehen von Gross Instrumente zur Verfügung, um solche Klassen zu beschreiben, die nicht durch eine begriffliche Unterordnung erfaßt werden können, und die nur distributionelle Eigenschaften von Lexemen erfassen.

Die nähere Betrachtung beider semantischen Beschreibungen deutet auf eine prinzipielle Problematik bei einer semantischen Gliederung des Nominalwortschatzes hin: Während die Klassifikation bestimmter Wortschatzbereiche relativ einfach vorzunehmen ist, weil hier sowohl sprachliche als auch ontologische Kriterien eine Einordnung erlauben, sind andere Bereiche sehr schwierig aufgrund dieser Kriterien zu klassifizieren. Dies gilt für beide Lexika insbesondere im Bereich der nicht-konkreten Nomina, für deren Einteilung ontologische Überlegungen von vornherein fraglich sind und deren kontextuelles Verhalten in vielen Fällen ebenfalls keine Zusammenfassung zu größeren Bedeutungsgruppen erlaubt. Es ist von daher zu erwarten, daß auch bei der Kodierung der Semantik im CISLEX die semantische Einteilung der Nomina mit abstrakter Bedeutung die größten Probleme bereiten wird.

2.4.11 Anwendungserfordernisse

Im Bereich der Lexika für die maschinelle Sprachverarbeitung ist zu unterscheiden zwischen elektronischen Wörterbücher im Sinne des CISLEX auf der einen und Maschinenlexika für spezielle Anwendungen (z. B. in Systemen zur maschinellen Übersetzung) auf der anderen Seite. Die Kodierung von Information ist in letzteren zugeschnitten auf die jeweilige Anwendung und auf den zu verarbeitenden Sprachausschnitt. Die Kodes sind, insbesondere im Bereich der Semantik, nur bedingt übertragbar und damit kaum für andere als die ursprüngliche Anwendung nutzbar. Für die Kodierung eines elektronischen Wörterbuches wie

des CISLEX ist die Betrachtung der Anforderungen an eine semantische Kodierung in Maschinenlexika insofern von Interesse, als es gilt, Gemeinsamkeiten zwischen den wichtigsten Anwendungstypen in der Sprachverarbeitung herauszuarbeiten und die gemeinsamen Eigenschaften zu einer Grundlage der semantischen Kodierung zu machen. Im folgenden werden exemplarisch die Anforderungen an die Bedeutungskodierung von Nomina für zwei Anwendungen untersucht, die gänzlich verschiedene Typen von semantischer Kodierung verlangen. Dies sind einerseits die maschinelle Übersetzung und andererseits das Information Retrieval.

2.4.12 Semantische Beschreibung von Nomina für die maschinelle Übersetzung

Ein einsprachiges Lexikon allein, kombiniert mit einem einsprachigen Lexikon einer anderen Sprache, ist sicher nicht geeignet, den lexikalischen Transfer für die maschinellen Übersetzung eines Texts zu gewährleisten. Hierzu sind zweisprachige Lexika mit entsprechenden Transferregeln notwendig. Eine geeignete einsprachige Kodierung kann aber doch wichtige Informationen über Regularitäten in der Quell- oder der Zielsprache liefern und damit in der maschinellen Übersetzung als ein Modul eingesetzt werden. Zudem kann ein einsprachiges elektronisches Lexikon als eine Entwicklungsgrundlage für ein zweisprachiges Lexikon dienen.

Die Eignung eines elektronischen Wörterbuchs als ein Modul in einem Übersetzungssystem kann eine Kodierung gewährleisten, die es einerseits in der Quellsprache erlaubt, polyseme Lexeme zumindest teilweise zu disambiguieren und damit die Zahl der möglichen Übersetzungen von vornherein zu restringieren und die andererseits in der Zielsprache Entscheidungskriterien zur Auswahl des geeigneten Lexems im Rahmen der Regularitäten dieses Sprachsystems zur Verfügung stellt. Auch für den Bereich der Syntax kann eine geeignete Kodierung der Wortsemantik sowohl in der Quell- als auch in der Zielsprache Hinweise auf die Auflösung ambiger Konstruktionen bieten. Die Anforderungen an ein einsprachiges Lexikon entsprechen damit im Wesentlichen den Ansprüchen an ein elektronisches Lexikon, wie es auch für die Sprachgenerierung oder die Sprachanalyse notwendig ist[38].

Wie lassen sich nun die genannten Anforderungen konkret für die angestrebte Kodierung im CISLEX umsetzen? Zum Zweck der maschinellen Übersetzung und der darin eingeschlossenen Teilaufgaben ist es meiner Ansicht nach realistisch, bei der Konzeption einer Nominalsemantik vor allem die drei folgenden, eng verwandten Aufgaben im Auge zu behalten:

- Unter verschiedenen Möglichkeiten der Übersetzung eines Nomens ist die adäquate Übersetzung zu selegieren. Diese wird in hohem Maße abhängen von den Selektionspräferenzen des/ der zu dem Nomen gehörigen Operator/ Operatoren (vgl. Gross 1992, 1994), also vor allem von der Kombinatorik mit Adjektiven bzw. Verben, aber auch mit nominalen Köpfen in semantisch transparenten Komposita. Zur Beschreibung dieser Selektionspräferenzen ist eine Unterteilung des Nominalwortschatzes in Klassen notwendig, deren Elemente von einer möglichst großen Anzahl häufiger Operatoren selegiert werden. Dabei ist es utopisch, von vornherein alle möglichen selektionsrelevanten Klassen von Nomina aufstellen zu wollen, da die meisten Verben und auch viele der anderen Operatoren in dieser Hinsicht Idiosynkrasien aufweisen. Zunächst müssen bei der Klassifizierung des Nominalwortschatzes die für ein Maximum an Operatoren relevanten größeren Klassen im Auge behalten werden. Diese Klassifizierung

[38] Eine Aufarbeitung der Literatur zum Thema Semantik in der maschinellen Übersetzung würde hier zu weit führen. Die folgenden Annahmen sind weitgehend unumstritten. Man vergleiche z. B. zur maschinellen Übersetzung und semantischen Merkmalen im Zusammenhang mit Selektionsrestriktionen Lehrberger (1988: 103ff) oder die Aussagen in Behrens (1993: 257).

kann dann allmählich verfeinert und auch auf kleinere Klassen ausgedehnt werden, die nur für wenige Operatoren relevant sind. Eine große Zahl von Verb-Nomen und Adjektiv-Nomen Kollokationen werden allerdings überhaupt nicht von dieser Art von Generalisierung erfaßt werden können, sondern müssen als idiosynkratische Verbindung direkt ins Lexikon eingetragen werden. Für eine Kodierung gilt hierbei das Prinzip der Verhältnismäßigkeit der Mittel: eine relativ große Klasse von Nomina, die für die Selektionspräferenzen einer relativ großen Klasse von Operatoren relevant ist, ist wesentlich interessanter für die Beschreibung der Kombinatorik und damit für die maschinelle Übersetzung als kleine Klassen, die nur für wenige und seltene Operatoren eine Rolle spielen. Werden die Klassen zu klein und sind sie für die Beschreibung von Selektionspräferenzen nur einiger weniger Operatoren relevant, ist eine direkte Aufnahme der Kollokationen in ein entsprechendes Teillexikon sowohl kodiertechnisch als auch anwendungstechnisch effizienter.

- Unter verschiedenen möglichen Übersetzungen eines Operators, also eines Verbs, eines Adjektivs oder eines als Operator fungierenden anderen Nomens von der Quell- in die Zielsprache ist eine geeignete Übertragung auszuwählen. Bezüglich dieses Punktes gelten prinzipiell dieselben Aussagen wie für den zuvor genannten. Eine Beschreibung von Selektionspräferenzen mittels semantischer Klassen ermöglicht auch hier die Auswahl des geeigneten Lemmas, in diesem Fall des Operators.

- Zusätzlich können Selektionspräferenzen von Prädikaten über semantische Klassen teilweise zur syntaktischen Disambiguierung herangezogen werden. Man betrachte etwa das folgende Satzmuster, in dem X eine Leerstelle bezeichnet:

  7 Er sah den Mann mit dem X

- Eine Beschreibung des semantischen Typs von X kann hier herangezogen werden, um zu erkennen, ob es sich um ein Instrumentaladverbial, wie im Falle der Füllung der Leerstelle durch *Fernrohr* und andere optische Geräte, oder um ein Attribut zum Nomen *Mann* handelt.

Die genannten Anforderungen lassen sich prinzipiell für den gesamten Nominalwortschatz verallgemeinern. Eine semantische Einteilung des Fachwortschatzes[39] kann allerdings in der ersten semantischen Kodierung des CISLEX schon aufgrund des nicht für alle Fachgebiete vorhandenen Spezialwissens kaum angestrebt werden. Fachspezifische Lemmata werden zunächst nur mit einem Hinweis auf das entsprechende Gebiet versehen, so daß gegebenenfalls ein rascher Zugriff möglich ist, um sie dann mit Hilfe eines Fachwörterbuchs kodieren zu können. Es bleibt als realistisches Zielsetzung, was die umittelbare Einsetzbarkeit des CISLEX als Modul für die maschinelle Übersetzung betrifft, eine Einteilung des nichtfachsprachlichen Wortschatzes in selektionsrelevante semantische Klassen.

2.4.13 Semantische Beschreibung von Nomina für das Information Retrieval

Die Zuordnung eines Textes zu thematischen Gruppen von Dokumenten erfordert zunächst eine Beurteilung des Themas des Dokuments. Hierzu gilt es, "den Inhalt eines Textes/

[39] Die Diskussion der Abgrenzung von Fach- und Gemeinsprache soll hier nicht referiert werden. Vgl. z. B. Fluck (1985: Kap. 1). Es geht bei unserem Gebrauch des Begriff zunächst um die Ausgrenzung eines Teils des Wortschatzes, der für fachliche Diskurse spezifisch ist, und der nicht unbedingt von Nicht-Fachleuten verstanden wird, der also auch nicht ohne weitere Erklärungen etwa im Zeitungstext einer allgemeinen Tageszeitung vorkommt.

Dokuments auf wesentliche Begriffe abzubilden“ (Panyr 1989: 699). Die Indexierung mittels der in einem Text enthaltenen Nomina spielt für diesen Vorgang eine herausragende Rolle, denn diese Wortart verhält sich wesentlich bereichs- und damit themensensitiver als etwa die Verben (vgl. die statistischen Untersuchungen in Ott 1992).

Die Nomina, die in einem Text vorkommen, lassen erkennen, von was für Typen von Entitäten der Text handelt. Um eine brauchbare Klassifizierung eines Dokuments leisten zu können, muß zusätzlich zur bloßen Indizierung eine Gewichtung der gefundenen Nomina aufgrund der Relevanz für die Einordnung erfolgen (vgl. Panyr/ Zimmermann 1989: 700f). Hierzu ist es zunächst notwendig, die Nomina zu identifizieren, die besonders textsorten- und themenspezifisch sind, und sie von jenen abzuheben, die über verschiedene Texte relativ gleichmäßig verteilt sind.

Zur inhaltlichen Klassifizierung eines Textes auf Grundlage der themenspezifischen Nomina ist eine Klassifikation des Nominalwortschatzes in einem Thesaurus erforderlich. Thesauri im Information Retrieval[40]sind ontologische Begriffshierarchien, deren Begriffen Lemmata zugeordnet sind; sie sind also eine auf thematischen Kriterien beruhende Einteilung des Wortschatzes oder eines Teilwortschatzes einer Sprache. Die Kriterien zur Erstellung eines solchen Thesaurus richten sich nach dem Fachgebiet der zu klassifizierenden Texte. Zahlreiche Nomina haben in unterschiedlichen Fachbereichen verschieden Bedeutungen. Das macht es meist unmöglich, die Thesauri für verschiedene Fachbereiche zusammenzuführen.

Das Ziel der Kodierung des gesamten deutschen Wortschatzes der einfachen Nomina kann somit nicht sein, eine Einteilung des Wortschatzes zu liefern, die für sämtliche mögliche Fachbereiche Relevanz besitzt; ebensowenig jedoch kann das Ziel sein, einen Thesaurus für eine bestimmte Subdomäne zu erstellen. Das Ziel muß eine Gliederung sein, die durch die Möglichkeit zur Identifikation der bereichsrelevanten Lemmata die rasche Erststellung eines fachbezogenen Thesaurus erlaubt, und die zusätzlich für den nicht fachgebundenen Diskurs eine Klassifizierung des Wortschatzes liefert, der die thematische Einordnung eines nicht fachspezifischen Textes ermöglicht. Diese Anforderungen lassen sich folgendermaßen umsetzen:

- Nomina, die ausschließlich oder in erster Linie einem Fachwortschatz zugehören, werden markiert als diesem Subwortschatz zugehörig. Diese Art der Kodierung ist eine Einteilung des Wortschatzes in Bereiche. Sie ermöglicht sofortigen Zugriff auf die Lemmata, die bei der Erstellung eines Thesaurus für ein Fachgebiet primär relevant sind.
- Für Nomina, die nicht nur für einen Fachwortschatz relevant sind, wird eine Einteilung in Klassen vorgenommen, die geeignet sind, für einen nicht fachgebundenen Text eine erste Zuordnung zu einem Themenfeld aufgrund der Art von Entitäten vorzunehmen, um die es im Text geht. Diese Art der Kodierung ist eine Einteilung in allgemeinsprachliche thematische Klassen.

2.4.14 Maschinelle Übersetzung - Information Retrieval: Gemeinsamkeiten der Anforderungen

Den Anforderungen durch beide genannten Typen von Anwendungen ist die Markierung fachspezifischer Wortschatzbereiche gemein, die in beiden Fällen aufgrund gleicher Kriterien

[40]Der Begriff 'Thesaurus' wurde bisher schon einige Male verwendet. Es ist jedoch Vorsicht geboten, denn unter Thesauri werden in unterschiedlichen Zusammenhängen Beschreibungseinheiten unterschiedlichen Typs und mit unterschiedlichen Strukturen verstanden: Die zuerst erwähnten Thesauri in Textverarbeitungsprogramm sind letztendlich nichts anderes als maschinenlesbare Synonymensammlungen; Thesauri als Untertyp des Wörterbuchs sind Begriffshierarchien, deren Einträge Sammlungen von thematisch/hyponymisch zusammenhängenden sprachlichen Ausdrücken sind; Thesauri in Systemen zum Information Retrieval sind thematisch strukturierte Begriffshierarchien, meist für ein eingeschränktes Themengebiet.

erfolgen kann. Eine spezielle Kodierung fachspezifischer Lexeme ist auch für beinahe alle anderen bereits im ersten Kapitel dieser Arbeit genannten Anwendungen eines elektronischen Wörterbuchs — bis hin zur automatischen Rechtschreibkorrektur — von Nutzen.

Anders verhält es sich dagegen mit den angesprochenen Voraussetzungen für die semantische Einteilung des Allgemeinwortschatzes. Auf den ersten Blick scheint zwischen einer Gliederung des nominalen Wortschatzes, wie sie das Information Retrieval von allgemeinsprachlichen Texten erfordert — also einer thematischen Einteilung — und einer solchen, die für die maschinelle Übersetzung verwendbar ist, ein prinzipieller Widerspruch zu bestehen:

- Die semantische Klassen, wie sie für die der maschinellen Übersetzung benötigt werden, sind Klassen von Nomina, deren Aufstellung und Abgrenzung auf geteilten Selektionskontexten beruhen. Sie beschreiben das sprachliche Verhalten lexikalischer Einheiten und haben damit nicht von vornherein eine ontologische Basis.

- Die semantischen Klassen für das Information Retrieval sind letztlich ontologisch begründete Klassen. Sie beschreiben eine durch die Sprechergemeinschaft geteilte Klassifizierung von Entitäten in der realen Welt und haben keine unmittelbare Beziehung zum distributionellen Verhalten der zu ihnen gehörigen Lexeme.

Es ist nun zu fragen, inwieweit beide Klassifizierungstypen sich trotz der unterschiedlichen Ansatzpunkte gleichen, d. h. inwiefern sich für eine Kodierung nach zunächst unterschiedlich wirkenden Kriterien doch noch ein gemeinsamer Nenner finden läßt. Meiner Ansicht nach ist zu erwarten, daß zwischen beiden Typen von Klassifizierungen tatsächlich eine große Übereinstimmung besteht: Der Tatsache, daß auf der Ausdrucksseite gleiche Prädikationen über verschiedene Ausdrücke gemacht werden können, liegt zugrunde, daß, was auf der Seite der Denotate als Ähnliches empfunden wird, ähnliche Eigenschaften aufzuweisen scheint, und in ähnliche Ereignisse involviert ist. Eine Brücke zwischen beiden Bereichen ist die im Abschnitt zu den Sinnrelationen bereits genannte Hyponymie bzw. Unterordnungsbeziehung, die eine Relation zwischen Lexemen ist, die auf der ontologischen Gliederung der Denotate basiert. Die bekannten sprachlichen Testrahmen, also die Kontexte *X ist ein Y* oder *X ist eine Art von Y*, die immer wieder für diese Sinnrelation genannt werden, geben inhaltlich betrachtet nichts wesentlich anderes wieder als eine Klassifizierung von Entitäten in Welt durch die Sprecher.

Diese Übereinstimmung kann selbstverständlich nicht für alle aufgestellten Klassen gelten. Es wird solche semantischen Klassen geben, die — auf der Basis sprachlicher Regularitäten aufgestellt — nur für die Beschreibung von Selektionspräferenzen Relevanz besitzen, und solche, die sich — auf der hyponymischen Struktur des Wortschatzes beruhend — in erster Linie als thematisch relevante Klassen erweisen, deren Eignung für die Disambiguierung von Operatoren sehr eingeschränkt ist.

Die im folgenden geschilderte Einteilung des Nominalwortschatzes für das CISLEX erhebt damit den Anspruch, bis zu einem gewissen Punkt sowohl distributionellen wie auch ontologischen Kriterien gerecht zu werden. Die Darstellung der Kodierung ausgewählter semantischer Klassen im folgenden Kapitel und die in Kapitel 4 geschilderten statistischen Untersuchungen werden noch zeigen, daß die Annahme bedingter Isomorphie durchaus praxistauglich ist.

2.5 Prämissen für die Konzeption einer Nominalsemantik

2.5.1 Formalisierung paradigmatischer Theorien

Nachdem ein rein mechanistischer Kollokationsansatz, der als einziger unter den in den Abschnitten zur Beschreibung der Wortsemantik untersuchten Ansätze sowohl voll formalisierbar als auch vorgehenstechnisch und von den Ergebnissen her eindeutig definierbar ist, für die Bedeutungskodierung in einem elektronischen Wörterbuch für unzureichend erachtet wird, bleibt die Frage, wie es nun mit möglichen Formalisierungen für andere der diskutierten Theorien der Bedeutung aussieht.

Wie man am Dargestellten erkennen konnte, gibt die Betrachtung der klassischen wortsemantischer Theorien keine Anlaß zu übertriebenen Zielsetzungen: Es gibt keine formalisierbare semantische Theorie, die nicht mit zahlreichen Problemen behaftet ist. Zu diesen Problemen gehören vor allem die Festlegung des Beschreibungsinventars und der Umgang mit Unschärfephänomenen. Die Ansätze der kognitiven Linguistik, die solche Phänomene berücksichtigen, legen zunächst keine Grundlage für eine formale semantische Beschreibung, sondern erschöpfen sich allzu oft in allgemeinen Aussagen. Die vorgestellten Ansätze zur semantischen Kodierung größerer Wortschatzausschnitte deuten ebenfalls darauf hin, daß keine operationalisierbaren kanonischen Verfahren zur Bedeutungsbeschreibung vorliegen.

Das Ergebnis unserer Übersicht deckt sich also mit folgender Aussage von Calzolari:

> "The net result of this is that we cannot assume we already know an optimal format or mechanism to at least represent lexical semantics. This must be remembered in any proposal for standards (albeit practical ones)." Calzolari (1994: 271)

Wie gezeigt werden konnte, hat diese Situation in der traditionellen Lexikographie vor allem dazu geführt, daß der Semantikteil eines Eintrags meist eine nicht eindeutig strukurierte Umschreibung der Bedeutung eines Lemmas ist, die ein nicht näher definiertes Hintergrundwissen bei den Benutzern voraussetzt. Eine solche Strategie ist für die maschinelle Lexikographie nicht gangbar.

Formal gesehen, d. h. unter Außerachtlassung natürlichsprachlicher Bedeutungsumschreibung, die in einem elektronischen Lexikon höchstens ergänzend zu einer formalen Beschreibung vorgesehen werden kann, gibt es prinzipiell nur zwei Methoden der Bedeutungskodierung für Lexeme:

- Den Lexemen werden Elemente einer Beschreibungssprache zugeordnet.
- Die Lexeme werden mittels eines Inventars von Bedeutungsrelationen anderen Lexemen zugeordnet.

Bei beiden Ansätzen können verschiedene Relationen zwischen den zu beschreibenden Lexemen und dem zugeordneten Element bestehen.

Die Menge der Elemente einer Beschreibungssprache kann wiederum strukturiert sein, indem die Elemente in eine Taxonomie eingebunden oder durch Folgerungsbeziehungen miteinander verknüpft werden.

Die klassische Dekompositionssemantik ist die am weitesten ausformulierte Methode der ersten Art: je nach den dahinterstehenden Annahmen wird eine offene oder geschlossene Menge von Merkmalen angenommen und eine Mehrfachzuordnung aus dieser Merkmalsmenge zu jedem Lexem vorgenommen. Zudem ist in den meisten Dekompositionstheorien

eine Strukturierung der Beschreibungssprache vorgesehen — d. h. eine Hierarchisierung der Merkmale oder eine andere Form des Bezugs der Merkmale aufeinander.

Eine deutliche Verbindung einer Merkmalsbeschreibung zur Theorie der Sinnrelationen ergibt sich dadurch, daß häufig, wenn es um die Auswahl der Merkmale geht, als Deskriptoren Zeichenketten gewählt wurden, deren Interpretation als direkte Bedeutungswiedergabe bestimmter Lexeme - für Nomina sind dies insbesondere Adjektive - sich geradezu aufdrängt (z. B. die schon genannten Merkmale ¡männlich¿ oder ¡unverheiratet¿ für Junggeselle). Es ist zumindest für diese Fälle — und sie scheinen bei der semantischen Dekomposition den überwiegenden Teil der Beschreibungselemente auszumachen — zu fragen, wo die Vorteile eines Dekompositionsansatzes gegenüber der Darstellung von Sinnrelationen zwischen objektsprachlichen Einheiten zu finden sein sollen, wenn die angebliche Unterscheidung der Merkmale von den entsprechenden Lexemen nur postuliert wird, aber keine praktischen Konsequenzen hat. Ohne eine weitere Bedeutungszuordnung zu den semantischen Deskriptoren[41], die ja Elemente einer im Grunde nicht bedeutungshaltigen Beschreibungssprache sind, hat man an sich nichts anderes erreicht als eine vermittelte Strukturierung des Wortschatzes[42]. Dies mag die Aufstellung metasprachlicher Beschreibungseinheiten auf den ersten Blick unnötig erscheinen lassen. Der genannte Sachverhalt muß jedoch nicht unbedingt negativ aufgefaßt werden. Ich glaube nämlich, daß eine solche vermittelte Strukturierung durchaus einen sinnvollen Weg zur Bedeutungskodierung in einem elektronischen Wörterbuch darstellt, die zumindest für manche Beschreibungsaufgaben gewisse Vorteile gegenüber einer direkten Kodierung semantischer Relationen zwischen objektsprachlichen Einheiten hat:

- Es gibt linguistisch relevante Bedeutungsklassen von Lexemen, die keinen natürlichsprachlichen Bezeichner haben. Dies zeigt sich auch in der WordNet-Klassifikation — die ja primär als Strukturierung aufgrund der Hyponymierelation konzipiert ist — indem einige der Klassen relativ künstlich wirkende Kollokationen als Bezeichner haben. Umso problematischer wäre eine solche direkte Kodierung für einen morphologisch begründeten Wortschatzausschnitt wie die 'einfachen Nomina'. Hier müßte ständig auf Lexeme zurückgegriffen werden, die im zu kodierenden Wortschatz eigentlich nicht enthalten sind.

- Lexeme sind potentiell polysem, Klassenbezeichner und Merkmalsbezeichner nicht; den metasprachlichen Einheiten kann eine semantische Umschreibung oder eine Menge von klassenkonstituierenden Kontexten beigegeben werden.

Eine indirekte lexikographische Darstellung zumindest der taxonomischen Strukturierung des Wortschatzes erscheint mir also gegenüber einer direkten Kodierung mittels Relationen zwischen Lexemen von Vorteil zu sein. Insbesondere gilt dies für die vorzunehmende Kodierung der 'einfachen Nomina' im CISLEX, welche ja nur einen Ausschnitt aus dem gesamten Nominalwortschatz darstellen. Ein Zugang etwa zum objektsprachlichen Hyperonym einer Klasse von Kohyponymen kann dann immer noch durch eine besondere Kenntlichmachung dieses Lexems erreicht werden.

Die Zuordnungssequenzen, die für eine solche vermittelte Strukturierung verwendet werden, können dabei verschiedener Art sein. Es sind letztlich die Zuordnungskriterien, die für sie dann einen Typ festlegen, insbesondere, ob sie Klassen von Lexemen zusammenfassen, die aufgrund der hyponymischen Struktur des Wortschatzes konstituiert werden (taxonomische

[41] Ich meine hier eine Bedeutungszuordnung etwa in Form der Interpretation in bezug auf ein Modell.

[42] Dies wurde bereits von D. Lewis (1970) der Merkmalssemantik vorgeworfen. Er prägte für einen nicht weiter semantisch interpretierbaren merkmalssemantischen Beschreibungsformalismus den abfälligen Ausdruck “Markerese“.

Klassen), oder solche Klassen, denen eine Menge geteilter Selektionskontexte gemeinsam ist (Selektionsklassen).

2.5.2 Taxonomische Klassen und Selektionsklassen

Im nächsten Kapitel dieser Arbeit werden die Konzeption und die Ausführung der Kodierung der Nominalsemantik im CISLEX dargelegt werden. Zur Hinführung will ich nun nach der Option für die vermittelte Methode der Beschreibung die wichtigsten inhaltlichen Grundbedingungen nennen, wie sie sich aus den bisherigen Erörterungen ergeben haben:

- Eine vollständige Beschreibung der Semantik eines Lexems ist nicht möglich. Die Bedeutungskodierung sollte an den zentralen Bedeutungsaspekten eines Wortes ansetzen.
- Kein Ansatz zur Bedeutungsbeschreibung kann ein Wort isoliert vom umgebenden Wortschatz betrachten oder isoliert von den sprachlichen Kontexten, in denen es vorkommt.
- Wesentliches Ziel einer Kodierung sollte die Einteilung des Nominalwortschatzes in Klassen von bedeutungsähnlichen Lexemen sein. Ohne eine solche Einteilung fehlt jegliche Grundlage für eine Kodierung anderer Bedeutungsrelationen und anderer semantischer Eigenschaften. Als Kriterien für Bedeutungsähnlichkeit kommen zunächst begriffliche (hyponymische Struktur des Wortschatzes) und sprachlich-distributionelle Kriterien in Betracht.

Tabelle reftabelle1 stellt kurz die Eigenschaften von semantischen Klassifizierungen nach unterschiedlichen Kriterien dar:

Wie bereits oben festgestellt, kommt für eine erste Kodierung der rein statistische Ansatz der Kollokationsklassen nicht in Frage. Eine zentrale Vorüberlegung für die Festlegung der Bedeutungskodierung im CISLEX war, daß die Überschneidungen zwischen den Klassen einer hyponymischen Hierarchie und von Selektionsklassen wesentlich umfangreicher sind, als es zunächst aufgrund der unterschiedlichen Zuordnungskriterien scheinen mag. Für die Durchführbarkeit einer Kodierung aufgrund dieser Annahme gab es durch die WordNet-Klassifizierung und die Kodierung der Objektklassen, wie in Gross (1992, 1994) beschrieben, einige Anhaltspunkte:

- Für die Aufstellung der hyponymischen Struktur in WordNet wurden Selektionspräferenzen von Adjektiven ausgewertet (vgl. Miller u.a. 1990, 1993). Außerdem ist die Klassifikation relativ erfolgreich zur Beschreibung von Selektionsrestriktionen eingesetzt worden (Resnik 1993, Ribas 1994, 1995).
- Eine große Zahl der in Gross (1994) genannten Selektionsklassen sind gleichzeitig Klassen von Kohyponymen.

Dabei ist es nicht legitim, die beiden Kriterien — hyponymische Struktur eines Teilwortschatzes und geteilte Selektionskontexte — unmarkiert zu vermischen. Insbesondere bei einer Hierarchisierung der Klassen könnten dabei Inkonsistenzen entstehen. Es ist also sinnvoll, die Kriterien zur Klassenaufstellung in der Kodierung festzuhalten. Nicht sinnvoll ist es aber, vom Kodierungsaufwand her betrachtet, beide Klassifizierungen vollständig zu trennen. Für die Einteilung des Nominalwortschatzes sind für bestimmte Bereiche Selektionspräferenzen als Kodierungskriterien wichtiger, für andere Bereiche ist Kohyponymie das entscheidende Kriterium. Bei einer großen Zahl von Klassen werden allerdings beide Kriterien zu übereinstimmenden Einordnungen führen, wie bei der detaillierten Beschreibung einzelner Klassen

	ist_ein Hierarchie/ taxonomische Klassen	**Selektionsklassen/ 'Objekt'klassen**	**Kollokationsklassen**
Grundlage für die Klassifizierung	Klassen von Kohyponymen. Zuordnung zu gemeinsamem Hyperonym	Durch Operatorenkombinationen bestimmte Selektionsklassen	Distribution, statistisch ausgewertet
Vorgehen zur Klassenerstellung	Lexeme vorhanden, Auswahl subjektiv, zusätzliche Klassen	subjektiv, Auswahl bestimmter Operatoren	objektiv nach statischen Werten
Vorgehen zur Einordnung in Klassen	Intuition, Testrahmen, intersubjektiv	gemeinsame Operatoren, Intuition	
Linguistische Kriterien	*ist-ein*-Test, Ersetzungen	ling. Tests, Kombinatorik mit Operatoren	statistische/ syntaktische Auswertung
Quellen	Intuition, Eigenbelege, Lexika, Weltwissen	Intuition, Eigenbelege, Texte, Lexika	große Textkorpora
Objekte	Lexeme, Klassenbezeichner	Lexeme, Klassenbezeichner	Lexeme, Lexemmengen
Relationen	Unterordnungsrelation zwischen Lexemen	Zuordnung von Lexemen zu Klassen, Hierarchisierung der Klassen	Zuordnung von Lexemen zu Klassen
Beschreibung der Klassen	Hyperonym, Umschreibung	Sprachlicher Kontext, Hyperonym	Statistische Nähe, Klassenelemente
Umgang mit Prototypikalitätseffekten	Heckenausdrücke, graduelle Hyponymie	nicht vorgesehen	Statistische Zwischenwerte
naheliegende Anwendungen	Information Retrieval	Automatische Übersetzung	Automatische Übersetzung?
Bereiche, in denen das Verfahren gut funktioniert	Konkreta, v. a. wo wissenschaftliche Taxonomien Gemeingut geworden sind (Tierwelt)	bestimmte Einzelklassen, die in eindeutiger sprachlicher Umgebung auftreten	funktioniert überall, Ergebnisse nicht immer intuitiv nachvollziehbar/ anwendbar
problematische Wortschatzber.	"Abstrakta"	vorwiegend ontologisch motivierte Klassen	
ausgew. Literatur	Lyons (1977) Cruse (1986)	Gross (1992,1994)	Hindle (1990) Church/Hanks (1990)

Tabelle 2.2: Vergleich unterschiedlicher Möglichkeiten zur semantischen Klassifizierung des nominalen Wortschatzes

im CISLEX im nächsten Kapitel noch deutlich werden wird. Die Uneinheitlichkeit für die Aufstellungs- und Zuordnungskriterien für die Klassen machen es notwendig, sie für jede Klasse durch gesonderte Merkmale zu beschreiben. Diese legen die verschiedenen Typen von semantischen Klassen fest und ermöglichen letztendlich auch eine konsistente Hierarchisierung.

Die (in der Beschreibung der meisten Kodierungen stillschweigend übergangene) Vermischung verschiedener Kriterien soll in der semantischen Klassifizierung im CISLEX explizit ausgedrückt werden, indem zunächst zwei Merkmale von Klassen festgelegt werden, von denen mindestens eines positiv spezifiziert sein muß:

- Taxonomische Klassen: Dieses Merkmal tragen semantische Klassen, die in erster Linie auf der hyponymischen Struktur des Wortschatzes beruhen, d. h. begriffliche motivierte Lexemgruppen. Klassen, die für dieses Merkmal positiv spezifiziert sind, erfordern keine Beschreibung der typischen Kontexte, in denen die in ihnen enthaltenen Lexeme vorkommen, schließen solche Beschreibungen aber auch nicht explizit aus. Typische Klassen dieser Art liegen bei der Untergliederung der Bezeichner für Tiere in Ausdrücke für Säugetiere, Reptilien etc. vor, die in hohem Maße der biologischen Klassifikation folgt.

- Selektionsklassen: Dies sind semantische Klassen, welche in erster Linie aufgrund einer relevanten Menge von geteilten Selektionskontexten aufgestellt werden. Reine Selektionsklassen sind nicht in jedem Fall durch ein Hyperonym gedeckt und spiegeln damit nicht die begriffliche Unterteilung des Wortschatzes wieder.

- Eine große Zahl der aufgestellten Klassen werden allerdings sowohl unter dem Aspekt der begrifflichen Gliederung als auch des Selektionsverhaltens aufgestellt. Ein typisches Beispiel sind die Wasserfahrzeuge, die einerseits eine begriffliche Klasse konstituieren und andererseits durch eine Reihe typischer Selektionskontexte ausgezeichnet sind (die Lexeme treten typischerweise als Kopf der Nominalphrase *X* in den folgenden Kontexten auf: *an Bord von X gehen*, *X sinkt*). Dieser Typ von Klasse ist der Default-Fall. Ist eine semantische Klasse nicht explizit als taxonomische Klasse oder als Selektionsklasse markiert, ist sie unter beiden Aspekten relevant.

2.5.3 Unschärfephänomene

Das verbleibende Problem ist nun die Einbindung der Erkenntnisse über Unschärfephänomene, wie sie Prototypenansätze beschreiben, in eine Klassifikation des Wortschatzes. Weder in der Lexikondatenbank WordNet noch in der semantischen Klassifikation durch G. Gross werden diese Phänomen explizit berücksichtigt.

Jegliche Art der semantischen Einteilung des Wortschatzes sollte aber meines Erachtens immer unter dem Gesichtspunkt durchgeführt werden, daß sie auf scharfen Grenzziehungen beruht, die der Realität der Sprachverwendung nur teilweise adäquat sind. Idealerweise sollte eine Einteilung in Klassen die damit zusammenhängenden Eigenschaften dieser Klassen beschreiben — also auch Information darüber enthalten, ob eine Klasse eine prototypische Struktur hat und ob sie scharfe oder unscharfe Ränder besitzt. In diesem Fall sollten extrem periphere oder zentrale Mitglieder einer semantischen Klasse als solche markiert werden.

Auch bei einer formalisierten Beschreibung ist es möglich, einer semantischen Klasse Eigenschaften zuzusprechen, die sich auf die interne Struktur der Lexemgruppe beziehen, die durch Zuordnung zu dieser Klasse beschrieben wird. Folgende Eigenschaftsparameter kommen dabei in Betracht:

Prototypische Struktur: Die Kriterien, die bei semantischen Klassen mit prototypischer Struktur zu einer Zuordnung führen, ermöglichen es, von einem typischen Vertreter der Gruppe zu sprechen. Dieses Merkmal ist somit positiv zu spezifizieren, wenn die zugrunde liegende Kategorie prototypische Vertreter hat.

Unscharfe Ränder: Dieses Merkmal beschreibt die Gradierbarkeit von Klassenzugehörigkeit. Ist es negativ spezifiziert, erlauben es die Kriterien, die zu einer Zuordnung führen, zwischen Lexemen, die der Klasse zugeordnet werden und solchen, die nicht in ihr enthalten sind, klar zu unterscheiden, d. h., es gibt es keine Grenzfälle der Zuordnung.

Kapitel 3

Semantische Kodierung der Nomina im CISLEX

Im vorigen Kapitel wurden die theoretischen Voraussetzungen der Bedeutungskodierung für ein elektronisches Wörterbuch erörtert und einige Projekte zur semantischen Kodierung größerer Wortschatzausschnitte vorgestellt. Das nun folgende Kapitel ist der Darstellung der praktischen Umsetzung dieser Überlegungen gewidmet. In den folgenden Abschnitten werden systematisch die Durchführung und die Ergebnisse der semantische Kodierung für die einfachen Nomina im CISLEX dargestellt.

Unter den gegebenen theoretischen und praktischen Voraussetzungen ist es nicht möglich, eine semantische Kodierung zu konzipieren, welche die Anforderungen von Seiten aller denkbaren Anwendungen eines elektronischen Wörterbuchs vollkommen zufriedenstellt. Es wurde von daher entschieden, für die Nomina im CISLEX zunächst eine Kodierung vorzunehmen, die als eine erste Grundlage für die spätere Ausrichtung des Lexikons auf verschiedene Anwendungen dienen kann. Berücksichtigt wurden bei der Konzeption insbesondere Überlegungen zu den Bereichen des Information Retrieval und der automatischen Übersetzung, die beide auf den ersten Blick ganz unterschiedlich wirkende Anforderungen stellen, für die es aber möglich ist, einen gemeinsamen Nenner zu finden.

Neben einer weitgehenden Anwendungsneutralität mußte bei der Konzeption der Nominalsemantik stets im Auge behalten werden, daß der Formalismus mit einer folgenden Kodierung der Semantik anderer Wortarten kompatibel sein sollte. Vor allem sollte er sich zur Beschreibung der Argumentstruktur von Verben und Adjektiven eignen. Als gangbare Strategie wurde herausgearbeitet, daß als erster Schritt eine Klassifikation des Nominalwortschatzes in semantische Klassen zu erfolgen hat. Diese Einteilung wurde für das CISLEX in zwei Stufen vorgenommen: Zunächst wurde die semantische Grobklassifikation des Wortschatzes angegangen, daraufhin wurden die Grobklassen in feinere semantische Klassen unterteilt.

Neben der Anwendbarkeit der erzielten Einteilung des Wortschatzes in Klassen — man vergleiche hierzu Kapitel 4 — kann sie in weiteren Kodierungsabschnitten als Grundlage für die feinere semantische Beschreibung dienen. Bisher wurde v. a. die Kodierung verschiedener Sinnrelationen in Angriff genommen.

In den folgenden Abschnitten wird die konkrete Klasseneinteilung für die Nomina beschrieben. Kodiert wurden zunächst die ca. 37 000 Nomina im Teillexikon der 'einfachen Formen' (EF, s. Einleitungskapitel). Dies ist meines Wissens das erste Mal, daß (beinahe) sämtliche nicht-präfigierten und nicht-komplexen Nomina des Deutschen für ein elektroni-

sches Wörterbuch vollständig und systematisch semantisch kodiert wurden. Die Kodierung wird also auch aus der Sicht der quantitativen Distribution dieses grundlegenden Teilwortschatzes über die aufgestellten semantischen Klassen jenseits des Anwendbarkeitskriteriums — dem bei einem elektronischen Lexikon die Hauptaufmerksamkeit gilt — Interesse in der theoretischen Linguistik beanspruchen können.

Die Beschreibung der Kodierung und der dabei auftretenden Fragestellungen folgt im wesentlichen dem tatsächlichen Kodierungsverlauf:

- Zunächst werden die Kodierungskriterien und die benutzten Hilfsmittel vorgestellt und erläutert.
- Es folgt die Beschreibung der Eigenschaften verschiedener Typen semantischer Klassen. Diese Eigenschaften dienen vor allem dazu, die Relevanz einer Klasse für unterschiedliche Anwendungen festzulegen.
- Dann wird zunächst die Grobklassifikation des gesamten Nominalwortschatzes vorgestellt. Diese Einteilung ging der Aufstellung detaillierter semantischer Klassen voran.
- Der Begriff der semantischen Klasse im CISLEX wird vom Begriff des semantischen Relators abgegrenzt.
- Es wird die Kombinatorik der Klassen erörtert. Formalismen zur Kodierung von mehrfacher Klassenzugehörigkeit und von Polysemie werden vorgestellt.
- Die Beschreibung einiger ausgewählter semantischer Klassen verschiedenen Typs soll die konkrete Kodierung exemplarisch darstellen.
- Verschiedene Unterordnungsrelationen, die für die Hierarchisierung von semantischen Klassen eine Rolle spielen, werden definiert.
- Abschließend wird die Kodierung von Sinnrelationen beschrieben. Dies sind Hyponymie, Synonymie (diskutiert wird auch deren Abgrenzung zur Variantenbildung), Opposition und Meronymie.

3.1 Kodierungskriterien und Hilfsmittel

Bei einem nicht-objektiven Vorgehen — und dies sind abgesehen von einer rein maschinellen Kollokationsanalyse in unterschiedlichem Maße alle Methoden zur semantischen Beschreibung in einem Lexikons — ist es vor allem notwendig, ein Vorgehen zu definieren, bei dem für die Kodierung in unterschiedlichen Phasen und durch unterschiedliche Kodierer möglichst einheitliche Kriterien für die semantische Beschreibung herangezogen werden.

Wie auch die Beurteilung der Grammatikalität im Bereich der Syntax sind Urteile über semantische Eigenschaften — und meist in noch höherem Maße als Grammatikalitätsurteile — immer von der Intuition einzelner Sprecher abhängig. Wie im vorigen Kapitel ausführlich erörtert, ist es beim jetzigen Stand der quantitativen Linguistik, d. h. der statistischen Analyse natürlicher Sprache auf Basis von Korpora, nicht möglich, die semantische Kodierung durch einen Lexikographen durch ein objektives Kodierungsverfahren auf Grundlage maschineller Auswertung von konkreten Texten zu ersetzen, da die Anzahl der zu kontrollierenden Parameter für eine statistische Analyse zu groß, bzw. der Umfang verfügbarer Korpora für die Kodierung seltener Wörter wesentlich zu klein ist. Zudem liegen fast alle vorhandenen Korpora für das Deutsche nicht linguistisch aufbereitet vor. Darüber hinaus ist überhaupt

fraglich, inwiefern Korpora die für eine semantische Kodierung relevante Information überhaupt liefern können.

Das soll nicht heißen, daß die Auswertung von sprachlichen Belegen nicht eine wichtige Grundlage zur Beschreibung der Semantik für ein Lexikon sein sollte. Die Konsultation von Textstellen, die in der traditionellen Lexikographie von großer Wichtigkeit ist, muß auch für die semantische Kodierung eines elektronischen Wörterbuchs erfolgen; hierbei erleichtern in elektronischer Form vorliegende Korpora die Arbeit beträchtlich. Die schlußendliche Auswertung des Materials ist allerdings nicht maschinell zu bewältigen.

Bei der Kodierung der Nomina spielten die Belege aus konkreten Texten vor allem eine große Rolle bei der Festlegung typischer gemeinsamer Kontexte für die Lexeme in den semantischen Klassen, die als Selektionsklassen ausgezeichnet sind. Als Korpora wurden dabei in erster Linie die Texte geschriebener Sprache herangezogen, die das IDS in Mannheim unter dem COSMAS-System zur Verfügung stellt[1]. Zudem wurden einige am CIS vorhandene Korpora als Belegquelle herangezogen, die hauptsächlich aus den Ausgaben der 'Süddeutschen Zeitung' von 1993-1995 bestehen.

Ein ebenso wichtiger Fundort für Textbelege zum Zweck der semantischen Kodierung sind bereits existierende Lexika in verschiedener Form. Zusätzlich zu den dort aufzufindenden Textbelegen wird die semantische Information noch in anderer Form aufbereitet angeboten, meist durch die Angabe von Synonymen, Oppositionen und Hyperonymen sowie durch sogenannte "Definitionen". Die Frage der Auswertbarkeit von herkömmlichen Lexika für ein elektronisches Wörterbuch wurde bereits im vorigen Kapitel ausführlich erörtert. Dabei wurde herausgestellt, daß es hierfür aufgrund der nicht-formalisierten und nicht vollständig systematischen Darstellung der Semantik in solchen Lexika stets der Vermittlung durch den Kodierer der formalisierten, maschinenlesbaren Semantik bedarf. Für die Bedeutungskodierung der Lemmata im CISLEX wurden der sechsbändige Duden Drosdowski u. a. (1977) und Wahrigs 'Deutsches Wörterbuch' (1986) herangezogen.

Zur Auswertung von Textbelegen und Lexika kommt die sprachliche Kompetenz des Kodierers. So ist es etwa nicht nötig, für die Einordnung des Lexems *Reh* in die taxonomische Klasse Tier irgendwelche Hilfsmittel heranzuziehen, denn dieser Einordnung liegt eine allen Sprechern bekannte hyponymische Teilstruktur des Wortschatzes zugrunde; in einem solchen Fall ist auch ohne die Angabe zusätzlicher Kriterien intersubjektive Nachprüfbarkeit gegeben.

Schwieriger wird die Angabe eindeutiger Kriterien für die Klassenzugehörigkeit bei vielen anderen Gruppen von Lexemen, für die keine hundertprozentige Einigkeit zwischen verschiedenen Sprechern zu erzielen ist. In diesen Fällen ist es unverzichtbar, daß der Kodierer eine Reihe von Kriterien auflistet, die zur semantischen Zuordnung führen. Dies gilt vor allem für die Aufstellung von und Zuordnung zu Selektionsklassen, in eingeschränktem Maße aber auch für taxonomische Klassen in bestimmten Wortschatzbereichen. Diese Entscheidungskriterien sind auch im Fall taxonomischer Klassen in erster Linie sprachlicher Natur — denn auch hier liegen nur in den seltensten Fällen vollständig objektive Zuordnungsgründe vor. Die Entscheidungskriterien lassen sich für beide Typen von Klassen im weitesten Sinne als kontextuelle Tests auffassen. Zum einen spielen Kontexte bei der Angabe von Sinnrelationen eine Rolle, da zu deren Bestimmung sprachliche Tests herangezogen werden. Zum anderen werden die Mitglieder von Selektionsklassen durch die direkte Angabe typischer oder definitorischer Kontexte charakterisiert. Die angewendeten sprachlichen Kriterien lassen sich grob in folgende Klassen aufteilen (vgl. auch Cruse 1986):

- Kontexte zum Testen von Sinnrelationen

[1]Zum COSMAS-System vgl. al-Wadi (1994). Die Auflistung der Korpora schriftsprachlicher Texte am IDS findet sich ebd. auf S. 257ff.

- Zuordnungskontexte (Hyponymie), die bei Klassen mit unscharfen Rändern mit Heckenausdrücken kombiniert werden können, wie in Beispiel (10):

  8 X ist ein Y,

  9 X ist eine Art von Y,

  10 X ist ein typisches Y

- Austauschbarkeitstest (zum Testen von Synonymie und Hyponymie):

  11 Karl besitzt ein Klavier

  12 Karl besitzt ein Piano

  13 Karl besitzt ein Musikinstrument

- sonstige Testrahmen (z. B. für Meronymie):

  14 X ist ein Teil von Y

- Selektionskontexte durch Operatoren (Verben, Adjektive, Nomina)

- Bei der Prüfung von Selektionskontexten handelt es sich darum, zu testen, ob ein Nomen als Kopf eines bestimmten Komplements des Verbs oder des Adjektivs auftreten kann; teilen verschiedene Nomina einen oder eine Menge von solchen Kontexten, konstituieren sie eine Selektionsklasse. Bei der Darstellung der Kodierung werden diese Kriterien noch oft demonstriert werden, weswegen an dieser Stelle auf Beispiele verzichtet werden kann.

- Spezielle Konstruktionen, die ausschließlich oder typischerweise mit bestimmten semantischen Klassen auftreten, z. B. vokativische Kontexte für Schimpfwörter.

Für die kodierten Klassen, insofern sie nicht von vornherein unproblematisch einzugrenzen waren, wurden mit ihrer Aufstellung jeweils solche linguistischen Kriterien festgelegt und aufgelistet. Diese kontextuellen Kriterien konnten im Kodierungsprozeß z. T. anhand von Korpora, zum Teil anhand der Kompetenz des Kodierers abgeprüft werden[2]. Ihre Angabe macht die semantische Kodierung des Wortschatzes überprüfbarer und damit etwas weniger subjektiv.

3.2 Eigenschaften semantischer Klassen

Bereits in der Übersicht zur Wortsemantik in der theoretischen Linguistik (Kapitel 2) hat sich gezeigt, daß für unterschiedliche Klassen von Nomina die Kriterien für die Klassenabgrenzung und -konstituierung unterschiedlich gewichtet werden. Es wurde festgestellt, daß die verschiedenen Kriterien für die semantische Klassifikation, die herangezogen werden müssen, um einen semantisch unrestringierten Wortschatzausschnitt zu kodieren — namentlich die taxonomische Strukturierung aufgrund der Hyponymiebeziehung und die Einteilung in Selektionsklassen — auch zu Unterschieden im Stellenwert der sprachlichen Kriterien für die Klassifikation führen, d. h. zu unterschiedlicher Wichtigkeit und zu unterschiedlichen Arten

[2]Selbst ein großes Korpus enthält nur knapp ein Drittel der Nomina, die in einem vollständigen Nominallexikon auftauchen. Um ein Wort aufgrund der sprachlichen Umgebungen, in denen es auftaucht, klassifizieren zu können, reicht zudem ein Beleg nicht aus.

von typischen sprachlichen Kontexten für Nomina, die zu unterschiedlichen Klassentypen gehören.

Diesen verschiedenen Aufstellungskriterien trägt die erste Gruppe von Merkmalen für semantische Klassen Rechnung — dies sind ¡taxonomische Klasse¿, ¡Selektionsklasse¿, ¡thematische Klasse¿ und ¡Sammelklasse¿. Diese Merkmale implizieren dabei auch bestimmte Verwendungsmöglichkeiten.

Die zweite Gruppe von Merkmalen trägt den Unschärfe- und Typikalitätsphänomenen bei der Zuordnung zu semantischen Klassen Rechnung. Die Merkmale ¡unscharfe Ränder¿ und ¡prototypenzentriert¿ beziehen sich auf die Möglichkeit der Gradierung der Klassenzugehörigkeit von Lexemen und der Klassenzugehörigkeit der Unterklassen zu einer semantischen Klasse.

Eine dritte Gruppe von Merkmalen schließlich — das sind ¡natürliche Art¿, ¡soziale Kategorie¿, ¡Denotatseigenschaften¿ und ¡funktional¿ — bezeichnen Typen von Kategorien, die semantischen Klassen zugrundeliegen[3].

3.2.1 Merkmale für Aufstellungskriterien und Anwendungsbereiche

Am Ende von Kapitel 2 wurde die unterschiedliche Relevanz verschiedener Typen von Klassen zur Beschreibung von Selektionsrestriktionen erläutert. Wie festgestellt wurde, sind einige der taxonomischen Klassen über sprachliche Kontexte kaum zu fassen und eignen sich damit auch nur bedingt zur Beschreibung von Selektionsrestriktionen. Dies gilt vor allem für solche Klassen, die mit einer wissenschaftlichen Taxonomie assoziiert sind, so etwa einige Unterklassen der Tiere (etwa Reptilien, Amphibien). Solche Gruppen von Lexemen, denen ontologische Kategorien zugrundeliegen, und die aufgrund der Voraussetzungen für ihre Aufstellung für die Beschreibung von Selektionspräferenzen nicht von vorherein geeignet sind[4], werden als ¡taxonomische Klasse¿ ausgezeichnet.

Klassen, die aufgrund von geteilten Selektionseigenschaften von Lexemen konstituiert wurden, und die damit auch primär für die Beschreibung von Selektionsrestriktionen relevant sind, jedoch nicht zur begrifflichen Klassifizierung der in ihnen enthaltenen Nomina herangezogen werden können, werden als ¡Selektionsklassen¿ ausgezeichnet.

Eine große Zahl von semantischen Klassen wird allerdings unter beiden Aspekten relevant sein, d. h. bei ihrer Aufstellung und Abgrenzung gegenüber anderen Klassen spielen sowohl die begriffliche Gliederung des Wortschatzes als auch typische Kontexte für die Lexeme in der Klasse eine entscheidende Rolle — was dann auch die entsprechenden Implikationen für ihre Anwendbarkeit hat.

Bisher wurde nicht auf die thematische Zuordnung von Lexemen eingegangen. Zahlreiche der Klassen, die auf den bisher genannten Aufstellungskriterien beruhen, haben jedoch klare thematische Bezüge — d. h. ihr Vorkommen in Texten liefert einen Hinweis auf mögliche thematische Zuordnungen dieser Texte. Für solche Klassen wurde das Merkmal ¡thematische Klasse¿ eingeführt. Die Relevanz dieses Merkmals wird noch deutlich werden, wenn es im Rahmen eines Tests über die Eignung der aufgestellten Klassen für Aufgaben im Bereich des Information Retrievals erforderlich sein wird, bezüglich bestimmter Themengebiete eine Auswahl von themenrelevanten Lexemen zu erstellen (s. Kapitel 4): Hier müssen nur Klassen berücksichtigt werden, für die das Merkmal ¡thematische Klasse¿ positiv spezifiziert ist.

[3] Vgl. zu dieser Einteilung auch die Eigenschaften der in Browns (1990) dargestellten Merkmale von Kategorien.

[4] Dies schließt weder grundsätzlich aus, daß diese Klassen in bestimmten Kontexten präferiert selegiert werden, noch, daß sie sich im Verlauf der Kodierung der Argumentstruktur von Verben oder Adjektiven zusätzlich zu ihrer Eigenschaft, taxonomische Klasse zu sein, als Selektionsklassen erweisen.

Einige Klassen, etwa die Grobklasse Diversa, sind nicht als einheitliche semantische Klassen im Sinne von Selektionsklassen oder taxonomischen Klassen zu verstehen, sondern als durch die Kodierungsökonomie bedingte Sammelklassen, die eine Anzahl von semantischen Klassen zusammenfassen. Solche Klassen entstehen v. a. in den Wortschatzbereichen, für die eine Einteilung anwendungsgesteuert erfolgen sollte, und in solchen, wo die Entwicklung von Unterteilungskriterien noch nicht abgeschlossen wurde. Diese als ¡Sammelklassen¿ ausgezeichneten semantischen Klassen sind also solche Lexemgruppen, die entweder in weiteren Kodierungsstufen noch feiner unterteilt werden müssen, oder solche, die überhaupt nur anwendungsgesteuert unterteilt werden können.

Die Lexeme dieser Klassen verfügen über kein gemeinsames Hyperonym und teilen nicht eine größere Anzahl typischer sprachlicher Kontexte. Die Angabe von Kriterien für die Einordnung von Lexemen erfolgt hier in Form einer Umschreibung der Klasse, wie sie auch in Einträgen in herkömmlichen Lexika üblich ist.

Die Bezeichnung 'Sammelklasse' impliziert, daß die bisher aufgezählten Merkmale negativ spezifiziert sind. Eine wichtige Eigenschaft dieser Klassen ist ferner, daß sie nie eine prototypische Struktur besitzen. Für ihre Abgrenzung gelten dagegen keine speziellen Bedingungen. Sie können unscharfe Ränder haben oder auch klar abgrenzbar sein.

3.2.2 Berücksichtigung von Unschärfephänomenen

Die beiden folgenden Eigenschaften ¡Prototypenzentriertheit¿ und ¡Unscharfe Ränder¿ als Merkmale von semantischen Klassen beziehen sich auf deren interne Struktur. Relevant für Anwendungen werden sie dann, wenn es darum geht, semantische Defaultmerkmale an die Mitglieder einer Klasse zu vergeben.

Defaultmerkmale von Lexemen werden dabei als solche Merkmale definiert, die im Fall von prototypisch strukturierten Klassen auf die prototypischen Lexeme zutreffen, bei Klassen mit unscharfen Rändern nicht (unbedingt) auf die peripheren Mitglieder.

Im Kapitel 2 wurde die Prototypentheorie ausführlich diskutiert. Dabei wurde im Anschluß an Wierzbicka (1990) herausgestellt, daß prototypische Subkategorien und Instanzen nur für einen Teil von Kategorien eine wichtige Rolle spielen. Dies läßt sich auf die semantischen Klassen in der Form übertragen, daß nur eine Teil von ihnen eines oder mehrere Lexeme enthält, die prototypennahe Kategorienmitglieder bezeichnen. Ich sehe im folgenden als prototypenzentrierte Lexemklassen solche an, deren gemeinsames Hyperonym — d. h. das Lexem, das als Überbegriff für die anderen Lexeme der Klasse gelten kann — über ein umschreibbares, zentrales Konzept verfügt. Eine semantische Klasse, die einer prototypischen Kategorie entspricht, erhält das Merkmal ¡prototypenzentriert¿.

Eine positive Spezifizierung des Merkmals der Prototypenzentriertheit geht nicht in jedem Fall mit dem Merkmal ¡unscharfe Ränder¿ einher. So weist etwa die Klasse Vögel Prototypenzentriertheit auf, ohne daß diese Klasse unscharfe Ränder besitzt. Allerdings gibt es zahlreiche Klassen, bei denen beide Eigenschaften gleichzeitig gegeben sind. Ein Beispiel hierfür die Berufe.

Die prototypische Struktur schlägt sich auch in typischen sprachlichen Umgebungen der Lexeme der entsprechenden Klassen nieder — prototypennahe Konzepte bezeichnende Lexeme teilen eine größere Menge von Kontexten untereinander und mit dem Klassenbezeichner als prototypenferne. Prototypische Struktur von Kategorien schlägt sich also auch auf sprachlicher Seite in der Kombinatorik mit Operatoren wieder. So sind etwa typische Kontexte für das Hyperonym der Klasse Vögel — also das Lexem *Vogel* — nicht im Zusammenhang mit den peripheren Lexemen der Klasse verwendbar.

15 Vögel singen

16 Vögel fliegen

17 ? Pinguine fliegen/singen

Die prototypische Struktur der Klasse der Berufe wird weiter unten noch ausführlich erörtert werden.

Bei den Klassen mit diesem Merkmal sollten im weiteren Verlauf der Kodierung die Nomina oder Subklassen markiert werden, die prototypennahe Konzepte bezeichnen. Bei einer Merkmalskodierung des Wortschatzes würde dies eine Zuordnung von Default-Merkmalen erleichtern.

Das Merkmal ¡unscharfe Ränder¿ legt für eine semantische Klasse von Nomina fest, ob die Klassenzugehörigkeit eindeutig entscheidbar oder ob sie gradierbar ist[5].

Eine typische Klasse mit unscharfen Rändern sind die Flüssigkeiten (vgl. eine Einordnung der Lexeme *Wasser/ Öl /Sahne/ Honig / Lava ...*), weil hier das zentrale Attribut *flüssig* eine unscharfe Bedeutung besitzt. Dies läßt sich an der Anwendbarkeit typischer Kontexte auf unproblematische im Vergleich zu peripheren Mitglieder der Klasse zeigen:

18 Das Wasser fließt den Fluß hinunter

19 Der Honig fließt aus dem Glas

20 Paul verschüttet das Wasser

21 ? Paul verschüttet den Honig

Für die Flüssigkeiten ist das Merkmal ¡Unscharfe Ränder¿ korreliert mit der prototypischen Strukturierung der Klasse — *Wasser* ist eindeutig der Bezeichner für die prototypische Flüssigkeit. Dies ist nicht immer der Fall. Eine semantische Klasse wie Möbel besitzt unscharfe Ränder, ohne daß die Spezifizierung der typischen sprachlichen Kontexte für ein prototypisches Mitglied der Klasse ohne weiteres möglich wäre. Ein Beispiel für Klassen mit scharfen Rändern sind die Schienenfahrzeuge.

Die Klassen mit unscharfen Rändern sind diejenigen, für die bei einer detailliert ausgearbeiteten semantischen Kodierung Grade der Klassenzugehörigkeit kodiert werden sollten. Im Falle einer prototypisch strukturierten Klasse geht dies einher mit einer Kodierung des Prototypikalitätsgrades.

3.2.3 Kategorienmerkmale

Die im folgenden zitierte Aussage aus Dahlgrens 'Naive Semantics' (1988) weist auf die Bedeutung des Typs von Kategorie, auf die eine Lexembedeutung referiert, für die Bedeutungsbeschreibung von Lexemen hin:

> In summary, some words do not have criterial attributes in their meaning representations. Apparently the representation of word meaning varies across the lexicon. Some words may have criterial verbal features in their representations. Obvious examples come form mathematics. A *triangle* can be defined as a 'three-sided figure'. Others have features which correspond to naive theories, or stereotypes of the extensions, but are used with the intention of referring to kinds, that is, to classes of objects with some sort of stable essence [...]. Still others are descriptive. Human viewpoint determines the classes (*weed, junk, witch, game*). Other words are represented in terms of visual and other perceptual features which are not readily translated into verbal predicates.“ (Dahlgren 1988: 17)

[5]Es muß prinzipiell unterschieden werden zwischen gradierbarer Klassenzugehörigkeit und mangelndem Konsens über die Klassenzugehörigkeit innerhalb der Sprechergemeinschaft.

Bei den folgenden Merkmalen geht es um Typen von ontologischen Kategorien, die einem Teil der semantischen Klassen zugrunde liegen (reine Selektionsklassen haben keine derartige ontologische Basis), d. h. es geht um die Kategorie, die der Oberbegriff zu einer taxonomischen semantischen Klasse denotiert. Zu beachten ist wiederum, daß es bei unserer Klassifikation letztlich stets um Eigenschaften von Klassen sprachlicher Ausdrücke geht und der Rückgriff auf die Beschreibung der Denotate nur eine teilweise Motivation für die sprachliche Einteilung liefern soll; es geht also nie primär um eine Einteilung der von den Ausdrücken bezeichneten Objekte, wenn diese auch oft eng mit der Klassifikation der sprachlichen Ausdrücke verbunden ist. Es wird daher im Folgenden versucht werden, die sprachlichen Äquivalente — soweit vorhanden — der teilweise zunächst ontologisch motivierten Kategorieneigenschaften herauszuarbeiten. Dabei wird darauf geachtet, terminologisch zwischen Kategorien, das sind im folgenden Mengen von Denotaten, und Klassen sprachlicher Ausdrücke zu unterscheiden.

3.2.4 Natürliche Art und soziale Art

Als natürliche Arten[6] werden Kategorien angesehen, die von einer Kulturgemeinschaft als in der Natur vorkommende, somit nicht vom Menschen festgelegte Arten von Entitäten angesehen werden. Häufig geht die Festlegung des Umfangs der Unterbegriffe einer natürlichen Art durch die Sprechergemeinschaft parallel zu einer wissenschaftlichen Klassifikation. Ausdrücke für 'natürliche Arten' werden sich innerhalb unserer Grobklassifizierung v. a. im Bereich der Tiere, Pflanzen, Stoffe und Konkreta(Objekte) finden.

Besonders relevant für die Klassifikation der Ausdrücke, die Kategorien natürlicher Arten denotieren, ist der starke Einfluß wissenschaftlicher bzw. quasiwissenschaftlicher Kriterien für die Zuordnung einer Unterkategorie (vgl. Putnam 1975, Carlson 1991); diese wissenschaftliche Klassifikation ist im Endeffekt bei natürlichen Arten wichtiger als der Einfluß perzeptueller Kriterien der Kategorisierung (vgl. Thagard 1991 für die Veränderung des Konzepts von *Wal* aufgrund der wissenschaftlichen Klassifikation). Dies hängt auch damit zusammen, daß Ausdrücke für natürliche Arten ähnlich den Eigennamen sogenannte starre Bezeichner sind, d. h. sie würden selbst dann als Bezeichner für eine Kategorie verwendet werden, wenn sich die Kategorienmitglieder als etwas völlig anderes erweisen, als angenommen (vgl. Cruse 1986: 141). Die Klassen mit dem Merkmal ¡natürliche Art¿ sind also von allen beschriebenen Klassen diejenigen, die am wenigsten durch die Angabe typischer sprachlicher Kontexte abzugrenzen sind.

Allerdings ist das Fehlen eindeutiger sprachlicher Kriterien für diesen Typ von semantischer Klasse nicht sehr problematisch, da für die Lexemklassen der 'natürlichen Arten' stabile und von den Sprechern als objektiv angesehene Klassifikationen existieren, die sich in hyponymisch strukturierten Taxonomien ohne weiteres ausdrükken lassen. Es sind dies auch die Klassen, bei denen die gängigen Thesauri (z. B. Dornseiff 1970 und Roget's Thesaurus — Kirkpatrick 1987) den wissenschaftlichen Klassifizierungen weitgehend folgen — man vergleiche die Klassifizierung der Bezeichnungen für den Bereich der Tier- und Pflanzenwelt in diesen Werken.

Zahlreiche Lexemklassen mit dem Merkmal ¡natürliche Art¿ werden auch als thematische Klassen für das Information Retrieval — neben ihrer Bedeutung für die Klassifizierung allgemeinsprachlicher Texte — im Bereich naturwissenschaftlicher oder populärwissenschaftlicher Beschreibung von besonderer Relevanz sein.

Als ¡soziale Kategorie¿ ("social kind") werden in Anlehnung an Dahlgren (1988: 66f) solche Klassen bezeichnet, die durch die Struktur des Sozialwesens festgelegt sind. Im Fall der Unterklassen von Menschenbezeichnungen liegen in erster Linie über soziale Rollen (z. B.

[6] Vgl. zum Begriff "natürliche Art" ("natural kind") Cruse (1986: 140ff) und Dahlgren (1988).

Denotate der Handwerksberufe), bzw. Typen von sozialen Rollen (Beruf) definierte Kategorien zugrunde. Doch auch für Klassen außerhalb des Bereichs der Menschenbezeichnungen kommen als Kriterien für die Abgrenzung der zugrundeliegenden Kategorie soziale Kriterien vor (Verbrechen). Meist lassen sich die sozialen Kategorien kaum durch einige wenige Kriterien umschreiben; sie ähneln damit den natürlichen Arten.

Auf sprachlicher Seite hat dies zur Konsequenz, daß die entsprechenden Klassen nicht immer durch einige wenige Kontexte abgegrenzt werden können. Meist liegt aber als Überbegriff für die Lexeme in den semantischen Klassen mit der Eigenschaft ¡soziale Kategorie¿ ein semantisch eindeutiges Hyperonym vor.

3.2.5 Perzeptuelle Eigenschaften und Funktion

Für die Abgrenzung fast aller Kategorien spielen perzeptuelle Eigenschaften eine Rolle; bei bestimmten Kategorien sind solche Eigenschaften jedoch das zentrale Kriterium für die Einteilung. Browns (1990: 18) nennt diese Eigenschaft einer Kategorie "Gestalt motivation". Sprachlich schlagen sich solche Kategorien in Klassen von Lexemen nieder, auf die eines oder ein relativ kleines Bündel von Attributen zutrifft. Eine typische Klasse der ersten Art sind beispielsweise die Flüssigkeiten. Das Hyperonym dieser Klasse, *Flüssigkeit*, beinhaltet bereits als Derivationsbasis das Adjektiv *flüssig*, dessen Denotat das zentrale Kriterium für die Zuteilung zur entsprechenden Kategorie ist.

Typische Kontexte für diesen Typ von Klasse lassen sich meist relativ leicht finden. Häufig sind dies solche verbalen Kontexte, die wesentliche Eigenschaften voraussetzen, welche konstituierend für die Klasse sind. Für Flüssigkeiten sind solche Kontexte etwa *fließen*, *verschütten*, *ausgießen*, *verdampfen*.

Ausdrücke für Artefakte — die in hohem Maße aufgrund funktionaler Kriterien weiter zu klassifizieren sind — wurden im CISLEX zunächst in die Klasse Konkreta (Objekte) eingeordnet (z. B. *Mikroskop*), zum Teil auch unter die Stoffbezeichnungen (*Styropor*). Bei Kriterien für ihre Klassifikation werden vor allem verbale Kontexte eine Rolle spielen: Ausdrücke für Artefakte werden meist danach klassifiziert, wie die bezeichneten Objekte hergestellt werden oder noch häufiger darüber, was mit ihnen gemacht werden kann, also zu welchem Zweck sie dienen. Solche funktionalen Kriterien schlagen sich auf sprachlicher Seite vor allem in der Kombinatorik mit entsprechenden Verben nieder. So wird etwa die Ausdrucksklasse optische Geräte typischerweise in einem Kontext auftreten wie *A betrachtet B durch X*, wobei X das optische Gerät bezeichnet. Ein Test, ob funktionale Kriterien für die Aufstellung einer Klasse in Frage kommen, liegt in Kontexten wie *A benutzt B als X* vor, wobei X ein Lexem der entsprechenden Klasse ist, B zu einer beliebigen semantischen Klasse gehören kann. Die Einsetzung in einen solchen Kontext ist nur für Lexeme möglich, in deren Bedeutung der funktionale Aspekt eine große Rolle spielt, nicht aber bei Lexemen, bei denen er in der Bedeutung eine unwichtige oder keine Rolle spielt:

22 Maria benutzt den Dosenöffner als Messer

23 Hans verwendet sein Schwert als Besteck

24 Paul verwendet seinen Wacholderbusch als Besen

25 ? Paul verwendet seinen Besen als Busch

Für funktional bedingte Klassen können zur Abgrenzung sprachliche Kontexte angegeben werden, die sich auf die Funktion beziehen. So dienen etwa die Denotate von Verkehrsmitteln der Fortbewegung:

26 Hans reist mit der Bahn/dem Auto/dem Flugzeug nach Amsterdam

Außer im Bereich der Artefakte treten funktional bedingte Klassifizierungen auch in anderen Klassen auf. So sind etwa die Klassen der Nutztiere und Haustiere, die quer zur biologischen Klassifikation der Tierwelt liegen, kaum anders als über ihre Funktion für den Menschen zu definieren.

3.2.6 Defaultzuweisungen und Abhängigkeiten unter Klassenmerkmalen

Als Defaultzuweisungen für die genannten Merkmale von semantischen Klassen wurden festgelegt:

- Alle drei Kriterien für Klassenaufstellung und Verwendbarkeit sind positiv spezifiziert. Die explizite Nennung eines Merkmals hebt die positive Spezifikation für die anderen Merkmale auf.
- Der Typ von Kategorie ist per Default unspezifiziert (keine Zuweisung).
- Die Merkmale für Klassenstruktur und Abgrenzung sind per Default negativ spezifiziert.

Eine weitere Festlegung zur Erleichterung der Kodierung sind die Abhängigkeiten unter den verschiedenen Klassenmerkmalen. Aufgrund des bisher gesagten ist leicht einzusehen, daß für die Kombinatorik der Merkmale zumindest die folgenden Einschränkungen gelten müssen:

- Kombinationen zwischen den Merkmalen, die sich auf Aufstellungs- und Anwendungskriterien beziehen, sind problemlos möglich.
- Alle Kombinationen zwischen den Merkmalen, die sich auf die Kategorientypen beziehen, sind ausgeschlossen. Eine semantische Klasse kann nur einem der Typen angehören.
- Kategorieneigenschaften sind nur vorhanden, wenn einer Klasse eine Kategorie zugrundeliegt. Dies ist bei reinen Selektionsklassen nie der Fall. Aus der negativen Spezifikation der Merkmals ¡taxonomische Klasse¿ ergibt sich damit die Irrelevanz der Kategorieneigenschaften.

Als weitere Abhängigkeiten ergeben sich: Sammelklassen können natürlicherweise keine prototypische Struktur haben, natürliche Arten keine unscharfen Ränder. Sonst sind zwischen den drei Gruppen von Merkmalen alle Kombinationen möglich.

3.3 Grobklassifikation

3.3.1 Ziele der Grobklassifikation

Die Grobklassifizierung des Wortschatzes verfolgt in erster Linie zwei Absichten: Einerseits gilt es, den sehr weiten Selektionsrestriktionen mancher Operatoren Rechnung zu tragen, andererseits soll diese grobe Einteilung eine erste Grundlage für eine feinere Unterteilung

Merkmalstyp	**Bezeichner**	**Merkmal**	**Implikationen**
Aufstellungskriterien und	+/-ta	¡taxonomische Klasse¿	-ta -¿ keine Kategorienmerkmale
Anwendungsbereiche	+/-se	¡Selektionsklasse¿	-
	+/-th	¡thematische Klasse¿	-
Aufstellung kodierungsökonomisch bedingt	+/-sk	¡Sammelklasse¿	-fk, -na, -sa, -de, -pz
Klassenstruktur und	+/-pz	¡prototypenzentriert¿	-sk
-abgrenzung	+/-ur	¡unscharfe Ränder¿	-na
	+/-na	¡natürliche Art¿	ta,-fk, -sk, -de, -sa, -ur
Kategorientypen	+/-sa	¡soziale ART¿	ta,-na, -sk, -de, -fk
	+/-de	¡Perzeptuelle Denotatseigenschaften¿	ta,-na, -sa, -sk, -fk
	+/-fu	¡funktional¿	ta,-na, -sa, -de, -sk

Tabelle 3.1: Merkmale semantischer Klassen

liefern[7]. Neben der Verankerung der Grobklassen in der hyponymischen Struktur des Wortschatzes ist die Relevanz der meisten dieser Klassen für die Beschreibung von verbalen und adjektivischen Argumentstrukturen unmittelbar einsichtig; einige der im folgenden genannten semantischen Distinktionen spielen als selektionsrelevante Merkmale eine so wichtige Rolle, daß selbst dem syntaktischen System zugerechnete Regularitäten, wie etwa die Wahl des richtigen Typs von Relativpronomen für ein Antezedens oder die Auswahl des Fragepronomens (*wer* vs. *was* durch menschlich/nicht-menschlich) von der Zugehörigkeit bzw. Nichtzugehörigkeit eines Ausdrucks zu bestimmten von diesen semantischen Klassen berührt werden.

Es ist sicher falsch, zu erwarten, daß die im folgenden genannten Grobklassen den nominalen Wortschatz des Deutschen so strukturieren, daß diese Klassifikation unter allen denkbaren Aspekten die einzig relevante für eine weitere Unterteilung ist. Bei der Aufstellung der feineren Klassen wird sich vielmehr zeigen, daß einige von ihnen quer zur zunächst vorgenommenen Einteilung liegen. Dies gilt etwa für die Klasse der Nahrungsmittel bezeichnenden Nomina, die sowohl Stoffbezeichnungen (z. B. *Honig*) als auch Individualnomina (z. B. *Orange*) enthält. Ziel für die Aufteilung in die Grobklassen war es aber nicht, sämtliche mögliche Generalisierungen zu erfassen, sondern die Klassen zu ermitteln, die für die Beschreibung des distributionellen Verhaltens und für die weitere Kodierung am wichtigsten sind.

Ebensowenig kann erwartet werden, daß für die Darstellung der Grobklassifikation alle in der theoretischen Linguistik geführten Diskussionen über die Distinktionen der verschiedenen Typen von Nomina (wie etwa in Carlson 1991 oder Krifka 1991 referiert) komplett aufgearbeitet werden können. Ohne eine zumindest teilweise unhinterfragte Übernahme von in der Lexikographie bewährten Unterscheidungen ist die semantische Beschreibung eines so großen Wortschatzausschnitts praktisch nicht durchführbar.

Im vorigen Kapitel wurde die semantische Grobklassifizierung zweier umfangreicher elek-

[7]Dieser zweite Motivationsgrund sollte vor allem für die Bewertung der Einteilung der weniger leicht intuitiv nachvollziehbaren Klassen vom Leser im Auge behalten werden. Die linguistische Motivation ist für einige der Grobklassen recht schwach; sie sind aber auch teilweise nur als Durchgangsstationen zu einer brauchbaren Feingliederung zu betrachten und damit ohnehin bei fortgeschrittener Kodierung obsolet.

tronischer Lexika vorgestellt: die "unique beginners" der Einteilung in der WordNet Datenbasis und die acht umfassendsten Objektklassen des Gross-Lexikons. Für die Grobklassifizierung im CISLEX, wie ich sie im folgenden darstelle, wurde zunächst die Einteilung durch Gross (1992, 1994) für den französischen Wortschatz zugrunde gelegt. Ich gehe nur dort, wo das CISLEX von dieser Klassifizierung abweicht — denn genau an diesen Punkten wird die Unterteilung letztlich erklärungsbedürftig, da hier kein einfacher Konsens zu erzielen ist — auf die Einteilung in WordNet sowie teilweise auf die Darstellungen in Dahlgren (1988) und auf die Klassifizierung in Aarts/Calbert (1979) ein.

Während alle vier genannten Arbeiten die Wichtigkeit des Kriteriums Kontext für die Einteilung des Nominalwortschatzes betonen, wird doch in allen vier Klassifikationsbeschreibungen davor zurückgescheut, für alle oder auch nur eine Mehrzahl der einzelnen Klassen der Einteilung die jeweiligen Kriterien für Aufstellung und Zuordnung zu konkretisieren[8]. Nur für die ohnehin unproblematischen Fälle werden einige Anhaltspunkte zur sprachlichen Umgebung der klassifizierten Ausdrücke gegeben. In der semantischen Beschreibung innerhalb der Linguistik und in der Lexikographie hat ein solches Vorgehen Tradition. Insbesondere Unterscheidungen wie ¡+/- belebt¿ ¡+/- abstrakt¿ werden häufig unhinterfragt verwendet, obwohl die Abgrenzung der Nomina mit diesen Merkmalen oft mehr als problematisch ist. Mir erscheint diese Vermeidung der Angabe teilweise recht vager Kriterien zwar verständlich, da es sich meist um recht unscharfe und damit leicht angreifbare Eigenschaftsbeschreibungen handelt. Andererseits aber wird die Darstellung einer solchen Wortschatzklassifizierung erst diskussionswürdig, wenn auch die linguistische Motivation für die Abgrenzung einzelner Klassen explizit gemacht wird. Ich werde von daher im Folgenden versuchen, jeweils einige Kriterien für die Aufstellung aller für das CISLEX angenommenen Grobklassen, und später auch für die feineren Klassen zu nennen.

Die folgende Auflistung der Grobklassifikation spiegelt in ihrer Reihenfolge auch die Ausprägung von Schwierigkeiten bei der Einteilung des Wortschatzes wieder. Waren die zuerst genannten Klassen relativ einfach aufzustellen und abzugrenzen, fiel dies bei den weiter unten genannten teilweise außerordentlich schwer.

Bei Grobklassen, auf deren Gliederung später nicht im Detail eingegangen wird, finden sich z. T. bereits Hinweise auf später kodierte Unterklassen.

3.3.2 Beschreibung der Grobklassifikation

Lebewesen

- Menschen
- Tiere
- Pflanzen

Fast völlig unproblematisch lassen sich diese vier Klassen festlegen und abgrenzen. Diese Distinktionen hat das CISLEX mit allen oben genannten semantischen Klassifizierungen und den meisten traditionellen Lexika gemein. Bezüglich dieser Klassen ist bei einer Betrachtung der Klassenbezeichner, die den Hyperonymen der Lexeme der Klasse entsprechen, sofort eindeutig klar, welche Nomina wo zuzuordnen sind. Die genannten Grobklassen lassen sich damit ohne weiteres als taxonomische Klassen beschreiben. Zusätzlich lassen sich die Lexeme

[8]Ein besonders krasses Beispiel für eine fehlende Begründung ist die semantische Einteilung der Nomina in Dixon (1991: 76f). Hier wird postuliert: "there are five major types associated with the grammatical class noun in English" (ebd. S. 76), aber keinerlei Motivation für die vorgenommene Klassifizierung genannt. Aber auch die anderen genannten Arbeiten versäumen es, die für die referierten Einteilungen relevanten Kontexte zu nennen, obgleich deren Bedeutung immer wieder betont wird.

der drei genannten Klassen auch mittels einer Reihe typischer Operatoren von den anderen Lexemkategorien und auch voneinander abgrenzen. Zahlreiche Operatoren selegieren Menschen, Teilklassen der Tiere oder der Pflanzen. Zu diesen Operatoren gehören beispielsweise die Verben *sterben* für alle Lebewesen, *lachen*, *weinen* etc. für Menschen, *wachsen*, *verwelken* und *verblühen* für größere Untergruppen der Pflanzen. Somit handelt es sich bei diesen primär als taxonomische Klassen definierten Lexemmengen auch um Selektionsklassen.

Zusammengenommen konstituieren diese drei Klassen die Klasse der Lebewesen. Aufgrund der relativ leichten Abgrenzbarkeit dieser Lexeme von den anderen Nomina des kodierten Wortschatzes ist auch das Komplement der Gesamtklasse der Lebewesen eindeutig definiert. Etwas problematisch werden die Distinktionen zwischen den einzelnen Unterklassen nur im Bereich der niederen Pflanzen und Tiere (v. a. Einzeller) und der Bezeichnungen für nur in fiktionalen Kontexten auftretende Fabelwesen (z. B. *Troll*), die jedoch quantitativ bei der Kodierung keine große Rolle spielen.

Die Grobklasse der Menschenbezeichnungen allein macht mit ca. 7000 Lemmata die größte semantische Grobklasse unter den einfachen Nomina aus. Ihre weitere Unterteilung wird noch im Detail dargestellt werden. Die Klasse Tiere kommt auf ca. 800, die Klasse Pflanzen auf ca. 500 Einträge.

Auf die Lebewesen folgen in der Auflistung der Grobklassen von Gross (1994: 20) *inanimé abstrait* und *inanimé concret* ('unbelebt-abstrakt', 'unbelebt-konkret'). Da die Dichotomie Konkreta vs. Abstrakta trotz ihrer weiten Verbreitung in der Literatur meines Wissens nirgendwo befriedigend definiert ist, wird sie für die Klassifizierung in dieser einfachen Form nicht übernommen. Es werden dagegen zwei Klassen gebildet, die einerseits die Bezeichner von Konkreta im engeren Sinne (Objekte) und andererseits die Stoffbezeichnungen umfassen — das Komplement dieser Klassen sind jedoch keineswegs Abstrakta.

Die Kategorie Abstrakta (bei Gross: *inanimé abstrait*) erscheint mir als eine Gesamtklasse nicht sinnvoll, da einerseits der Begriff des Abstraktums sehr schwammig und ungenau ist, und andererseits die Nomina, die üblicherweise hierunter gefaßt werden, weder eine taxonomische Klasse konstituieren noch im sprachlichen Verhalten irgendwelche weitergehenden Gemeinsamkeiten aufweisen[9]. Die traditionell unter Abstrakta abgehandelten Nomina sind in verschiedene der weiter unten genannten Klassen einzuordnen.

Konkreta (Objekte)

Die erste der Klassen der Konkreta im engeren Sinne umfaßt Begriffe, die für festumrissene Objekte mit räumlicher Ausdehnung stehen.

Typische Operatoren sind hier diejenigen, die Maße (z. B. Verben wie *wiegen*, *umfassen*) und Aussehen (Adjektive für Farben und Formen) beschreiben. Die bisher genannten Klassen der Lebewesen sind in diesem Sinne ebenfalls als Konkreta (Objekte) zu beschreiben. Aufgrund ihrer zahlreichen zusätzlichen Eigenschaften und der quantitativen Relevanz für den kodierten Wortschatz erscheint es allerdings sinnvoll, die Lebewesen von vornherein als separate Kategorie zu kodieren.

Für die Klassifizierung eines großen Teil der Konkreta sind neben den äußeren Eigenschaften der Denotate auch funktionelle Kriterien relevant. Dies gilt insbesondere für die

[9]Dahlgren (1988: 55) hat die abstrakt-konkret-Unterscheidung als einen der obersten Knoten in ihrer Ontologie, ohne jedoch zu explizieren, was sie unter Abstrakta versteht. Die Ablehnung der Zusammenfassung verschiedener Kategorien zu "Abstrakta" scheint mir auch sprachlich leicht begründbar: während für Konkreta (im engeren Sinne) und Stoffbezeichnungen gemeinsame Hyperonyme (*Ding*, *Stoff*, *Zeug* ...) existieren, existieren für eine Gruppe wie die angeblichen Abstrakta keine übergeordneten sprachlichen Ausdrücke. Man betrachte auch den Artikel zum Stichwort ABSTRAKTA in Bußmanns 'Lexikon der Sprachwissenschaft' (1990), der als einziges greifbares Kriterium für den Begriff des Abstraktums die Auflistung verschiedener Unterklassen (Eigenschaften, Beziehungen u.a.) anbietet, und ansonsten auf den Eintrag KONKRETA verweist, der allerdings hier auch nicht weiterführt, da er wiederum nur einige Subgruppen auflistet, ohne einen brauchbaren Ansatz zur Distinktion beider Begriffe zu liefern.

Subklasse der Artefakte. Erklärlicherweise ist damit für diese Unterklasse eine Einteilung aufgrund verbaler Kontexte relativ einfach vorzunehmen.

Gemeinsame Hyperonyme für die Lexeme der Klasse Konkreta(Objekte) sind *Ding, Objekt, Sache*. Somit läßt sie sich — neben ihrer eindeutigen Relevanz für Selektionsregularitäten — auch also taxonomische Klasse auffassen. Die genannten hyperonymischen Ausdrücke können allerdings jeweils auch in übertragener — und damit abstrakterer — Bedeutung verwendet werden.

Stoffe

Von allen bisher aufgezählten Klassen sind die Stoffbezeichnungen abzugrenzen. Im Gegensatz zu den Individualnomina bezeichnen diese Ausdrücke keine Entitäten mit bestimmter räumlicher Ausdehnung, sondern Substanzen. Für die Abgrenzung lassen sich ohne weiteres sprachliche Kriterien angeben: Die Nomina, die Stoffe bezeichnen, können im Gegensatz zu Ausdrücken für Individualnomina in bestimmten generischen Kontexten auch im Singular ohne Artikel verwendet werden, während dies bei Individualnomina eine hochmarkierte bzw. textsortenbeschränkte Konstruktion ist, bei der ein Artikel stets ergänzt werden kann:

27 Wasser verdampft bei 100 Grad

28 ?Löwe beißt bei Annäherung

29 Löwen beißen bei Annäherung

30 Der Löwe lebt in den Wüsten Afrikas

Stoffbezeichnungen treten zudem mit bestimmten Maßbezeichnungen zusammen und als Komplemente bestimmter Präpositionen (*aus* X, *voll* X) auf. Von den Kollektiva, die üblicherweise mit ihnen zu den Massennomina zusammengefaßt werden, unterscheiden sich Stoffbezeichnungen durch einige der bisher genannten Eigenschaften und dadurch, daß sie nicht-zählbare, homogene Entitäten bezeichnen (Kollektiva werden im CISLEX durch einen semantischen Relator bezeichnet, s.u.)[10].

Das gemeinsame Hyperonym für diese Klasse wird bereits durch den Klassenbezeichner Stoff festgelegt; in Frage kommen zusätzlich, allerdings mit eingeschränkterem Gebrauch, *Substanz* und *Material*.

Weiter unterteilt werden können die Stoffbezeichnungen aufgrund der Konsistenz der Denotate in Gase, Flüssigkeiten und Feststoffe. Diese Distinktionen schlagen sich auch deutlich in den typischen Kontexten der Nomina wieder (etwa *ausschütten* oder *fließen* für Flüssigkeiten). Eine quantitativ wichtige funktionelle Untergruppe sind die Kleiderstoffe, die mit ca. 150 Lemmata eine relativ große homogene Klasse von Lexemen ausmachen.

Ich schließe damit die Beschreibung des Bereich der eindeutig abgrenzbaren Klassen der Konkreta (im engeren Sinne) ab. Fast alle der folgenden Klassen beinhalten Nomina, deren Bedeutung mehr oder minder abstrakte Aspekte umfaßt. Mit den bisher aufgezählten Einteilungen sind bis auf *événement*, *locatif, temps* und *inanimé abstrait* alle bei Gross auftretenden Grobklassen übernommen. Die Klassifikation des CISLEX weicht im folgenden von derjenigen im Lexikon von Gross ab, indem von vornherein eine etwas feinere Unterteilung vorgenommen wird.

Ereignisse

Zustände

Eigenschaften

[10] Vgl. zur Abgrenzung von Massennomina von Individualnomina und der Abgrenzung von Stoffbezeichnungen und Kollektiva auch Krifka (1991) und die in Oesterle (1994: 5ff) aufgelisteten Kriterien.

Ereignisse, Zustände und Eigenschaften werden typischerweise nicht durch Nomina ausgedrückt, sondern durch Verben und Adjektive. Dennoch sind diese Distinktionen auch für den Nominalwortschatz relevant, denn von fast allen Verben und Adjektiven existieren Nominalisierungen, teils in Form von Konversionen, teils in Form von Suffix- und Präfixderivationen. Im kodierten Wortschatz spielten quantitativ v. a. die suffigierten Nomina eine wichtige Rolle; es fanden sich unter den 37 000 Lemmata allein ca. 3500 Ableitungen auf *-ung*, 1100 Ableitungen auf *-ion* und 2400 auf *-heit* und *-keit*.

Der Klasse der Ereignisse (*événements*) bei Gross (1994: 20) stehen in WordNet die Klassen "act/action/activity", "event/happening" und "process" gegenüber. Ihnen ist gemein, daß die meisten der enthaltenen Nomina von Verben abgeleitet sind. Die Abgrenzung und die Kriterien für die Unterscheidung untereinander werden in den Artikeln zu WordNet (Miller u. a. 1993) jedoch nicht näher beschrieben. Die Kodierung im CISLEX faßt diese Klassen daher — wie auch die Klassifikation von Gross — zunächst unter einer Grobklasse Ereignisse zusammen.

Unter Ereignissen sind dabei Vorkommnisse zu verstehen, die Anfang und Ende aufweisen (können), also typischerweise mit den entsprechenden Verben (*beginnen, dauern*, *enden*, ...) kombinierbar sind. Daneben lassen sich einige andere Verben zur Eingrenzung verwenden (*stattfinden*, *sich ereignen*). Die Klasse umfaßt Lexeme, die sowohl Aktionen — also Ereignisse, die einen willenhaften Verursacher einschließen — als auch Ereignisse im engeren Sinne und Prozesse bezeichnen. Eine Unterteilung dieser Grobklasse in Unterklassen setzt Instrumente zur Beschreibung der Semantik und Argumentstruktur von Verben voraus.

Zwei Klassen, die ebenfalls in erster Linie Nomina umfassen, die Derivationen von Lexemen anderer Wortarten sind, finden sich in WordNet als "state/condition" und "attribute/property" (Miller u a. 1993: 16). Diese Klassen werden für die Einteilung am CISLEX als Zustände und Eigenschaften übernommen. Zustände werden sprachlich in der Regel durch Adjektive oder durch Partizipien inchoativer oder durativer Verben ausgedrückt. Die meisten der Nomina, die Zustände denotieren, sind also deadjektivisch, einige auch deverbal. Gegenüber den Ereignissen fehlt den Zuständen die Eigenschaft, eine temporale Struktur zu besitzen. Bezeichnungen für Zustände beinhalten nicht deren Ende bzw. den Anfang. Die entsprechenden Lexeme können von daher nur teilweise in Kontexten wie *X beginnt/endet* u. ä. auftreten; ausgeschlossen sind sie als Subjekt verbaler Ausdrücke wie *stattfinden*.

Abgrenzen lassen sich Zustände schließlich auf der anderen Seite von den Eigenschaften. Die Lexeme dieser Klasse sind ebenso häufig deadjektivisch, bezeichnen jedoch unveränderbare Attribute der Entitäten über die prädiziert wird. Sie sind weniger häufig von solchen Adjektiven abgeleitet, die Partizipien von Verben formgleich sind. Die Abgrenzung von Zuständen und Eigenschaften ist allerdings etwas problematisch[11]. Während eine Reihe von Nomina klar Eigenschaften (z. B. *Rechtschaffenheit*) und eine Reihe eindeutig Zustände (z. B. *Trunkenheit*) bezeichnen, sind zahlreiche andere doppelt zu klassifizieren (z. B. *Schläfrigkeit*).[12]. Die vorzunehmenden weiteren Aufteilungen der genannten Grobklassen in hyponymische Unterklassen entsprechen weitgehend den Unterscheidungen, die für die Basislexeme der derivierten Nomina vorzunehmen sind. Zusätzlich zu diesen Übereinstimmungen hat der größte Teil der Ereignisse, Zustände und Eigenschaften bezeichnenden Nomina eine Argumentstruktur, die sich — mehr oder weniger systematisch — von der Argumentstruktur der zugrundeliegenden Lexeme herleitet. Letztlich hängt die detaillierte Unterteilung dieser Klassen von Nomina also eng mit der vorausgehenden semantischen

[11] Aarts/Calbert (1979: 23) fassen STATE und PROPERTY ebenso wie Dixon (1991: S. 76) unter dem Merkmal +STATE zusammen als "concepts referring to a mental or physical condition".

[12] Abgrenzungskriterien sollten zunächst für die zugrundeliegenden Adjektive entwickelt werden.

Beschreibung der zugrundeliegenden Adjektive und Verben zusammen[13].

Die Angabe von Kriterien zur Festlegung größerer Klassen von Lexemen und zu deren Abgrenzung wird für den restlichen Nominalwortschatz zunehmend schwieriger. Dies wirkt sich auch auf die Kodierung aus und bringt mit sich, daß die meisten der im folgenden genannten Klassen wohl eher als Sammlungen von Unterklassen zu verstehen sind, die nur noch wenige gemeinsame Bedeutungsaspekte besitzen und von den Sprechern im Gegensatz zu den bisher genannten Kategorien nicht mehr als einheitliche Lexemgruppen gesehen werden, was sich auch im Fehlen von Hyperonymen für größere Lexemgruppen und in einem uneinheitlichen Verhalten in sprachlichen Kontexten ausdrückt. Der gesamte Wortschatz der einfachen Nomina[14] weist somit keine einheitliche Struktur auf. Während für die bisher genannten Klassen gilt, daß ihre Aufstellung und Abgrenzung intersubjektiv nachvollziehbar bleibt, durch zahlreiche sprachliche Regularitäten gedeckt wird und zudem zu einem großen Teil mit wissenschaftlichen/ sozialen Klassifikationen übereinstimmt, umfaßt der in die folgenden Klassen fallende Wortschatzausschnitt den weniger systematisierbaren Teil des Gesamtwortschatzes. Die wesentlichen Strukturen sind hier Mikrostrukturen, d. h. umfassendere sprachliche Gemeinsamkeiten gelten oft nur für einige wenige Lexeme; Hyponymiestrukturen lassen sich hier — wenn überhaupt — oft nur über zwei Ebenen hinweg ermitteln, während bei einigen der oben genannten Klassen (z. B. den Lebewesen) eine konsistente hyponymische Struktur über mehr als drei Ebenen hin aufstellbar ist (man betrachte etwa die Reihe (*Ara - Papagei - Vogel - Tier - Lebewesen*). Der Nominalwortschatz kann also als eine Struktur gesehen werden, aus der sich, mit umfangreichen Klassen von Lexemen mit zahlreichen gemeinsamen semantischen Eigenschaften beginnend, zunehmend kleinere Klassen herausschälen lassen, bis dann am Ende ein Bodensatz von idiosynkratischen Nomina übrigbleibt, die für eine Einteilung aufgrund von Kohyponymie oder geteilten Selektionskontexten nur mehr wenige und schwache Anhaltspunkte liefern, und bei denen die Einteilung in solche Klassen auch kaum mehr die Festlegung thematischer Zugehörigkeit erlaubt.

Temporalia

Leicht abgrenzbar von anderen Nomina sind noch die Ausdrücke, die Zeitintervalle und -punkte beschreiben.

Sprachliche Kontexte sind hier Verben wie *dauern* (für Zeitintervalle) temporal gebrauchte Präpositionen (*bis am* X); die typische Funktion von Nominalphrasen und Präpositionalgruppen, die Nomina mit temporaler Bedeutung enthalten, ist die eines temporalen Adverbials. Dies wird deutlich an den Ausdrücken *Minute*, *Moment*, *Sonntag* und *Morgen* in den folgenden Beispielsätzen:

31 Die Fahrt dauerte eine Minute/einen Moment

32 Otto bleibt bis am Sonntag

33 Otto wäscht sich nie am Morgen

Die Klasse der Temporalia ist relativ klein. Sie umfaßt im kodierten Wortschatzbereich nur ca. 200 der 37 000 Einträge. Sie wurde als eine Sammelklasse ausgezeichnet, da die wenigen genannten Regularitäten keine Klassifizierung als Selektionsklasse zulassen und bisher keine detailliertere Untersuchung dieser Ausdrücke stattgefunden hat.

Lokativa

[13]Ein ausführliche semantische Beschreibung dieser Klassen des Nominalwortschatzes würde von daher den Rahmen dieser Arbeit sprengen.

[14]Diese Aussage läßt sich wohl ohne weiteres für den gesamten Nominalwortschatz verallgemeinern.

Eine recht uneinheitliche Klasse sind die in Gross' Klassifikation genannten Lokativa, die auch in WordNet unter der Bezeichnung "location/place" auftreten. In keiner der beiden Arbeiten wird diese Klasse näher charakterisiert. Problematisch für die Einteilung von Ortsbezeichnungen ist zunächst, daß praktisch jeder Begriff, welcher der Klasse Konkreta (Objekte) zugeordnet wurde, auch relativ unproblematisch lokativ verwendet werden kann, wie folgende Beispiele zeigen, in denen das auftretende Komplement der Präpositionalphrase jeweils als Ort gesehen wird:

34 Die Maus sitzt im Computer

35 Die Laus sitzt auf dem Stein

Die Funktion als Komplement lokativer Präpositionen ist von daher als Kriterium für die Zuordnung eines Nomens zu den Lokativen nicht zu gebrauchen.

Ich sehe als Lokativa einerseits Objekte an, die in erster Linie die Funktion haben, als Ort für den Aufenthalt von Menschen oder anderen Objekte zu dienen, wie etwa Gebäude (die als ein Sondertyp von Behältern — prototypischerweise für Menschen — anzusehen sind), Behälter für Objekte und Stoffe, Verkehrswege u. ä. Lokativa sind somit zunächst eine — relativ große — Unterklasse der Konkreta. Sie sind damit in der ersten Klassifikation bereits zugeordnet und können nach einer erfolgten Feinklassifikation der Konkreta klassenweise zusätzlich den Lokativen zugeordnet werden.

Andererseits sind Lokativa Bezeichner von Teilen von Objekten (*Gipfel*, *Ende*) oder Räumen (*Decke*, *Boden*), die primär auf deren räumliche Struktur Bezug nehmen, und die unterschiedliche Abstraktionsgrade aufweisen. Sie treten typischerweise mit ihren Bezugsbegriffen in NPn auf, wie hier an den Beispielen *Ende* und *Decke* demonstriert wird:

36 Das Ende der Schnur

37 Die Decke des Zimmers/Zimmerdecke

Nomina dieser Art wurden primär in die Grobklasse Lokativa eingeordnet.

Eine weitere, allerdings relativ kleine Gruppe von Lexemen, die ebenfalls bereits bei der Grobklassifizierung den Lokativa zugeordnet wurden, sind die Richtungsangaben — hierzu gehören primär die Himmelsrichtungen. Hinzu kamen einige Lexeme wie *Ferne*, *Nähe*, die sich relativ idiosynkratisch verhalten, die aber eindeutig primär auf eine räumliche Situation abheben.

Die Klasse der Lokativa ist eine Gruppe von Lexemen, die aufgrund der mangelnden Kohärenz als Gesamtklasse nur ein Übergangsstadium zur Feinklassifizierung der zugeordneten Lemmata darstellen kann. Sie ist weder eine wirklich relevante Selektionsklasse noch eine taxonomische Einheit.

Formen

Dasselbe gilt für die folgende Klasse, die von den restlichen Nomina nur aufgrund geteilter Bedeutungsaspekte abgrenzbar ist. Es handelt sich um die Formen (in WordNet existiert die Klasse "shape/form"), die allerdings wiederum zwei ganz unterschiedliche Subklassen umfassen: auf der einen Seite die ein- zwei- und dreidimensionalen Abstraktionen von Gegenständen aufgrund der Form, die auch Objekte verschiedener Art in erster Linie aufgrund von deren äußerer Form bezeichnen können (z.B. *Quader*), und auf der anderen Seite Teil- oder Oberflächenstrukturen von Gegenständen (*Rundung ...*) bezeichnende Ausdrücke. Diese Klasse ist vor allem durch den gemeinsamen Beschreibungsterm *Form* motiviert; ihre Feingliederung wird in einem speziellen Abschnitt dargestellt. Diese Klasse ist als Ganzes ebenso wie die vorher genannten Lokativa als eine kodiertechnisch bedingte Übergangsklasse

zu betrachten, und beansprucht nicht für sich, eine durch die hyponymische Struktur des Wortschatzes oder durch sprachliche Regularitäten gedeckte Einheit darzustellen.

Diversa

Die letzte Gruppe von Einträgen bilden die Lexeme, die zunächst nicht in eine der genannten semantischen Grobklassen eingeordnet werden konnten.

Sie gehören zu teilweise klar definierbaren kleineren Klassen (z. B. Krankheiten), lassen sich aber nicht in eine der bisher genannten Grobklassen einordnen, d. h. es gibt für diese Gruppen von Wörtern keine sinnvollen Oberklassen mehr, die — vergleichbar den bisher genannten — eine große Zahl von Lexemen enthalten, welche ein gemeinsames Hyperonym besitzen oder die gemeinsamen Selektionsrestriktionen gehorchen. Für die Gliederung dieses Wortschatzbereichs scheinen nur wesentlich kleinere Klassen relevant zu sein Dies gilt auch für die thematische Stukturierung.

Einige der Subklassen in dieser Gruppe seien genannt, um die Unmöglichkeit ihrer Zuordnung zu einer größeren Klasse von Lexemen zu demonstrieren:

Krankheiten: z.B. *Cholera*, *Pest*.

Neben der unproblematischen Zuordnung zum Überbegriff *Krankheit* sind die Bezeichner für Krankheiten durch eine große Reihe von typischen Operatoren eingrenzbar; die Gruppe ist damit eine der kontextuell einheitlichsten Klassen des Wortschatzes. Viele der typischen Operatoren (*sich mit X anstecken, X bekommen, an X leiden*) werden von keiner anderen Lexemgruppe geteilt. Eine Einordnung in eine übergeordnete Klasse ist aufgrund des speziellen kontextuellen Verhaltens der Nomina dieser Klasse nicht sinnvoll.

Maße: z. B. *Gramm*, *Kilogramm*, *Liter*. Diese sind weiter unterteilbar in verschiedene Typen von Maßen (Hohlmasse, Gewichtsmasse etc.), die jeweils einige gemeinsame sprachliche Eigenschaften besitzen[15].

Abschlüsse: z. B. *Abitur*, *Matura*, *Magister*. Für diese Lexeme bestehen einige typische Kontexte, u. a. die im folgenden genannten:

38 Karl hat seinen Magister 1956 gemacht

39 Maria hat gestern die Abitursprüfung abgelegt

Weder die Zuordnung zur Grobklasse der Ereignisse noch zu den Zuständen ist m. E. für diese Klasse angemessen.

Gezeiten: *Ebbe*, *Flut*. Dies ist eine sehr kleine Klasse von Lexemen, die sich aber kaum einer übergeordneten Klasse zuordnen läßt. Als Hyperonyme lassen sich *Gezeiten* und *Tide* angeben.

Eine aufgrund völlig anderer Kriterien abzugrenzende Gruppe von Lexemen bilden die fachsprachlichen Lexeme. Sie konstituieren keine semantische Klasse im bisher definierten Sinn. Auf ihre Behandlung wird später noch detaillierter eingegangen.

3.3.3 Tabelle der semantischen Klassen der Grobklassifikation

Zusammenfassend ergeben sich für die Grobklassifikation des Wortschatzes die inTabelle reftabelle3 aufgeführten Klassen[16].

[15] Vgl. zu den Maßbezeichnungen Oesterle (1994) und die Ausführungen zur Beschreibung der Klasse der Formen in der Darstellung der Feinklassifikation.

[16] Für A in der Spalte 3 ist eine beliebige passende Nominalgruppe zu ergänzen. X steht für eine Nominalgruppe mit einem Nomen aus der genannten semantischen Klasse.

Klassenbezeichnung	**Hyperonym/ Beschreibung**	**Typische Kontexte/ linguistische Eigensch.**
Lebewesen	Lebewesen	*X lebt/stirbt/vergeht*
Menschen	Alle Bezeichnungen für Menschen	*X lacht/weint*
Tiere	Tiere	*A tötet X*
Pflanzen	Pflanzen	*X wächst, wird gepflanzt*
Konkreta(Objekte)	Gegenstände	*X ist groß/klein/schwer...*
Stoffbezeichnungen	Bezeichnungen für Stoffe	*ein A aus X, ein Kilo X* (X im Singular)
Ereignisse	Ereignis	*X beginnt /endet*, entsprechende verbale Umschreibung möglich
Zustände	Zustand	*X herrscht, A verharrt im X*
Eigenschaften	Eigenschaft	*X ist eine Eigenschaft*
Zeitbezeichnungen	Zeitpunkte/räume	*an X, ein X lang*
Formen	Abstrakte Begriffe für Formen	*ein X an A, ein X-förmiges A* (wobei A ein Gegenstand ist)
Lokativa	Ortsbezeichnungen	*in/an/auf* X
Diversa	Einzelklassen	—

Tabelle 3.2: Grobklassifizierung

3.4 Voraussetzungen zur Feinklassifizierung

Im Kodierungsprozeß mußte vor der Inangriffnahme der detaillierten semantischen Beschreibung geklärt werden, ob neben den semantischen Klassen noch andere Bedeutungsdeskriptoren zu definieren waren, deren Kodierung nicht auf die Beschreibung der semantischen Klassen folgen sollte, sondern günstiger mit dieser parallel zu laufen hatte. Für eine solche zusätzliche Beschreibung wurde der Beschreibungstyp des semantischen Relators eingeführt.

Eine weitere notwendige Vorarbeit für die detaillierte semantische Klassifizierung war die Festlegung der Behandlung verschiedener Fälle von semantischer Mehrfachklassifizierung eines polysemen Lexems.

3.4.1 Semantische Relatoren

Manche Lexeme versagen sich der direkten Einordnung in die hierarchische Strukturierung der Wortschatzes. So läßt sich das Lexem *Flotte*, das eindeutig Wasserfahrzeuge bezeichnet, nicht im normalen Testrahmen für Hyponymie bezüglich der zunächst anzunehmenden Hyperonyme *Schiff* oder *Wasserfahrzeug* verwenden:

40 *Eine Flotte ist eine Art von Schiff

41 *Eine Flotte sind Schiffe

Bei anderen Lemmata erfolgt eine hyponymische Zuordnung indirekt über ein semantisch eng verwandtes Lexem:

42 Ein Schiffchen ist ein kleines Schiff

43 Ein Schiff ist ein Wasserfahrzeug

Dies ist keine singulären Fälle. Eine nicht geringe Zahl von Nomina gehören zu semantischen Bedeutungsgruppen, die sich in erster Linie durch ihr Verhältnis mittels einer bestimmten semantischen Relation — z. B. der Meronymie — zu anderen semantischen Einheiten bestimmen lassen. In gedruckten Lexika wird für diese Nomina meist keine Beschreibung mittels Hyponymen geliefert, sondern eine andere semantische Relation spezifiziert. Diese Art der Bedeutungsbeschreibung wird für die Lemmata herangezogen, für deren semantische Struktur eine andere semantische Beziehung — etwa die Teil-Ganzes-Relation — wichtiger wird als die Hyponymiebeziehung[17]. So ist etwa das primäre Einordnungskriterium für das Lexem *Dach*, Teile eines Gebäudes zu sein; das primäre Kriterium für die semantische Einordnung der Nomina *Kalb*, *Ferkel* oder *Kitz,* das Jugendstadium einer Tierart zu bezeichnen. Dies drückt sich auch in der Definition dieser Lexeme in herkömmlichen Lexika aus, die meist durch einen Rückgriff auf andere Nomina erfolgt. So wird die Bedeutung von *Dach* in Wahrigs 'Deutschen Wörterbuch' (1986) beschrieben als "oberer Abschluß eines Gebäudes", das Lemma *Kitz* erhält dort die Umschreibung "Junges der Ziege ... von Reh-, Gams-, Steinwild". *Wrack* wird umschrieben als "durch Beschädigung unbrauchbar gewordenes, zerschelltes Schiff". Die Umschreibungen in solchen Einträgen lassen sich teilweise systematisch auf einige grundlegende Relationsbezeichner zurückführen, wobei bei einer solchen Rückführung notwendigerweise einige Information verlorengeht, die in der voll ausformulierten Definition vorhanden ist.

Die allgemeine Form der Kodierung von semantischen Relatoren ist Relator(X). Innerhalb der Klammern kann dabei eine semantische Klasse oder ein Lexem stehen. Die folgenden Beispiele illustrieren diese beiden Möglichkeiten je anhand eines Eintrags im CISLEX:

44 Ferkel;JUV("Schwein");

45 Dach;PAR(GEB);

Die Zuordnung der oben genannten Ausdrücke für Jungtiere mittels der Relatoren ist der primäre Kode dieser Lemmata. Die Zuordnung zu semantischen Klassen erfolgt dann zweckmäßigerweise sekundär über das durch die Relation zugeordnete Lexem. Der Meronymieoperator im zweiten Beispiel operiert auf einer semantischen Klasse, d. h. auf einer Menge von Kohyponymen, in diesem Fall auf den Bezeichnern von Gebäuden.

Die semantische Klasse eines Lexems, das mit einem semantischen Relator kodiert wurde, ergibt sich in vielen Fällen aus bestimmten Regeln, die für den Operator gelten. Die Zugehörigkeit zur Klasse Säugetier ergibt sich für das Beispiel *Ferkel* aus der Implikation:

```
semantische_Klasse(JUV(X)) = semantische_Klasse(X)
semantische_Klasse(PAR(Artefakt)) = Artefakt
```

3.4.2 Kollektiva

Eine Einteilung des Wortschatzes in semantische Klassen spiegelt die hyponymische Struktur des Wortschatzes wider. Mit der Klasse "*group, collection*", Kollektiva, in WordNet liegt dort eine Klasse vor, die zur sonstigen Klassifikation vollständig quer liegt und die mit fast allen anderen der genannten Klassen, zumindest jenen, deren Einheiten zählbare Entitäten bezeichnen, kompatibel ist. Zudem liegt für diese Gruppe von Nomina kein übergeordneter Begriff vor, auch wenn relativ neutrale Ausdrücke für Ansammlungen von Objekten zur Verfügung stehen. In den folgenden Belegen zeigt sich aber, daß Nomina wie *Gruppe* und

[17] Vgl. zu einer Erörterung der Bereiche des Lexikons, für deren Einteilung die Meronymiebeziehung relevanter ist als die Hyponymie, die Ausführungen in Tversky (1990).

Ansammlung keineswegs Hyperonyme zu allen Kollektiva sind; hier scheint in den zuerst genannten Beispielen jeweils ein syntaktisches Komplement zu den scheinbaren Hyperonymen zu fehlen:

46 ?Eine Herde ist eine Gruppe/Ansammlung

47 Eine Herde ist eine Gruppe/Ansammlung von Tieren

48 ?Eine Flotte ist eine Ansammlung

49 Eine Flotte ist eine Ansammlung von Schiffen

Kollektiva insgesamt können somit nicht als eine semantische Klasse in einer hyponymischen Struktur angesehen werden. Vielmehr scheint es sich um eine Eigenschaft von Lexemen aus verschiedensten semantischen Klassen zu handeln: So ist etwa das Lexem *Beamtenschaft* als kollektiver Menschenbezeichner einzuordnen, das Lexem *Flotte* als eine Kollektivbezeichnung für Wasserfahrzeuge, *Wild* als Sammelbezeichnung für bestimmte Tiere. Es ist dabei zu beachten, daß einige Bedeutungsbestandteile und damit auch einige kontextuelle Eigenschaften, die von der semantischen Klasse, denen die Kollektiva zugeordnet sind, eigentlich impliziert werden, durch die Eigenschaft Kollektiv aufgehoben werden. In den folgenden Beispielen zeigt sich dies deutlich am Vergleich zwischen dem Individualnomen *Schiff* und dem Kollektivum *Flotte* in Verbindung mit zwei für Wasserfahrzeuge typischen Kontexten:

50 Das Schiff sinkt

51 Die Flotte sinkt

52 Das Schiff kentert

53 ? Die Flotte kentert

Hier tritt das Kollektivum *Flotte* ebenso wie das Individualnomen *Schiff* als Subjekt von *sinken* auf; andererseits ist es allerdings nicht möglich, den typischen Kontext *kentern* auf das Kollektivum zu übertragen[18]. Dahlgren (1988: 49) behandelt die Unterscheidung "INDIVIDUAL/ COLLECTIVE" in Form semantischer Klassen wie auch etwa die Unterscheidung "REAL/ ABSTRACT", und unterscheidet sie nicht von den anderen Klassen. Gross (1992, 1994) und die Klassifikation in WordNet (Miller u. a. 1994) unterscheiden nicht zwischen orthogonalen Bedeutungsaspekten und hyponymischen semantischen Klassen. Gross stellt unter einzelnen Grobklassen Unterklassen mit Kollektiva zu Verfügung (so etwa bei Menschenbezeichungen relationale Kollektiva mit dem Beispiel *famille,* vgl. Gross 1994: 20). Mir scheint die Behandlung dieser Eigenschaft in Form eines eigenen Klassentyps sinnvoller als eine solche über alle anderen Klassen distribuierte gesonderte Behandlung, weil dies ermöglicht, über den gesamten Wortschatz hinweg die Gemeinsamkeiten der Kollektiva zu erfassen[19] und gleichzeitig die Gemeinsamkeiten mit den nicht-kollektiven Nomina durch die automatische Zuordnung zur Operandenklasse zu beschreiben.

Die Eigenschaft 'kollektiv' wird aus den genannten Gründen als semantischer Relator kodiert. Ein Lexem wie *Flotte* erhält damit — in einer Bedeutung — den Kode kollektiv(Wasserfahrzeug). Für die Kollektiva gilt die Regel:

[18] Eine Untersuchung darüber, welche Kontexte übertragbar sind und welche nicht, setzt eine Berücksichtigung der Verbsemantik voraus und würde an dieser Stelle zu weit führen.

[19] Es wird für eine detailliertere semantische Beschreibung sicher notwendig werden, zwischen verschiedenen Typen von Kollektiva zu unterscheiden. Bestimmte Kollektiva wie etwa *Klerus* sind nicht oder kaum pluralfähig, während andere, wie *Familie, Regierung*, die eher Funktionseinheiten bezeichnen, ohne weiteres einen Plural bilden können.

`semantische_Klasse(KOL(X)) = semantische_Klasse(X)`

Ein weiterer, für den gesamten Nominalwortschatz relevanter Relator ist diminutiv. Für die Diminutivbildung besteht eine fast hundertprozentige Übereinstimmung zwischen Morphologie und Semantik. Im Deutschen treten hauptsächlich die Suffixe -*chen* und -*lein* auf, wobei das erstere wesentlich häufiger ist; andere Diminutivsuffixe wie -*le*, -*li*, -*ke* sind regional/dialektal gebunden und haben nur in einigen wenigen Fällen, etwa bei *Steppke* oder *Müsli*, Eingang ins Standarddeutsche gefunden (vgl. Fleischer/Barz 178ff). Nur bei einigen wenigen Lexemen mit Diminutivsuffix ist dessen diminutive Bedeutung ganz verlorengegangen. Teilweise hat jedoch eine so starke Veränderung der Semantik der suffigierten Form stattgefunden, daß die Bedeutung des Simpliziums + Suffix nicht mehr kompositionell ermittelt werden kann (*Weibchen*, *Männchen*). In den meisten Fällen des Auftretens eines Diminutivsuffixes ist die Bedeutung von Nomen + Suffix jedoch kompositionell zu ermitteln. Die Kodierung dieses Merkmals kann somit auf Basis der morphologischen Kodierung weitgehend automatisch erfolgen. Das Verhalten von Derivationsbasis und Diminutivform bezüglich Selektionspräferenzen von Operatoren und bezüglich der Hyponymiebeziehungen zu anderen Lexemen ist in den kompositionellen Fällen kaum verschieden. Somit kann die semantische Klasse der Suffigierungsbasis in diesen Fällen für die Diminutivform übernommen werden. Ein Diminutivausdruck wie *Fischchen* erhält den Operator diminutiv() mit der Derivationsbasis *Fisch* als Operand zugewiesen, was in diesem Falle zum Kode diminutiv("Fisch") führt. Die Zuordnung zur semantischen Klasse Fische erfolgt indirekt über die Derivationsbasis aufgrund der Regel:

`semantische_Klasse(diminutiv(X)) = semantische_Klasse(X)`

Das Lexem *Dach* wird als PAR(Gebäude) kodiert, wobei PAR für die Meronymiebeziehung (s. Kapitel 2) steht. Diese Zuordnung als primären semantischen Kode anzugeben läßt sich leicht rechtfertigen, wenn man die üblichen Testrahmen für Hyponymie und Meronymie betrachtet:

54 Ein Dach ist eine Art von ...

55 Ein Dach ist Teil eines Gebäudes

Es ist nicht möglich, für das Lexem *Dach* ein brauchbares Hyperonym zu finden, während die Nennung des Holonyms *Gebäude* unproblematisch ist. Die direkte Zuordnung zu einer semantischen Klasse ist auch aufgrund von typischen Kontexten kaum möglich — *Dach* teilt typische Kontexte praktisch nur mit seinen Hyponymen wie *Satteldach* usw.

Kriterium für die Kodierung der Meronymierelation sind die üblichen Testrahmen. Die so kodierte Meronymiebeziehung kann als Prototyp für die Kodierung der Teil-Ganzes-Beziehung auch bei Lexemen gesehen werden, wo sie nicht zentraler Aspekt der Bedeutung ist. Wie bereits oben erwähnt, läßt sich für den Meronymierelator keine allgemeine Regel für die Zuordnung zu einer semantischen Klasse formulieren.

3.4.3 Sexus

Sexus wird üblicherweise als semantisches Merkmal kodiert. Es ist jedoch fast stets möglich, zu den für Sexus spezifizierten Ausdrücken die sexusneutralen Terme anzugeben, wie die folgenden Beispiele zeigen:

56 Arbeiterin - Arbeiter

57 Löwin - Löwe

58 Stier - Rind

Im Sinne einer Vernetzung des Wortschatzes ist es effizient, die Merkmale männlich und weiblich als semantische Relatoren zu definieren. Die Einträge für die genannten Beispiellexeme sind daher:

59 Arbeiterin;FEM(“Arbeiter“)

60 Löwin;FEM(“Löwe“)

61 Stier;MSK(“Rind“)

Die gewählten semantischen Kodes “FEM“ und “MSK“ dürfen nicht mit dem entsprechenden Kode für das Genus der Lexeme gleichgesetzt werden. Sexus ist nicht in jedem Fall mit den Genusmerkmalen für die entsprechenden Nomina zu identifizieren, denn häufig sind maskuline und feminine Nomina nicht sexusmarkiert. In einigen Fällen widerspricht die Genusmarkierung sogar der Sexusmarkierung. Selbst bei den Menschenbezeichnungen läßt sich nicht in allen Fällen von Genus auf Sexus schließen (z. B. *Mädchen*).

Sexus-Merkmale sind nur für die Grobklassen der Menschen und der Tiere relevant. Bei den Menschenbezeichnungen wurden als männlich bzw. weiblich die Nomina markiert, die ausschließlich für Menschen eines Geschlechts verwendet werden können. Dies führt dazu, daß der semantische Relator weiblich im Wörterbuch wesentlich häufiger vorkommt als der Relator männlich, da die männliche Form häufig mit der neutralen Form zusammenfällt[20], während die weibliche Wortform morphologisch und semantisch fast immer markiert ist. Diese Markiertheit gilt nicht oder nur in einigen Fällen für die Verwandtschaftsbezeichnungen (*Mutter - Vater*, *Neffe - Nichte*, aber: *Enkel - Enkelin*) und die Bezeichnungen für Menschen allgemein (*Mann*, *Frau, Junge*, *Mädchen*). In diesen Fällen steht auch häufig kein sexusneutraler Begriff zur Verfügung. In diesem Fall wird beim semantischen Kode für die entsprechenden Lexeme als Operand die semantische Klasse direkt angegeben. So wird *Nichte* markiert als FEM(Verwandte).

Besonders ausgeprägt ist der Zusammenfall von männlicher und neutraler Form für die Berufe. Hier sind die Nomina mit dem Sexusmerkmal weiblich auch fast durchweg morphologisch durch ein Movierungssuffix markiert. Nur einige wenige nicht-suffigierte Nomina (etwa *Amme*) tragen das Merkmal weiblich.

Das häufigste Movierungssuffix im Deutschen ist *-in*. Hinzu kommen einige aus anderen Sprachen entlehnte Ableitungsmöglichkeiten von der männlichen Form (etwa *-euse* zu *-eur*). Die morphologische Markierung erleichtert vor allem die semantische Kodierung der weiblichen Berufsbezeichungen wie *Arbeiterin*, *Verkäuferin* u.ä., da diese fast durchgehend automatisch auf Basis der maskulinen Pendants erfolgen kann. Ebenso wie bei den Diminutivformen wird auch hier die semantische Klasse der zugrundeliegenden maskulinen Form an die sexusmarkierte weibliche Form vererbt. Dies gilt durchgehend für die Relatoren männlich und weiblich. Die Regeln lauten:

```
semantische_Klasse(FEM(X)) = semantische_Klasse(X)
semantische_Klasse(MSK(X)) = semantische_Klasse(X)
```

3.4.4 Pejorativa

Quer durch die Grobklassen gibt negativ konnotierte Ausdrücke. Für diese Nomina als Gesamtheit existiert kein Hyperonym (das Lexem *Schimpfwort* ist eine Metabezeichnung). Es

[20] Dies gilt noch weitgehend, obwohl die Sexus-Neutralität der maskulinen Form zunehmend in Frage gestellt wird durch das Vordringen der *-Innen* Formen in schriftsprachlichen und der Verwendung von Konjunkten wie *Studenten und Studentinnen* in mündlichen Diskursen.

handelt sich hier eindeutig um eine Eigenschaft von Lexemen, die nach unserer Unterscheidung von semantischen Klassen vs. semantische Relatoren als Relator zu kodieren ist. Eine typische sprachliche Eigenschaft ist für die Nomina mit dem Relator Pejorativ, daß sie häufig in exklamativen Kontexten verwendet werden und dabei auch alleine stehen können; daneben gibt es auch andere typische sprachliche Kontexte, insbesondere die Kombination mit Adjektiven wie *ziemlich* oder *recht*:

62 Idiot

63 So ein Mist

64 Was für ein Unsinn

65 Das ist ein ziemlicher /rechter Mist

Alle derartigen Nomina erhalten im CISLEX den Relator pejorativ zugewiesen. So ist etwa der Ausdruck *Schwein* einerseits als Tier kodiert, andererseits aber auch als Pejorativausdruck für Menschen.

Die Tierbezeichnungen sind einer der Bereiche, in denen solche doppelten Bedeutungen — wobei eine davon stets eine Übertragung in die Klasse Menschen ist — sehr häufig vorkommen; interessanterweise werden dabei vor allem die Bezeichnungen für Nutztiere — und bei diesen ist dies beinahe ausnahmslos möglich — auch als Pejorativausdrücke verwendet. Neben diesem systematisch zur Pejorativbildung verwendeten Wortschatzbereich gibt es im Deutschen einige Affixe, die typischerweise zur Bildung von Pejorativausdrücken verwendet werden, so das Circumfix *Ge-e* (z. B. *Gesinge*, vgl. Fleischer/Barz 1992, 207f) zur Bezeichnung von Ereignissen/Aktionen auf verbaler Basis und das Suffix *-ling* zur Bildung von Menschenbezeichnungen (besonders auf adjektivischer Basis: z. B. *Schwächling*, vgl. Fleischer/Barz 165). Keines dieser Affixe wird jedoch ausschließlich pejorativ verwendet, was heißt, daß das Merkmal in diesen Fällen manuell kodiert werden muß.

3.4.5 Andere semantische Relatoren

Der semantische Relator Jugendform ist nur für die semantische Klasse der Tiere[21] relevant. Der Testrahmen für die Zuweisung diese Relators ist, an einem Beispiel demonstriert:

66 Ein Kitz ist ein junges Reh

Das Lexem Kitz erhält damit folgenden Eintrag:

67 Kitz;JUV(“Reh“);

Die so kodierten Lexeme gehören wie die mit den Relatoren männlich und weiblich kodierten Lemmata derselben semantischen Klasse wie der Operand an. Es gilt die Regel:

```
semantische_Klasse(JUV(X)) = semantische_Klasse(X)
```

68 Aus X wird ein A

69 X ist der Überrest von A

70 X ist ein zerstörtes A

Diese beiden Relationen dienen als Sammelbecken für relativ heterogene Phänomene. Sie könnten im weiteren Verlauf der Kodierung in verschiedene Unterrelationen aufgespalten werden. Für diese beiden Relatoren läßt sich keine allgemeine Zuordnungsregel zu einer semantischen Klasse formulieren.

[21] Die Bezeichner für junge Menschen sind so zahlreich und enthalten in ihrer Bedeutung so viele zusätzliche Implikationen, daß sie eine eigene semantische Klasse konstituieren.

Bezeichner	Erklärung	Relevant für:	Beispiel
KOL	kollektiv	alle	Flotte;KOL(“Schiff“)
DMV	diminutiv	alle	Hündchen;DMV(“Hund“)
PAR	Meronym	alle	Dach;PAR(Gebäude)
MSK	männlich	MEN TIE	Stier;MSK(“Rind“)
FEM	weiblich	MEN TIE	Kuh;FEM(“Rind“)
PEJ	pejorativ	alle	Gesinge;PEJ(Singen)
JUV	Jugendform	LEB	Kalb;JUV(“Rind“)
PRO	Vorstufen	alle	Saat;PRO(Pflanze)
RES	Überrest	alle	Wrack;RES(Schiff)

Tabelle 3.3: Semantische Relatoren im CISLEX

3.4.6 Tabelle der semantischen Relatoren

Abschließend eine tabellarische Übersicht über die bisher eingeführten semantischen Relatoren.

3.4.7 Andere Beschreibungselemente im Zusammenhang mit der semantischen Kodierung

3.4.8 Semantische Merkmale

Für die aufgestellten semantischen Klassen gilt mindestens eines der folgenden Kriterien:

- Die Untergliederung ist gedeckt durch die hyponymische Struktur des Wortschatzes.
- Die Untergliederung ist durch typische Kontexte für die einzelnen Klassen motiviert.

Es gibt nun Bedeutungskomponenten von Lexemen, welche die genannten Eigenschaften nicht besitzen; d. h. sie sind nicht durch die hyponymische Struktur des Wortschatzes gedeckt, und die entsprechenden Lexeme teilen nur in geringem Maße typische Kontexte. Sie sind aber auch nicht als Ableitungen von Bedeutungen anderer Lexeme oder Lexemklassen zu erfassen, wie dies durch die semantischen Relatoren formalisiert wurde. Solche Eigenschaften sollen als semantische Merkmale bezeichnet werden. Sie können wiederum in verschiedene Typen (analytische Merkmale wie ¡unverheiratet¿ für *Junggeselle*, Default-Merkmale wie ¡flugfähig¿ für Vögel) unterteilt werden.

Semantische Merkmale wurden bisher für das CISLEX nicht kodiert. Zu einem Überblick über mögliche Merkmalstypen s. Dahlgren (1988: 61f). Ebenda (S. 69) finden sich auch Überlegungen zu Beschränkungen der Merkmalsvergabe, die zeigen, daß eine taxonomische Einteilung des Wortschatzes — wie sie die Strukturierung durch semantische Klassen darstellt — die Möglichkeit bietet, Redundanzen in der Merkmalsvergabe zu vermeiden.

3.4.9 Relationale Nomina

Relationale Nomina sind solche, die mindestens eine Argumentstelle haben, an die sie eine thematische Rolle vergeben. Dies führt dazu, daß sie häufig mit Komplementen einer bestimmten Form auftreten. Wenn solche Nomina als Köpfe in Komposita auftreten, besteht häufig eine von ihnen determinierte Relation zwischen Kopf und Erstglied, während bei nicht-relationalen Köpfen die Bestimmung der Relation zwischen den Gliedern auf allgemeinen Regeln beruht.

Wie schon Vater (1978) bezüglich der Unterscheidung von Komplementen und Adjunkten zeigte und wie Jacobs (1994) durch die Problematisierung des Valenzbegriffs ebenfalls demonstrierte, ist es nicht einfach, eindeutige Kriterien für das Vorhandensein und die Zahl von Argumentstellen festzulegen. Für Nomina gilt dies vielleicht in noch höherem Maße als für Verben. Die gängigen Test, so der Ablösungstest (vgl. z. B. die in Meyer 1993: 106f aufgezählten Testreihen) sind oft nur wenig zuverlässig und es ist bei einigen der genannten Tests unklar, ob hier wirklich das Vorhandensein einer Argumentstelle getestet wird. Als Kriterien für die Kodierung im CISLEX wurde herangezogen, ob Nomina in Konstruktionen auftauchen, die auf die Vergabe einer Thetarolle und die Präferenz einer bestimmten Relation zwischen Komplement und Nomen hindeuten. Paraphrasierung mittels verbalen und adjektivischen Konstruktionen war hierbei ein wichtiger Test.

Relativ eindeutige Fälle von Klassen relationaler Nomina sind Verwandtschaftsbezeichnungen (*Mutter*, *Onkel*) und deverbale Nomina mit einer Argumentstelle, die einer Argumentstelle des zugrundeliegenden Verbs entspricht (*Untersuchung*). Auch einige nicht-deverbale Nomina, die eindeutig semantisch einem Verb zugeordnet werden können (z. B. *Dieb* zu *stehlen*) gehören zu den relationalen Nomina.

In der semantischen Kodierung des CISLEX hat bisher nur eine Markierung der eindeutig relationalen Nomina stattgefunden. Wie einige der Beispiele deutlich machen, wird das Merkmal der Relationalität in einigen Fällen semantischen (Verwandtschaftsbezeichnungen) oder morphologischen (deverbale Nomina auf *-ung*) Klassen im ganzen zumindest als Defaultwert zugeordnet und muß von daher nicht für jedes Nomen gesondert kodiert werden[22].

Die Eigenschaft relational ist für die weitere Kodierung insofern relevant, als für die relationalen Nomina die Syntax und die Semantik der Argumentstruktur spezifiziert werden müssen.

3.4.10 Fachsprachenbezeichner

Wie bereits in Kapitel 2 festgestellt, läßt sich für das Deutsche ein nicht unbedeutender Teil der einfachen Nomina nicht dem Allgemeinwortschatz zurechnen, sondern gehört einer Fachsprache an. Ein großer Teil der fachgebundenen Lexeme läßt sich nicht in die aufgestellten allgemeinsprachlichen semantischen Klassen einordnen. Die Einteilung dieser Sonderwortschatzbereiche kann nur aufgrund von Fachwissen bzw. durch Hinzuziehung eines Fachlexikons erfolgen. Dies konnte nicht das Ziel einer ersten semantischen Kodierung des deutschen Nominalwortschatzes sein. Ich begnügte mich damit, diese Lemmata als einem Fachwortschatz zugehörig zu markieren. Zu diesem Zweck wurden Bezeichner für Fachwörter verschiedener Bereiche festgelegt. Dabei wurde die gängige Klassifikation aus Standardwörterbüchern des Deutschen größtenteils übernommen (vgl. Wahrig 1986: 22f). Eine weitere Klassifikation dieser Wortschatzbereiche fand zunächst nicht statt.

3.4.11 Kombinatorik semantischer Klassen

Zur Kodierung der semantischen Klassen gehört auch die Berücksichtigung der Möglichkeit der Zuweisung mehrerer semantischer Deskriptoren zu einem Lemma. Solch eine Mehrfachkodierung wird nötig, wenn eine der folgenden drei semantischen Eigenschaften eines Nomens

[22]Zu trennen vom Vorhandensein einer Argumentstelle ist die Frage, ob diese Stelle obligatorisch gefüllt sein muß. Hier gibt es zwischen vollständiger Optionalität (*Koch*) und beinahe obligatorischer Füllung (*Nachbar*) alle Zwischenstufen. Diese Abstufungen wurden bei der Kodierung bisher nicht berücksichtigt. Ebenso unberücksichtigt bleibt bisher die Zahl der Argumentstellen. Insbesondere deverbale Nomina haben häufig zwei oder mehr Argumente.

in bezug auf die aufgestellten Klassen gegeben ist:

- Ein Nomen ist im Sinne der Klassifizierung mehrdeutig, und die verschiedenen Bedeutungen unterscheiden sich so, daß sie die Zuweisung unterschiedlicher semantischer Klassen erforderlich machen.
- Ein Nomen gehört in einer Bedeutung verschiedenen kompatiblen semantischen Klassen an. Bei einem einzelnen Vorkommen des Nomen kann die Zuordnung zu beiden Klassen relevant sein.
- Einem Nomen wird zusätzlich zu seiner Zuordnung zu einer semantischen Klasse ein Relator zugewiesen.

Diese Phänomene müssen in der Konzeption der Bedeutungskodierung berücksichtigt werden. Ihre Kodierung erfolgt durch zwei unterschiedliche Operatoren, die im Falle der Mehrfachzuweisung zwischen den semantischen Deskriptoren bei einem Lemma stehen.

Zunächst zum ersten Phänomen, der Polysemie. Bei einer Betrachtung der lexikographischen Praxis — sowohl für einsprachige wie für mehrsprachige Wörterbücher — wird schnell klar, daß es unmöglich ist, vollständig objektiv festzulegen, wieviele verschiedene Bedeutungen ein Lexem tatsächlich hat. Die Festlegung der Zahl der Bedeutungen hängt vielmehr ab von der Tiefe der semantischen Beschreibung und den Anwendungsgebieten des Wörterbuchs. Bei einer formalisierten Einteilung in semantische Klassen, wie sie für das CISLEX vorgenommen wird, ist die Kodierung von Polysemie relativ einfach zu regeln: Einem Nomen werden mehrere Bedeutungsbeschreibungen zugeordnet, wenn es unter zwei inkompatible semantische Klassen fällt. Dies heißt natürlich, daß die Zahl der Bedeutungen eines Nomens von der erzielten Kodierungstiefe abhängt. So wird etwa das Nomen *Clown* nicht polysem kodiert, solange nicht eine Unterscheidung getroffen wird zwischen Berufen und abfällig konnotierten, allgemeinen Ausdrücken für Menschen, da es in beiden Bedeutungen als Menschenbezeichnung einzuordnen ist.

Eine Abgrenzung von Polysemie und Homonymie[23] findet nicht statt. Das CISLEX ist so aufgebaut, daß Nomina mit verschiedenen morphologischen Eigenschaften von vornherein mehrere Einträge erhalten, so daß sich das Problem der Kodierung partieller Homonymie bei den Semantikeinträgen nicht stellt, denn dem jeweiligen Formenparadigma wird die entsprechende Semantik zugeordnet. Bei vollständiger — d. h. sich über das gesamte Formenparadigma eines Nomens erstreckender — Homonymie werden die verschiedenen Bedeutungen ebenso kodiert wie die eines polysemen Lexems.

Die Kodierung der Polysemie erfolgt durch den Operator '—' zwischen den zugeordneten semantischen Klassen. Dieser kann bezüglich der Klassenzugehörigkeit eines in einem konkreten Text auftretenden Nomens als Disjunktion, also als ein logisches 'oder' interpretiert werden. Bei Anwendungen des Lexikons sollte das Auftreten eines so kodierten Lexems semantisch disambiguiert werden.

Eine zweite Möglichkeit des Zusammenfalls verschiedener semantischer Klassen bei einem Eintrag ist die gleichzeitige Zuweisung kompatibler semantischer Klassen. Ein Beispiel: Das Lexem *Gans* wird sowohl der Klasse VÖGEL als auch der Klasse NUTZTIERE zugeordnet, also einmal unter biologischen, ein anderes mal unter funktionalen Aspekten klassifiziert. Diese Aspekte sind jedoch kompatibel: eine Gans ist zugleich Nutztier und Vogel; selbst in ein und demselben Kontext können beide Aspekte relevant werden.

[23] Die in den meisten Untersuchungen zur Abgrenzung von Polysemie und Homonymie vorgeschlagenen synchronen Eigenschaften sind ohnehin problematisch (vgl. Bußmann 1990: 314) und die eindeutigen Fälle vollständiger Homonymie aufgrund der Möglichkeit zur formalen Differenzierung im gesamten Deklinationsparadigma im Deutschen sehr selten.

Eine solche Mehrfachklassifizierung aufgrund unterschiedlicher Klassifizierungsaspekte wird anders kodiert als die oben diskutierte Polysemie. Im Falle von kompatibler Mehrfachklassifizierung werden die Klassen durch den Operator '&' verbunden, der bezüglich der Klassenzugehörigkeit des Nomens wie ein logisches 'und' interpretiert wird.

Der Operator '&' verbindet auch Klassen und Relatoren. So wird etwa das Lexem *Gans* für seine zusätzliche Bedeutung als Schimpfwort für einen weibliche Menschen zusätzlich zur Kodierung als Vogel und Nutztier die Kodierungssequenz FEM(MEN)&PEJ(MEN) erhalten. Diese Bedeutung ist jedoch nicht kompatibel mit der zuerst genannten und wird von daher von dieser durch den Operator '—' abgetrennt.

Eine Sequenz von durch '&' verbundenen Klassen kann auch als eine komplexe semantische Klasse gesehen werden; so steht etwa die Sequenz NUTZTIER&VOGEL für Nutzgeflügel. Eine komplexe Klasse ist jeweils den Hyperonymen von beiden durch die Konjunktion verbundenen Klassen untergeordnet. Liegt eine kontextuelle Abgrenzung der durch Konjunktion verbundenen Klassen vor, ergeben sich die Zuordnungskriterien für die Gesamtklasse aus der Summe der Kontexte für beide Klassen.

Der vollständige Eintrag (unter Auslassung der Morphologie) für das Lemma *Gans* sieht damit im CISLEX folgendermaßen aus:

71 Gans;f;TVO&TNU—FEM(MEN)&PEJ(MEN)

Dieser Eintrag zeigt beide Typen von Operatoren und demonstriert die Behandlung von kompatiblen und inkompatiblen Bedeutungen.

3.4.12 Formale Grammatik eines Semantikeintrags

Ich habe bisher die Typen von semantischen Deskriptoren im CISLEX und die vorhandenen Operatoren dargestellt. Die Kombinatorik der Grundeinheiten läßt sich formal beschreiben. Jeder Semantikeintrag im CISLEX gehorcht der folgenden Syntax (hier in Form einer kontextfreien Grammatik formuliert; der Startknoten ist S, Terminale sind unterstrichen):[24]

```
/*Syntax der Semantikeinträge im CISLEX*/
S -> SKonjunktion
S -> SKonjunktion|S
Skonj -> Klasse
Skonj -> Klasse&SKonjunktion1
Klasse -> Klassenbezeichner
Klasse -> Relator(Klassenbezeichner)
Klasse -> Relator(\Lemma\)
/*Grundeinheiten der Beschreibungssprache*/
Klassenbezeichner -> X /* z.B. VOG */
Relator -> X /* z.B. PAR */
Lemma -> X /* z.B. Amsel */
```

Operatorenpräzedenz ist '&' ¿ '—', d. h. der Operator '—' bindet weniger stark als der Operator '&'.

[24] Nicht berücksichtigt sind in dieser Syntax die Informationen zu Synonymen, Varianten und Oppositionen (s.u.), die in einem separaten Feld kodiert werden.

3.5 Automatische Kodierung aufgrund morphologischer Kriterien

Bereits bei der Beschreibung der semantischen Deskriptoren diminutiv und weiblich wurde darauf hingewiesen, daß auf bestimmte semantische Eigenschaften aus der Morphologie eines Wortes geschlossen werden kann. Für die Nomina handelt es sich bei den quantitativ (für die Kodierung) relevanten und semantisch kohärenten Korrelationen zwischen Morphologie und Semantik ausschließlich um semantische Implikationen von Suffixen. Bereits genannt wurde die große Zahl von als weiblich markierten Menschenbezeichnungen, die auf *-in* enden, und die morphologische Markierung für Diminutive durch die Suffixe *-chen* und *-lein*.

Außer diesen Suffixen kommen für eine automatische semantische Vorklassifizierung des Wortschatzes in Betracht:

- *-ung*, *-ion* : Diese sehr häufigen Suffixe dienen in erster Linie zur Bildung von Nomina auf Basis von Verbstämmen. Die meisten der so gebildeten Nomina gehören in die Klasse Ereignisse. Leider ist diese semantische Klassifikation nur eine Annäherung an die tatsächliche Semantik, da die *-ung* Formen gleichzeitig auch für die Denotate verschiedener Argumente der zugrundeliegenden Verben stehen können (vgl. Fleischer/Barz 1992: 174-177). Ein Beispiel ist das Lexem *Bebauung* vom Verb *bebauen*, das u. a. für die Aktion und für das effizierte Objekt stehen kann, wie die beiden folgenden Beispiel zeigen:

  72 Die schnelle Bebauung der Geländes

  73 Die Bebauung wird abgerissen

  Eine vollständige automatische Klassifizierung dieser Nomina ohne eine Berücksichtigung der Verbsemantik ist damit nicht möglich

- *heit*, *-keit*, *-igkeit*: Diese Suffixe, die in erster Linie zur Bildung von Nomina auf adjektivischer Basis dienen[25], bezeichnen mit wenigen Ausnahmen Eigenschaften oder Zustände (Ausnahmen sind beispielsweise *Dreifaltigkeit*, *Gerichtsbarkeit*). Nomina mit diesen Suffixen verhalten sich semantisch relativ einheitlich. Diese Suffixe können damit für eine semantische Vorklassifizierung der Nomina eingesetzt werden.

- *-itis* u.ä.: Im Fachwortschatz treten einige semantisch eindeutige Suffixe auf. So bezeichnet *-itis* im medizinischen Fachwortschatz entzündliche Erkrankungen. Allerdings kann dieses Suffix auch zu scherzhaften Bildungen herangezogen werden (*Telephonitis, Institutionitis* [26], was eine automatische Klassifizierung der entsprechenden Wörter zumindest überarbeitungsbedürftig macht. Andere Suffixe griechischer Herkunft, die sich eindeutig dem Fachwortschatz zuordnen lassen, sind z. B. *-ose* und *-phobie*.

Die anderen Nominalsuffixe [27] geben entweder aufgrund ihrer Seltenheit oder aufgrund der mangelnden Bedeutungsähnlichkeit zwischen den suffigierten Nomina keine ausreichenden Anhaltspunkte für die automatische semantische Einordnung. So tritt etwa das sehr häufige

[25] Es treten auch einige Bildungen auf substantivischer Basis auf, z. B. *Menschheit*, Kindheit (cf. Fleischer/ Barz 1992: 162).

[26] Letzterer Ausdruck aus der 'Süddeutschen Zeitung' vom 26./27. 8. 1995.

[27] Auf nominale Präfixe wird hier nicht eingegangen, da präfigierte Nomina erstens im Verhältnis zum gesamten Nominalwortschatz quantitativ keine sehr große Rolle spielen und zweitens denominale präfigierte Nomina aufgrund der Kriterien für die Zuordnung zu den 'einfachen Formen' im CISLEX ohnehin bisher noch nicht kodiert wurden.

Suffix -*er*, das oft zur Bildung deverbaler Nomina agentis verwendet wird, zu häufig auch bei solchen Nomina auf, wo die semantische Ableitungsbeziehung synchron verdunkelt ist (z. B. *Richter* zu*richten*) oder die zumindest so stark lexikalisiert sind (z. B. *Lehrer* zu*lehren*), daß die vollständige Bedeutung nicht aufgrund allgemeiner Regeln von der Bedeutung der Basis abgeleitet werden kann.

3.6 Feinklassifizierung: Ausgewählte Beispiele

Zur exemplarischen Darstellung der Feinklassifizierung [28] wurden drei größere semantische Klassen aus dem kodierten Wortschatz ausgewählt.

- Die Menschenbezeichnungen machen quantitativ einen wesentlichen Teil des nominalen Wortschatzes im Deutschen aus. An der Feinklassifizierung dieser Grobklasse — vor allem an der Untergliederung der Berufsbezeichnungen — zeigt sich zudem beispielhaft die Überlagerung verschiedener Klassifizierungskriterien.

- Die Nomina, die Fahrzeuge bezeichnen, sind prinzipiell eine relativ einfach zu behandelnde Klasse. Die hier erkennbaren Klassifizierungsprobleme treten bei der Unterteilung von beinahe allen anderen semantischen Klassen ebenfalls auf.

- Die Formen sind ein relativ kleiner, aber auch recht uneinheitlicher Wortschatzbereich. Es liegt für ihn keine kanonische Klassifizierung vor. Die Darstellung seiner Kodierung soll exemplarisch die Schwierigkeiten bei der Aufstellung von Unterteilungskriterien für semantische Klassen zeigen, für die weder typische sprachliche Kontexte noch eine stabile zugrundeliegende Ontologie zu ermitteln sind.

Es wurde bewußt darauf verzichtet, die Einteilung einer in erster Linie in taxonomische Unterklassen aufzuteilenden Klasse wie der TIERE darzustellen, bei der eine wissenschaftliche Aufteilung zugrundeliegt [29]. Eine solche Klassifikation liegt in brauchbarer Form in jedem Thesaurus und sogar in jedem größeren Wörterbuch vor.

3.6.1 Menschenbezeichnungen

Die Grobklasse der Menschenbezeichnungen macht mit ca. 7 500 der 37 000 Lemmata die größte semantische Grobklasse unter den 'einfachen Nomina' des CISLEX aus. Ihre semantische Unterteilung ist insbesondere für das Information Retrieval von großer Bedeutung, da in den meisten Textsorten die Menschenbezeichnungen eine quantitativ herausragende Rolle spielen. Die Benennungsmotivation für menschliche Referenten in einem Text kann in hohem Maße zur seiner thematischen Einordnung beitragen.

Für die Menschenbezeichnungen wurden folgende semantischen Klassen festgelegt:

Verwandtschaftsbezeichnungen. Dies ist eine relativ kleine und abgeschlossene Gruppe von Lexemen. Sie sind durchweg relational. Zu jeder Relation, die Verwandtschaftsbezeichnungen zugrundeliegt, lassen sich die konverse Relation und die dazugehörigen Lexeme spezifizieren. Die Angabe eines einzelnen konversen Paars von Ausdrücken ist jedoch oft nicht möglich, da die meisten Verwandtschaftsbezeichnungen zusätzlich für Geschlecht spezifiziert

[28] Die Beschreibung dient in erster Linie zur Darstellung der Vorgehensweise; es werden dabei aus Platzgründen einige semantische Klassen nicht berücksichtigt, insbesondere dann, wenn ihre Beschreibung nichts zur Klärung der Kriterien für die Klassifizierung beiträgt.

[29] Hinzu kommt in diesem Fall (Tierwelt) eine funktionale Einteilung in Nutztiere, Haustiere etc.

sind. Eine Gruppe von Lexemen, denen zwei konverse Relationen zugrundeliegen, ist beispielsweise *Eltern/ Vater/ Mutter — Tochter/ Sohn/ Kind* [30].

Etwas anders verhalten sich einige der Kollektiva in dieser Lexemklasse. So ist das Lexem *Geschwister* nur optional relational, da die zugrundeliegende Relation auch als Reziprokrelation zwischen den Referenten des Terms gelesen werden kann. Die zugrundeliegenden reziproken Relationen der Kollektiva lassen sich allerdings den Relationen bei den entsprechenden nicht-kollektiven Verwandtschaftsbezeichnungen zuordnen.

Die Klasse der Verwandtschaftsbezeichnungen hat weder unscharfe Ränder, noch eine prototypische Struktur; es handelt sich um eine taxonomische Klasse, für die auch eine Reihe von Selektionsregularitäten Gültigkeit haben.

Verwandtschaftsähnliche. Dies sind Nomina, die wie Verwandtschaftsbezeichnungen Relationen von Menschen zu anderen Menschen ausdrücken, jedoch im Gegensatz zu diesen nicht-genetische, bzw. nicht vollständig institutionalisierte Relationen. Zwar lassen sich teilweise ebenfalls die Bezeichner der konversen Relationen angeben (z. B. *Geliebte — Liebhaber*), doch gilt dies nicht für alle Lexeme dieser Klasse, und es besteht hier nicht in jedem Fall eine sichere Folgerungsbeziehung. Da es sich ebenso wie bei den Verwandtschaftsbezeichnungen um relationale Nomina handelt, treten die verwandtschaftsähnliche Relationen bezeichnenden Nomina häufig mit dem Possessivartikel auf (*sein Freund* ,*ihr Geliebter* etc.).

Im Gegensatz zu den reinen Verwandtschaftsbezeichnern hat diese Klasse keine scharfen Ränder. Für Lexeme wie *Nachbar*, die sowohl räumliche Situierung beinhalten als auch Relationen zwischen Menschen bezeichnen, ist die Zuordnung zu dieser Klasse problematisch. Es handelt sich nicht um eine taxonomische Klasse.

Altersstufen. Eine relativ kleine Gruppe von Nomina (ca. 100) bezeichnet Menschen aufgrund ihrer Altersstufe. Zu dieser semantischen Klasse gehören beispielsweise *Kind*, *Erwachsener*, *Baby*. Viele von ihnen sind zusätzlich bezüglich des Sexus festgelegt (z. B. *Junge*, *Greisin*). Ein typischer Kontext für diese Nomina ist ihr Auftreten als Komplement einer temporalen *als*-Phrase wie in den folgenden beiden Beispielen:

74 Als Kind wirkte er noch begabt

75 Noch als Greis bestieg er allmonatlich den Wendelstein

Zudem treten die Nomina dieser Klasse in Komposita mit dem Kopf *-alter* auf (*Greisenalter*, *Kindesalter*).

Einige Nomina, bei denen andere Bedeutungsaspekte primär sind, so etwa *Schüler* (als Mensch in einem bestimmten Ausbildungsabschnitt) oder *Pensionär* (als Bezieher einer Pension) haben typische — wenn auch aufhebbare — Implikationen in Hinsicht auf das Lebensalter, ohne ausschließlich auf dieses Bezug zu nehmen. Sie wurden zusätzlich in diese Klasse eingeordnet und damit doppelt klassifiziert.

Eine weitere Unterteilung dieser Klasse aufgrund der verschiedenen Lebensaltersstufen ist problemlos möglich.

Herkunftsbezeichnungen. Diese Nomina (ca. 400) bezeichnen Menschen aufgrund ihrer Herkunft oder Abstammung. Sie können in der Mehrzahl ebenso wie die im nächsten Abschnitt dargestellten Berufsbezeichnungen ohne Artikel nach der Kopula *sein* auftreten:

76 Sie ist Französin

77 Er ist Südländer

[30] Zudem lassen sich alle Verwandtschaftsrelationen auf einige wenige einfache Relationen zurückführen. Es gibt zahlreiche Untersuchungen zur Semantik der Verwandtschaftsbezeichnungen in verschiedenen Sprachen. Für eine Übersicht siehe Bußmann (1990: 836).

Viele der Nomina dieser semantischen Klassen befinden sich im Grenzbereich zwischen Gattungs- und Eigennamen (*Maure*, *Ossi, Wessi*). Die Herkunftsbezeichnungen lassen sich in folgende Subgruppen unterteilen:

- Formal mit Landes- oder Regionennamen verbundene Nomina (*Französin*, *Afrikaner*). Diese Gruppe ist einfach abzugrenzen und verhält sich sprachlich sehr einheitlich. Sie kann selbstverständlich weiter unterteilt werden.
- Ethnische Zugehörigkeit. Dies ist eine Gruppe, die sich teilweise mit der zuvor genannten überschneidet. Zu ihr gehören Klassifizierungen aufgrund von Hautfarben und Ethnien. Beispiele sind *Zigeuner*, *Germane*. In diesem Wortschatzbereich wäre es für die automatische Übersetzung wichtig, eine zusätzliche Klassifizierung nach wertenden Konnotationen der Ausdrücke vorzunehmen.
- Vage regionale oder ethnische Herkunftsbezeichnungen. Diese Lexemgruppe ist nicht sehr groß. Sie enthält Nomina wie *Exot*, *Südländer*.
- Religiöse Zugehörigkeit. Diese Ausdrücke werden den Herkunftsbezeichnungen zugerechnet, weil sich bei ihrem Gebrauch häufig Kriterien der Abstammung und Religionszugehörigkeit vermischen (*Muslim*, *Jude*). Sie treten häufig mit Adjektiven wie *praktizierend*, *gläubig* auf, was sie von den anderen genannten Untergruppen unterscheidet.

Wie insbesondere aus der letzten Gruppe ersichtlich, ist die Klasse der Herkunftsbezeichnungen ebenfalls eine Klasse mit unscharfen Rändern. Das gilt auch für die Unterklassen.

Berufsbezeichnungen haben gegenüber anderen Menschenbezeichnungen einige zusätzliche kontextuelle Eigenschaften, die zur Abgrenzung dieser Lexemklasse verwendet werden können. Einige Kontexte grenzen nur bestimmte Unterklassen der Berufsbezeichnungen ab und können somit auch zur weiteren Unterteilung dieser Gruppe verwendet werden. Die Berufsbezeichnungen machen die größte Untergruppe der Menschenbezeichnungen aus. Ca. 2 700 der rund 7 500 kodierten Ausdrücke für Menschen gehören im weitesten Sinn [31] zu dieser Klasse.

Eine Klassifizierung der Berufsbezeichnungen ist relevant für die maschinelle Textverarbeitung. In zahlreichen Texttypen werden Menschen über ihren Beruf eingeführt und charakterisiert. Oft werden bereits eingeführte Diskursreferenten mittels der Berufsbezeichnung anaphorisch aufgegriffen. Diese herausragende Stellung macht bei der großen Menge der dazugehörigen Lexeme eine weitere Unterteilung in Subklassen sinnvoll. In der Menge der Unterklassen sollten dabei sowohl themenspezifische als auch selektionsrelevante Aspekte berücksichtigt werden.

Sowohl die Oberklasse Beruf als auch die meisten der Unterklassen besitzen unscharfe Ränder und eine prototypenzentrierte Struktur; es ist dabei jedoch nicht möglich, ein einzelnes Nomen als Prototypenbezeichner für die Gesamtklasse Berufe festzulegen. Auch in den meisten Unterklassen existiert mehr als ein prototypennaher Bezeichner. Es lassen sich allerdings einige Eigenschaften prototypischer Berufsbezeichnungen festlegen.

[31] Die Einteilung in WordNet konnte im Falle der Menschenbezeichnungen nicht als Referenz herangezogen werden. So fehlt etwa ein Knoten *Beruf* als allgemeiner Überbegriff: Der erste gemeinsame übergeordnete Knoten von je zweien der Einträge *carpenter*, *teacher, priest* und *writer* ist stets *person*, *individual*, der Knoten, der in unserer Taxonomie den Menschenbezeichnungen entspricht. Auch die den Lexemen direkt übergeordneten Klassen zu den Berufsbezeichnungen sind nicht sehr weit differenziert. Ähnliches gilt für die anderen Unterklassen der Menschenbezeichnungen in WordNet. Ein Nomen zur Herkunftsbezeichnung wie *african* hängt beispielsweise direkt unter dem Knoten *person*, *individual*. Die Einteilung der Berufsbezeichnungen in Gross' Klassifikation ist — falls vorhanden — in Gross (1994) nicht vollständig dokumentiert. Einige der anderen Klassen in der dargestellten Klassifikation der CISLEX haben Entsprechungen in Gross' Objektklassen (vgl. Gross 1994: 20f).

- Sie bezeichnen eine Tätigkeit, die ganztägig und regelmäßig ausgeübt wird; diese Implikation kann durch Präfixe und Attribute wie *Teilzeit-*, *nebenberuflich* aufgehoben werden. Bei nicht prototypischen Berufen ist eine solche Modifikation ungebräuchlich (z. B. ?*Teilzeitschriftsteller,* ?*Teilzeitpriester*).

- Das Berufsbild eines prototypischen Berufs ist eindeutig identifizierbar, d. h. es lassen sich typische Tätigkeitsszenarien angeben.

- Die denotierte Tätigkeit ermöglicht üblicherweise, den Lebensunterhalt zu bestreiten; dies schlägt sich in Kontexte wie den folgenden nieder:

  78 Er verdient sein Geld als Bauarbeiter

  79 Sie bestreitet ihren Lebensunterhalt als Sekretärin

- Diese Implikation kann durch bestimmte Kontexte aufgehoben werden (z. B. durch das Erstglied *Hobby* im Kompositum *Hobbytischler*)

Sämtliche Berufsbezeichnungen lassen sich auch abstrakt, d. h. ohne jegliche Referenz auf Menschen, verwenden:

80 Tischler ist ein Handwerksberuf

81 Dichter ist kein Beruf sondern eine Berufung

Zunächst seien eine Reihe typischer Kontexte zur Unterscheidung zentraler und weniger zentraler Lexeme in der semantischen Klasse Beruf genannt[32]:

82 Karl **ist** Lackierer (**von Beruf**)

83 Hans **möchte** Schaffner **werden**

84 Maria **arbeitet als** Kassiererin

85 Karla **hat eine Anstellung als** Erzieherin

Der erste Kontext umfaßt auch die berufsähnlichen Ausdrücke — allerdings können noch zahlreiche andere Nomina, die habituelle Tätigkeiten, Menschen nach Herkunft oder Menschen aufgrund anderer Eigenschaften bezeichnen, ohne Artikel nach der Kopula auftreten.

Zur Gruppe der Berufsähnlichen lassen sich vor allem Bezeichnungen aufgrund der Haupttätigkeit einer Person in einer temporär eingeschränkten Lebenssituation rechnen; diese Lexeme sind allerdings nicht als Berufe im eigentlichen Sinne anzusehen:

86 Otto ist Schüler/Pensionär

Ausdrücke dieser Art, die nur aufgrund weniger geteilter Eigenschaften mit den eigentlichen Berufsbezeichnungen verglichen werden können, werden gesondert klassifiziert. Sie konstituieren zwei größere Gruppen: Bezeichnungen für Menschen in der Ausbildung (*Schüler*, *Student*) und für Menschen im Ruhestand (*Pensionär*, *Rentner*). Sie mit den Berufsbezeichnungen in Verbindung zu bringen, läßt sich außer durch die mit prototypischen Berufsbezeichnungen geteilten Eigenschaften durch Anforderungen der Informationserschließung begründen: üblicherweise werden diese Ausdrücke in Texten und Formularen in derselben

[32] In der Arbeit von Bohnhof (1993) wird eine große Zahl von typischen Kontexten für Berufsbezeichnungen aufgeführt.

Weise zur Kurzcharakterisierung von Personen verwendet wie die Berufsbezeichnungen für erwerbstätige Personen.

Der zweite Kontext (refmoechtwerd) ist zur Abgrenzung von möglichen Hauptlebenstätigkeiten, wie sie Berufe darstellen, von temporären Ausbildungsphasen wie den oben genannten verwendbar. Zu den Berufen in diesem Sinne gehören auch Ämterbezeichnungen und Bezeichnungen für Künstler:

87 Maria möchte Äbtissin werden

88 Hans möchte Minister werden

89 Otto möchte Dichter werden

Doch nicht alle genannten Ausdrücke lassen sich ohne weiteres mit Kontext (84) kombinieren:

90 ? Hans arbeitet als Abt

91 ? Maria arbeitet als Ministerin

92 ? Otto arbeitet als Dichter

Wesentlich für Berufe im engsten Sinne scheint also die Möglichkeit, den Beruf zu wählen. Die bereits oben genannten Beispiel *Abt* und *Minister* sind deshalb in einige Kontexte kaum einsetzbar, weil es sich hier um Ämter handelt, in die man berufen wird. Ein ähnlicher Ausschluß bestimmter Kontexte gilt für Tätigkeiten, die als Berufung und weniger als Beruf angesehen werden, wie *Pfarrer*, *Abt* etc.

Für diese Gruppen wurden zwei Klassen festgelegt: Kirchenämter und politische Ämter. Hinzu kam mit einigen ähnlichen Eigenschaften die Klasse der Künstler, die Lexeme wie *Dichter*, *Poet* und *Maler* einschließt. Sie ist zu unterscheiden von der Klasse der künstlerischen Berufe, die Lexeme wie *Schauspieler* beinhaltet, die mit Kontext (84) kompatibel sind und damit zu den Berufen im engeren Sinne gehören.

Die nun noch zu klassifizierende Ausdrücke für Berufe im engsten Sinne läßt sich unter verschiedenen Aspekten weiter unterteilen. Eine relevante Beschreibungsdimension ist die Ausbildung für einen Beruf; diese läßt sich unter den folgenden Teilaspekten betrachten:

Ausbildungsnotwendigkeit

Die Bedeutungskomponenten Grad der Notwendigkeit und Typ einer Ausbildung haben großen Einfluß auf die Kombinatorik von Berufsbezeichnungen. Sie lassen sich primär über die Kombinationsmöglichkeiten mit verschiedenen Adjektiven abprüfen, etwa mit *promoviert*, *studiert*, *diplomiert*, *ausgebildet* oder *gelernt*. Mit abnehmender Notwendigkeit einer Ausbildung nimmt auch die Akzeptabilität von Sätzen ab, die eine Person allein aufgrund ihrer Ausbildung bezeichnen, aber eine Tätigkeit im selben Bereich explizit verneinen:

93 Maria ist Biologin arbeitet aber zur Zeit als Kassenkraft

94 Hans ist Fräser arbeitet aber zur Zeit als Wagenwäscher

95 ? Ophelia ist Kassiererin arbeitet aber zur Zeit als Straßenkehrerin

Die Art der Ausbildung korreliert zum Teil mit anderen Unterteilungsaspekten.

subsubsection*Ausbildung und Berufsbild

Der Grad der Berufsrelevanz einer Ausbildung und das zu einer Ausbildung gehörige Berufsbild sind ebenfalls wesentliche Bedeutungskomponenten von Berufsbezeichnungen.

Es existieren Ausbildungsbezeichnungen, v. a. Abschlüsse universitärer Art, die nur schwerlich als Tätigkeitsbezeichnungen zu lesen sind. So ist etwa der folgende Satz kaum akzeptabel:

96 ? Hans arbeitet als Germanist

Neben Ausbildungen, denen kein einzelnes typisches Tätigkeitsszenario entspricht, gibt es solche, denen mehrere Tätigkeitsbilder zuzuordnen sind, und die ihrerseits keine oder nur in eingeschränktem Maße Berufsbezeichnungen sind. Dies gilt z. B. für die Bezeichnung *Jurist*, der verschiedene Berufsbilder entsprechen, die durch Lexeme wie *Anwalt* und *Richter* ausgedrückt werden.

Nicht relevant für diese reinen Ausbildungsbezeichnungen sind die weiter unten genannte Parameter für die Unterklassifizierung von Berufen. Für diese Gruppe kommt eher eine weitere Einteilung nach Ausbildungstypen wie Lehre und akademische Ausbildung in Betracht.

subsubsection*Hobby

Der Grad der Möglichkeit, eine berufliche Tätigkeit auch als Freizeitaktivität auszuüben, läßt sich daran messen, inwiefern die Bezeichnung allein auf den Berufsstatus schließen läßt, und inwiefern Bildungen mit Erstgliedern wie *Hobby-*, *Freizeit-* auf der einen und *Profi-* oder *Berufs-* auf der anderen Seite akzeptabel sind:

97 Maria ist Violinistin

98 Hans ist Straßenkehrer

99 Maria ist Hobbyviolinistin

100 ? Hans ist Hobbystraßenkehrer

Aus den bisher genannten Distinktionen innerhalb der Berufe im engsten Sinn lassen sich mindestens vier Klassen herausarbeiten. Zu den semantischen Klassen werden in der folgenden Auflistung jeweils Beispiele und nach dem Semikolon Gegenbeispiele angegeben:

Ausbildung: handelt es sich um eine über die Ausbildung definierte Bezeichnung? Zu dieser Klasse gehören beispielsweise: *Germanist, Jurist*; nicht *Kassierer, Bonne.*

Tätigkeit: Handelt es sich um eine Bezeichnung, die auf die konkrete Tätigkeit Bezug nimmt? Zu dieser Klasse gehören beispielsweise: *Übersetzer, Anwalt, Straßenkehrer;* nicht *Germanist, Jurist.*

Ausbildungsberuf: Ist für eine tätigkeitsorientierte Berufsbezeichnung eine Ausbildung nötig? Zu dieser Klasse gehören beispielsweise: *Anwalt, Richter;* nicht *Kassierer, Straßenkehrer.*

Freizeitaktivität: Läßt sich die Berufsbezeichnung auch zur Bezeichnung einer Person aufgrund einer Freizeitbeschäftigung verwenden? Zu dieser Klasse gehören beispielsweise: *Violinist, Schreiner*; nicht *Richter, Straßenkehrer*

Damit sind die wichtigsten semantischen Klassen von Berufsbezeichnungen zur Beschreibung von Distinktionen entlang der Dimensionen Ausbildung und Tätigkeit aufgezählt.

Die zweite wichtige Dimension für die semantische Klassifikation der Berufsbezeichnungen ist die Tätigkeitsart. Diese Distinktionen lassen sich aufgrund des sprachlichen Kontexts allerdings weniger einfach abgrenzen als die bisher aufgelisteten Klassen. Sie besitzen aber

Relevanz für die hyponymische Strukturierung des Wortschatzes und sind für die thematische Zuordnung von Texten und Textabschnitten von Belang. Es geht hier in erster Linie um Unterscheidungen, die aufgrund von Oberbegriffen wie *Handwerker*, *Bankangestellte*, *Arbeiter* und *Verkäufer* zu legitimieren sind. Dieser Typ von Distinktion ist die Grundlage für die Klassifikation der Bundesanstalt für Arbeit (1992).

Eine Klassifizierung dieser Art ist insbesondere relevant für das Information Retrieval, wenn es etwa darum geht, Texte oder Textabschnitte über bestimmte Berufsgruppen automatisch zu identifizieren. Für das CISLEX war die von der Bundesanstalt vorgenommene Klassifikation allerdings zunächst zu technisch und zu fein[33]. Es galt, die dort vorhandenen Unterscheidungen auf einige wenige — und allgemeinsprachlich relevantere — Klassen zurückzuführen. Durch die Reduktion ergaben sich u. a. folgende Gruppen, hier jeweils mit Beispielen:

Handwerker:	*Schreiner*, *Klempner*
Industriearbeiter:	*Schweißer*, *Schmelzer*
Büroangestellte:	*Sekretärin*, *Sachbearbeiter*
Heilberufe:	*Ärztin*, *Neurologe*
Polizeiberufe:	*Carabiniere*, *Polizist*
Soldatenberufe:	*Artillerist, Fähnrich*

Es kam hierbei häufig zu Doppelklassifikationen.

Nicht in der Klassifikation des Arbeitsamtes fanden sich die im CISLEX als Statusbezeichner klassifizierten, abstrakten Metabezeichnungen (*Arbeiter*, *Angestellter*, *Beamter*).

Eine weitere Gruppe sind die Hierarchiebezeichnungen (*Boß*, *Manager*, *Untergebener*). Ebensowenig fanden sich dort die Ausdrücke, die im CISLEX in die provisorische Kategorie Halbweltberufe eingeordnet wurden (*Prostitutierte* etc.).

Eine weitere Sondergruppe innerhalb der Berufe waren noch die historischen Berufe. Das sind diejenigen Lexeme, die für die Beschreibung der modernen Berufswelt nicht mehr relevant sind. Für sie wurde eine gesonderte Klasse vorgesehen, da sie in modernen, nichtfiktionalen Texten weder in gleichen Kontexten vorkommen wie die übrigen Berufsbezeichnungen, noch eine ähnliche Relevanz für das Information Retrieval besitzen.

Menschen aufgrund von Tätigkeiten. Diese Klasse umfaßt in erster Linie die Ausdrücke, die herkömmlicherweise als Nomina agentis bezeichnet werden. Zu dieser Klasse gehören deverbale Nomina wie *Raucher* und *Trinker,* aber auch einige denominale (*Urlauber*) und nicht abgeleitete Nomina (*Dieb*). Kriterium für die Zugehörigkeit eines Lexems zu dieser Klasse ist die eindeutige semantische Zuordnung zu einem Verb oder einem Funktionsverbgefüge. Die modifizierenden Adverbiale in den entsprechenden Umschreibungen können dabei zur Unterklassifizierung dieser Gruppe herangezogen werden:

101 Ein Raucher ist jemand der (regelmäßig) raucht

102 Ein Trinker ist jemand der (zuviel) trinkt

103 Ein Urlauber ist jemand der (gerade) Urlaub macht

104 Ein Fahrer ist jemand der ein Fahrzeug steuert

[33]Die Klassifikation enthält ca. 350 Klassen, darunter so detaillierte wie *Krankenversicherungsfachleute.*

Wie das letzte Beispiel exemplarisch zeigt, gehören zahlreiche Ausdrücke aus dieser semantischen Klasse zusätzlich zu den Berufen. Als Sondergruppe innerhalb der Nomina agentis sind außerdem diejenigen herauszugreifen, die habituelle Tätigkeiten bezeichnen oder bezeichnen können, wie *Raucher*, *Trinker* und *Schnarcher*.

In beiden Gruppen finden sich ambige Lexeme. So ist *Fahrer* als Beruf oder als bloße Tätigkeitsbeschreibung zu deuten. Auch ein Wort wie *Trinker*, das als einfaches Nomen weitgehend mit *Alkoholiker* synonym ist, kann zumindest in Komposita wie *Biertrinker*, *Wassertrinker* eine weitere, nicht zwingend habituelle Bedeutung annehmen.

Ein weiteres Identifikationskriterium für die habituelle Lesart ist, daß ein Nomen nach der Kopula ohne Artikel verwendet werden kann. In diese Konstruktion sind Lexeme, bei denen eine habituelle Lesart schwierig zu erzielen ist, nicht einsetzbar. Zumindest wirkt die Konstruktion dann stark markiert, wie das zweite der folgenden Beispiele zeigt:

105 Otto ist Raucher

106 ? Otto ist Antragsteller

Da diese Mehrdeutigkeiten nicht ohne weiteres aus der Verbbedeutung ableitbar sind, müssen sie kodiert werden. Lexeme mit habitueller Lesart werden im CISLEX der Klasse für Menschen aufgrund habitueller Tätigkeiten zugeordnet. Die resultierende Subklassen sind somit:

Menschen aufgrund von Tätigkeiten: *Lacher*, *Redner*

Menschen aufgrund habitueller Tätigkeiten: *Trinker*, *Schnarcher*

Menschen aufgrund von Eigenschaften. Wichtigstes Kriterium für die Nomina dieser Klasse ist die eindeutige Zuordnung zu einem Adjektiv, einem lexikalisierten Partizip oder einer Prädikatsgruppe der Form *X hat A*.

107 Ein Narr ist verrückt/ närrisch

108 Ein Albino hat eine weiße Haut

109 Ein Zwerg ist klein

Teilweise ist es problematisch, diese Gruppe von Lexemen von den Tätigkeitsbezeichnungen abzugrenzen. So läßt sich etwa das Lexem *Spinner* umschreiben mit:

110 Ein Spinner ist jemand der spinnt

111 Ein Spinner ist verrückt

Hier wird bei der Einordnung weitgehend Rücksicht genommen auf die synonymische Struktur des Wortschatzes; sind *Narr* und *Verrückter* den Eigenschaftsnomina zugeordnet, erfolgt diese Zuordnung auch für das Lemma *Spinner* .

In diese Gruppe gehört eine große Menge von Schimpfwörtern mit dem Merkmal ¡pejorativ¿ (z. B. *Narr*, *Egoist*). Die meisten von ihnen nehmen auf negative psychische Eigenschaften bezug.

Diese semantische Klasse läßt sich aufgrund des Eigenschaftstyps noch weiter unterteilen. Wichtigste Untergruppen sind die Bezeichnungen aufgrund von psychischen und physischen Eigenschaften, sowie solche aufgrund stereotyper Verhaltenseigenschaften[34]. Letztere befinden sich dabei in einem Übergangsbereich zwischen Menschenbezeichnungen aufgrund von Tätigkeiten und denen aufgrund von Eigenschaften:

[34]Eine Kodierung der entsprechenden Adjektive müßte mit der Klassifizierung der Nomina dieser Klasse abgestimmt werden.

Mensch_phys_eig.: *Zwerg*, *Riese*
Mensch_psych_eig.: *Narr*, *Genie*
Mensch_verhaltenseig.: *Don Juan*, *Workoholic*

Anthropomorphe. Dies ist eine klar abgrenzbare Gruppe, welche im Wortschatz der einfachen Nomina ca. 230 Lexeme umfaßt, die mythische Erscheinungen, Sagen- und Märchenwesen bezeichnen, deren reale Existenz zumindest hinterfragbar ist, und die damit in Kontexten wie den folgenden auftauchen können:

112 Es gibt keine Götter

113 Trolle existieren nicht

Den Menschenbezeichnungen wurden sie zugeordnet, weil alle diese Ausdrücke als Argumente von Prädikaten auftreten können, die nur Menschenbezeichnungen selegierten, wie *X lacht*, *X weint*, *X spricht* etc. Die Klasse der anthropomorphen Sagenwesen ist abzugrenzen von der entsprechenden Unterklasse bei den Tieren, zu der beispielsweise das Lexem *Drache* gehört, die also Bezeichnungen für Wesen umfaßt, die keine oder kaum menschenähnliche Züge tragen.

Eine weitere Unterteilung der semantischen Klasse der Anthropomorphen wurde bisher nicht vorgenommen, obwohl sie so unterschiedliche Lexeme enthält wie *Gott* und *Zombie* [35]. Eindeutig männliche Wesen bezeichnende Ausdrücke wurden mit dem Relator männlich (z. B. *Nöck*), Ausdrücke für eindeutig weibliche Geschöpfe mit der entsprechenden Auszeichnung weiblich (z. B. *Nymphe*) versehen.

Diese Gruppe von Lexemen ist relativ uneinheitlich und als eine Sammelklasse zu verstehen.

Die folgende Tabelle faßt die Klassifizierung der Menschenbezeichnungen noch einmal übersichtlich zusammen [36].

3.6.2 Fahrzeuge

Die Klasse Fahrzeuge, als Unterklasse der Artefakte klassifiziert, umfaßt ca. 300 Lexeme unter den einfachen Nomina. Berücksichtigt wurde bei ihrer Klassifizierung in erster Linie die Hauptfunktion von Fahrzeugen als Transportmittel, die sich darin zeigt, daß Fahrzeugbezeichnungen vor allem in Kontexten vorkommen, die den Gedanken ausdrücken, daß etwas oder jemand mit Hilfe eines Fahrzeugs den Ort verändert. Folgende Dimensionen erwiesen sich bei der Unterteilung als wichtig:[37]

Verkehrsweg

Der Typ von Verkehrsweg, auf dem sich ein Fahrzeug fortbewegt, ist für die Wahl des Fortbewegungsverbs ausschlaggebend und spielt auch für zahlreiche andere Konstruktionen eine Rolle, in denen Fahrzeugbezeichnungen üblicherweise verwendet werden. In den folgenden Belegen wird dies deutlich:

114 Karl **fährt** morgen mit der **Bahn**/dem **Schiff** nach Hamburg

[35] In der Frage der tatsächlichen Existenz bzw. Nichtexistenz der Denotate der in diese Klasse aufgenommenen Nomina möchte sich der Autor vorsichtshalber nicht festlegen.

[36] Klammern um ein Lexem in der Hyperonymspalte deuten an, daß das entsprechende Nomen nur bedingt als Hyperonym für die Lexeme der Klasse zu werten ist, weil es entweder nicht die ganze Klasse abdeckt oder nur unter bestimmten Aspekten als übergeordneter Begriff zu werten ist.

[37] Die Klassifikation erfolgt teilweise in Anlehnung an G. Gross' Einteilung der entsprechenden Lexeme im Französischen (vgl. Gross 1994: 22ff). Auch in der Einteilung der Fahrzeuge in WordNet haben die beschriebenen Klassen teilweise Entsprechungen.

Klasse	**Hyperonym**	**Klassenmerkm.**	**Beispiel**
Verwandtschaft	*Verwandte*	+na, -ur	*Mutter, Onkel*
Verwandtschaftsähnliche	-	+sk, +ur	*Freund, Geliebte*
Altersstufen	-	+sk, +ur	*Baby, Greis*
Herkunft	-	+sk, +ur	*Europäer, Moslem*
Menschen in Ausbildung	*Auszubildende*	+sa, +ur	*Schüler, Student*
Menschen im Ruhestand	*(Rentner)*	+sa, +ur	*Pensionär, Rentner*
Beruf	*(Beruf)*	+sa, +ur	*Schaffner*
Kirchenämter	*(Seelsorger)*	+sa, +ur	*Priester*
Pol. Ämter	*Politiker*	+sa, +ur	*Minister*
Ausbildung	*(Akademiker)*	+fu, +ur	*Germanist*
Berufstätigkeit	-	+fu, +ur	*Anwalt*
Ausbildungsberuf	-	+fu, +ur	*Anwalt*
Heilberufe	*(Arzt)*	+sa, +ur	*Arzt, Pfleger*
Industrieberufe	*Arbeiter*	+sa, +ur	*Schmelzer*
Handwerksberufe	*Handwerker*	+sa, +ur	*Tischler*
Polizeiberufe	*Polizist*	+sa, +ur	*Flic, Carabiniere*
Soldatenberufe	*Soldat*	+sa, +ur	*Artillerist*
Freizeittätigkeit	-	+fu, +ur	*Violinist*
Tätigkeitsbez.	-	+sk, +ur	*Raucher, Fahrer*
habituelle Tätigkeiten		+sk, +ur	*Trinker*
Eigenschaftsbez.	-	+sk, +ur	*Zwerg, Genie*
Psychische Eigenschaft	-	+sk, +ur	*Genie*
Physische Eigenschaft	-	+sk, +ur	*Zwerg*
Verhaltenseigenschaft	-	+sk, +ur	*Held, Erotomane*
Anthropomorphe	-	+sk	*Nymphe, Gott*

Tabelle 3.4: Unterklassen der Menschenbezeichnungen

115 Karl **setzt** mit dem **Schiff** nach Oslo **über**

116 Das **U-Boot tauchte ab**

117 Karl **fliegt** morgen mit dem **Jet** nach Moskau

118 Karl **geht an Bord** des **Dampfers/Flugzeugs**

119 Das **Schiff sank**

120 Der **Zug entgleiste**

Aufgrund von Kontexten dieser Art wurden die folgenden Klassen aufgestellt:

Luft-, Wasser-, Landfahrzeuge, wobei letztere noch die spezifischen Unterklassen Schienenfahrzeuge und Schlitten beinhalten. Eine relativ kleine Sondergruppe innerhalb der Wasserfahrzeuge konstituieren die Unterwasserfahrzeuge. Die Klassenbezeichner lassen auch auf die Hyperonyme der entsprechenden Klassen schließen.

Die dieser Dimension folgenden Klassen sind zwar prototypenzentriert, weisen aber keine unscharfen Ränder auf.

Benutzungmodalitäten

Die bezüglich dieser Beschreibungsdimension aufgestellten Klassen sind funktionale Klassen.

Individualverkehrsmittel (wie *Auto* oder *Katamaran*) bilden unter diesem Aspekt die größte Gruppe. Sie sind weitgehend definierbar als solche Verkehrsmittel, welche die Kontexte ausschließen, die für die folgenden beiden Klassen genannt werden.

Öffentliche Verkehrsmittel sind an Fahrpläne gebunden und tauchen somit häufig in Verbindung mit Zeitspezifikationen auf:

121 Der Zug um fünf.

122 Das Flugzeug nach Moskau

123 Karl hat das Schiff verpaßt

Mietfahrzeuge (z. B. *Taxi*) sind keine öffentlichen Verkehrsmittel im vorigen Sinn. Signifikante Kontexte für diese Unterklasse von Fahrzeugen sind:

124 Maria ruft/bestellt ein Taxi

125 Jean bestellt das Taxi ab

126 Maria bezahlt das Taxi

Transportgut

Gütertransportfahrzeuge dienen nicht dem Transport von Menschen. Zu dieser semantischen Klasse gehören Lexeme wie *Laster* und *Frachter*. Die so kodierten Nomina treten nicht typischerweise in Kontexten auf, die den Transport von Menschen ausdrücken, im Gegensatz zu der Mehrzahl der Ausdrücke für Fahrzeuge, deren Denotate in erster Linie dem Transport von Menschen dienen:

127 ?Maria verreist mit dem Laster

128 Jean verreist mit dem Bus

Hingegen treten Gütertransportfahrzeuge typischerweise in Kontexten auf, welche die Beförderung von Waren betreffen:

129 Dieser Laster transportiert Atommüll

130 Der Güterzug wird entladen

Viele Simplizia zur Bezeichnung von Fahrzeugen sind bezüglich der Eigenschaft Gütertransport versus Menschentransport unterspezifiziert, wie man an den folgenden Determinativkomposita erkennen kann, wo jeweils das Erstglied diese Eigenschaft ausdrückt:

131 Zug — Güterzug — Personenzug

132 Wagen — Lastwagen — Personenwagen

Antriebsart

Sie spielt für die Auswahl zahlreicher Kontexte eine Rolle. Teilweise ist die Antriebsart sogar für die Selektion des Fortbewegungsverbs ausschlaggebend, wie im ersten der folgenden Beispiele deutlich wird. Im Falle von motorgetriebenen Fahrzeugen kann der Ausdruck für das Fahrzeug als Ganzes in Kontexten auftreten, die sich eigentlich auf die Antriebsmaschine beziehen[38], wie das letzte Beispiel zeigt.

133 Peter und Maria segeln mit dem Zweimaster nach Ribe

134 Johann spannte die Kalesche an

135 Otto ließ den Wagen an

136 Maria würgte das Auto ab

Diese Überlegungen führten zu den Klassen Motorfahrzeuge, Pferdefahrzeuge, Segelboote und durch Menschenkraft angetriebenen Fahrzeuge.

Zweirädrige Fahrzeuge

Eine Sonderklasse innerhalb der Landfahrzeuge bilden noch die zweirädrigen Fahrzeuge (*Mofa*, *Rad*), die häufig in Kontexten mit Verben wie *umkippen*, *aufstellen* u.ä. vorkommen.

Fahrzeugteile

Einige Lexeme bezeichnen keine vollständigen Fahrzeuge, sondern weitgehend selbständige Teilfahrzeuge. Hierzu gehören Ausdrücke wie *Traktor, Lok* und *Schlepper* auf der einen und *Waggon* und *Hänger* auf der anderen Seite. Diese bilden die semantischen Klassen der Zugmaschinen und der Fahrzeuganhänger.

Hier nun einige Beispielklassifikationen und ihre Implikationen für exemplarische Aufgabenstellungen aus der maschinellen Übersetzung vom Französischen ins Deutsche:

Helikopter wird klassifiziert als Luftfahrzeug und Motorfahrzeug

Auto wird klassifiziert als Landfahrzeug auf Rädern und Motorfahrzeug. Bezüglich der Funktion ist dieses Lexem nicht positiv für die Klasse Öffentliches Verkehrsmittel spezifiziert.

Zug[39] wird klassifiziert als Schienenfahrzeug und öffentliches Verkehrsmittel.

Die Relevanz der vorgenommenen Klassifizierung zeigt sich bei der Übersetzung aus dem Französischen ins Deutsche. Man betrachte folgende Übersetzungsbeispiele unter dem Aspekt der Auswahl des Hauptverbs des jeweiligen Satzes in der Zielsprache Deutsch, gesteuert von der semantischen Klasse des Arguments:

[38] Vgl. auch den Ausdruck *Maschine* für ein Motorrad.

[39] Das Lexem ist selbstverständlich polysem und nicht ausschließlich als Fahrzeug kodiert.

Klasse	**Hyperonym**	**Klassenmerkm.**	**Beispiel**
Luftfahrzeuge	*Luftfahrzeug*	+sk, -ur	*Jet, Ballon*
Wasserfahrzeuge	*Schiff*	+sk, -ur	*Schiff, Barke*
Unterwasserfahrzeuge	*U-Boot*	+sk, -ur	*Bathyscaphe*
Landfahrzeuge Auf Mehr als Zwei Rädern	*(Auto)*	+sk, +ur	*Auto, Schlitten*
zweirädrige Landfahrzeuge	*(Räder)*	+de	*Rad, Mofa*
Schienenfahrzeuge	*Schienenfahrzeug*	+de, -ur	*Zug, Tram*
Schlitten	*Schlitten*	+sk, +pz, -ur	*Schlitten, Ackja*
Öffentliche Verkehrsmittel	-	+fk, +pz, +ur	*Tram, Bus*
Mietfahrzeuge	-	+fk, +pz, +ur	*Taxi*
Gütertransportfahrzeuge	*Lastfahrzeug*	+fk	*Laster,*
Motorfahrzeuge	*Motorfahrzeug*	+de	*Auto, Jet*
Durch Menschenkraft Getriebene Fahrzeuge	*Pedalfahrzeug*	+de	*Rad*
Pferdefahrzeuge	*(Kutsche, Schlitten)*	+de	*Kutsche*
Segelboote	*Segelboot*	+de	*Katamaran*
Fahrzeuganhänger	*Hänger*	+fu	*Hänger, Tender*
Zugmaschinen	*Zugmaschine*	+fu	*Lok, Traktor*

Tabelle 3.5: Unterklassen der Fahrzeuge

137 Jean va a Marseille en hélicoptère

137a *Jean fährt mit dem Helikopter nach Marseille

137b Jean fliegt mit dem Helikopter nach Marseille

138 Marie a emprunté la voiture de Jean

138a Marie hat das Auto von Jean ausgeliehen

139 Marie a emprunté le train

139b Marie hat den Zug genommen

Im Beispiel (137) wird die Selektion des Verbs im Deutschen gesteuert vom Typ des Fortbewegungsmittels. Das Verb im Französischen ist diesbezüglich unterspezifiziert — die Semantik von *aller* bietet keinen Anhaltspunkt für die Auswahl des richtigen Verbs im Deutschen. Diese Unterspezifikation des Fortbewegungsverbs ist im Deutschen nicht direkt nachzubilden[40], es hat im Generierungsabschnitt des Übersetzungsprozesses daher auf Basis der semantischen Klasse des Komplements der instrumentalen Präpositionalphrase eine Auswahl des Vollverbs zu erfolgen — im obigen Beispiel des Verbs *fliegen* im Zusammenhang mit *Hubschrauber*.

An den Beispielen (138) und (139) zeigt sich die Auswahl des richtigen Ziellexems im Deutschen für ein polysemes französisches Verb. Nur die Berücksichtigung der Distinktion Öffentliches Verkehrsmittel vs. Individualverkehrsmittel erlaubt die Selektion des Ziellexems *ausleihen* oder *nehmen*.

[40]Die Unterspezifikation des Verbs im Französischen bezüglich des verwendeten Verkehrsmittels ließe sich mit der Semantik des Verbs *reisen* im Deutschen vergleichen; dieses hat aber wieder andere semantische Implikationen (in diesem Fall eine größere Entfernung von Ursprungs- und Zielort, bzw. die Reiselänge).

Die dargestellte Klassifikation der Verkehrsmittel zeigt exemplarisch, daß eine taxonomische Einteilung der Wortschatzes und eine in erster Linie durch typische Kontexte gestützte Klassifikation von Lexemen keineswegs zu unvereinbaren Ergebnissen führen müssen. Aufgrund beider Typen von Kriterien war die Unterteilung der Fahrzeuge relativ unproblematisch vorzunehmen. Auch die Relevanz der erreichten Klassifizierung für die maschinelle Übersetzung ließ sich ohne weiteres an einigen Beispielen demonstrieren[41].

Die im folgenden beschriebenen Einteilungsversuche für die Klasse der Formen werden den entgegengesetzten Fall aufzeigen: es handelt sich hier um eine semantische Teilgruppe der Nomina, die sowohl einer ontologisch basierten als auch einer auf den sprachlichen Kontext gestützten Einteilung erhebliche Widerstände entgegensetzt.

3.6.3 Formen

Wie bereits bei der Beschreibung dieser Klasse in der Auflistung der Grobklassen erwähnt, sind die Formen eine recht uneinheitliche Klasse von Lexemen. In sie wurden bei der ersten Kodierung all jene Lexeme eingeordnet, die Objekte oder Objektteile primär aufgrund des Kriteriums der äußeren Form bezeichnen, oder die direkt Objektformen oder Abstraktionen von solchen als Denotate haben. Es fallen unter diese — zugegebenermaßen etwas vage — Umschreibung so unterschiedliche Lexeme wie *Dreieck, Pyramide, Delle* und*Kurve.*

Die Formen genannte Klasse umfaßt ca. 400 Lemmata in der Datei der einfachen Nomina. Im Gegensatz zu den Klassen der Menschen und Fahrzeuge, in denen sich rasch einige große Subklassen herauskristallisierten, besteht die Klasse der Formen aus vielen, sich überlagernden, kleinen Subklassen, und enthält einen hohen Prozentsatz an idiosynkratischen Ausdrücken.

Wie man im folgenden sehen wird, bleibt nach der ersten Klassifikation dieser semantischen Klasse eine große Restklasse von Nomina zurück, deren sprachliches Verhalten nur noch in sehr eingeschränktem Maße parallele Eigenschaften zeigt, und die damit weitgehend individuell beschrieben werden müssen. In einem System zur automatischen Übersetzung würde die Kodierung der Kombinatorik solcher Nomina im Verhältnis zu ihrer Häufigkeit relativ aufwendig sein. Eine eins-zu-eins Zuordnung von Lexemen verschiedener Sprachen ist aber gerade in diesem Wortschatzbereich nicht möglich. Dies wird an einigen Beispielen der Übersetzung aus dem Französischen im Verlauf der Erörterungen gezeigt werden.

Für das Information Retrieval spielt diese Klasse nur sehr begrenzt eine Rolle, da in ihr weder eine größere Menge von Fachtermini enthalten ist, noch anderweitig eine klare thematische Zuordnung der Unterklassen vornehmbar scheint. Eine Ausnahme machen diesbezüglich die geometrischen Formen, die sich primär dem Fachbereich Mathematik zuordnen lassen.

Die nachfolgend geschilderten semantischen Gliederungsversuche zu dieser Klasse zeigen wiederum, daß die drei Kriterien für die Gliederung des Nominalwortschatzes, also 1. hyponymische Struktur, 2. thematische Einteilung und 3. typische sprachliche Kontexte, sehr stark korreliert sind. Während bei den bisher dargestellten Teilgliederungen eine Aufteilung anhand aller drei Kriterien relativ unproblematisch war und sich aus einer hyponymisch gestützten Ontologie unter Zuhilfenahme des Kontextes eine brauchbare Gliederung entwickeln ließ, bieten sich für einen Großteil der Ausdrücke in den Formen bezüglich aller drei Kriterien weit weniger Anhaltspunkte zu einer konsistenten Einteilung.

Als wichtigste Dimensionen und Parameter zur Beschreibung von Formen stellten sich heraus:

[41] Die Möglichkeit der Heranziehung der kodierten semantischen Klassen für das Information Retrieval liegt auf der Hand und muß wohl nicht demonstriert werden.

Selbständigkeit als Objekt

Es handelt sich hier um die Möglichkeit des Denotats eines Ausdrucks, als unabhängiges Objekt aufzutreten versus Teil oder Eigenschaft eines Objekts zu sein. Eine solche Einteilung der Denotate ist auf sprachlicher Seite für zahlreiche Kontexte relevant; insbesondere können Formbezeichner, die selbständige Objekte bezeichnen, in allen sprachlichen Umgebungen auftreten, die auch für Konkreta (Objekte) typisch sind. Bezeichner von gebundenen Formen treten in solchen Kontexten nicht auf:

140 In der Ecke liegt ein Quader aus Holz

141 ? In der Ecke liegt eine Rille aus Holz

Dieses Kriterium ermöglicht die Abgrenzung der großen Menge von Nomina in der Klasse Formen, die geometrische Formen bezeichnen. Diese Nomina können auch verwendet werden, um auf ganze Objekte mit den entsprechenden Formeigenschaften zu referieren. Hierher gehören die dreidimensionalen geometrischen Formen wie *Pyramide*, *Quader* und die zweidimensionalen[42] Formen wie etwa *Quadrat* und *Trapez*. Abgesehen von speziellen, lexikalisierten Nebenbedeutungen einiger von ihnen (*Pyramide* etwa als Gebäude) verhalten sich diese Nomina dieser Gruppen kontextuell relativ einheitlich.

Aufgrund dieses Kriteriums sind die Klassen Zweidimensionale Geometrische Formen und Dreidimensionale Geometrische Formen aufzustellen. Diese Klassen könnten aufgrund der Reihenbildung (*Dreieck*, *Viereck*, ... *N-Eck*) weiter gegliedert werden.

Eine große Anzahl weiterer Lexeme in der semantischen Klasse der Formen gehört zu den Ausdrücken, die selbständige Objekte bezeichnen können. Diese Restgruppe kann aufgrund einiger der im folgenden beschriebenen Kriterien weiter untergliedert werden; andere der genannten Kriterien gelten auch oder ausschließlich für solche Ausdrücke, die nichtselbständige Objekte bezeichnen.

Material

Bei den selbständigen Formen kann ein Lexem mit einer Materialbezeichnung assoziiert sein. So besteht ein *Spritzer* oder ein *Tropfen* aus einer Flüssigkeit und ist inkompatibel mit Ausdrücken anderer Klassen:

142 Ein Spritzer Wasser/Alkohol

143 ?Ein Spritzer Sand

Eine *Garbe* ist ein Bündel aus Heu oder Stroh. Somit ist es möglich, das Kompositum *Strohgarbe* zu bilden, nicht aber **Zweiggarbe*, (wohl aber *Zweigbündel*). Nicht immer ist diese Implikation jedoch so klar. So hat Klecks als Kopf von Komposita eine starke Präferenz für *Tinte/Farbe* als Erstglied, es ist jedoch auch möglich, Komposita wie *Senfklecks*, *Leimklecks* zu bilden.[43]

Typen von Objekten, die Träger einer Teilform sein können

Für die vorher genannten Bezeichner von selbständigen Formen gilt, daß sie fast auf alle Objekte anwendet werden können, welche die entsprechende Formeigenschaft besitzen. Doch für die Subgruppe der unselbständigen Formen ist der Parameter 'betroffenes Objekt' wesentlich für die Bedeutung. Um ihn abzuprüfen, kann man den Testrahmen *A hat ein Y* verwenden, wobei die Möglichkeiten der Einsetzungen von A untersucht werden:

[42] Wird etwa mit *Quadrat* ein Objekt bezeichnet, ist natürlich ein flaches **drei**dimensionales Objekt gemeint, bei dessen Benennung die dritte Dimension vernachlässigt wird.

[43] Die selegierte semantische Klasse ist hier idiosynkratisch für das Nomen *Klecks*, d. h. es ist nicht möglich, eine auch für andere Selektionsregularitäten relevante semantische Klasse von Lexemen zu bestimmen, die als Erstglied zu *Klecks* treten kann.

144 Das Auto hat ein Beule

145 Karl hat einen Buckel

146 Das Auto hat einen Kratzer

Ist der Typ von Objekt nicht adäquat, kommt es zu Inkompatibilität:

147 ? Der Computer hat einen Buckel

148 ? Das Papier hat einen Kratzer

Sehr eingeschränkt möglich ist eine Verwendung in diesem Kontext für Ausdrücke für Landschaftsformen; üblich ist hier die Verwendung eines Genitivattributs:

149 Die Hügel der Provence

150 Die Täler Norwegens

Die Landschaftsformen konstituieren also eindeutig ein Sondergruppe innerhalb der Formen.

Der Typ von Objekt, auf dem die Form auftritt, kann wesentlich für die Übersetzung einer Formbezeichnung sein. Dies zeigt sich etwa an der Übersetzung des französischen Lexems *rayure* ins Deutsche, das im Zusammenhang mit Kleiderstoffen mit *Streifen*, bei Möbelstücken u. ä. aber mit *Kratzer* oder *Schramme* übersetzt werden muß.

Formtypen

Die wichtigste Gruppe von Parametern für die unselbständigen Formen sind die Formtypen selbst.

Ausdrücke für Formen können Einbuchtungen, Ausbuchtungen, komplexe oder flache Strukturen bezeichnen. Diese Eigenschaften, die für eine große Zahl von Formen distinktiv sind, lassen sich durch eine Reihe von verbalen und adjektivischen Kontexten verifizieren:

151 Karl füllt die Senke auf

152 Die Ritze füllt sich mit Wasser

153 Karl trägt die Erhebung ab

154 Eine hohe Beule

155 Eine tiefe Scharte

156 ?Der Streifen füllt sich mit Wasser

157 ?Ein hoher Streifen

Die Lexeme *Senke*, *Scharte* und *Ritze* bezeichnen also Einbuchtungen, *Erhebung* und *Beule* Ausbuchtungen, während das Lexem *Streifen* keiner dieser beiden Klassen zugeordnet werden kann.

Einige Formen beinhalten in ihrer Semantik den Aspekt der Längenausdehnung, andere nicht:

158 Ein lange Rille

159 ?Eine lange Delle

160 Eine längliche Delle

Weitere Formtypen sind durch den Parameter Größe bestimmt: bestimmte Formbezeichner haben eine Implikation bezüglich der Ausdehnung. Distinktionen bezüglich der implizierten Größe, auf den gleichen Typ von Objekten angewandt, beinhaltet die Bedeutung von einigen, z. T. nur in bestimmten Kontexten konkurrierenden Lexemen wie *Mulde* vs. *Senke*, *Rille* vs. *Scharte*, *Hügel* vs. *Berg*, *Delle* vs. *Beule*. Dieser Parameter kann nur innerhalb einer Gruppe als Rangfolge beschrieben werden; unter diesem Aspekt ist eine Klassenbildung nicht sinnvoll.

Wertung/Entstehung

Einige Formbezeichner, etwa *Hügel*, *Rille* und *Erhebung,* bezeichnen Formen, ohne daß die Bezeichnung selbst eine Wertung enthält. Anders ist dies bei *Beule*, *Kratzer* oder *Riß*. Diese Lexeme beinhalten, daß die bezeichnete Form nicht den Normalfall darstellt, sondern die Veränderung eines Normalzustandes, d. h. sie implizieren, daß die bezeichnete Form ein Defekt ist[44]. Bestimmte Ausdrücke implizieren auch die Art ihrer Entstehung (z. B. *Kratzer* durch kratzen oder ritzen).

Öffnungen

Eine relativ umfangreiche Untergruppe der Formen sind die Lexeme, die Öffnungen bezeichnen. Wesentlicher Kontext für die Klasse Öffnungen ist die Präposition *durch*.

161 Maria schlüpfte durch die Öffnung/Türe/den Durchgang

Die Nomina dieser Klasse können häufig auch für das eine Öffnung verschließende Objekt stehen. Bei einigen Lexemen dieses Typs, wie etwa *Tür,* scheint das Verhältnis umgekehrt zu sein: Das die Öffnung verschließende Konkretum ist semantisch primär, kann aber auch für die Öffnung selbst stehen. Wie ein Vergleich der folgenden Beispiele zeigt, ist bei einigen Lexemen der Klasse diese Doppeldeutigkeit allerdings nicht gegeben.

162 Das Fenster/die Türe ist aus Holz

163 ? Der Durchgang ist aus Holz

164 ?? Die Öffnung ist aus Holz

Die Lexeme mit der Eigenschaft, neben der Öffnung selbst auch das dazugehörige Objekt zu bezeichnen, müssen doppelt klassifiziert werden, auf der einen Seite als Subklasse der Formen, auf der anderen als Subklasse der Konkreta.

Maßbezeichnungen

Einige Nomina aus der Klasse der Formen sind als Maßbezeichnungen (quantifizierende Nomina im weitestem Sinne[45]) verwendbar. Dieser Parameter scheint zunächst mit dem vorher genannten Parameter Material zusammenzuhängen, erweist sich aber letztlich, wie Nr. (166) und (167) der folgenden Beispiele zeigen, als unabhängig von ihm, da Lexeme in dieser Gruppe vorkommen, die in Verbindung mit Stoffbezeichnungen verschiedenster Art auftreten können. So ist das Denotat von *Haufen* bezüglich des Materials, aus dem ein solcher besteht, kaum eingeschränkt. Ein Testrahmen für diese Eigenschaft ist die enge Apposition mit einer Stoffbezeichnung oder mit einem zählbaren Konkretum im Plural:

[44] Häufig können solche Lexeme formal und semantisch in Verbindung mit Präfixverben auf *zer-* und *ver-* gebracht werden (z. B. *zerkratzen*, *verbeulen*)

[45] Zu den Maßbezeichnungen gehören auch zahlreiche Lexeme, die keinen Formaspekt beinhalten (so *Liter*, *Kilo*). Zu Maßbezeichnungen im allgemeinen vgl. Oesterle (1994, insbesondere 11ff).

165 Ein Spritzer Eau de Cologne

166 Ein Haufen Bücher

167 Ein Haufen Stroh

168 ? Eine Ritze Wasser

Wie an dem Beispiel mit dem Lexem *Ritze* deutlich wird, tritt diese Konstruktion nicht auf im Zusammenhang mit Formbezeichnern, die nicht zu den Maßbezeichnungen gehören.

Abstraktheit

Einige der behandelten Ausdrücke sind Abstraktionen von einer Teilgruppe der Formen aufgrund eines der bisher genannten Kriterien oder Metabezeichnungen, die auf das Kriterium selbst Bezug nehmen; so ist eine *Deformation* eine Form, die als Defekt gesehen wird. *Proportion* bezieht sich auf das Kriterium der äußeren Form. Die Lexeme dieser Gruppe verhalten sich distributionell nicht einheitlich und ermöglichen damit nicht die Aufstellung einer Selektionsklasse. Auch thematische Implikationen besitzen sie nicht. Sie gehören somit zu den Lexemen, deren Semantik individuell beschreiben werden muß.

Einteilung der Formen in Klassen

Die genannten Beschreibungsparameter operieren auf ganz unterschiedlichen Eigenschaften. Ihre Gewichtung für eine Einteilung fällt für einige Lexeme äußerst schwer, besonders, da viele Ausdrücke in der Klasse FORMEN in hohem Grade polysem und die verschiedenen Parameter keineswegs unabhängig voneinander sind. So ist es etwa nicht ausreichend, die Bedeutung von *Beule* als nach außen gehende Verformung zu fassen, da diese Beschreibung z. B. in den folgenden beiden Belegen nur im ersten Fall greift:

169 Maria hat eine Beule

170 Marias Auto hat eine Beule.

Eine Einteilung durch Zuteilung der verschiedenen Eigenschaften ist in diesem Fall schwierig, da diese sich gegenseitig bedingen: eine *Beule* ist etwa auf einem metallenen Gegenstand etwas anderes als eine *Beule* auf einem menschlichen Körper (zentraler Aspekt der Übertragung scheint hier die Entstehung als eine Art von Verletzung bzw. Beschädigung zu sein).

Die abgrenzbaren semantischen Klassen, die zu den Formen gehören, gingen schon aus der Diskussion der relevanten Beschreibungsdimensionen und Parameter teilweise hervor. Zunächst wurden bei der Einteilung dieses Wortschatzes die geometrischen Formen — wiederum unterteilt in ein- zwei- und dreidimensionale geometrische Formen (Beispiele sind *Linie*, *Quadrat*, *Kubus*) — und die Landschaftsformen als relativ einfach separat zu beschreibende Klassen ausgegrenzt. Als gesonderte Klasse wurden ferner die Öffnungen (s. o.) und einige andere definiert (s. die folgende Tabelle). Daraufhin wurden Lexeme, die auf Formtypen Bezug nehmen, in Gruppen zusammengefaßt. Einige Beispiele für Formtypenklassen sind *Scharte*, *Rille*, *Kerbe* vs. *Buckel*, *Wulst*, *Hubbel*, also Aus- vs. Einbuchtungen. *Wellung*, *Riffelung*, *Draperie* konstituieren eine Gruppe, die komplexe Formstrukturen beschreibt.

Nach der Einbeziehung der anderen Kriterien ergab sich eine Klassifikation der Formen, die sich dadurch auszeichnet, daß die verschiedenen Distinktionen den Wortschatz der Formen nach verschiedensten Kriterien in vergleichsweise kleine Gruppen aufteilen, und die Bündelung verschiedener Kriterien zu Klassen führt, die immer nur eine begrenzte Anzahl semantischer Ähnlichkeiten aufweisen. Keines der Klassifizierungskriterien allein ermöglicht es jedoch, das kontextuelle Verhalten der eingeordneten Lexeme in einer größeren Menge

von Belegen zu beschreiben, ganz im Gegensatz zu einigen anderen der in den vorigen Abschnitten beschriebenen Klassen der Menschenbezeichnungen und der Fahrzeuge, wo die Klassifizierung zu großen, relativ einheitlichen Gruppen führte.

Vorliegender Versuch einer Einteilung dieses Teilwortschatzes aufgrund sprachlicher wie ontologischer Kriterien zeigt die Grenzen einer semantischen Klassifizierung auf. Die vorgenommene Einteilung kann wohl einige Generalisierungen erfassen, erlaubt es aber nicht, eine größere Menge von kombinatorischen Regularitäten vorherzusagen. Für die Klasse der Formen liegen nicht — abgesehen von den Gruppen der geometrischen Formen und der Landschaftsformen — wie bei den anderen beschriebenen Einteilungsversuchen kanonische Klassifizierungskriterien für die zugrundeliegenden Kategorien vor, welche die Leitlinien für die semantische Einteilung in semantische Klassen vorgeben könnten. Ebensowenig kann man für diesen Teilwortschatz von einer durchgehenden hyponymischen oder meronymischen Struktur ausgehen. Dies erkennt man deutlich an den Problemen, auf die man stößt, wenn man die gängigen Testverfahren für Hyponymie anwendet:

171 ? Eine Beule ist eine Ausbuchtung

172 Hans hat eine Beule

173 ? Hans hat eine Ausbuchtung

Am ehesten läßt dieser Teilwortschatz sich noch aufgrund teilweiser Synonymie — d. h. Austauschbarkeit in spezifischen Kontexten — in Gruppen zusammenfassen. Das folgende Beispiel zeigt zunächst die (weitgehende) Synonymie der Lexeme *Beule* und *Delle* im Zusammenhang mit Gegenständen, dann aber auch, daß diese Synonymiebeziehung nur unter bestimmten Aspekten gilt.

174 Das Auto hat eine Beule

175 Hans hat eine Beule

176 Das Auto hat eine Delle

177 ? Hans hat eine Delle

Bei einer Gliederung aufgrund von teilweiser Synonymie ergibt sich dann das Problem der Zusammenfassung von Kontexten, in denen Bedeutungsgleichheit vorliegt, zu kohärenten Kontextgruppen. Dabei kommen wiederum alle oben für die Klassifizierung der Formen aufgelisteten Parameter ins Spiel.

Es bleibt zu fragen, ob für einen Großteil der Formen eine statistische Modellierung von kontextuell-semantischen Eigenschaften zumindest für die häufigeren Lexeme nicht einen effizienteren, und letztlich im Ergebnis besseren Ausgangspunkt für Anwendungen liefern kann. Das hieße etwa im Bereich der maschinellen Übersetzung, die Auswahl der richtigen Ausdrücke in der Zielsprache nach einer Vorauswahl durch das zweisprachige Lexikon aufgrund der statistischen Wahrscheinlichkeiten im Kontext vorzunehmen. Dies sollte allerdings für eine brauchbare Übersetzung auch heißen, nicht nur Wortformen oder Lemmata in der Umgebung des Ziellexems, sondern auch im Kontext auftretende semantische Klassen zu berücksichtigen.

Für die thematische Gliederung des Wortschatzes und damit für Anwendungen im Bereich des Information Retrieval spielen die meisten der genannten Distinktionen mit Ausnahme einiger weniger Klassen (Landschaftsformen, geometrische Formen, Frisuren) keine wesentliche Rolle.

Die Tabelle 3.6 der Unterklassen der Formen ist unter den formulierten Vorbehalten zu betrachten.

Klasse	**Hyperonym**	**Merkmale**	**Beispiel**
Ausbuchtungen	*Ausbuchtung*	+de, +ur	*Beule, Buckel*
Einbuchtungen	*Einbuchtung*	+de, +ur	*Beule, Rille*
Verzweigungen	*Verzweigung*	+de, +ur	*Gabelung*
Falten	*(Falte)*	+de, +ur	*Falte*
Ecken	*Eck*	+de, +ur	*Ecke*
Komplexe Formen	-	+de, +ur	*Riffelung*
Muster	*Muster*	+de, +ur	*Streifen*
Flecken	*Fleck*	+de, +ur	*Fleck, Klecks*
Öffnungen	*Öffnung*	+de, +ur	*Tür, Luke*
Frisuren	*Frisur*	+de, +ur	*Pony, Locke*
Landschaftsformen	—	+na	*Berg, Tal*
Geometrische Formen	—	+sk	*Quadrat, Kubus*
Abstrakte Formbezeichner	-	+sk	*Rundung, Form*

Tabelle 3.6: Unterklassen der Formen

3.6.4 Zusammenfassung

Wie die genannten Darstellungen der Kodierung der Menschenbezeichnungen, der Fahrzeuge und der Formen exemplarisch aufzeigen, lassen sich aus der Kodierung detaillierter semantischer Klassen folgende generelle Schlüsse ziehen:

- Der Nominalwortschatz des Deutschen — und sicher auch anderer Sprachen — weist keine einheitliche Struktur auf. Umfassenden Klassen von Lexemen mit gemeinsamen semantischen Eigenschaften stehen einzelne Lexeme oder minimale Lexemgruppen gegenüber, die sich in ihrer Kombinatorik und bezüglich der Zuordnung zu Überbegriffen extrem idiosynkratisch verhalten, und die sich einer weiteren Einordnung in eine hierarchische Strukturierung des Wortschatzes entziehen.
- Bei der Einordnung des Wortschatzes in Klassen von bedeutungsähnlichen Wörtern läßt sich kein monokriteriales Vorgehen definieren. Bedeutungsähnlichkeit ist eine komplexe Eigenschaft, die sich aus zahlreichen, überlagernden Teilkriterien ergibt. Es sollte darauf geachtet werden, die verschiedenen Kriterien für die Aufstellung der Klassen bestmöglichst auseinanderzuhalten, indem Typen von semantischen Klassen festgelegt werden.

Die Kriterien für die semantische Einteilung des Wortschatzes lassen sich zu zwei Hauptgruppen zusammenfassen:

a 'außersprachliche Kriterien'

  - (angenommener) Ähnlichkeiten der Referenten (ontologisierende Kriterien)
  - wissenschaftliche oder quasiwissenschaftliche Taxonomien

b 'sprachliche Kriterien'

  - hyponymische Struktur des Wortschatzes
  - von Lexemen geteilte typische Kontexte

Dabei ist die Hyponymiebeziehung bedingt durch die Weltsicht einer Sprachgemeinschaft. Die sprachliche Einteilung beeinflußt wiederum deren Einteilung der Welt. In den Fällen in denen eine klare Ontologie (wie im Falle der Verkehrsmittel) zugrundeliegt, ist die hyponymische Einteilung des Wortschatzes leicht zu ermitteln und die sprachlichen Kriterien für die Einteilung sind relativ einfach zu benennen. Dies gilt häufig auch umgekehrt: in den Fällen wo eine klare ontologische Einteilung fehlt, ist die hyponymische Einteilung der Lexeme ebenfalls problematisch und auch eine Gliederung des Wortschatzes aufgrund geteilter typischer Selektionskontexte kaum zu erreichen. Ein Sonderfall sind die Wortschatzbereiche, bei denen eine wissenschaftliche Taxonomie die größte Bedeutung für die Einteilung in semantische Klassen besitzt, wie vor allem im Bereich der Tierwelt; hier ist die Nennung typischer Kontexte zur Eingrenzung semantischer Klassen kaum möglich, wohingegen eine Strukturierung aufgrund der Hyponymierelation ohne weiteres zu erreichen ist.

Für die lexikographische Arbeit ergeben sich bezüglich der semantischen Unterteilung des Nominalwortschatzes die folgenden Konsequenzen:

- Selektionsklassen und taxonomische Klassifizierung laufen in vielen Fällen parallel. Dies gilt allerdings nicht für alle Wortschatzbereiche.

- Für die diskutierten Feinklassifizierungen ergaben sich aus sprachlichen Selektionskriterien und der hyponymischen Struktur des Wortschatzes Klassen, die auf Denotatsebene als ontologische Klassen verstanden werden und umgekehrt. Dies wurde besonders deutlich bei der dargestellten Unterteilung der Verkehrsmittel.

- Beim Kodierungsprozeß sollte für die aufgestellten Klassen ermittelt werden, wo der Schwerpunkt der Kriterien für die Einteilung liegt, da diese Metaeigenschaften die Relevanz von Klassen für bestimmte Typen von Anwendungen steuern. Dies geschah bei der beschriebenen Kodierung durch die Benennung der Eigenschaften ¡taxonomische Klasse¿, ¡Selektionsklasse¿ und ¡thematische Klasse¿.

- Ferner sollte der Typ von Struktur der semantischen Klassen explizit kodiert werden. Sowohl die Möglichkeit der Markierung von Lexemen, die zentrale Konzepte einer Kategorie denotieren, als auch die Möglichkeit zur Modellierung gradierbarer Klassenzugehörigkeit wurden berücksichtigt durch die Merkmale ¡prototypenzentriert¿ und ¡unscharfe Ränder¿.

3.7 Tiefe der Kodierung und Hierarchisierung

Bei der semantischen Klassifikation stellte sich als zusätzliche Aufgabe die Hierarchisierung der kodierten Klassen. Das praktische Vorgehen in diesem Punkt wurde schon weitgehend durch die oben beschriebenen Unterteilungen einiger Grobklassen demonstriert. Nicht erwähnt wurden dabei einige wichtige Unterscheidungen von Typen von Unterordnung.

Die Kodierung der semantischen Klassen wurde bisher nur so weit durchgeführt, wie sie für denkbare Anwendungen für nicht-fachsprachliche Texte zweckdienlich sein könnte. Dies impliziert auch, daß keine Fachlexika herangezogen wurden, um eine feinere Aufteilung der fachsprachlichen Lexik vorzunehmen. Eine solche Gliederung des Wortschatzes ist erst dann zweckdienlich, wenn die Domäne einer Anwendung feststeht. Für diesen Fall ermöglichen die existierenden semantischen Klassen eine rasche Vorauswahl solcher Lexemgruppen, die für die Domänenmodellierung relevant sind. Die ausschließlich fachspezifisch relevanten Lexeme sind bereits durch die Fachsprachenbezeichner (s. o.) markiert.

Während zunächst eine Strukturierung aufgrund einer einzigen Unterordnungsrelation vorgesehen war, ergab sich nach genauerer Betrachtung die Notwendigkeit, zwischen mindestens drei Typen von Unterordnung zu unterscheiden[46]:

- Strikte Unterordnung
- Bedingte Unterordnung
- Metonymische Unterordnung

Sie werden im Folgenden genauer charakterisiert.

3.7.1 Strikte Unterordnung

Im Falle der Hierarchisierung von taxonomischen Klassen handelt sich hierbei um eine unbedingte[47] begriffliche Unterordnung. Ihre Kodierung führt zu einer Taxonomie von Klassen und damit der darin enthaltenen Lexeme. Die Hyponymierelation in diesem Sinne wurde im Kapitel 2 bereits diskutiert. Typisches Beispiel für diesen Typ von Unterordnung ist die Unterteilung der Tiere in Säugetiere, Vögel etc. Der strikten Unterordnung zweier semantischen Klassen entspricht in diesem Fall die Hyponymiebeziehung zwischen den mit dem Kode =Semantische Klasse (s. u.) versehenen Lexemen der Klassen (etwa zwischen *Vogel* und *Tier*, die als Kode =Vogel bzw. =Tier erhalten). Eine strikte Unterordnung kann auch im Fall der Mehrfachunterordnung einer Klasse vorliegen. Ein Beispiel hierfür ist die Zuordnung von den Haushunden zu den Säugetieren (biologische Einordnung) einerseits und zu den Haustieren (funktionale Einordnung) andererseits.

Bei reinen Selektionsklassen, die als Einheiten nicht in der hyponymischen Struktur des Wortschatzes verankert sind, wird Unterordnung über geteilte typische Kontexte mit der Oberklasse geregelt. Eine Selektionsklasse ist einer anderen Selektionsklasse dann untergeordnet, wenn sie alle deren typischen Kontexte teilt und für die untergeordnete Klasse zusätzliche sprachliche Kontexte aufgezeigt werden können, die allen Lexemen dieser Klasse gemein sind.

Eine Selektionsklasse kann auch einer taxonomischen Klasse untergeordnet sein. Solche Selektionsklassen haben den Sinn, Klassen von Kohyponymen, die aufgrund der Hyponymiebeziehung — und damit aufgrund der begrifflichen Struktur des Wortschatzes — nicht weiter unterteilt werden können, aufgrund von typischen Kontexten, die nur für einige von den Lexemen in der übergeordneten Klasse gelten, weiter zu untergliedern. Für diese Unterordnung von Selektionsklassen unter taxonomische Klassen gilt, daß die Menge der Lexeme in der Selektionsklasse eine Untermenge der Lexeme in der taxonomischen Klasse sein muß.

3.7.2 Bedingte Unterordnung

Es handelt sich hierbei um einen Typ von Unterordnung, der für taxonomische Klassen phänomenologisch gewisse Aspekte der hyponymischen Unterordnung beibehält[48]. Bedingte

[46] Wichtige Anregungen hierzu verdanke ich Dietmar. Zaefferer (persönliches Gespräch).

[47] Diese Unterordnung ist meiner Ansicht nach nicht analytisch: vgl. den Abschnitt zum Beschreibungsgegenstand der Wortsemantik in Kapitel 2.

[48] Ich unterscheide diese Art der Unterordnung von der metonymischen Unterordnung wie im folgenden Abschnitt beschreiben, da sie mir a) wesentlich stärker lexikalisiert erscheint, und b) eine größere phänomenologische Ähnlichkeit mit der Hyponymiebeziehung im engeren Sinne besteht als bei eindeutig als metonymisch identifizierbarer Bedeutungsübertragung. Damit soll nicht ausgeschlossen werden, daß hier letztlich auch metonymische Prozesse zugrunde liegen.

Unterordnung soll nicht heißen, daß die Zuordnung zur Oberklasse nur für einen Teil der Lexeme der untergeordneten Klasse gilt. Bedingte Unterordnung liegt vielmehr vor bei der Zuordnung von Klassen von systematisch polysemen Lexemen zu Oberklassen. Ein Beispiel ist die Zuordnung von Ausbildungsinstitutionen (z. B. *Schule*, *Universität*) zu den Gebäuden[49]. Alle Lexeme der Klasse der Ausbildungsinstitutionen haben systematisch einen lokativen Bedeutungsaspekt, der den diskutierten herkömmlichen Hyponymietests standhält. So folgt in der folgenden Belegreihe die zweite Aussage logisch aus der ersten; zudem ist auch der*ist-ein*-Test für Hyponymie anwendbar.

178 Eine Schule wird erbaut/abgerissen

179 Ein Gebäude wird erbaut/abgerissen

180 Eine Schule ist ein Gebäude

Auch kann der Ausdruck *Schule* in allen für Gebäude typischen Kontexten auftreten. Im Unterschied zur echten Hyponymie ist die Lesart als Hyponym der übergeordneten Klasse jedoch nicht prinzipiell möglich. So ist *Schule* im folgenden Beispiel nicht als Hyponym von *Gebäude* zur interpretieren:

181 Die Schule wurde 1890 gegründet

Es handelt sich hier demnach nicht um eine begriffliche Unterordnung im gleichen Sinn, wie von der Klasse Vögel zur Klasse Tiere. Dem wird Rechnung getragen, indem die Beziehung in anderer Form kodiert wird. Im Gegensatz zur metonymischen Bedeutungsübertragung, die der folgenden Unterordnungsbeziehung zugrundeliegt, ist die Lesart der Lexeme, die zu einer bedingten Unterordnung führt, mit anderen nicht-metonymischen Lesarten kompatibel, wie der Zeugma-Test (vgl. Cruse 1986) im folgenden Beispiel zeigt:

182 Die Schule deren Gründung erst 3 Jahre zurückliegt ist heute abgebrannt

183 ??Die Schule deren Gründung erst 3 Jahre zurückliegt ist heute von einem Ausflug nicht mehr zurückgekehrt

Im Falle von reinen Selektionsklassen ist die bedingte Unterordnung so zu verstehen, daß die Lexeme der Klasse im Falle eines Typs von Lesart, der ihnen allen gemeinsam ist, alle typischen Kontexte der Oberklasse teilen.

3.7.3 Metonymische Unterordnung.

Alle Lexeme der genannten Klasse Ausbildunginstitutionen haben noch eine weitere Lesart als Gruppe von Menschen. Sie wird in den ersten beiden der folgenden Belege deutlich. Der dritte Beleg zeigt allerdings, daß es sich hier um eine Zuordnung handelt, die nicht mehr als Hyponymie gedeutet werden kann:

184 Die Schule machte einen Ausflug

185 Die ganze Schule war heute schlecht gelaunt

[49]Die systematischen lokativen Bedeutungsaspekte der Lexeme dieser Klasse, die durch die quasihyponymische Zuordnung zu den Gebäuden kodiert wurde, läßt sich auch daran erkennen, daß als Hyperonym für die Lexeme dieser Klasse auch das Wort *Ausbildungsstätten* existiert. Man vergleiche auch die Zuordnung des Lexems *school* in WordNet, das in der Lesart als Synonym zu *schoolhouse* auch den Gebäuden untergeordnet wird. In WordNet wird allerdings keine Unterscheidung zwischen verschiedenen Typen von Unterordnung gemacht.

186 ?Eine Schule ist eine Gruppe von Menschen

Auch ist in diesem Falle unter dem Aspekt vererbter Selektionskontexte nicht mehr die eindeutige Zuordnung zu den Menschenbezeichnungen gegeben. Das Lexem *Schule* verhält sich, wie die folgenden Beispiele demonstrieren, auch in der Lesart als Menschenbezeichnung signifikant anders als andere kollektive Menschenbezeichner:

187 ? Die Schule war heute schlecht gelaunt

188 Die Lehrerschaft war heute schlecht gelaunt

Die Unterordnungsbeziehung kommt in diesem Falle durch eine metonymische Bedeutungsübertragung zustande, die für die gesamte Klasse der Ausbildungsinstitutionen möglich ist. Ihre Kodierung als Typ von Unterordnungsbeziehung ist deshalb geboten, weil die Relation für Disambiguierungsaufgaben als Unterordnung gedeutet werden muß, um bestimmte Lesarten von Operatoren nicht auszuschließen. So werden bei der Untersuchung zu Selektionspräferenzen von Köpfen in Nominalkomposita in Kapitel 4 alle drei Typen von Unterordnungsrelation gleichermaßen berücksichtigt.

3.7.4 Thematische Hierarchie

Neben der Möglichkeit einer hyponymischen Hierarchisierung der Klassen besteht die Option zur Erstellung einer thematischen Hierarchie. Für eine solche Aufstellung müssen diejenigen semantischen Klassen, die einen eindeutigen thematischen Bezug haben — also durch das Merkmal ¡thematische Klasse¿ markiert sind — herausgegriffen und den entsprechenden Themenbereichen zugeordnet werden. Zusätzlich wird dabei zunächst die strikte Unterordnungsbeziehung relevant, d. h. die Klassen, die strikte Subklassen einer thematisch relevanten Klasse sind, werden bei der thematischen Klassifizierung berücksichtigt. Bezüglich der Verwendung der bedingten Unterordnung für eine thematische Hierarchie gilt die Einschränkung, daß sie, als thematische Zuordnung betrachtet, Disambiguierung der entsprechenden Ausdrücke wünschenswert macht.

Eine solche thematische Hierarchie ist Grundlage für den Einsatz der semantischen Kodierung für das Information Retrieval. Die thematische Gliederung des Wortschatzes sollte jedoch aufgrund der Vielzahl möglicher Themen zielgesteuert stattfinden, d. h. vor dieser Gliederung müßten zunächst die zu bestimmenden Themenbereiche festgelegt und bezüglich des für sie typischen Wortschatzes untersucht werden. In Kapitel 4 wird ein entsprechender Probelauf anhand von vier Themenbereichen vorgestellt werden, denen Texte aus dem Teil “Vermischtes“ der “Süddeutschen Zeitung“ zugeordnet werden.

3.7.5 Vorgehen zur Hierarchisierung

Die Hierarchisierung der semantischen Klassen ergab sich, wie gezeigt werden konnte, innerhalb des Kodierungsprozesses, indem von oben nach unten — d. h. von den groben Klassen zu feineren Klassen kodiert wurde.

Wie bei der Darstellung der Grobklassifizierung angedeutet wurde, handelt es sich bei der entstandenen Struktur nicht um eine Baumstruktur mit einer einzigen gemeinsamen Wurzel, sondern um eine Menge disjunkter Strukturen, deren oberste Knoten die Grobklassen sind. Das Phänomen disjunkter Einzelklassen ist auch auf niedrigerer Ebene festzustellen, wenn etwa eine Klasse wie Krankheiten nicht weiter eingeordnet werden kann — abgesehen von einer kodierungsökonomisch bedingten Zuordnung zu einer Restklasse, deren Teilklassen aber keine klassenkonstituierenden Eigenschaften teilen. Doch auch die zusammenhängenden Strukturen sind nicht unbedingt Baumstrukturen. In der hierarchischen Struktur können

nicht-polyseme Lexeme in zwei, nach unterschiedlichen Kriterien aufgestellte Klassen eingeordnet werden (so wurde etwa *Huhn* als Vogel und als Nutztier klassifiziert). Bei der Hierarchisierung der semantischen Klassen aufgrund der strikten Unterordnung ergaben sich nur wenige Mehrfachzuordnungen. Im Zusammenhang mit der bedingten Unterordnung tritt diese häufiger auf. So werden die Ausbildungsinstitutionen und einige Klassen, deren Lexeme ähnliche systematische Polysemien zeigen, im strikten Sinne den Institutionen und bedingt hyponymisch den Gebäuden untergeordnet. Nicht möglich ist in der entstandenen Struktur ein zyklischer Pfad. Eine semantische Klasse kann also nicht einer anderen gleichzeitig über- und untergeordnet sein, gleich welcher Typ von Unterordnungsrelation vorliegt.

Die Angabe der Hierarchie erfolgt in Regeln der folgenden Form (*typ* steht für den Typ von Unterordnungsrelation) — hier mit dem Beispiel der Unterordnung der Klasse Vögel unter die Klasse der Tiere:

`untergeordnete Klasse -(`*typ*`)-> übergeordnete Klasse`
`Vogel -(strikt)-> Tier`

3.8 Weitere semantische Angaben

3.8.1 Kodierung von Sinnrelationen

Im Kapitel 2 der vorliegenden Arbeit wurden die wichtigsten Sinnrelationen aufgezählt und charakterisiert.

Die Hyponymiebeziehung ist bereits durch die Kodierung der Klassen und durch deren Hierarchisierung berücksichtigt. Allerdings mußte noch innerhalb der Klassen das Hyperonym zu den Klassenmitgliedern gegenüber den anderen Lexemen ausgezeichnet werden.

Auch für die Kodierung der Meronymiebeziehung ist mittels des semantischen Relators PAR (s. o.) bereits die Grundlage gelegt, wenn diese Relation auch bisher nur für die Lexeme kodiert wurde, bei denen die Meronymierelation zu einem anderen Lexem oder einer Klasse von Lexemen einen zentralen Aspekt der Bedeutung ausmacht.

Unter den wichtigsten Sinnrelationen bleibt die Kodierung der Synonymie und der Opposition. Im Gegensatz zu den vorher genannten Bedeutungsbeziehungen handelt es sich bei diesen beiden Sinnrelationen um binäre Beziehungen ausschließlich zwischen Lexemen und nie zwischen Lexemklassen[50]. Beide Relationen sind symmetrisch. Dies ermöglicht, die Synonymie bzw. Oppositionsbeziehung jeweils nur in einem Eintrag zu spezifizieren — sie kann dann für das andere Lexem deduziert werden.

3.8.2 Hyponymie

Die Aufstellung der semantischen Klassen basiert in hohem Maße auf der hyponymischen Struktur des Nominalwortschatzes. Die Lexeme, die innerhalb einer taxonomischen Klasse das Hyperonym der anderen in die Klasse eingeordneten Lemmata bezeichnen, haben dabei in mehrerer Hinsicht eine Sonderstellung:

- Sie befinden sich in der hyponymischen Struktur des Wortschatzes nicht auf derselben Ebene wie die anderen Lexeme der Klasse.
- Ihre Bedeutung ist allgemeiner als die der anderen Lexeme. Bei einer Untergliederung der Klassen in Teilklassen können sie keiner Teilklasse zugeordnet werden.

[50] S. a. Miller u. a. (1993: 6f), die Synonymie und Antonymie als Relationen zwischen Wort**formen** beschreiben, Hyponymie und Meronymie als Relationen zwischen Wort**bedeutungen**.

- In Texten können sie als Oberbegriff für die anderen Nomina in der Klasse verwendet werden.

Von daher wurden diese Lexeme — insofern für eine semantische Klasse ein solches Hyperonym überhaupt existiert und es im kodierten Wortschatzauschnitt enthalten war — durch den Operator '=' ausgezeichnet. Das Lexem *Vogel* erhielt somit nicht den Kode Vogel sondern =Vogel, das Lexem *Schiff* nicht den Kode Wasserfahrzeug sondern den Kode =Wasserfahrzeug.

Eine wordNet-ähnliche hyponymische Strukturierung des Wortschatzes kann somit leicht aus der semantischen Kodierung der Nomina im CISLEX extrahiert werden, allerdings mit der Einschränkung, daß bei taxonomischen Klassen, die kein solches ausgezeichnetes Lexem enthalten — etwa weil es nicht im bisher kodierten Wortschatzbereich enthalten war — bei der Darstellung der hyponymischen Struktur auf die Klassenbezeichner oder eine Umschreibung der semantischen Klasse zurückgegriffen werden muß.

3.8.3 Varianten und Synonyme

Die Sinnrelation der Bedeutungsgleichheit ist wesentlich zur Strukturierung des Lexikons auf Mikroebene, d. h. innerhalb einer semantischen Klasse.

Wichtig erscheint es mir zunächst, eine Unterscheidung zwischen Synonymen und Formvarianten zu machen.

Kriterium für die Kodierung als Variante ist in erster Linie ausreichende formale Ähnlichkeit bei identischer Semantik. Im Unterschied zu Synonymen, die bei formaler Ungleichheit semantische Ähnlichkeiten aufweisen, sind Varianten in ihrer Bedeutung identisch, so daß von zwei Varianten nur eine semantisch kodiert werden muß und die andere dann automatisch dieselbe Bedeutungsbeschreibung erhält.

Unter Varianten eines Wortes sind im einzelnen zu verstehen:

- Schreibvarianten wie *Telephon/ Telefon*: Dieser Typ von Variante taucht besonders häufig bei Wörtern aus dem Griechischen, dem Lateinischen und aus den romanischen Sprachen auf. Relativ häufig alternieren *f/ph* und *c/z* bzw. *k*. Die Aussprache ist nicht betroffen. In diesen Fällen liegt keine signifikante semantische Differenzierung vor, sondern höchstens ein leichter stilistischer Unterschied in der Konservativität des Sprachgebrauchs[51]

- Endungsvarianten kommen vor allem in Fremdwörtern vor — es gibt hier häufig eine fremdsprachennähere Variante und eine andere Variante mit Endung auf den Laut Schwa, und damit auf das Graphem *-e* (z. B. *-is* vs. *-e* in *Hypotaxis/ Hypotaxe*). Teilweise gehen die verschiedenen Endungen allerdings auch mit einer semantischen Differenzierung einher (*Basis/Base*), oft auch, indem eine der beiden Formen in den allgemeinen Sprachgebrauch übergegangen ist, während die andere fachsprachlich eingeschränkt verwendet wird (vgl. z. B. *Thesis* vs. *These*). In diesem Fall handelt es sich nicht um Varianten im Sinne der Kodierung.

- Phonologisch motivierte Varianten tauchen relativ systematisch bei Derivationen von Verben und häufig auch bei denominalen Nomina mit bestimmten Endungen auf. Sie betreffen oft den Schwa-Laut in bestimmten phonologischen Kontexten. Bei diesen

[51] Im Deutschen spielt die Auswahl der verwendeten Lexemvarianten keine besonders große Rolle für die Einordnung eines Textes, anders als etwa im Norwegischen (v. a. in der Schriftsprachenform Bokmål), wo eine solch große Anzahl von — sprachhistorisch/-politisch bedingten — Varianten innerhalb der normierten Schriftsprache gestattet ist, daß eine Analyse ihres Gebrauchs eine Zuordnung etwa einer Zeitung u einer politische Richtung erlaubt.

Lexemen existiert dann eine Variante mit dem Graphem *e* und eine andere ohne es (*Wanderer/ Wandrer*, *Vogeler/ Vogler*). Daneben gibt es noch andere Typen von phonologischen Varianten, die bei der Entlehnung von Lexemen aus Fremdsprachen entstehen, da hier Anpassung der Aussprache an das phonologische System des Deutschen auf verschiedene Weisen erfolgen kann, was dann auch graphemisch verschieden ausgedrückt wird (z. B. *Dschonke/ Dschunke*).

- Kurzwörter (vgl. Fleischer/Barz 1992: 218-223) ohne beträchtliche semantische Differenzierung gegenüber der Ursprungsform bilden die letzte Gruppe von Varianten (*Limonade/ Limo, Telefax/ Fax*). Dies ist der Typ von Variante, für den eine Abgrenzung zur Synonymie am schwersten fällt. Handelt es sich um Varianten, sind beide Formen in beliebigen Kontexten füreinander substituierbar, und die Verwendung der anderen Form bringt allenfalls leichte stilistische Unterschiede mit sich. Für einige Kurzwörter gilt dieses Austauschbarkeitskriterium nicht (so etwa *Sozialist/ Sozi*, wo das letztere Lexem eine pejorative Konnotation hat). Es handelt sich in diesen Fällen dann nicht um Varianten.

Existieren von einem Lexem Varianten, ist häufig eine von ihnen wesentlich verbreiteter als die anderen. Die Häufigkeit der Varianten wurde bei der Kodierung zunächst nicht berücksichtigt. Die Markierung der gebräuchlicheren Form kann ohne weiteres aufgrund der Analyse von Korpora erfolgen[52] Ein weiteres Desideratum ist die Zuordnung von Varianten zu regional oder sprachsoziologisch gebundenen Sprachformen sowie die Markierung von veralteten Formen. Da eine systematische Festlegung möglicher diasystematischer Angaben für das CISLEX bisher noch nicht erfolgt ist, wurde auch eine diesbezügliche Zuordnung von Varianten noch nicht vorgenommen.

Synonyme können im Gegensatz zu Varianten Unterschiede in der Bedeutung — in erster Linie von deren konnotativen Aspekten — aufweisen; somit ist auch keine vollständige Austauschbarkeit in allen Kontexten gegeben. Auch kann bei polysemen Lexemen die Synonymiebeziehung zu einem anderen Lexem auf eine der Bedeutungen beschränkt sein.

Es wurden bei der Kodierung zunächst keine unterschiedlichen Grade der Synonymie berücksichtigt (wie etwa in Cruse 1986: 265-291 differenziert[53]), da es kaum möglich ist, operationalisierbare Kriterien für eine solche Abstufung zu finden. Wenn ein Lexem polysem war und dies sich in der Zuweisung zu mehreren semantischen Klassen niederschlug, wurde bei Synonymangaben die semantische Klasse bezeichnet, innerhalb derer die Synonymierelation gilt.

Die Kodierung von Synonymen bedarf dringend der Ergänzung durch diasystematische Angaben, um allgemeingültige Synonyme von geographisch gebundenen Lexemen abzugrenzen. Besonders auffällig wurden bei der Kodierung der Synonymik solche Lexeme, die auf das österreichische (z. B. *Paradeiser* für *Tomate*) oder schweizerische Sprachgebiet (z. B. *Anstößer* für *Nachbar*) beschränkt sind. Auch veraltete Lexeme sind als solche zu bezeichnen[54].

[52] Sind aufgrund der Auswertung von Korpora gänzlich ungebräuchliche Nebenformen von Lexemen als solche markiert, könnte diese Information eingesetzt werden, um im Rahmen eines Rechtschreibkorrekturprogramms oder eines Stilkorrekturprogrammes Ersetzungsvorschläge zu machen.

[53] Cruse (1986) unterscheidet zwischen 'absoluter Synonymie', 'kognitiver Synonymie' (bei Austausch bleiben die Wahrheitsbedingungen gleich; ebd. 270) und 'Plesionymie', wobei ich letztere nicht mehr als Synonymie ansehe. Da absolute Synonymie sehr selten ist, läßt sich die im CISLEX kodierte Relation weitgehend mit der von Cruse beschriebenen 'kognitiven Synonymie' identifizieren.

[54] Während der semantischen Kodierung entstanden aus dieser Notwendigkeit heraus erste Überlegungen und Ansätze zur Kodierung diasystematischer Angaben. Die Darstellung der Festlegung dieser Angaben kann jedoch nicht Thema dieser Arbeit sein.

3.8.4 Oppositionen

Im Gegensatz zu den Adjektiven und Verben, bei denen die verschiedenen Typen von Opposition eine große Rolle spielen, ist die Relation des Bedeutungsgegensatzes bei Nomina in wesentlich geringerem Maße von genuiner Relevanz. Bei deadjektivischen und deverbalen Nomina tritt Opposition häufig sekundär auf. Daher kommen alle bei diesen Wortarten vorhandenen Untertypen der Opposition auch im Nominalwortschatz vor. Diese Fälle sollten auf der Semantik der Derivationsbasen aufbauend kodiert werden. Opposition bei nicht abgeleiteten Nomina beruht häufig auf dem Gegensatz eines Merkmals, das durch Umschreibung mit Adjektiven ausdrückbar ist (so die 'Opposition' *Mann* - *Frau*, die sich letztlich auf *männlich* - *weiblich* reduzieren läßt[55]). Einige nicht ohne weiteres auf einen adjektivischen Kontrast reduzierbare Oppositionspaare lassen sich allerdings finden, so *Ebbe* - *Flut*, *Orient* - *Okzident*, *Himmel* - *Hölle*. Sprachlich sind sie insofern relevant, als die beiden gegensätzlichen Lexeme meist zahlreiche gemeinsame Kontexte bis hin zu geteilten idiomatischen Wendungen haben:

189 Maria kommt in den Himmel

190 Karl kommt in die Hölle

191 Der Himmel auf Erden/die Hölle auf Erden

Einige dieser in Opposition stehenden Lexempaare konstituieren minimale semantische Klassen, die kaum oder nur unter Schwierigkeiten weiter einzuordnen sind und deren kontextuelles Verhalten sich mit dem anderer Lexeme nicht vergleichen läßt. Ein solches Paar ist etwa *Ebbe* - *Flut* mit den Hyperonymen *Gezeiten* und *Tide*.

Bei Paaren von Lexemen, die eindeutig in einer Oppositionsbeziehung stehen, welche nicht direkt von der entsprechenden Relation eines zugrundeliegenden Verbs oder Adjektivs abgeleitet werden kann, wurde diese Opposition kodiert.

3.9 Tabellarische Zusammenfassung der verwendeten semantischen Beschreibungselemente

Die folgende Tabelle gibt eine Überblick über die bisher definierten semantischen Beschreibungselemente.

[55] Allerdings ist in diesem Fall schwer zu entscheiden, welche Opposition grundlegender ist.

Bezeichnung	Typ	Gegenstand	Selektions-präferenzen	thematische Einteilung des Wortschatzes	sonstige Testrahmen
semantische Klassen	Klassen von Lexemen	hyponymische Struktur// des Wortschatzes und geteilte Selektionskontexte	+	+	+
semantische Relatoren	Klassen v. Lexem-relationen	Bedeutungsrelationen und abgeleitete Bedeutungen	+	-	+
Merkmale	Klassen v. Lexemen	nicht-hyponymische Bedeutungsbestandteile	+	-	+
Synonyme	binäre Beziehungen	synonymische Struktur	+	+	+
Varianten	zwischen	Lexemvarianten	+	-	-
Oppositionen	Lexemen	Opposition	+	+	+

Tabelle 3.7: Semantische Beschreibungselemente für die Nomina im CISLEX

Kapitel 4

Anwendungen der semantischen Kodierung

Die Kodierung der Nomina wurde im Sinne der in der Einleitung genannten möglichen Anwendungen der maschinellen Übersetzung und des Information Retrievals zu zwei Tests herangezogen:

- Zum einen wurde anhand eines größeren Korpus von Nominalkomposita die Relevanz der semantischen Klassen zur Beschreibung von Selektionseigenschaften und von Aspekten der Semantik von Kompositaköpfen ermittelt.
- In der zweiten Untersuchung wurde anhand eines kleinen Testkorpus mit diversen Texten (Zeitungstexte aus dem "Vermischten" der Süddeutschen Zeitung) die Tauglichkeit der Semantik für die Identifizierung des thematischen Bereichs von Texten überprüft.

Bei beiden Untersuchungen kamen dieselben statistischen Maße zur Anwendung, namentlich die Transinformation (MI) und das Tschebyschow-Risiko.

4.1 Anwendungen 1: Selektionspräferenzen in Nominalkomposita

4.1.1 Motivation

Am CIS stand für die folgende Untersuchung ein Korpus aus ca. 200 000 segmentierten Nomen-Nomen-Komposita zur Verfügung[1], die zum größten Teil aus der Süddeutschen Zeitung der Jahre 1992, 1993 und 1994 stammen - z. T. ergänzt durch andere Quellen - Lexika des Deutschen und maschinenlesbare Dokumente Die statistische Untersuchung dieses Korpus von Nominalkomposita anhand der kodierten semantischen Klassen wurde aus mehreren Gründen vorgenommen:

- Für die aufgestellten semantischen Klassen im CISLEX wurde eine distributionelle Evaluationsbasis benötigt. Es lag von daher nahe, die bisher festgelegten semantischen Klassen anhand von Selektionsrestriktionen von Nominalkomposita zu bewerten und die vorgenommene Einteilung in der Folge unter Umständen teilweise zu revidieren.

[1] Das Korpus wird ständig erweitert und umfaßte beim Abschluß der Arbeit bereits mehr als eine Million Komposita.

- Zur Beschreibung von Nominalbedeutung gehört auch die Beschreibung der Semantik von Nominalkomposita. Die Beschreibung der Simplizia sollte von daher von vornherein auf eine nachfolgende Kodierung der komplexen Ausdrücke Bezug nehmen, in denen sie enthalten sind. Es sollte mit dieser Untersuchung herausgefunden werden, welche Aspekte der Semantik der Zweitglieder in Komposita besonders wichtig werden, und was die statistische Verteilung der semantischen Klassen von Erstgliedern zum Verständnis der semantischen Relationen zwischen Gliedern von Nominalkomposita und damit zur Beschreibung der Semantik von Komposita beitragen kann.

4.1.2 Zu deutschen Nominalkomposita

Im Deutschen ist die Möglichkeit zur Bildung von Nominalkomposita besonders stark ausgeprägt (vgl. Augst 1975, 1975a). Nominalkomposita sind aus mehreren, auch selbständig auftretenden Teilen zusammengesetzte Lexeme, deren Kopf ein Nomen ist[2]. Der nominale Kopf steht im Deutschen stets am rechten Rand des Kompositums. Als Erstglieder können Lexeme verschiedener Kategorien auftreten. Für die nachfolgend beschriebene Untersuchung interessierten nur solche Nominalkomposita, deren Erstglied ein nicht-präfigiertes Nomen ist, da nur diese Nomina im CISLEX bisher semantisch kodiert wurden. Außerdem wurde die Untersuchung auf zweigliedrige Komposita eingeschränkt.

Bei den sogenannten Determinativkomposita[3] bestimmt das Erstglied das Zweitglied näher. Dabei können zahlreiche Relationen zwischen den Gliedern auftreten, die jedoch in der Lexemform nicht in Erscheinung treten. Während die meisten Komposita von muttersprachlichen Sprechern des Deutschen eindeutig interpretiert werden können — d. h. die Relation zwischen den Gliedern wird erkannt — bringt die Erkennung der nicht-ausformulierten Relation Schwierigkeiten für die automatische Interpretation mit sich. Dies wird für eine Reihe von computerlinguistischen Anwendungen zum Problem. Für die maschinelle Übersetzung etwa ins Französische muß die Relation zwischen den Gliedern eines Kompositums bis zu einem gewissen Grad determiniert sein, um eine brauchbare Übersetzung zu gewährleisten, da die Relation hier meist in Form einer Präposition ausformuliert wird. So wird etwa *Eisengehalt* ins Französische übersetzt als *teneur en fer*, *Eisenerzeugung* als *production de fer*. Um für dieses Beispiel in der Zielsprache eine Auswahl zwischen den Präpositionen *en* und *de* treffen zu können, muß die Relation zwischen den Erstgliedern und dem Zweitglied *Eisen* bei der Analyse des deutschen Ausdrucks determiniert werden. Zwar läßt sich kaum eine abschließende Liste möglicher Relationen zwischen Kompositagliedern aufstellen (vgl. Meyer 1993: 7-9), eine Auflistung bestimmter wichtiger Haupttypen ist aber durchaus möglich[4].

Man weiß nun, daß unter dem Aspekt einer bestimmten Relation zwischen den Gliedern als Erstglieder zu einem Kopf in erster Linie Nomina bestimmter semantischer Klassen in Frage kommen; in Verbindung mit der Meronymierelation beispielsweise können nur Begriffe als Erstglieder von *-fell* auftreten, deren Denotate ein Fell als Teil, d. h. in den meisten Fällen als Körperbedeckung, haben. Man kann also einerseits als Erstglieder zu bestimmten Köpfen Lexeme aus bestimmten semantischen Klassen erwarten und andererseits aus

[2]In diesem Abschnitt findet sich nur eine kurze Skizze der für die Untersuchung wesentlichen Eigenschaften von deutschen Nominalkomposita. Eine ausführliche Darstellung des Forschungsstandes würde den Rahmen dieser Arbeit sprengen. Eine neuere Arbeit zum Thema, in der die wesentliche Literatur zum Thema aufgeführt wird, ist Meyer (1993).

[3]Nominalkomposita werden traditionell unterschieden nach Kopulativkomposita und Determinativkomposita (vgl. Fleischer/ Barz 1992: 45f). Die weitaus häufigsten Nominalkomposita im Deutschen sind die Determinativkomposita. Die Kopulativkomposita fallen statistisch kaum ins Gewicht.

[4]Ein Beispiel sind die Ausführungen in Fanselow (1981), der eine relativ exhaustive Liste möglicher Relationen innerhalb von Komposita vorlegt.

statistisch herausfallenden semantischen Klassen der Erstglieder Relationen zwischen den Gliedern deduzieren.

Allerdings kann beinahe jedes Nomen im Deutschen mit jedem anderen kombiniert werden, um ein interpretierbares Kompositum zu bilden. Es ist von daher zu erwarten, daß die Auswahl der Köpfe bezüglich ihrer Erstglieder sich nicht in hundertprozentiger Selektion einer oder mehrerer semantischer Klassen durch einen bestimmten Kopf ausdrückt; vielmehr ist es wahrscheinlich, daß beinahe jede häufigere Klasse bei jedem häufigeren Kopf auftaucht.

4.1.3 Die statistische Untersuchung

Die vorgenommene Untersuchung berücksichtigt die unterschiedlichen Relationen nicht von vornherein; ebensowenig geht in die Prämissen eine Unterscheidung zwischen Kopulativ- und Determinativkomposita ein. Dies würde eine Vorsortierung des Korpus nötig machen. Für eine solche Vorsortierung würden allerdings einige Grundlagen fehlen, z. B. gibt es keine mit ausreichenden Kriterien versehene Liste aller semantischen Relationen zwischen Kompositagliedern. Zudem ist die Vorsortierung eines so großen Korpus kaum in angemessener Zeit durchführbar. Die statistische Analyse kann somit für die semantische Beschreibung teilweise zu etwas unscharfen Ergebnissen führen. Dies gilt umso mehr, als bisher die lexikalisierten und nicht mehr vollständig semantisch transparenten Komposita (wie etwa *Bahnhof*) noch nicht aus dem Korpus entfernt sind. Letzteres dürfte allerdings im Rahmen der Kodierung des CISLEX bald erfolgen und dann eine weniger unscharfe Auswertung ermöglichen.

Die Untersuchung ging folgendermaßen vor sich[5].

1. Alle Erstglieder der segmentierten Komposita im Korpus wurden mit den semantischen Klassen getaggt, wie sie sich aus der Kodierung der einfachen Nomina ergeben hatten.

2. Die Korpusgröße N wurde ermittelt.

3. Alle semantischen Klassen aller Erstglieder wurden gezählt. Dabei wurden auch die in der Hierarchie übergeordneten Klassen berücksichtigt. Alle drei Typen von Hierarchiebeziehung, d. h. strikte, bedingte und metonymische Unterordnung wurden in die Zuordnung zu Oberklassen einbezogen.

4. Für Köpfe k mit Häufigkeit f(k) wurden nun ebenfalls alle semantischen Klassen aller Erstglieder gezählt, wenn f(k) über einem bestimmten Schwellenwert (20) lag.

Dabei wurde keine semantische Disambiguierung der Erstglieder vorgenommen. Trat ein polysemes Lexem als Erstglied auf, wurden alle im CISLEX-Eintrag enthaltenen semantischen Klassen berücksichtigt.

Mit diesen Zahlen kann man sich nun folgendermaßen über die semantischen Selektionspräferenzen der Köpfe informieren - bezogen auf die semantischen Klassen der Erstglieder:

Es gibt eine gewisse durchschnittliche Frequenz einer Klasse bei allen Erstgliedern im ganzen Korpus. Bei einem speziellen Kopf wäre diese Klasse bei zufälliger Verteilung über den Korpus mit einer gewissen Häufigkeit zu erwarten. Wird diese Häufigkeit beträchtlich über- bzw. unterschritten, heißt das, daß der entsprechende Kopf die Klasse präferiert selegiert oder eben mit ihr inkompatibel ist. Das verwendete Maß für die Selektionspräferenzen

[5]Nicht beschrieben wird an dieser Stelle die Kompositasegmentierung. Zu den Voraussetzungen hierzu siehe Langer (in Vorbereitung), wo ausführlich die Kodierung der Fugenformen von Nomina für das CISLEX beschrieben wird.

ist die Transinformation (Mutual Information) [6], die sich in unserem Fall nach folgender Formel berechnet:

$$MI(s;k) = log\frac{f_k(s)/\sum f_k(S)}{f(s)/\sum f(S)}$$

MI ist der Mutual-Information Wert. Ist er Null, so liegt keine Selektion f
grqqur die Klasse s beim Kopf k vor, ist sie gr
grqqosser Null, selegiert der Kopf diese Klasse pr
grqqaferiert, ist sie kleiner Null, gibt es Inkompatibilit
grqqaten zwischen dem Kopf und der Klasse.

s ist die semantische Klasse, deren Mutual-Information-Wert MI bez
grqquglich des Kopfes k ermittelt werden soll.

f ist die H
grqqaufigkeit von Token der semantischen Klasse bei einem bestimmten Kopf ($f_k(s)$), bzw. im Gesamtkorpus (f(s)); die Summenzeichen summieren
grqquber alle semantischen Klassen(S) bei einem Kopf bzw. im gesamten Korpus.

$r_T = Pg * (1 - Pg)/\sum fk(S) * (P1 - Pg)2$

Pg: Wahrscheinlichkeit von einer semantischen Klasse im gesamten Korpus : $f(s)/\sum f(S)$

P1: Wahrscheinlichkeit von einer semantischen Klasse bei Kopf k : $fk(s)/\sum fk(S)$

Nachfolgend einige Einzelergebnisse, die verschiedene Aspekte der Semantik von Nominalkomposita und die Bedeutung dieser Untersuchung für die semantische Kodierung demonstrieren.

Es finden sich in den Statistiken folgende Werte:

l: lokale Häufigkeit einer semantischen Klasse bei allen Erstgliedern eines Kopfes

g: Gesamthäufigkeit einer semantischen Klasse bei allen Erstgliedern des Korpus

erw: bei zufälliger Verteilung erwarteter Wert für die lokale Häufigkeit

MI: Transinformation (s.o.)

TS: Tschebyschow-Risiko (s.o.)

Liste der auftretenden semantischen Klassen der Erstglieder in den folgenden Beispielen:
ABS: Bildungsabschlüsse, AKT: Aktionen, ASP: Sportarten, DIS: Diskursobjekt, EIG: Eigenschaft, ERE: Ereignis, FES: Feste, FRU: Früchte, GED: Druckerzeugnisse, GMI: Genußmittel, KLE: Kleidung, KSF: Kleiderstoffe, KTE: Körperteil, MIN: Musikinstrument, NHG: Nahrungsgrundstoffe, NTI: Nutztier, PBA: Bäume, PBL: Blumen, PFL: Pflanzen, SFS: Feststoffe, SPO: Sport, STI: Säugetiere, TIE: Tiere, VEK: Verkehrsmittel, VOG: Vögel, WER: Werkzeug, WET: Wettererscheinungen, WIS: Wissenschaften, ZUS: Zustände.

4.1.4 Identifizierung von Relationen

```
Komposita auf -fell: 60 Auftreten im Korpus
   ERE: l: 0 g: 41141 erw: 15 MI : -inf TS :0.0596
   TIE: l: 36 g: 31226 erw: 11 MI : 1.125 TS :0.0185
   KTE: l: 9 g: 5301 erw: 1 MI : 1.512 TS :0.0399
   STI: l: 33 g: 2404 erw: 0 MI : 3.602 TS :0.0009
   NTI: l: 9 g: 1056 erw: 0 MI : 3.126 TS :0.0053
   MIN: l: 3 g: 855 erw: 0 MI : 2.238 TS :0.0445
```

- Über die Hälfte der Erstglieder fällt in die semantische Klasse SÄUGETIERE. Diese präferierte Selektion zeigt sich auch bei den NUTZTIEREN und — via Vererbung — an dem in der Taxonomie übergeordneten Knoten TIER.

[6] Zur Anwendung des Maßes Transinformation zur Erkennung von Kollokationen s. Church/ Hanks (1990). Zur Anwendung auf Selektionsrestriktionen s. Hindle (1990), Breidt (1993). Zur Untersuchung von Selektionsrestriktionen aufgrund der Klassifikation in WordNet s. Resnik (1993).

- Eine weitere selegierte Klasse sind die MUSIKINSTRUMENTE (*Trommelfell*, *Paukenfell* ...).
- Zudem zeigt sich eine eindeutige negative Selektionspräferenz: Als Erstglied zu -*fell* taucht im gesamten untersuchten Korpus nie ein EREIGNIS auf; eine Zufallsverteilung würde für diesen Kopf zu 15 Komposita mit einem solchen Erstglied führen.

-*schutz*: 242
WET: l: 7 g: 1428 erw: 1 MI : 1.400 TS :0.0619
VOG: l: 6 g: 1212 erw: 1 MI : 1.410 TS :0.0710

- Der Schutz vor WETTERERSCHEINUNGEN
- Schutz von VÖGELN

Erstere Gruppe entspricht einer semantischen Selektion des Verbs *schützen* bezüglich seines Präpositionalkomplements (*schützen vor*), letztere entspricht der Relation zwischen dem zugrundeliegenden Verb und seinem Akkusativobjekt.

-*hose*: 80
SPO: l: 6 g: 3379 erw: 1 MI : 1.308 TS :0.0841
KLE: l: 17 g: 1723 erw: 0 MI : 3.023 TS :0.0032
KSF: l: 13 g: 1107 erw: 0 MI : 3.197 TS :0.0034

- *für*-Relation, im Zusammenhang mit Erstgliedern, die SPORTLICHE TÄTIGKEITEN bezeichnen (z. B. *Trainingshose*)
- Meronymierelation, innerhalb von Komposita, die kombinierte KLEIDUNGSSTÜCKE bezeichnen (z. B. *Anzugshose*)
- *gemacht-aus*-Relation, bezüglich eines KLEIDERSTOFFS aus dem die Hose besteht (z. B. *Flanellhose*)

4.1.5 Polysemie des Zweitglieds

-*blatt*: 279
PFL: l: 41 g: 5738 erw: 8 MI : 1.526 TS :0.0086
PBA: l: 12 g: 1698 erw: 2 MI : 1.515 TS :0.0300
GED: l: 17 g: 3843 erw: 5 MI : 1.046 TS :0.0488
WER: l: 8 g: 1169 erw: 1 MI : 1.483 TS :0.0474
DAK: l: 11 g: 2197 erw: 3 MI : 1.170 TS :0.0591

- *Blatt* als PFLANZENTEIL (sehr häufig zusammen mit Bäumen)
- *Blatt* als Teil eines DRUCKWERKES
- *Blatt* als Teil diverser WERKZEUGE
- *Blatt* im Sinne von 'Zeitung' im Zusammenhang mit Diskursobjekten.

DAK bezeichnet eine Sammelklasse, in die Diskursobjekte verschiedener Art eingeordnet wurden (*Nachrichten, Witz, Propaganda*); sie erweist sich hier trotz ihrer Uneinheitlichkeit als selektionsrelevant — ebenso bei anderen Lexemen wie *Brief*, *Formel* u. v. a. In der Statistik schlagen sich die Bedeutungsvarianten von Blatt in *Rotorblatt* oder *Ruderblatt* nicht nieder . Für diese Spezialbedeutung ist keine Reihenbildung erkennbar.

-*ball*: 147
ASP: l: 12 g: 3443 erw: 2 MI : 1.460 TS :0.0327
FES: l: 8 g: 1121 erw: 0 MI : 2.176 TS :0.0180
PBL: l: 3 g: 297 erw: 0 MI : 2.524 TS :0.0316
ABS: l: 2 g: 175 erw: 0 MI : 2.647 TS :0.0410

- *Ball* in der Bedeutung 'Ball für Ballspiele' selegiert als Erstglied Sportarten (*Tennisball* etc.).
- In der Bedeutung von 'Tanzveranstaltung' selegiert das Nomen als Erstglieder Feste (Silvesterball etc.), Blumen (*Magnolienball* etc.) und Ausbildungsabschlüsse (*Abiturs-ball*, *Maturaball*).

Obwohl diese Klassen im Korpus im Zusammenhang mit diesem Kopf absolut gesehen nicht übermäßig häufig sind, fallen sie doch statistisch sehr stark heraus.

4.1.6 Köpfe mit schwach ausgeprägte Selektionspräferenzen

-*problem*: 322
ERE: l: 95 g: 31671 erw: 46 MI : 0.725 TS :0.0182
ZUS: l: 64 g: 14595 erw: 21 MI : 1.104 TS :0.0113
EIG: l: 26 g: 5122 erw: 7 MI : 1.251 TS :0.0214

Die wenig spezifischen Selektionspräferenzen des Zweitglieds -*problem* zeigen sich darin, daß nur eine Gruppe von in der Hierarchie weit oben liegenden Klassen, namentlich Ereignisse, Zustände, und Eigenschaften, sich in der Statistik niederschlägt, und dies zudem mit nicht allzu hohen MI-Werten. Da diese Klassen zudem insofern benachbart sind, als sich zwischen ihnen teilweise Überschneidungen und Abgrenzungsschwierigkeiten ergeben (vgl. die Überlegungen zur Grobklassifikation in Kapitel 3), ergibt sich aus der Statistik in diesem Fall kaum ein Hinweis auf die Semantik der Zweitglieds: weder liegen Indizien für Polysemie des Lexems *Problem*, noch Hinweise auf klar abgrenzbare Typen von Relationen zwischen Erst- und Zweitglied vor.

-*produktion*: 237
KNK: l: 149 g: 68607 erw: 94 MI : 0.457 TS :0.0281
ERE: l: 16 g: 31671 erw: 43 MI : -1.001 TS :0.0545
MEN: l: 13 g: 26313 erw: 36 MI : -1.023 TS :0.0645
AKT: l: 14 g: 25831 erw: 35 MI : -0.931 TS :0.0736
STO: l: 55 g: 14649 erw: 20 MI : 1.005 TS :0.0162
NAH: l: 24 g: 6622 erw: 9 MI : 0.969 TS :0.0406
FAZ: l: 19 g: 6220 erw: 8 MI : 0.798 TS :0.0776
SFS: l: 17 g: 4270 erw: 5 MI : 1.063 TS :0.0471
GMI: l: 8 g: 1948 erw: 2 MI : 1.094 TS :0.0943
KSF: l: 6 g: 1186 erw: 1 MI : 1.303 TS :0.0852
FRU: l: 4 g: 676 erw: 0 MI : 1.459 TS :0.0985
NHG: l: 5 g: 627 erw: 0 MI : 1.758 TS :0.0503

Ein ebenso unspezifisches Profil wie für -*problem* ergibt sich für das Zweitglied -*produktion*. Im Endeffekt läßt sich hier nur eine Präferenz für Konkreta im weiteren Sinne (Stoffe und Konkreta(Objekte)) ermitteln, sowie eine deutliche negative Präferenz für Ereignisse und Menschenbezeichnungen. Die zunächst vielleicht erwartete Präferenz für Artefakte ergibt sich aus den Zahlen nicht — dies läßt sich auf die große Zahl von Komposita mit -*produktion* zurückführen, die Konkreta als Erstglied haben, die nicht den Artefakten zugerechnet werden (*Eisenproduktion*, *Getreideproduktion* etc.).

4.1.7 Zusammenfassung der Ergebnisse

Die angesetzte Schwelle beim Tschebyschow-Risiko (0,1) führte dazu, daß nur die wirklich signifikant abweichenden semantischen Klassen in den Statistiken auftauchten. Bei einem Test ohne Schwellenwerte für dieses Maß stellt man fest, daß trotz der Selektionspräferenzen beinahe alle semantischen Klassen als Erstglieder beinahe jedes Kopfs möglich sind. Als wichtigstes Ergebnis kann also festgehalten werden: Die Kombinatorik von Nominalkomposita ist nicht völlig beliebig; es handelt sich aber beim semantischen Selektionsverhalten der Köpfe tatsächlich nur um Präferenzen, nicht um kategorische Selektionsrestriktionen.

Bezüglich dieser Selektionspräferenzen läßt sich festhalten:

- Es gibt bei einer großen Menge der untersuchten Köpfe in den verschiedenen mit ihnen gebildeten Komposita mehrere unterschiedliche Relationen zwischen Erst- und Zweitglied, die sich bei einem Teil der Komposita in der präferierten Selektion mehrerer, semantisch unterschiedlicher Klassen widerspiegeln.

- Auch Polysemie des Kopfes spiegelt sich häufig in den Selektionspräferenzen bezüglich der Erstglieder wieder. Das typische Selektionsprofil sowohl eines Kopfes, bei dem verschiedene, klar abgrenzbare Relationen auftreten als auch eines polysemen Zweitgliedes ist dabei folgendes: Es werden mehrere relativ spezifische, klar voneinander abgrenzbare semantische Klassen selegiert, die keinen unmittelbaren gemeinsamen Mutterknoten haben. Die kleinste gemeinsame übergeordnete Klasse der selegierten Klassen wird nicht selegiert.

- Köpfe, die in einer unspezifischen Relation zu ihren Erstgliedern stehen (eine Relation, die sich etwa als 'in bezug auf' paraphrasieren ließe), wie *-frage*, *-problem*, *-idee*, zeigen gering ausgeprägte Selektionspräferenzen, die stets nur in bezug auf Klassen deutlich werden, die in der Taxonomie weit oben stehen. Feinere Klassen werden nicht statistisch auffallend als Erstglieder selegiert.

- Köpfe, die eine nicht als Klasse kodierte Untermenge einer Grobklasse selegieren, wie *-produktion*, das u. a. eine Untermenge der Konkreta als Erstglieder hat, zeigen Selektionspräferenzen, die für die Grobklasse und eine Teilmenge ihrer Unterklassen deutlich werden, wobei die Werte statistisch nicht allzu herausfallend sind.

- Bei deverbalen Nomina lassen sich anscheinend in vielen Fällen im Selektionsverhalten der Rektionskomposita die Selektionspräferenzen der zugrundeliegenden Verben wiedererkennen[7].

Aus einigen der genannten Punkte ergaben sich direkte Konsequenzen für die semantische Kodierung. So wurde der Bestand an Klassen aufgrund der Ergebnisse der statistischen Untersuchung neu beurteilt und erweitert; bestimmte Nomina wurden aufgrund der erkannten Selektionspräferenzen neu kodiert. Dies betraf sowohl solche Nomina, die durch ihr Selektionsverhalten als Köpfe in Komposita Polysemie gezeigt hatten, insofern diese noch nicht kodiert war, als auch Erstglieder, die bei einem Durchgehen der Ergebnisse als mangelhaft klassifiziert erkannt wurden. Zudem ließen sich aus fehlenden Selektionspräferenzen bei bestimmten Köpfen in einigen Fällen Lücken in der semantischen Klassifizierung der einfachen Nomina ableiten, d. h. bestimmte selektionsrelevante Klassen waren nicht berücksichtigt worden. Einige Köpfe zeigten allerdings so idiosynkratische Selektionspräferenzen, daß die

[7]Dies müßte noch anhand einer entsprechenden Untersuchung des Selektionsverhaltens der erbalen Derivationbasen verifiziert werden.

Aufstellung einer Klasse aufgrund dieser Selektionseigenschaften allein nicht geeignet war, weitere Generalisierungen zu beschreiben.

Die erfolgte statistische Untersuchung von Nominalkomposita wies, trotz einiger interessanter Ergebnisse, noch erhebliche Mängel auf. Vor einem weiteren Durchgang müßten vor allem die Komposita, die synchron nicht mehr semantisch durchsichtig sind, aus dem Korpus entfernt werden, da sie das Ergebnis der Statistiken verfälschen. Wünschenswert wäre ferner ein ausgereifterer Segmentieralgorithmus, da die vorgenommene Segmentierung teilweise zu mehreren Abtrennungen führte, von denen nur eine akzeptabel war. Ein solcher Algorithmus könnte teilweise sicher auf den bisherigen Ergebnissen aufbauen, indem er statistisch weniger präferierte Aufspaltungen eines komplexen Nomens hintanstellt.

Eine weitere Untersuchung wäre die Kombinatorik der semantischen Klassen wert, mit dem Ziel, herauszufinden, welche semantischen Klassen in Komposita präferiert mit welchen anderen semantischen Klassen auftreten. Auch diese Ergebnisse könnten Aufschlüsse über Relationen zwischen den Gliedern geben und zur Verbesserung eines Segmentierungsalgorithmus für Komposita verwendet werden.

Die vorliegende Untersuchung hat weitgehend bestätigt, daß die kodierten semantischen Klassen geeignet sind, um Selektionspräferenzen von Operatoren zu beschreiben. Ein weiterer Test soll nun die Frage beantworten, inwiefern die semantischen Klassen im CISLEX auch zur Beschreibung von thematischen Bereichen geeignet sind.

4.2 Anwendungen 2: Thematische Bereichszuordnung von Texten

Die Kodierung der semantischen Klassen der einfachen Nomina erlaubt bereits erste Tests bezüglich ihrer Anwendbarkeit für die Erkennung des thematischen Bereichs eines Textes — also der Eignung der Klassifikation für das Information Retrieval. Zu diesem Test wurde folgende Frage formuliert:

Inwiefern können anhand der semantischen Klassen der auftretenden Nomina die thematischen Bereiche ermittelt werden, denen ein nicht fachspezifischer Text zuzuordnen ist?

Für eine solche Eignungsprüfung lieferte eine thematisch bedingte Auswahl der kodierten semantischen Klassen die notwendige Grundlage. Die vorgenommene Untersuchung soll dabei nicht ein ausgereiftes Information-Retrievalverfahren dokumentieren, sondern darf nur als Test zur prinzipiellen Eignung der semantischen Klassen für die automatische Klassifizierung von Texten verstanden werden. Bei dem Testverfahren handelte es sich um ein einfaches Oberflächenverfahren ohne Berücksichtigung syntagmatischer Strukturen im Text[8]. Die Untersuchung ging folgendermassen vonstatten: Als Referenzkorpus wurden die Texte aus dem "Vermischten" der Süddeutschen Zeitung aus allen Ausgaben von November 1994 bis zum Juni 1995 gewählt. Dann wurden vier Themenbereiche ausgewählt, denen Texte automatisch zugeordnet werden sollten.

Die ausgewählten Themenbereiche sind:

- Kriminalität (Verbrechen)
- Raumfahrt
- Kirche und Klerus
- Medizin

[8]Zur Problematik solcher rein statistischer Ansätze ohne die Berücksichtigung von syntagmatischer Regularitäten s. Kuhlen (1989: 692f).

Entsprechend dieser Themenbereiche wurde eine thematische Auswahl von semantischen Klassen im CISLEX vorgenommen.

Diese Klassen wurden dann mit ihren Unterklassen in eine thematische Hierarchie eingeordnet. Eine solche thematische Hierarchie weicht von einer hyponymischen Taxonomie ab, indem sie bestimmte Subbäume einer hyponymischen Struktur zu thematischen Bereichen zusammenfaßt. So wurden für das Themengebiet "Kirche und Klerus" die Klassen Seelsorgerische Berufe, religiöse Ereignisse und Riten, kirchliche Gebäude mit ihren Unterklassen zu einer Bereichsklasse zusammengefaßt. Die semantischen Klassen, die einem Themenbereich zugeordnet wurden, erhielten bei der Untersuchung zunächst keine Gewichtung bezüglich ihrer Relevanz für den Themenbereich, sondern wurden alle in gleicher Weise statistisch berücksichtigt.

Als statistisches Maß wurde wiederum die Transinformation in Verbindung mit dem Tschebyschow-Risiko gewählt (genauere Erläuterungen zu den statistischen Maßen im Abschnitt zur statistischen Analyse von Nominalkomposita oben). Die Ermittlung der Werte erfolgte nach dem gleichen Prinzip wie bei den Nominalkomposita:

1. Alle einfachen Nomina in den Texten wurden mit ihren semantischen Klassen getaggt.

2. Alle semantischen Klassen aller getaggten Nomina in sämtlichen Texten wurden gezählt. Dabei wurden auch die in der Hierarchie übergeordneten Klassen berücksichtigt. Es wurden neben den thematischen Unterordnungen nur die strikten und bedingten Zuordnungen, nicht aber die metonymischen Unterordnungsbeziehungen einbezogen.

3. Für alle zu klassifizierenden Texte ti mit der jeweiligen Anzahl von getaggten Nomina f(ti) wurden nun ebenfalls alle semantischen Klassen sämtlicher Nomina gezählt.

Es gibt nun eine durchschnittliche Frequenz einer semantischen Klasse im Gesamtkorpus. Bei einem ausgewählten Text wäre diese Klasse, bei zufälliger Verteilung über den Korpus, mit einer gewissen Häufigkeit zu erwarten. Wird diese Häufigkeit beträchtlich über- bzw. unterschritten, heißt das, daß im entsprechenden Text diese semantische Klasse präferiert auftritt oder eben thematisch mit ihr inkompatibel ist. Das verwendete Maß ist wiederum die Transinformation, die sich in diesem Falle nach folgender Formel berechnet:

$$MI(s;k) = log\frac{f_t(s)/\sum f_t(S)}{f(s)/\sum f(S)}$$

MI ist der Mutual-Information Wert. Ist er Null, so liegt keine Selektion f
grqqur die Klasse s beim Kopf k vor, ist sie gr
grqqosser Null, selegiert der Kopf diese Klasse pr
grqqaferiert, ist sie kleiner Null, gibt es Inkompatibilit
grqqaten zwischen dem Kopf und der Klasse.

s ist die semantische Klasse, deren Mutual-Information-Wert MI bez
grqquglich des Textes t ermittelt werden soll.

f ist die H
grqqaufigkeit von Token der semantischen Klassein einem Text ($f_k(s)$), bzw. im Gesamtkorpus (f(s)); die Summenzeichen summieren
grqquber alle semantischen Klassen(S) in einem Text bzw. im gesamten Korpus.

$r_T = Pg * (1 - Pg)/\sum ft(S) * (P1 - Pg)2$

Pg: Wahrscheinlichkeit von einer semantischen Klasse im gesamten Korpus : $f(s)/\sum f(S)$

P1: Wahrscheinlichkeit einer semantischen Klasse in Text t : $fk(s)/\sum fk(S)$

4.2.1 Probleme

Im Gegensatz zu der geschilderten Untersuchung von Nominalkomposita, die eine Statistik über den Bestand im Lexikon war — d. h. jedes Nomen wurde nur einmal als Erstglied bei einem bestimmten Kopf berücksichtigt und die tatsächliche Häufigkeit in Texten spielte keine Rolle bei der Auswertung — wird bei den nachfolgenden Statistiken auch mehrfaches Auftreten eines Nomens in den zu klassifizierende Texten gezählt.

Da keinerlei Disambiguierung der Nomina aufgrund des Kontextes vorgenommen wurde, verfälschten verschiedene Einflüsse die Statistiken geringfügig:

- Zu einfachen Nomina homonyme Eigennamen (etwa *Münster*), die in einem Text sehr häufig auftauchten, führten zu einer scheinbaren Präferenz für eine Klasse, die im Text u. U. überhaupt nicht auftrat, da die Wortform im Kontext keiner semantischen Klasse zuzuordnen war.
- Polyseme Nomina, die häufig in einem Text zu finden waren, führten dazu, daß für diesen Text statistisch zwei Klassen stark präferiert erschienen, wobei nur eine von ihnen mit dem tatsächlichen Textinhalt zu tun hatte (etwa *Schwestern* als Verwandte oder Religiöser Beruf). In einem elaborierten System zum Information Retrieval könnte hier oftmals eine einfache Präferenzregel für die häufigere Bedeutung aufgrund der Textart (Zeitungstext) Abhilfe schaffen. In vielen Fällen wären allerdings sicher aufwendigere Disambiguierungsalgorithmen nötig.

Weitere Einschränkungen für die Aussagekraft dieser Untersuchung sind aufgrund der Restringiertheit der semantischen Kodierung zu machen, da ja nur ein morphologisch definierter Wortschatzausschnitt in der thematischen Zuordnung berücksichtigt wurde. Auf der jetzigen Kodierungsstufe steht damit nur ein Teil der für eine zuverlässige Zuordnung notwendigen Information zur Verfügung:

- Weder präfigierte Nomina noch Nominalkomposita sind bisher semantisch kodiert; somit werden nur die semantischen Klassen eines Teils (ca. der Hälfte) der Nomina in einem Text bei der Auswertung berücksichtigt.
- Die bisher fehlende Kodierung der Verben und Adjektive erlaubt noch keine kontextuelle Disambiguierung der polysemen Nomina.

Nach einer vollständigen semantischen Kodierung des Gesamtwortschatzes sind also wesentlich aussagekräftigere Ergebnisse zu erwarten, da dann sowohl eine Disambiguierung polysemer Formen stattfinden kann, als auch eine wesentlich größere Teilmenge der Inhaltswörter eines Textes berücksichtigt werden wird.

4.2.2 Ergebnisse

Die Ergebnisse der Untersuchung werden in zwei Abschnitten dargestellt. Zunächst werden anhand von vier Texten exemplarisch Zuordnungen und Probleme der Zuordnung zu den vorgegebenen Themenbereichen dargestellt, um einen Eindruck von der Ermittlung der statistische Werte zu geben, die Grundlage für die folgende quantitative Klassifikation sind. Bei dieser werden die Trefferquoten für eine größere Zahl von Texten aus dem Korpus statistisch dargestellt.

Es wurden nur solche Texte klassifiziert, in denen mindestens 30 semantisch getaggte Nomina auftraten.

Die Schwellenwerte für die thematische Zuordnung waren:

- `Transinformation > 1,0`
- `Tschebyschowrisiko < 0,1`

4.2.3 Text 1 (Süddeutsche Zeitung 2.11.1994): Unproblematische Klassifikation

> Perfekter Start ins All. Europäische Rakete bringt neuen Funksatelliten in den Weltraum. Kourou (Eigener Bericht) - Mit einem an Präzision nicht zu überbietenden Start

einer 54 Meter hohen Rakete des Typs Ariane 42P ist in der Nacht zum Dienstag ein weiterer Fernsehsatellit in den Weltraum gebracht worden. Auf die Sekunde genau um 21.37 Uhr Ortszeit (0.37 Uhr MEZ) hob die Ariane vom europäischen Raumfahrtbahnhof in Kourou (Französisch-Guayana) ab. Sie setzte exakt 20 Minuten und 56 Sekunden danach einen Astra-Satelliten der Luxemburger Betreibergesellschaft Societe Europeenne des Satellites (SES) aus. Astra 1D ist der vierte Satellit der SES zur Übertragung von Fernseh- und Hörfunkprogrammen. Er hat 18 Transponder, die über vier verschiedene Frequenzbänder bis zu 66 Kanäle nutzbar machen können; 14 davon werden auf dem D-Band auch als Reserve für bestehende Kanalkapazität benutzt; vier weitere dienen der Erprobung digitaler Übermittlungstechniken. Die SES gab noch nicht bekannt, wer die neue Kanalkapazität wird nutzen können. Die Digitaltechnik gilt als das Übertragungsverfahren der Zukunft. Die SES wird voraussichtlich im Juni 1995 mit Astra 1E einen ausschließlich für diese Technik reservierten Satelliten auf der Astra-Position 19,2 Grad Ost im Weltraum in rund 36 000 Kilometer Höhe geostationär positionieren. Auf dieser Position befinden sich alle Satelliten dieses Typs der SES. Um die Digitaltechnik zu forcieren, hat die Luxemburger Gesellschaft, wie sie in Kourou bekanntgab, mit dem französischen Sender Canale Plus eine Zusammenarbeit gegründet. Der technische Direktor der SES, Azevedo, sagte, es beginne hier eine neue Ära des Fernsehens. Für die Betreibergesellschaft Ariane Space, die die Ariane in den Weltraum brachte, sagte ihr Präsident Bigot: Einen solchen Bilderbuchstart hat das Unternehmen bisher nicht verzeichnet. Es war der 69. Start einer Ariane, und der sechste allein in den vergangenen fünf Monaten. Bis Ende 1995 sind weitere 14 Starts geplant; Ariane Space ist die weltweit am meisten gebuchte Raketenbetreibergesellschaft.

4.2.4 Statistik und Erläuterungen zu Text 1

```
Raketenstart: 39 getaggte Nomina
   Raumfahrt: l: 7 g: 159 erw: 0 MI : 3.885 TS :0.0031
```

Die Klassifikation erfolgt problemlos aufgrund der Schlüsselwörter (*Satellit*, *Rakete*) im Text. Es gibt keine größere Gruppe von Nomina, die — bezogen auf die vorgegebenen vier Themenbereiche — auf ein anderes Thema als Raumfahrt hinweist. Die Werte für Transinformation und Tschebyschowrisiko sind hier hochsignifikant. Eine Hinzuziehung der im Text vorkommenden präfigierten Nomina, Nominalkomposita und Eigennamen würde das Ergebnis zusätzlich bestätigen.

4.2.5 Text 2 (Süddeutsche Zeitung, 24.3.1995): Doppelklassifikation

Fall Orlandi: Vatikan sollte erpreßt werden. Rom (dpa) - Im mysteriösen Fall um die Entführung der Tochter eines Vatikanangestellten sollte der Kirchenstaat angeblich um 40 Milliarden Lire (etwa 35 Millionen Mark) erpreßt werden. Schlüsselfiguren in der Affäre sind nach Berichten vom Donnerstag der Priester und Caritasdirektor Don Tonino Intiso sowie der Anwalt Matteo Starace. Beide waren festgenommen worden. Den unbestätigten Angaben zufolge versuchte der Priester, den Vatikan mit der Behauptung zu erpressen, die vor zwölf Jahren verschwundene Emanuela Orlandi sei die Tochter eines Kardinals. Das Mädchen war am 22. Juni 1983 nicht mehr nach Hause zurückgekehrt und ist bis heute spurlos verschwunden. Seitdem ranken sich unzählige Gerüchte und Spekulationen um den Fall. Die Tageszeitung Il Messaggero berichtete, der Vatikan verhandle mit den Entführern des Mädchens, das angeblich noch am Leben sei und ein fünfjähriges Kind habe. Als Beleg für diese Theorie wurde der jetzt festgenommene Don Tonino zitiert. Nach anderen Gerüchten lebt Emanuela heute in Kolumbien und hat zwei Kinder. Sie sei nicht entführt worden, sondern habe vielmehr wegen eines Verhältnisses mit einem Geistlichen freiwillig ihr Elternhaus verlassen. Das Verbrechen wird immer wieder auch in Zusammenhang mit dem Attentat auf den Papst

im Mai 1981 gebracht, da Unbekannte in Anrufen bei den Eltern die Freilassung des Türken Ali Agca gefordert hatten.

4.2.6 Statistik und Erläuterungen zu Text 2

```
Vatikangerüchte: 35 getaggte Nomina
   Kriminalität: l: 7 g: 2807 erw: 2 MI : 1.163 TS :0.0917
   Religion: l: 4 g: 412 erw: 0 MI : 2.522 TS :0.0236
```

Der Text ist beiden ermittelten Themenbereichen zuzuordnen. Es handelt sich sowohl um einen Text zum Vatikan, d. h. zum Themenbereich Kirche, als auch um einen Text aus dem Bereich Kriminalität. Der statistische Wert für Kriminalität ist nicht sehr hoch, bedingt durch die mehrmalige Verwendung von Verben zur Beschreibung der Verbrechensspekulationen, die bei der Auswertung nicht berücksichtigt wurden.

4.2.7 Text 3 (Süddeutsche Zeitung, 20.04.1995): Fragliche Klassifikation

Späte Ehre für eine noble Wissenschaftlerin. Marie Curie wird als erste Frau wegen ihrer Verdienste im Pariser Pantheon bestattet Paris (AFP) - Erstmals wird am heutigen Donnerstag eine Frau wegen ihrer eigenen Verdienste in das Pantheon, den Ruhmestempel der französischen Nation, aufgenommen: Mehr als 60 Jahre nach ihrem Tod findet die Wissenschaftlerin Marie Curie (1867-1934) einen Platz unter den illustren Franzosen. Die sterblichen Überreste der zweimaligen Nobelpreisträgerin und ihres Mannes Pierre Curie (1859-1906) werden in einer feierlichen Zeremonie in den Ehrentempel überführt, eine der letzten Amtshandlungen des scheidenden Präsidenten Francois Mitterrand. Zu Ehren der aus Polen stammenden Physikerin und Chemikerin kommt auch der polnische Staatschef Lech Walesa nach Paris. Bisher ist im Pantheon nur Sophie Berthelot als Ehefrau des Chemikers Marcelin Berthelot bestattet. Die in Warschau geborene Marie Sklodowka war als junge Studentin nach Frankreich gekommen, um ihr Studium an der Sorbonne-Universität fortzusetzen. Kurze Zeit später heiratete sie den französischen Physiker Pierre Curie, mit dem sie vor allem wegen ihrer Arbeiten über die Radioaktivität Weltruhm erlangte. Für die Entdeckung des Radiums erhielt das Ehepaar 1903 zusammen mit Maries Lehrer Henri Becquerel den Nobelpreis für Physik. Fünf Jahre nach dem Tod ihres Mannes wurde Marie Curie 1911 für ihre Arbeiten auf dem Gebiet der Radiochemie auch mit dem Chemie-Nobelpreis ausgezeichnet. Die Wissenschaftlerin durfte vom Jahr 1903 an als erste Frau an der Sorbonne-Universität lehren, was damals als gesellschaftliche Revolution galt. Die Aufnahme ins Pantheon ist bereits die zweite posthume Ehrung für das Ehepaar innerhalb kurzer Zeit: Erst vor einem Monat brachte die Bank von Frankreich neue 500-Franc-Geldscheine mit den Porträts von Marie und Pierre Curie heraus. Vom ersten Vorschlag, die Wissenschaftler ins Pantheon aufzunehmen, bis zur Umsetzung dauerte es mehrere Jahre. Den Beschluß verkündete Mitterrand am 8. März 1994, dem Internationalen Frauentag. Der Bau im Studentenviertel Quartier Latin war im 18. Jahrhundert ursprünglich als Grabkirche für die Heilige Genoveva geplant. Bereits während der Revolution wurde er zum Ehrentempel für die großen Männer Frankreichs bestimmt, danach aber wieder als religiöser Bau genutzt. Erst seit der Überführung des Schriftstellers Victor Hugo 1885 ist das Pantheon ausschließlich dem Ruhm großer Franzosen gewidmet. In dem Ruhmestempel wird bislang 69 bedeutender Männer gedacht. Nicht alle liegen tatsächlich auch dort begraben. Am Donnerstag werden die beiden Särge von Studenten der Sorbonne in einer kurzen Prozession vor das Pantheon getragen. Dort stehen Reden des PhysikNobelpreisträgers von 1993, Pierre-Gilles de Gennes, und des Staatschefs auf dem Programm. Anschließend werden die Särge in den Ruhmestempel getragen. Dort sind sie für die Öffentlichkeit zu sehen, bevor sie am Wochenende ihren endgültigen Platz in

der Krypta erhalten. An einem Festessen, zu dem Mitterrand die Familie Curie eingeladen hat, will auch die über 90 Jahre alte Tochter des Ehepaares teilnehmen, die in den USA lebt. Für den Staatschef, der im Mai nach 14 Jahren den Elysee-Palast verläßt, ist die Zeremonie vier Tage vor den Präsidentschaftswahlen von symbolischer Bedeutung: Mitterrand hatte kurz nach seiner Wahl im Mai 1981 im Pantheon rote Rosen auf drei Gräber niedergelegt - zur Erinnerung an den Sklavenbefreier Victor Schoelcher, den Widerstandsführer Jean Moulin und den 1914 ermordeteten Sozialistenchef Jean Jaures.

4.2.8 Statistik und Erläuterungen zu Text 3

```
Curie_im_Pantheon: 99 getaggte Nomina
   Religion: l: 6 g: 412 erw: 0 MI : 2.063 TS :0.0277
```

Text 3 ist keinem der vier berücksichtigten Themenbereiche zuzuordnen. Einige Nomina im Text (*Krypta*, *Prozession*, *Zeremonie*) entstammen aber dem Themenbereich Religion. Aufgrund dieser Nomina erfolgt die Zuordnung des Textes zu diesem Themenbereich, was trotz der kurzen Thematisierung der Geschichte des Pantheons als Kirchenbau im Text zumindest fraglich erscheint.

4.2.9 Text 4 (Süddeutsche Zeitung, 10.02.1995): Fehlerhafte Klassifikation

Das Monopol der sieben schönen Schwestern zeigt Risse. Internationale Modeschöpfer streiten sich um den Stellenwert ihrer Topmodels / Verächtliches über Kälber und Vampire. Die aktuellen Haute-Couture-Schauen in Paris und Rom haben den Streit über hochbezahlte Top-Models neu entfacht. Als Kälber bezeichnet sie verächtlich der französische Designer Paco Rabanne, für Pierre Cardin sind sie eine Schande und für Emanuel Ungaro gar Vampire. Karl Lagerfeld ist anderer Meinung: Schöne Frauen sind schön, schöne Kleider sind schön, aber beide zusammen sind noch schöner. Die amerikanische Presse verbannte die Idole der neunziger Jahre inzwischen auf die Out-Liste. Der italienische Modeschöpfer Valentino kaufte daraufhin eine ganze Anzeigenseite in der Tageszeitung International Herald Tribune und schrieb unter die Photos von Claudia Schiffer, Nadja Auermann und Ellen MacPherson: Suzy, Du hast Dich in allem geirrt, aber wir mögen Dich trotzdem. Adressatin war die Päpstin unter den Modejournalisten, Suzy Menkes. Nun holt auch Gianni Versace zum Gegenschlag aus. Unser Beruf sollte auch dazu dienen, Menschen träumen zu lassen, erklärt er in der jüngsten Ausgabe des italienischen Nachrichtenmagazins L'Espresso. Der italienische Star-Designer gilt als Vater des Model-Mythos. Nach Versaces Ansicht zeichnet sich ein neuer Typ von Frau ab: Eleganter, weiblicher, sinnlicher. Seine beiden Neuentdeckungen, Christen MacMenamy und Shalom Harlow, sollen zur Jahrtausendwende die Schönheitsidole der fünfziger und sechziger Jahre wiederaufleben lassen. Christen ist für Versace eine moderne Suzy Parker, und Shalom erinnert ihn an Audrey Hepburn. Noch ist eine Narbe auf dem Bauch von Linda Evangelista dem italienischen Fernsehen eine Meldung in den Hauptnachrichten wert, aber das Monopol der sieben Schwestern, wie die Top-Stars der Mode gerne genannt werden, zeigt Risse. So avancierte die 20jährige Berlinerin Nadja Auermann im Laufe von wenigen Monaten zum inzwischen meistgefragten Fotomodell. Linda Evangelista, Naomi Campbell, Claudia Schiffer, Helene Christensen, Christy Turlington, Cindy Crawford und Carla Bruni bekommen Konkurrenz. Hunderte junge Kolleginnen, ohne Starallüren und nicht so teuer, warten nur auf ihre Chance. Im Gepräch sind vor allem die Deutsche Georgia Göttmann, die Texanerin Bridget Hall, die Französin Chrystele, die Holländern Fanke Jannsen, die Amerikanerin Irish Goff und die Russin Natalia Samanova - sie wird im Jahr 2000 gerade 20 Jahre alt sein.

Themengebiet	**Klassifikationsart**			
	Manuell	Automatisch		
		richtig	falsch	nicht erkannt
Kriminalität	31	27	0	4
Religion	14	14	3	0
Raumfahrt	12	11	0	1
Medizin	11	9	0	2
GESAMT KLASSIFIZIERT	68	61	3	7

Tabelle 4.1: Ergebnisse der automatischen Klassifizierung von 600 Texten

4.2.10 Statistik und Erläuterungen zu Text 4

```
Modeschöpferstreit: 54 getaggte Nomina
   Religion: l: 3 g: 412 erw: 0 MI : 1.851 TS :0.0733
```

Die (statistisch nicht sehr signifikante, aber doch über den festgelegten Schwellenwerten liegende) Zuordnung des Textes zum Bereich Religion würde bei einer Definition des Themenbereichs Mode sicherlich gegenüber der richtigen Zuordnung in den Hintergrund treten. Die fehlerhafte Klassifikation tritt nur auf, weil dreimal das (ambige) Lexem *Schwester* im Text auftritt, das unter anderem als Religiöser Beruf kodiert ist.

4.2.11 Zuordnung von Referenztexten

Aus dem Korpus wurden die ersten 600 Texte ausgewählt, die mehr als dreißig getaggte Nomina enthielten, und von Hand klassifiziert. Diese Klassifizierung wurde mit der automatischen Zuordnung der Texte zu den genannten Themenbereichen verglichen, um herauszufinden, wie hoch die Fehlerquote bei der statistischen Zuordnung war. Die folgende Tabelle listet die Ergebnisse auf. Angegeben sind zu jedem Themengebiet:

- Die Zahl der zugeordneten Texte aufgrund der manuellen Klassifikation
- Die Ergebnisse der automatischen Klassifikation mit der Zahl der richtig und der fälschlich zugeordneten Texte sowie der nicht erkannten Texte, die aber zu dem Themengebiet gehören.

Wie deutlich wird, sind die Ergebnisse — verglichen mit dem Aufwand — relativ gut. Obwohl keinerlei Disambiguierung vorgenommen wurde und nur ein Bruchteil der Nomina der klassifizierten Texte semantisch getaggt war, kam es zu einer sehr hohen Trefferquote von ca. 90

4.2.12 Verwendung der semantischen Klassen für das Information Retrieval

Die Untersuchung zeigte, daß die semantischen Klassen im CISLEX als Grundlage für das Information Retrieval geeignet sind. Zur Erzielung brauchbarer Ergebnisse für ein Themengebiet sind, aufbauend auf der vorhandenen semantischen Kodierung, folgende Schritte notwendig:

- Inhaltliche Eingrenzung des fraglichen Themengebiets unter Vergleich mit angrenzenden Themengebieten
- Ermittlung der themenrelevanten Lexeme
- Aufbauend auf einer Liste von themenrelevanten Lexemen die Auswahl der für das Thema relevanten semantischen Klassen.

Zur Verbesserung der Ergebnisse könnten noch folgende zusätzliche Maßnahmen beitragen:

- Gewichtung der für ein Themengebiet ausgewählten Klassen bezüglich ihrer Relevanz für das Thema

- Kontextuelle Disambiguierung polysemer Nomina u. a. durch die Berücksichtigung der Argumentstruktur von Verben und Adjektiven
- Vermeidung der Berücksichtigung von zu Nomina homonymen Eigennamen durch einen Abgleich mit dem Eigennamenlexikon und durch die Berücksichtigung der Nominalgruppensyntax
- Aufstellung weiterer themenspezifischer semantischer Klassen.

4.3 Selektionsklassen und thematische Klassifizierung

Die weitgehend zufriedenstellenden Ergebnisse beider Untersuchungen zeigen, daß es möglich ist, die semantischen Klassen, die für die Nomina des CISLEX kodiert wurden, sowohl für die Beschreibung von Selektionsregularitäten von Operatoren als auch zur thematischen Einordnung eines Textes heranzuziehen. Somit sind diese Klassen sowohl auf der Mikroebene von Texten zur Beschreibung sprachlicher Regularitäten, als auch auf der Makroebene zur Inhaltsbestimmung von Dokumenten verwendbar. Dies muß nicht heißen, daß diese Multifunktionalität auf jede einzelne der semantischen Klassen zutrifft. Es gibt unter den kodierten Klassen solche, die sich eher für die Beschreibung von Selektionsregularitäten eignen, und andere, die in erster Linie Relevanz für die Ermittlung von thematischen Bezügen haben. Die Schnittmenge beider Typen von Klassen ist allerdings relativ groß. Dies läßt sich relativ deutlich an den statistisch signifikanten Klassen ablesen, die in der ersten Untersuchung auftraten: Die Möglichkeit zur Verwendung einer Mehrzahl von ihnen zur thematischen Einordnung von Texten läßt sich unschwer erkennen.

Andererseits ist es nötig, für zu ermittelnde Themen eine andere Strukturierung der semantischen Klassen vorzunehmen als dies für Selektionsklassen notwendig ist. Für die Beziehung der Selektionsklassen untereinander ist neben der Hierarchisierung reiner Selektionsklassen v. a. die hyponymische Struktur des Wortschatzes ausschlaggebend, da die meisten der taxonomischen Klassen auch in bezug auf Selektionspräferenzen relevant sind. Zwar kann diese hyponymische Struktur zumindest auch für die Ermittlung der Unterklassen einer themenrelevanten semantischen Klasse herangezogen werden, zusätzlich nötig für die thematischen Einordnung ist allerdings eine Strukturierung der Hyponymieklassen in der Art eines thematischen Thesaurus. Eine solche Strukturierung läßt sich nur auf Grundlage einer bestimmten Aufgabenstellung vornehmen — d. h. aufgrund einer Spezifikation der relevanten Themengebiete für die gewünschte Einordnung von Dokumenten.

Kapitel 5

Zusammenfassung und Ausblick

5.1 Ergebnisse

5.1.1 Struktur des Teilwortschatzes der einfachen Nomina

Der Ausgangspunkt zu den Ausführungen in dieser Arbeit war die Einteilung aller 'einfachen Nomina' des Deutschen in semantische Klassen. Die gewonnenen Erkenntnisse beziehen sich von daher zunächst auf die semantische Struktur dieses morphologisch definierten Ausschnitts des deutschen Nominallexikons.

Es konnte gezeigt werden, daß der Nominalwortschatz keine homogene Struktur besitzt, d. h. er läßt sich nicht in semantische Klassen gleichen Typs von etwa gleicher Größe und gleicher Relevanz für sprachliche Phänomene untergliedern. Relativ großen, bezüglich einer Auswahl linguistischer Kriterien (v. a. hyponymische Zuordnung und Selektionspräferenzen) homogenen Klassen stehen minimale Klassen von wenigen Lexemen und vollkommen idiosynkratische Lexeme gegenüber. Dasselbe gilt bei der Hinzuziehung der genannten Unterteilungskriterien für die interne Strukturierung der meisten größeren der kodierten semantischen Klassen. Hier lassen sich ebenfalls einige umfangreichere homogene Unterklassen festlegen, doch am Ende der Unterteilung bleibt ein Rest an Lexemen übrig, den weiter in semantische Klassen einzuordnen nicht möglich ist oder der nur die Bildung extrem kleiner kohärenter Klassen erlaubt . So konnte in der Klasse der Berufe gezeigt werden, daß hier neben Berufsbezeichnungen im engeren Sinn, welche die Mehrzahl der Lemmata mit diesem Kode ausmachen, eine Reihe von Nomina auftauchen, die nur gewissen der ermittelten Kriterien für typische Berufsbezeichnungen entsprechen und die kaum eine Zusammenfassung zu kohärenten Klassen erlauben.

Im Zusammenhang mit der inhomogenen Struktur des kodierten Wortschatzes konnte ferner gezeigt werden, daß es keine durchgehend gleichermaßen wichtigen Aufstellungs- und Abgrenzungskriterien für semantische Klassen gibt. Vielmehr hatten die einzelnen Teilaspekte aus dem Bündel der definierten Abgrenzungskriterien — geteilte Selektionskontexte und hyponymische Struktur — für unterschiedliche Klassen eine unterschiedlich gewichtete Relevanz. Während für eine Klasse wie die Tiere die biologische Klassifikation der Tierarten die Grundlage für eine erste Einteilung lieferte und es kaum möglich war, eine relevante Menge geteilter Selektionskontexte für diese Unterklassifizierung heranzuziehen, waren für größere Teile der Fahrzeuge sprachliche Kriterien — d. h. typische Kontexte, in denen die Bezeichner für Fahrzeuge auftraten — auch für die taxonomische Einteilung verwendbar. In anderen Wortschatzbereichen (z. B. den Formen) war nur mehr eine rudimentäre taxonomische Strukturierung zu erkennen, und eine Einteilung war ausschließlich aufgrund geteilter Selektionskontexte von Lexemen möglich — allerdings auch dies oft nur unter Schwierigkeiten. In der Klasse der Formen gingen diese Schwierigkeiten teilweise so weit, daß der Sinn einer Einteilung in semantische Klassen — gleich welchen Typs — überhaupt in Frage zu stellen war. Der unterschiedlichen Gewichtung der Kriterien zur Aufstellung und Abgrenzung semantischer Klassen wurde Rechnung getragen, indem den semantischen Klassen Metaeigenschaften zugewiesen wurden,

die sich auf die Aufstellungskriterien und auf ihre interne Struktur beziehen.

Es wurden verschiedene Typen der Unterordnung von semantischen Klassen unterschieden. Unterordnung im strikten Sinne wurde von der bedingten und der metonymischen Unterordnung getrennt. Die Einordnung der Lexeme in die semantischen Klassen und die Hierarchisierung der semantischen Klassen ergaben keine Baumstruktur. Zahlreiche Lexeme sind — in ein und derselben Bedeutung — in mehrere Klassen einzuordnen (so *Katze* zu Säugetieren und Haustieren), einige der kodierten Klassen haben mehrere übergeordnete Knoten. Zyklische Pfade können innerhalb der ermittelten Struktur hingegen nicht auftreten.

Einige auffallende Einzelergebnisse, die sich bei der Gliederung des Wortschatzes ergaben, sind die folgenden:

- Ca. 7 700 der 38 000 einfachen Nomina bezeichnen Menschen. Die größte Untergruppe davon, die durch einige typische Kontexte abgrenzbar war, sind mit ca. 2 500 Lemmata die Berufsbezeichnungen.
- Erwartungsgemäß treten im kodierten Wortschatz besonders viele einfache Nomina auf, die Dinge oder Sachverhalte bezeichnen, die der unmittelbaren sozialen oder natürlichen Umwelt des Menschen entstammen oder die für Menschen anderweitig relevant sind. So sind unter den 870 Tierbezeichnungen hauptsächlich Bezeichner von Säugetieren (269), Vögeln (145) und Fischen (62), während es für niedere Tiere (44) wesentlich weniger Ausdrücke gibt (obwohl es hier - biologisch gesehen - eine wesentlich größere Artenvielfalt gibt; für diese Tiere gibt es aber in der Mehrzahl keine morphologisch einfachen Ausdrücke). Es gibt unter den ca. 1500 Stoffbezeichnungen immerhin 148 Ausdrücke für Kleiderstoffe; unter den 181 Bezeichnungen für Flüssigkeiten sind 152 auch Bezeichnungen für Getränke, davon bezeichnen 108 (71
- Eine Grobklasse mit gemeinsamen linguistischen Eigenschaften, die man als Abstrakta bezeichnen könnte, konnte nicht ermittelt werden. Abstraktheit als graduelle Eigenschaft läßt sich für sprachliche Ausdrücke definieren als die Unmöglichkeit, den Ausdruck mit Operatoren zu kombinieren, die für den Kernbestand der Konkreta typische sprachliche Umgebungen ausmachen. Konkrete Nomina im engeren Sinne sind die Nomina, die Objekte mit räumlicher Ausdehnung bezeichnen.

5.1.2 Semantische Beschreibungselemente

Neben der Kodierung der hyponymischen Wortschatzstruktur und von selektionsrelevanten Lexemgruppen mittels des Beschreibungselementes 'semantische Klasse' wurden eine Reihe anderer Beschreibungseinheiten für die semantische Kodierung der Nomina eingeführt.

- Die semantischen Relatoren erlauben es, semantische Relationen und Ableitungsbeziehungen zwischen Bedeutungen zu kodieren. Insbesondere konnten so primär meronymisch strukturierte Wortschatzbereiche semantisch beschrieben, und die Bedeutung von Diminutiva und movierten Lexemen unter Berücksichtigung der semantischen Beziehung zum Basislexem effizient kodiert werden.
- Die Auszeichnung relationaler Nomina bietet die Grundlage für eine spätere Beschreibung der Argumentstruktur.
- Die Kodierung von Synonymen und Oppositionen ergänzt die Darstellung der hyponymischen Struktur des Nominalwortschatzes durch die Beschreibung der semantischen Relationen innerhalb kleinster Kohyponymklassen.
- Die Bezeichnung von Varianten erlaubt eine effiziente Kodierung von bedeutungsgleichen Formen.

5.1.3 Selektionsklassen und thematische Gliederung des Wortschatzes

Als ein wesentliches Ergebnis der Arbeit läßt sich festhalten, daß die Aufteilung des nominalen Wortschatzes in Klassen von Kohyponomen und in thematische Klassen sowie die Zuweisung zu

Selektionsklassen keineswegs unabhängig voneinander vorgenommen werden müssen. Bei der lexikographischen Beschreibung kann eine Aufteilung nach beiden Kriterien weitgehend parallel erfolgen. Die Brücke zwischen beiden Gliederungsaspekten ist hierbei die hyponymische Struktur des Wortschatzes. Einerseits operiert die Selektion durch Verben und andere Operatoren in hohem Maße auf kohyponymen Nomina. Andererseits läßt sich auch ein thematischer Thesaurus ausgehend von Klassen von Kohyponymen aufstellen. Die Festlegung der für ein Themengebiet relevanten Lexeme läßt sich aufgrund der semantischen Klassen leicht erzielen, indem eine themenrelevante Auswahl von ihnen anstatt einzelner Lexeme einem Themengebiet zugeordnet wird.

Der enge Zusammenhang zwischen thematischer und hyponymischer Struktur und Selektionsklassen hat sich vor allem an den positiven Ergebnissen der in Kapitel 4 dargestellten Untersuchung zur thematischen Einordnung von Zeitungstexten bestätigt.

5.1.4 Eignung der Kodierung für computerlinguistische Anwendungen

Die Zuordnung des nominalen Wortschatzes zu semantischen Klassen eignet sich wie gezeigt sowohl zur Beschreibung von Selektionspräferenzen verbaler und adjektivischer Operatoren, als auch zur thematischen Zuordnung von Lexemen. Damit kann ein mit semantischen Klassen versehenes Nominallexikon in verschiedenen computerlinguistischen Anwendungen als Lexikonmodul eingesetzt werden:

- Sowohl in der Sprachgenerierung als auch in der maschinellen Übersetzung mit der Zielsprache Deutsch können Informationen über die Zugehörigkeit von Nomina zu semantischen Klassen herangezogen werden, um die Auswahl geeigneter Verben, Adjektive und anderer Operatoren in der Synthese zu steuern.
- In der Analyse, z. B. im Prozeß der maschinellen Übersetzung mit der Quellsprache Deutsch, können die semantischen Klassen herangezogen werden, um polyseme Verben und Adjektive oder polyseme Nomina selbst zu disambiguieren.
- Im Information Retrieval kann eine geeignete Auswahl semantischer Klassen herangezogen werden, um die in einem Dokument auftretenden Nomina thematisch einzuordnen. Wie in Kapitel 4 gezeigt werden konnte, ist die Zusammenstellung der semantischen Klassen für ein Themengebiet relativ schnell und unkompliziert vorzunehmen, zumindest, soweit es sich nicht um die Klassifizierung fachsprachlicher Texte handelt.
- Im Bereich der Textverarbeitungsprogramme kann die Einteilung in semantische Klassen und die Kodierung von Varianten, Synonymen und Oppositionen zur Konzeption eines Thesaurus eingesetzt werden, der dem Benutzer die Möglichkeit gibt, alternative Ausdrücke zu einem Eingabelexem zu ermitteln.

Um das CISLEX im Zusammenhang mit der semantischen Kodierung für diese Anwendungen voll einsatzfähig zu machen, sind jedoch noch eine Reihe anderer Kodierungsschritte notwendig, eine Aufgabe, die ich im folgenden abschließend ansprechen möchte.

5.2 Weitere Kodierungsschritte

Die Kodierung der semantischen Klassen für die einfachen Nomina im CISLEX war nur ein erster Schritt zu einer umfassenderen semantischen Beschreibung der Lemmata in den verschiedenen Sublexika. Folgende Kodierungsschritte stehen an:

- Die manuelle Kodierung der präfigierten und lexikalisierten komplexen Nomina. Sie kann weitgehend auf Basis der gleichen Kriterien und durch die Zuordnung zu denselben semantischen Klassen erfolgen, wie sie für die einfachen Nomina angewendet wurden.
- Die automatische oder halbautomatische Kodierung der komplexen Nomina mit kompositioneller Gesamtbedeutung. Das gesamte Lexikon der komplexen Nomina läßt sich aufgrund

der (prinzipiell unbegrenzt) großen Zahl an Lemmata nicht manuell kodieren. Es müssen von daher Strategien für eine automatische semantische Kodierung entwickelt werden. Die Kodierung sollte unter Hinzuziehung der in Kapitel 4 geschilderten Untersuchung, einer Auflistung der stark reihenbildenden Kopfnomina und weiterer semantischer Analysen erfolgen.

- Eine feinere Aufteilung der kodierten semantischen Klassen aufgrund der Selektionspräferenzen von Verben und Adjektiven, u. U. unter Berücksichtigung konkreter Anforderungen von Seiten einer Anwendung im Bereich des Information Retrieval. Zudem sollte eine Untergliederung der nominalen Klassen Ereignisse, Zustände und Eigenschaften bei gleichzeitiger Kodierung der entsprechenden Adjektive und Verben erfolgen.
- Die Aufstellung eines Inventars von semantischen Merkmalen und deren Zuweisung an die einzelnen Lexeme der semantischen Klassen und an vollständige Klassen. Dabei sollten Vererbungsregeln für die Elemente einer semantischen Klasse unter Berücksichtigung der Klassenmerkmale aufgestellt werden.
- Die Beschreibung der Argumentstruktur relationaler Nomina. Dazu gehört die Angabe der Zahl der Argumente und die Kodierung der syntaktischen und semantischen Füllung der Argumentstellen.
- Die Kodierung der Wortarten Adjektiv und Verb.

Der erste der genannten Kodierungsschritte ist schon teilweise in Angriff genommen. Seine Durchführung bereitet auf Basis der bisherigen Kodierung keine weiteren Schwierigkeiten. Auch für die weiteren Phasen der Kodierung legt die bisher geleistete Arbeit einige Grundlagen: Die vorhandene Einteilung kann, wie gezeigt wurde, zur Analyse von Nominalkomposita herangezogen werden und erlaubt in der vorliegenden Form bereits die Beschreibung semantischer Teilaspekte der Argumentstruktur von Verben und Adjektiven. Die Metaeigenschaften semantischer Klassen bieten für die Zuweisung von Default-Merkmalen eine erste Grundlage.

Bibliographie

Bei Titeln mit Verlagsangabe Cmp-lg handelt es sich um Artikel aus dem elektronischen Archiv der Association for Computational Linguistics (ACL) (abrufbar unter *ftp: xxx.lanl.gov* oder unter *http: xxx.lanl.gov/form/cmp-lg*).

Aarts, Jan, Joseph P. Calbert (1979): Metaphor and Non-Metaphor. The Semantics of Adjective-Noun Combinations. Tübingen: Niemeyer. (= Linguistische Arbeiten 74).

Aarts, Jan, Willem Meijs (Hrsg.) (1990): Theory and Practice in Corpus Linguistics. Amsterdam: Editions Rodopi. (= Language and Computers: Studies in Practical Linguistics 4).

Agarwal, Rajeev (1995): Evaluation of Semantic Clusters. erscheint in: ACL Proceedings 1995.

Aitchison, Jean (1994): Words in the Mind. An Introduction to the Mental Lexicon. Oxford/Cambridge(Mass.): Blackwell.

al-Wadi, Doris (1994): COSMAS Ein Computersystem für den Zugriff auf Textkorpora. Version R.1.3-1: Benutzerhandbuch. Mannheim: Institut für Deutsche Sprache.

Amsler, Robert A. (1981): A Taxonomy for English Nouns and Verbs. In: Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics. Stanford: ACL, S. 133-138.

Amsler, Robert A. (1994): Research Toward the Development of a Lexical Knowledge Base for Natural Language Processing. In: Zampolli/Calzolari/Palmer: Current Issues in Computational Linguistics. Pisa/Dordrecht: Giardini/Kluwer, S. 155-175. (= Linguistica Computazionale IX, X).

Association for Computational Lingustics (ACL), (Hrsg.) (1992): Third conference on Applied Natural Language Processing (Trento). Proceedings of the Conference. Trient: ACL.

Atkins, Beryl T., Antonio Zampolli (Hrsg.) (1994): Computational Approaches to the Lexicon. Oxford: Oxford University Press.

Augst, Gerhard (1975): Lexikon zur Wortbildung. Morpheminventar. Tübingen: Narr.

Augst, Gerhard (1975a): Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache. Tübingen: Narr.

Basili, Roberto, Maria Teresa Pazienza, Paola Velardi (1991): Combining NLP and Statistical Techniques for Lexical Acquisition. In: Proceedings of the AAAI Fall Symposion on Probabilistic Approaches to Natural Language. Cambridge: Massachusetts.

Basili, Roberto, Maria Teresa Pazienza, Paola Velardi (1992): Computational Lexicons: The Neat Examples and the Odd Exemplars. In: Association for Computational Linguistics: Third Conference on Applied Natural Language Processing, S. 96-103.

Basili, Roberto, Maria Teresa Pazienza, Paola Velardi (1994): Lexicon Acquisition for Real Natural Language Processing Systems. Cambridge: Cambridge University Press.

Bátori, István, Winfried Lenders, Wolfgang Putschke (Hrsg.) (1989): Computational Linguistics/Computerlinguistik. An International Handbook on Computer Oriented Language Research and Applications. Berlin/New York: Walter de Gruyter.

Beckwith, Richard, Christiane Fellbaum, Derek Gross, George Miller (1991): WordNet: A lexical database organized on psycholinguistic principles. In: Uri Zernik: Lexical Acquisition. Hillsdale u.a.: Erlbaum, S. 211-232.

Behrens, Leila (1993): Lexikalische Amgiguität und Disambiguierung im Kontext der maschinellen Übersetzung. In: Lutzeier: Studien zur Wortfeldtheorie. Tübingen: Niemeyer, S. 251-268.
Berlin, Brent, Paul Kay (1969): Basic Color Terms. Berkeley: University of California Press.
Bierwisch, Manfred (1983): Semantische und konzeptuelle Repräsentation lexikalischer Einheiten. In: Ruzicka/Motsch: Untersuchungen zur Semantik. Berlin: Akademie Verlag, S. 91-99. (= Studia Grammatica 22).
Bierwisch, Manfred (1989): Event-Nominalizations. Proposals and Problems. In: Linguistische Studien Reihe A 194. Berlin: Akademie Verlag, S. 1—73.
Boguraev, Branimir (1994): Machine-Readable Dictionaries and Computational Linguistics Research. In: Zampolli/Calzolari/Palmer: Current Issues in Computational Linguistics. Pisa/ Dordrecht: Giardini/Kluwer, S. 119-154.
Boguraev, Branimir, Ted Briscoe (Hrsg.) (1989): Computational Lexicography for Natural Language Processing. London: Longman.
Bohnhof, Anne (1993): Klassifizierung von Berufsbezeichnungen. Abschlußarbeit zum Studiengang Computerlinguistik. München: CIS-Berichte.
Brachmann, Ronald J. (1985): An Overview of the KL-ONE Knowledge Representation System. In: Cognitive Science 9, S. 171-216.
Braden-Harder, Lisa, Wlodek Zadrozny (1991): Lexicons for Broad Coverage Semantics. In: Uri Zernik: Lexical Acquisition. Hillsdale u.a.: Erlbaum, S. 369-388.
Breidt, Elisabeth (1993): Extraction of V-N-Collocations from Text Corpora. A Feasibility Study for German. In: ACL: Workshop on Very Large Corpora.
Brown, Cecil H. (1990): A survey of category types in natural language. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 17-47.
Brown, Peter F. u.a. (1990): A Statistical Approach to Machine Translation. In: Computational Linguistics 16/2, S. 79-85.
Brown, R. (1958): How Shall a Thing Be Called. In: Psychological Review 65, S. 14-21.
Bundesanstalt für Arbeit, (1988): Klassifizierung der Berufe. Systematisches und alphabetisches Verzeichnis der Berufsbenennungen. Bonn: Bundesanstalt für Arbeit.
Bundesanstalt für Arbeit, (1992): Schlüsselsystem für die computergestützte Arbeitsvermittlung in den Arbeitsämtern. Bonn: Bundesanstalt für Arbeit.
Bußmann, Hadumod (1990): Lexikon der Sprachwissenschaft. Stuttgart: Kröner.
Byrd, Roy J. (1994): Discovering Relationships among Word Senses. In: Zampolli/Calzolari/Palmer: Current Issues in Computational Linguistics. Pisa/ Dordrecht: Giardini/Kluwer, S. 177-199.
Calzolari, Nicoletta (1991): Lexical Databases and Textual Corpora. Perspectives of Integration for a Lexical Knowledge Base. In: Uri Zernik: Lexical Acquisition. Hillsdale: Erlbaum, S. 191-208.
Calzolari, Nicoletta (1994): Issues for Lexicon Building. In: Zampolli/Calzolari/Palmer: Current Issues in Computational Linguistics. Pisa/ Dordrecht: Giardini/Kluwer, S. 267-281.
Carlson, Greg N. (1991): Natural Kinds and Common Nouns. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 370-398.
Chodorow, M.S., R.J. Byrd, G.E. Heidorn (1985): Extracting Semantic Hierarchies from a Large On-Line Dictionary. In: ACL: Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics. Chicago: ACL.
Church, Kenneth, William Gale, Patrick Hanks, Donald Hindle (1991): Using Statistics in Lexical Analysis. In: Uri Zernik: Lexical Acquisition. Hillsdale: Erlbaum, S. 115-164.
Church, Kenneth Ward, Patrick Hanks (1990): Word Association Norms, Mutual Information, and Lexicography. In: Computational Linguistics 16 (1), S. 22-29.
Church, Kenneth Ward u.a. (1994): Lexical substitutability. In: Atkins/Zampolli: Computational Approaches to the Lexicon. Oxford: Oxford University Press, S. 153-177.
Clas, André, H. Safar (Hrsg.) (1992): L'environnement traductionnel. Sillery: Presses de l'Université de Québec.
Copestake, Ann (1995): The Representation of Group Denoting Nouns in a Lexical Knowledge Base. In: Saint-Dizier/Viegas: Computational Lexical Semantics. Cambridge: Cambridge University Press, S. 207-230.

Courtois, Blandine (1990): Un système de dictionnaires électroniques pour les mots simples du français. In: Langue française 87. Paris: Larousse, S. 11—22.
Cowie, J. (1983): Automatic Analysis of Descriptive Texts. In: Proceedings on the Conference on Applied Natural Language Processing. Santa Monica (Kalif.)), S. 117-123.
Cruse, D. A. (1986): Lexical Semantics. Cambridge: Cambridge University Press.
Cruse, D. A. (1990): Prototype Theory and Lexical Semantics. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 382-402.
Cruse, D. A. (1995): Polysemy and Related Phenomena. In: Saint-Dizier/Viegas: Computational Lexical Semantics. Cambridge: Cambridge University Press, S. 33-49.
Dagan, Ido, Fernando Pereira, Lillian Lee (1994): Similarity-Based Estimation of Word Cooccurence Probabilities. Elektronisches Archiv der ACL: Cmp-lg 9405001.
Dahlgren, Kathleen (1988): Naive Semantics for Natural Language Understanding. Boston u.a.: Kluwer.
Dixon, Robert M. W. (1991): A New Approach to English Grammar, on Semantic Principles. Oxford: Clarendon Press.
Dornseiff, Franz (1970): Der deutsche Wortschatz nach Sachgruppen. Berlin/New York: de Gruyter.
Dowty, David R. (1979): Word Meaning and Montague Grammar. Dordrecht: Reidel.
Drosdowski, Günther u.a. (1977): Duden. Das große Wörterbuch der deutschen Sprache in sechs Bänden. Mannheim/Wien/Zürich: Dudenverlag.
Eglowstein, Howard (1991): Can a Grammar and Style Checker Improve your Writing. In: Byte, 8/1991, S. 238-242.
Ehrlich, Monika (1991): Nominalisierungen. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 441-458.
Fanselow, Gisbert (1981): Zur Syntax und Semantik von Nominalkomposita. Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung des Deutschen. Tübingen: Niemeyer.
Fanselow, Gisbert Staudacher Peter (1991): Wortsemantik. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 53-70.
Fass, Dan (1991): met*: A Method for Discriminating Metonymy and Metaphor by Computer. In: Computational Linguistics 17/1, S. 49-90.
Fetzer, James H. (Hrsg.) (1991): Epistimology and Cognition. Dordrecht u.a.: Kluwer. (= Studies in Cognitive Systems).
Fillmore, Charles J. (1982): Frame Semantics. In: Linguistic Society of Korea: Linguistics in the Morning Calm. Seoul: Hanshin.
Firth, J. R. (1957): Papers in Linguistics, 1934-1951. London: Oxford University Press.
Fleischer, Wolfgang, Irmhild Barz (1992): Wortbildung der deutschen Gegenwartssprache. Tübingen: Niemeyer.
Fluck, Hans-R. (1985): Fachsprachen. Tübingen: Franke.
Gale, William A., Kenneth W. Church, David Yarowsky (1992): A Method for Disambiguating Word Senses in a Large Corpus. In: Computers and the Humanities 26/5-6, S. 415-439.
Garza-Cuarón, Beatriz (1991): Connotation and Meaning. Berlin u.a.: de Gruyter.
Geckeler, Horst (1993): Strukturelle Wortfeldforschung heute. In: Lutzeier: Studien zur Wortfeldtheorie. Tübingen: Niemeyer, S. 11—22.
Geeraerts, Dirk (1990): The lexicographical treatment of prototypical polysemy. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 195-210.
Goddard, Cliff (Hrsg.) (1994): Semantic and Lexical Universals. Amsterdam: Benjamins.
Granger, R. (1977): Foulup: A program that figures out meanings of words from context. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence. Palo Alto: Morgan Kaufmann.
Gross, Gaston (1992): Forme d'un dictionnaire électronique. In: Clas/Safar: L'environnement traductionnel. Sillery: Presses de l'Université de Québec, S. 255-275.
Gross, Gaston (1994): Classes d'objets et description des verbes. In: Langages 115. Paris: Larousse, S. 15-30.

Guenthner, Franz, Petra Maier (1994): Das CISLEX-Wörterbuchsystem. München: Universität München. (= CIS Bericht 94-76).
Guenthner, Franz (1989): Discourse: Understanding in Context. In: Schnelle/Bernsen: Logic and Linguistics. Research Directions in Cognitive Science Vol 2. Hove/London/Hillsdale: Lawrence Erlbaum, S. 127-142.
Haimann, J. (1980): Dictionaries and Encyclopedias. In: Lingua 50, S. 329-357.
Harlass, Gertrude, Heinz Vater (1974): Zum aktuellen deutschen Wortschatz. Tübingen: Narr. (= Forschungsberichte des Instituts für Deutsche Sprache 21).
Hausmann, Franz Josef, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Hrsg.) (1990): Wörterbücher. Ein internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter.
Hearst, Marti A., Hinrich Schütze (1993): Customizing a lexicon to better suit a computational task. In: Proceedings of the ACL SIGLEX Workshop. Columbus (Ohio).
Heß, Klaus, Jan Brustkern, Winfried Lenders (1983): Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration. Tübingen: Niemeyer. (= Sprache und Information 6).
Heyn, Matthias (1992): Zur Wiederverwendung maschinenenlesbarer Wörterbücher. Eine computergestützte metalexikographische Studie am Beispiel der elektronischen Edition des 'Oxford Advanced Learner's Dictionary of Current English'. Tübingen: Niemeyer. (= Lexicographica Series Maior 45).
Hindle, Donald (1990): Noun classification from predicate-argument structures. In: ACL Proceedings, 28th Annual Meeting. Pittsburgh: ACL, S. 268-275.
Hindle, Donald, Mats Rooth (1993): Structural Ambiguity and Lexical Relations. In: Computational Linguistics 19/1, S. 103-120.
Hjelmslev, Louis (1959): Pour une sémantique structurale. In: Essais Linguistiques. Kopenhagen: Cercle Linguistique de Copenhague.
Hoppenbrouwers, Geer A. C., Pieter A. Seuren, Antonius J. M. M. Weijters (1985): Meaning and the Lexicon. Proceedings of the Second International Colloquium on the Interdisciplinary Study the Semantics of Natural Language held at Cleves. Dordrecht: Foris Publ..
Huckle, Christopher C. (1995): Grouping Words Using Statistical Context. Elektronsisches Archiv der ACL: Cmp-lg 9502034.
Hutchins, W. John (1978): Languages of Indexing and Classification. A Linguistic Study of Structures and Functions. Stevenage: Peregrinus.
Hutchins, W. John (1986): Machine Translation: Past, Present, Future. Chichester: Ellis Horwood.
Jacobs, Joachim (1994): Kontra Valenz. Trier: WVT.
Jacobs, Paul S. (1987): A Knowledge Framework for Natural Language Analysis. In: Proceedings of IJCAI 87. Mailand: IJCAI.
Jacobs, Paul S. (1991): Making Sense of Lexical Acquisition. In: Uri Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 29-44.
Jacobs, Paul S., U. Zernik (1988): Learning phrases from text. A case study. In: Proceedings of the 7th National Conference on Artificial Intelligence. St. Paul: Morgan Kaufmann.
Johnson-Laird, Philip D. (1984): Semantic Primitives or Meaning Postulates. Mental Models or Propositional Representations. In: B.G. Bara/ G. Guida: Computational Models of Natural Language Processing. Amsterdam: North Holland, S. 227-246.
Kamp, Hans, Uwe Reyle (1993): From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Dordrecht: Kluwer.
Katz, Jerrold J. (1972): Semantic Theory. New York: Harper&Row.
Katz, Jerrold J., Jerry A. Fodor (1963): The Structure of a Semantic Theory. In: Language 39, S. 170-210.
Keil, Frank C. (1979): Semantic and Conceptual Development. Cambridge/Mass.: MIT Press.
Kilgariff, Adam (1992): Dictionary Word Sense Distinctions. An Enquiry into their Nature. In: Computers and the Humanities 26/5-6, S. 365-387.
Kilgariff, Adam (1992): Polysemy. Brighton: University of Sussex. (= CSRP 261).
Kirkpatrick, Betty (1992): Roget's Thesaurus of English Words and Phrases. Harlow: Longman.
Kremer, Dieter (1990): Das Wörterbuch der Berufbezeichnungen. In: Hausmann u.a.: Wörterbücher. Berlin/New York: de Gruyter, S. 1248-1254.

Krifka, Manfred (1991): Massennomina. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 390-417.
Krovetz, Robert (1991): Lexical Acquisition and Information Retrieval. In: Uri Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 45-64.
Kuhlen, Rainer (1989): Information Retrieval. Verfahren des Abstracting. In: Batori/Lenders/Putschke: Computational Linguistics. Berlin/New York: Walter de Gruyter, S. 688-696.
Labov, William (1973): The Boundaries of Words and their Meaning. In: Baiey, C-J.N./ Shuy, R.: New ways of Nalyzing Variation in english. Washington: Georgetown University Press.
Lakoff, George (1973): Hedges: A study in meaning criteria and the logic of fuzzy concepts. In: Journal of Philosophical Logic 2, S. 458-508.
Lakoff, George (1987): Women, Fire and Dangerous Things. Chicago/ London: Chicago University Press.
Lakoff, George, Mark Johnson (1980): Metaphors we live by. Chicago: Chicago University Press.
Langacker, Ronald W. (1987): Foundations of Cognitive Grammar. Vol 1 Theoretical Prerequisites. Stanford: Stanford University Press.
Le Pesant, Denis (1994): Les compléments nominaux du verbe lire. Une illustration de la notion de 'classe d'objet'. In: Langages 115. Paris: Larousse, S. 31-46.
Leech, Geoffrey (1975): Semantics. Harmondsworth: Penguin Books.
Lehrberger, John (1988): Machine Translation. Linguistic Charcteristics of MT Systems and General Methodology of Evaluation. Amsterdam/Philadelphia: John Benjamins.
Lehrer, Adrienne (1990): Prototype theory and its implications for lexical analysis. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 368-381.
Lenders, Winfried (1989): Computergestützte Verfahren zur semantischen Beschreibung von Sprache. In: Bátori/Lenders/Putschke: Computational Linguistics/Computerlinguistik. Ein internationales Handbuch, S. 231-244.
Lerner, Jean-Yves, Thomas E. Zimmermann (1991): Eigennamen. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 349-369.
Lewis, David (1970): General Semantics. In: Synthese 22, S. 18-67.
Link, Godehard (1991): Plural. In: Stechow/Wunderlich: Semantik/Semantics. Berlin/New York: Walter de Gruyter, S. 418-440.
Ludewig, Petra (1993): Inkrementelle wörterbuchbasierte Wortschatzerweiterungen in sprachverarbeitenden Systemen . St. Augustin: Infix-Verlag.
Lüdi, Georges (1985): Zur Zerlegbarkeit von Wortbedeutungen. In: Schwarze/ Wunderlich: Handbuch der Lexikologie. Königstein: Athenäum, S. 64-102.
Lutzeier, Peter Rolf (1985): Linguistische Semantik. Stuttgart: Metzler.
Lutzeier, Peter Rolf (Hrsg.) (1993): Studien zur Wortfeldtheorie. Tübingen: Niemeyer.
Lyons, John (1977): Semantics. Cambridge: Cambridge University Press.
Maier, Petra (1995): Lexikon und automatische Lemmatisierung. Dissertation. CIS, Universität München. München: CIS-Bericht 95-84.
Martin, James H. (1991): Representing and Acquiring Metaphor-Based Polysemy. In: Uri Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 389-415.
Mathieu-Colas, Michel (1993): Dictionnaire électronique des mots français à trait d'union. Problèmes de lexicographie informatique. Thèse de doctorat en linguistique. Paris: Unversité Paris XIII Laboratoire de Linguistique Informatique.
McCawley, James D. (1974): Prelexical Syntax. In: Seuren, P.: Semantic Syntax. London: Oxford University Press, S. 29-42.
McMahon, John, F. J. Smith (1994): Structural Tags, Annealing and Automatic Word Classification. Belfast: Ms..
McMahon, John, F. J. Smith (1995): Improving Statistical Language Model Performance to Automatically Generated Word Hierarchies. Elektronisches Archiv der ACL: Cmp-lg 9503011.
McRoy, Susan W. (1992): Using Multiple Knowledge Sources for Word Sense Discrimination. In: Computational Linguistics 18/1, S. 1—30.

Meyer, Ralf (1993): Compound Comprehension in Isolation and in Context. The contribution of conceptual and discourse knowledge to the comprehension of German novel noun-noun compounds. Tübingen: Niemeyer.
Miller, George (1990): Nouns in Word Net. A Lexical Inheritance System. In: International Journal of Lexicography 3/4 (special issue), S. 245-265.
Miller, George A., Walter G. Charles (1991): Contextual Correlates of Semantic Similarity. In: Language and Cognitive Processes 6/1, S. 1—28.
Miller, George u.a. (1990): Wordnet: An on-line lexical database. In: International Journal of Lexicography 3/4 (special issue), S. 235-312.
Miller, George u.a. (1993): An Introduction to Wordnet: An on-line lexical database. Princeton University: Begleitpapier zum WordNet Programm.
Mittelbach, Henning (1992): Statistik. München/Wien: Oldenbourg.
Morris, Jane, Graeme Hirst (1991): Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. In: Computational Linguistics 17(1), S. 21-48.
Oesterle, Jürgen (1994): Syntaktische und semantische Aspekte von Maßkonstruktionen im Deutschen. München: Centrum für Informations- und Sprachverarbeitung. (= CIS-Berichte 94-78).
Osherson, Daniel N. (1981): On the Adequacy of Prototype Theory as a Theory of Concepts. In: Cognition 9, S. 35-58.
Ott, Nikolaus (1992): Statistische Untersuchungen an einsprachigen Zeitungstexten. Heidelberg: Wissenschaftliches Zentrum/ Institut für Wissensbasierte Systeme. (= IWBS Report 233).
Panyr, Jíri, Harald H. Zimmermann (1989): Information Retrieval. Überblick über aktive Systeme und Entwicklungstendenzen. In: Batori/Lenders/Putschke: Computational Linguistics. Berlin/New York: Walter de Gruyter, S. 696-708.
Parikh, Rohit (1994): Vagueness and Utility. The Semantics of Common Nouns. In: Linguistics and Philosophy 17/6, S. 521-535.
Pereira, Fernando, Naftali Tishby, Lillian Lee (1993): Distributional Clustering of English Words. Elektronisches Archiv der ACL: Cmp-lg 9408011.
Pollard, Carl, Ivan A. Sag (1987): Information-Based Syntax and Semantics. Menlo Park: CSLI.
Pollard, Carl, Ivan A. Sag (1994): Head Driven Phrase Structure Grammar. Chicago/London: University of Chicago Press.
Pottier, Bernard (1964): Vers une sémantique moderne. In: TraLiLi II, S. 107-137.
Pulman, S. G. (1983): Word Meaning and Belief. London: Croom Helm.
Pustejovsky, James (1989): Current Issues in Computational Lexical Semantics. In: ACL: Proceedings of the Third European Conference of the Association for Computational Linguistics.. Menlo Park: ACL.
Pustejovsky, James (1991): The Generative Lexicon. In: Computational Linguistics 17/4.
Pustejovsky, James (Hrsg.) (1993): Semantics and the Lexicon. Dordrecht u.a.: Kluwer. (= Studies in Linguistics and Philosphy 49).
Putnam, Hilary (1975): The Meaning of Meaning. In: K. Gunderson: Language, Mind and Knowledge. Minneapolis: University of Minnesota, S. 131-193.
Quine, Willard van Orman (1951): Two dogmas of empiricism. In: Philosphical Review 60, S. 20-43.
Radszuweit, Siegrid, Martha Spalier (1982): Knaurs Lexikon der sinnverwandten Wörter. 20 000 Stichwörter mit ihren Synonymen. München: Knaur.
Ravin, Yael (1993): Grammar Errors and Style Weaknesses in a Text-Critiquing System. In: Jensen/Heidorn/Richardson: Natural Language Processing: The PLNLP Approach. Boston/Dordrecht/London: Kluwer, S. 65-76.
Resnik, Philip Stuart (1993): Selection and Information. A Class-Based Approach to Lexical Relationships. Pennsylvania: Pennsylvania University.
Rhodes, R. (1985): Lexical taxonomies. In: Hoppenbrouwers/Seuren/Weijters: Meaning and the Lexicon.
Ribas, Francesc (1994): An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus. In: COLING-94 Proceedings, S. 769-774.
Ribas, Francesc (1995): On Learning More Appropriate Selectional Restrictions. Elektronisches Archiv der ACL: Cmp-lg/9502009 (erscheint in EACL 95, Proceedings).

Rickheit, Mechthild (1993): Wortbildung. Grundlagen einer kognitiven Wortsemantik. Opladen: Westdeutscher Verlag.
Rieger, Burghard (1984): Unscharfe Wortbedeutungen. Ein quantitatives Verfahren zur lexikalischen Analyse des verwendeten Vokabulars im Rahmen eines Strukturmodells unscharfer (Fuzzy) Semantik. In: Hellmann, Manfred: Ost-West-Wortschatzvergleiche. Tübingen: Narr, S. 293-339.
Rooth, Mats (1994): Two-Dimensional Clusters in Grammatical Relations. Universität Stuttgart: unveröffentlichtes Paper.
Rosch, Eleanor (1975): Cognitive Representation of Semantic Categories. In: Journal of Experimental Psychology-General 204, S. 192-233.
Rosch, Eleanor (1977): Linguistic Relativity. In: Johnson-Laird/Wason: Thinking. Cambridge (Engl.): University Press.
Rosch, Eleanor (1978): Principles of Categorization. In: Rosch E./LLoyd B.B.: Cognition and Categorization. New Jersey: L. Erlbaum.
Rosch, Eleanor u.a. (1976): Basic Objects in Natural Categories. In: Cognitiv Psychology 8, S. 382-439.
Saint-Dizier, Patrick, Evelyne Viegas (Hrsg.) (1995): Computational Lexical Semantics. Cambridge: University Press.
Schabes, Yves, Michael Roth, Randy Osborne (1993): Parsing the Wall Street Journal with the Inside-Outside Algorithm. In: EACL: EACL 1993: Proceedings.
Schwarze, Christoph, Dieter Wunderlich (Hrsg.) (1985): Handbuch der Lexikologie. Königstein: Athenäum.
Sinclair, John (1987): Looking up. An Account of the COBUILD Project in Lexical Computing. : Collins.
Slator, Brian M. (1991): Using Context for Sense Preference. In: Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 65-90.
Smadja, Frank (1991): Macrocoding the Lexicon with Co-occurrence Knowledge. In: Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 165-189.
Smadja, Frank (1993): Retrieving Collocations from Text: Xtract. In: Computational Linguistics 19/1, S. 143-177.
Sowa, John F. (1991): Principles of Semantic Networks. Explorations in the Representation of Knowledge. San Mateo: Morgan Kaufman.
Sparck Jones, Karen (1986): Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press.
Stechow, Arnim von, Dieter Wunderlich (Hrsg.) (1991): Semantik/Semantics. Ein internationales Handbuch der zeitgenössischen Forschung. Berlin/New York: Walter de Gruyter.
St-Onge, David (1995): Detecting and Correcting Malapropisms with Lexical Chains. : University of Toronto. (= Master Thesis).
Thagard, Paul (1991): Concepts and Conceptual Change. In: Fetzer: Epistimology and cognition. Dordrecht u.a.: Kluwer, S. 101-120.
Tsohatzidis, Savas L. (Hrsg.) (1990): Meanings and Prototypes. Studies in Linguistic Categorization. London/New York: Routledge.
Tversky, Barbara (1990): Where partonomies and taxonomies meet. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 334-344.
Vandeloise, Claude (1990): Representation, prototypes, and centrality. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 403-437.
Vater, Heinz (1978): On the Possibility of Distinguishing between Complements and Adjuncts. In: Abraham, Werner: Valence, Semantic Case and Grammatical Relations. Amsterdam: John Benjamins, S. 21-45. (= Studies in Language 1).
Velardi, Paola (1991): Acquiring a Semantic Lexicon for Natural Language Processing. In: Uri Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 341-367.
Velardi, Paola, Maria Teresa Pazienza, Michela Fasolo (1991): How to Encode Semantic Knowledge. A Method for Meaning Representation and Computer-Aided Acquisition. In: Computational Linguistics 17/2, S. 153-170.

Vendler, Zeno (1967): Linguistics in Philosophy. New York: Cornell University Press.
Wahrig, Gerhard (1986): Deutsches Wörterbuch. München: Mosaik Verlag.
Walker, Donald E., R. A. Amsler (1986): The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In: Grishman/Kittridge: Analyzing Language in Restricted Domains. Hillsdale: Erlbaum, S. 69-83.
Warren, Beatrice (1978): Semantic Patterns of Noun-Noun Compounds. Göteborg: Universität Göteborg. (= Gothenburgh Studies in English 41).
Wehrle, Hugo, Hans Eggers (1968): Deutscher Wortschatz. Stuttgart: Klett.
Wierzbicka, Anna (1990): Prototypes save. On the uses and abuses of the notion 'prototype' in linguistics and related fields. In: Tsohatzidis: Meanings and Prototypes. London/New York: Routledge, S. 347-367.
Wilks, Yorick (1986): An intelligent analyzer and understander of English. In: Grosz/Sparck Jones/Webber: Readings in Natural Language Processing: Morgan Kaufmann.
Wilks, Yorick u.a. (1989): A tractable machine dictionary as a resource for computational semantics. In: Boguraev/Briscoe: Computational lexicography for Natural Language Processing. London/NewYork: Longman, S. 193-228.
Winston, Morton E., R. Chaffin, D.J. Hermann (1987): A Taxonomy of Part-Whole Relations. In: Cognitive Science, S. 417-444.
Wittgenstein, Ludwig (1953): Philosophical Investigations. Oxford: Blackwell.
Woods, W. A. (1991): Understanding Subsumption and Taxonomy. A Framework for Progress. In: Sowa: Principles of Semantic Networks. San Mateo: Morgan Kaufman, S. 45-94.
Wunderlich, Dieter (1991): Arbeitsbuch Semantik. Königstein: Athenäum.
Yarowsky, David (1992): Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: Proceedings of COLING-92 (Nantes). Nantes, S. 454-460.
Zadeh, Lotfi A. (1975): Fuzzy Logic and Approximate Reasoning. In: Synthese 30, S. 407-428.
Zadrozny, Wlodek (1994): From compositional to systematic semantics. In: Linguistics and Philosophy 17, S. 329-342.
Zampolli, Antonio, Nicoletta Calzolari, Martha Palmer (Hrsg.) (1994): Current Issues in Computational Linguistics. In Honour of Don Walker. Pisa/ Dordrecht: Giardini/Kluwer.
Zernik, Uri (Hrsg.) (1991): Lexical Acquisition. Using On-line Resources to Build a Lexicon. Hillsdale: Lawrence Erlbaum.
Zernik, Uri (1991a): Train1 vs. Train2. Tagging Word Senses in Corpus. In: Uri Zernik: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale: Erlbaum, S. 91-112.