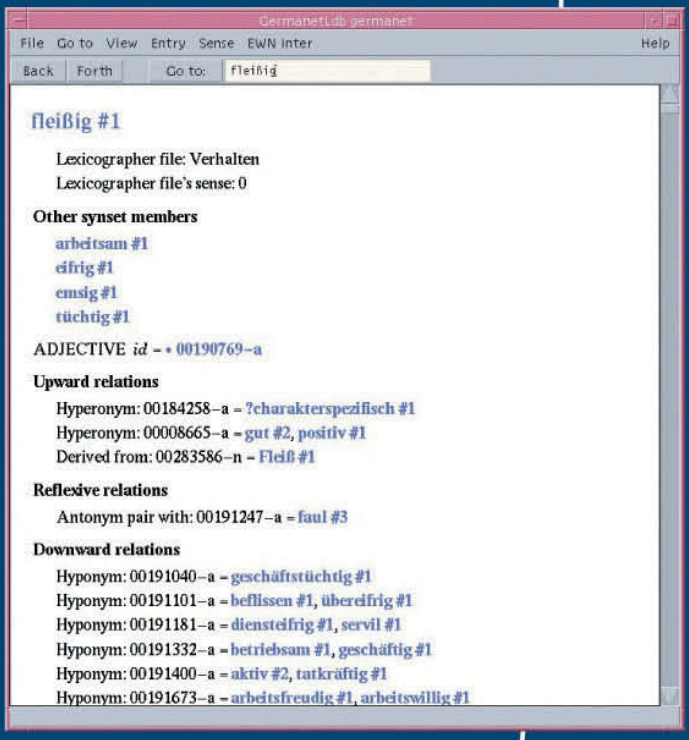
Claudia Kunze
Lothar Lemnitzer

Computer-lexikographie



gnv Gunter Narr Verlag Tübingen

Claudia Kunze / Lothar Lemnitzer

Computer-lexikographie

Eine Einführung

gnv Gunter Narr Verlag Tübingen

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Internet: http://www.narr.de
E-Mail: info@narr.de

ISBN 978-3-8233-6315-6

Inhalt

1

Einleitung

Für Jahrhunderte war das Handwerk der Wörterbucherstellung auf materielle Medien wie Pergament oder Papier und geeignete Schreibwerkzeuge angewiesen. Die Erfindung des Buchdrucks und, Jahrhunderte später, der Schreibmaschine, erleichterten das Handwerk selber und die Vervielfältigung und Verbreitung seiner Produkte. Dennoch blieb bis vor ca. dreißig Jahren der Zettelkasten der wesentliche Bezugspunkt der lexikographischen Arbeit: Wörterbücher entstanden durch die Kompilierung der von hunderten Beiträgern gesammelten Informationen. Der andere Bezugspunkt lexikographischer Arbeit war der Wörterbuchbenutzer, der in einem gedruckten Werk die Informationen finden sollte, die er benötigte, und das möglichst schnell. Platz war in gedruckten Werken das größte Problem. So entwickelte das lexikographische Handwerk über die Jahrhunderte ausgefeilte Techniken der Gewinnung und Darstellung der lexikographischen Informationen auf der zweidimensionalen Fläche der Buchseite. Es forderte und fordert dem Benutzer auch heute einiges an Kenntnissen und Fähigkeiten ab, die gerade benötigte Information in den meist nicht kleinen Druckwerken zu finden.

Vor etwa 30 Jahren kehrte der Computer dann auch in die Wörterbuchverlage ein und veränderte das Handwerk grundlegend:

- Der Zettelkasten ist nun nicht mehr die einzige Materialbasis lexikographischer Arbeit, ja oftmals nicht einmal die wichtigste, auch wenn er noch nicht aus den Redaktionsstuben vrschwunden ist. Statt dessen bezieht man sich heute selbstverständlich auf große digitalisierte Textsammlungen, aus denen die lexikographisch relevanten Informationen destilliert werden müssen. Dazu bedarf es ausgefeilter und effizienter, aber möglichst einfach zu bedienender Abfragetechniken;
- zum gedruckten Buch als Medium ist nun das digitalisierte, elektronische Wörterbuch getreten. Es ist zu erwarten, dass das elektronische Wörterbuch das Printwerk auf Dauer ablösen wird, jedenfalls in den zentralen Benutzergruppen, die heute über einen Computer oder über ein mobiles Endgerät verfügen; der Platz für die Präsentation der lexikographischen Informationen ist in diesem Medium kein Problem mehr. Dafür stellen sich andere editorische Herausforderungen, vor allem hinsichtlich der Präsentation der Informationen und ihrer effektiven und benutzerfreundlichen Erschließung;

- schließlich ist der Computer selbst zum „Konsumenten“ lexikographischer Daten geworden, genauer: sprachtechnologische Software, die umfassende linguistische und lexikalische Informationen benötigt. Diese Informationen sind für viele sprachtechnologische Anwendungen essenziell, und es gibt einen wachsenden Markt für lexikalische Daten, die für diese neue „Zielgruppe“ geeignet sind. Die Herausforderung liegt darin, die lexikographischen Daten in einer so strikt formalen Weise zu präsentieren, dass sprachtechnologische Anwendungen sie nutzen können. Computer sind nach wie vor weniger tolerant gegenüber Fehlern und Auslassungen als Menschen. Die zweite große Herausforderung besteht darin, die Daten so zu modellieren, dass sie von vielen sprachtechnologischen Anwendungen genutzt werden können.

Die soeben genannten Herausforderungen in der lexikographischen Praxis müssen von Spezialisten gelöst werden, die etwas von lexikographischen Prozessen, von Computern und insbesondere von Sprachtechnologie verstehen. Dieses Spezialgebiet wird COMPUTERLEXIKOGRAPHIE genannt.

Definition 1. *Als* **Computerlexikographie** *bezeichnen wir Lexikographie mit dem Computer und für den Computer. Die lexikographische Arbeit gestaltet sich umfangreich und datenintensiv, so dass maschinelle Unterstützung dieses Prozesses unerlässlich ist. Lexikographen werden bei der Erhebung, Bearbeitung, Darstellung und Verarbeitung lexikographischer Daten unterstützt. Wörterbuchbenutzern wird der elektronische Zugriff auf die für sie relevanten Daten ermöglicht. Schließlich benötigen sprachverarbeitende Systeme lexikalische Ressourcen, welche verarbeitungsrelevante lexikalische Informationen über ein Inventar von Wörtern einer oder mehrerer Sprachen zur Verfügung stellen.*

Computerlexikographie umfasst daher nicht nur die konkrete Erstellung von Wörterbüchern, also das Verfassen von Lexikonartikeln oder Einträgen, sondern auch die theoretische Auseinandersetzung mit Lexikonmodellen und Wörterbuchstrukturen sowie die Abschätzung der Anforderungen, die an einen bestimmten Lexikontyp gestellt werden.

Aufgrund der Dynamik und Wechselwirkung dieser Entwicklungen sind die Möglichkeiten der Computerlexikographie noch längst nicht ausgeschöpft. Dieses Buch gibt einen einführenden Überblick über das Feld und orientiert sich dabei am Stand der Forschung.

Der folgende Abschnitt stellt die relevanten Themen kurz vor und verweist auf die Kapitel, in denen sie ausführlich besprochen werden:

Die Darstellung des sprachbezogenen lexikalischen Wissens ist eingebettet in den Zusammenhang einer Theorie des lexikalischen Zeichens und der lexikalischen Semantik, für die in unserem Zusammenhang die Schlüsselbegriffe der Polysemie, Homonymie und Ambiguität relevant sind und erörtert

werden. Wir werden uns in Kapitel 3 auf die Aspekte der lexikalischen Semantik beschränken, die für die Computerlexikographie relevant sind. Unsere Darstellung wird sich aber nicht auf Einzelwörter beschränken. Es werden auch Fügungen von Wörtern, so genannte Mehrwortlexeme betrachtet, die besonders interessante Forschungsphänomene darstellen und spezifische Anforderungen an das Lexikonmodell sowie an Verarbeitungssysteme stellen.

Eine grundlegende theoretische Unterscheidung mit weitreichenden praktischen Konsequenzen ist die zwischen lexikalischem Wissen und allgemeinem Weltwissen. Auf der Ebene des Lexikonmodells spiegelt sich diese Unterscheidung in der Trennung von Sprachwörterbuch und Sachwörterbuch oder Enzyklopädie wider. Auf der Ebene der Bedeutungsbeschreibung lexikalischer Einheiten ist die Beschränkung auf sprachliche Aspekte der Wortbedeutung wesentlich, um lexikographische Beschreibungen handhabbar zu halten. Wir stellen die aktuelle Diskussion um diese Unterscheidung in Kapitel 4 dar.

Die Strukturierung lexikalischer Information im Wörterbucheintrag setzt voraus, dass auf der Basis einer Eintragsspezifikation, welche die Angabetypen und Klassifikationskriterien festlegt, eine geeignete Auswahl relevanter Worteigenschaften beschrieben wird. Diese Beschreibung unterliegt gewissen Anordnungs- und Aufbereitungskonventionen, welche zusammen die Artikelstruktur kennzeichnen. Eine standardisierte Artikelstruktur ist nicht nur für den menschlichen Benutzer hilfreich, sondern auch Voraussetzung für die Transformation der Angaben im Wörterbucheintrag in die Struktur einer lexikalischen Datenbank. Dies ist das Ziel des Wörterbuchparsings, das die Struktur standardisierter Wörterbuchartikel in Printwörterbüchern nutzt, um die dort vorhandenen Angaben in digitalisierte lexikalische Datenbanken zu übernehmen. Wir gehen in Kapitel 5 auf diese Aspekte ein und stellen eine Initiative vor, die das Ziel hat, Artikelstrukturen in elektronischen lexikalischen Ressourcen zu standardisieren.

Ausgehend von Organisationsmodellen lexikalischer Daten, die eher auf der konzeptuellen Ebene anzusiedeln sind, gelangt man zu technisch-organisatorischen Modellen, die unmittelbar die physikalische Speicherung und Repräsentation der Daten betreffen. Eine herausragende Rolle spielen hier Datenbankmodelle für die statische und dynamische Verwaltung der Daten und XML als Markupsprache für semistrukturierte Daten, die in Textdateien oder Datenbanksystemen gespeichert werden können. Auch hierauf gehen wir in Kapitel 5 ein.

Wortnetze und Ontologien sind Organisationsformen lexikalischer Ressourcen, die eine bedeutende Rolle für sprachtechnologische Anwendungen spielen. Diese lexikalischen Ressourcen ordnen den Wortschatz nach lexikalisch-semantischen Kriterien. Bedeutungsverwandte Wörter und Konzepte werden miteinander verknüpft. Ontologien sind dabei, wie wir sehen

werden, strenger formalisiert als Wortnetze. Beide Arten von Ressourcen eignen sich für jeweils andere sprachtechnologische Anwendungen. Kapitel 6 ist der Beschreibung dieser lexikalischen Ressourcen gewidmet.

Lexikalische Regeln sind Mechanismen, die man ausschließlich bei solchen lexikalischen Ressourcen findet, die von sprachtechnologischen Systemen verwendet werden. Sie erlauben die kompakte Darstellung von Regularitäten auf allen Ebenen der lexikographischen Beschreibung. Mit lexikalischen Regeln kann man linguistische Generalisierungen kodieren. Sie machen so die wiederholte Darstellung derselben Zusammenhänge bei vielen einzelnen Einträgen überflüssig. Menschliche Benutzer werden die Darstellung dieser Zusammenhänge beim einzelnen Wörterbucheintrag bevorzugen. Deshalb finden wir diese Regeln nicht in traditionellen Wörterbüchern. Tatsächlich ist das Design dieser Regeln eine der Fähigkeiten, die Computerlexikographen gegenüber traditionellen Lexikographen auszeichnet. Wir behandeln lexikalische Regeln ausführlich in Kapitel 7.

Mit der Verfügbarkeit sehr großer Textkorpora als Datenbasis für lexikalische und linguistische Arbeiten bekommt die lexikalische Statistik eine prominente Rolle. Quantitative Sprachmodelle können die (computer)lexikographische Arbeit in vielerlei Hinsicht unterstützen, wie wir in Kapitel 8 zeigen.

Für Sprachen mit einer reicheren Morphologie spielt die Berücksichtigung von Flexion und Wortbildung, im einfachsten Fall bei der Ermittlung von Grundformen für Textwörter (die Lemmatisierung), eine wichtige Rolle. Für die Lemmatisierung und morphologische Wortanalyse gibt es heute ausgereifte sprachtechnologische Werkzeuge, die wir in Kapitel 9 vorstellen werden. Ebenfalls dort werden wir auf die Präsentation morphologischer Informationen in Printwörterbüchern eingehen. Formbezogene elektronische Ressourcen sollten mindestens dieses Niveau der Beschreibung erreichen.

Die Akquisition lexikalischer Information aus Korpora ist dann wichtig, wenn ein Wörterbuch aufgebaut, erweitert, verbessert oder aktualisiert werden soll. Da schon viel lexikalisches Wissen in den unterschiedlichsten Quellen vorliegt, liegt der Gedanke nahe, bereits existierende Quellen auszubeuten, wie wir im Kapitel über das Wörterbuchparsing gezeigt haben. Daneben bieten sich große digitale Textsammlungen an. Diese sind heute für alle großen europäischen Sprachen verfügbar. Wichtig ist jeweils das zugrunde liegende Lexikonmodell, in das die neuen Informationen integriert werden sollen. Außerdem spielen Verarbeitungsprozesse, die zur Erkennung und Voranalyse der lexikalischen Einheiten führen, eine große Rolle, genau wie statistische Verfahren zur Ermittlung der Relevanz lexikalischer Information. In Kapitel 10 werden wir diese Aspekte diskutieren und abschließend ein allgemeines Vorgehen für Projekte der lexikalischen Akquisition vorstellen.

Den mehrgliedrigen lexikalischen Einheiten ist das letzte Kapitel gewidmet. Hier gehen wir vor allem auf Kollokationen und Phraseme ein. Beide Typen mehrgliedriger lexikalischer Einheiten standen in den letzten Jahren im Zentrum der computerlexikographischen Diskussion. Ausgangspunkt war die Erkenntnis, dass eine Herangehensweise an die automatische Sprachanalyse, die sprachliche Äußerungen als Kombinationen voneinander unabhängiger Einzelwörter auffasst, zu kurz greift. Mehrgliedrige Einheiten mit eingeschränkter Kombinierbarkeit und einer Bedeutung, die über die Summe ihrer Teile hinausgeht, durchziehen unsere sprachlichen Äußerungen. Es ist deshalb wichtig, sie bei der Textanalyse zu erkennen und angemessen lexikalisch zu beschreiben.

In einem Abschnitt von Kapitel 10 stellen wir einige Werkzeuge vor, mit denen Sie selbst Korpora unter den verschiedensten Aspekten analysieren können. Dies soll Ihnen helfen, die in diesem Buch vorgestellten Techniken und Methoden auszuprobieren und selbständig lexikalische Studien zu betreiben.

Das Buch wird durch eine Webseite – `http://www.lemnitzer.de/lothar/CoLex` – begleitet, auf der wir a) weiteres Material zur Verfügung stellen; b) wir Sie über die Computerlexikographie-Tagungen, neue Literatur etc. auf dem Laufenden halten und c) auf die nie ganz zu vermeidenden Tippfehler hinweisen[1].

[1] Ein großer Teil der Lehrmaterialien, aus denen dieses Buch entstand, wurde von uns im Rahmen des BMBF-geförderten Projekts *Medienintensive Lehrmodule für die Computerlinguistik-Ausbildung* (MiLCA) in den Jahren 2001-2004 entwickelt. Wir danken dem BMBF und seinem Projektträger für die materielle und ideelle Unterstützung des Vorhabens.

2

Das Lexikon

Nach der Lektüre dieses Kapitels werden Sie wissen, was im Kontext verschiedener Theorien unter einem Lexikon und unter einem lexikalischen Zeichen zu verstehen ist. Die meisten der hier angeschnittenen Themen werden in Kapitel 3 weiter vertieft.

1 Begriffsbestimmung

In der modernen Linguistik werden mit GRAMMATIK und LEXIKON zwei elementare Teilsysteme der Sprache unterschieden. Das grammatische Modul legt das Kategorieninventar der linguistischen Betrachtungsebenen und die Regularitäten ihrer Verknüpfungen fest, während das Lexikon den Wortschatz einer Sprache beisteuert. Idiosynkratische, d.h. nicht vorhersagbare Eigenschaften von Wörtern finden ihren Platz im Lexikon. Diese Eigenschaften sind auf allen linguistischen Ebenen von der Phonetik und Phonologie bis hin zur Pragmatik zu spezifizieren.

Der Begriff des Lexikons in der Sprachwissenschaft ist vielschichtig und ist ein gutes Beispiel für die Polysemie eines sprachlichen Zeichens, auf die wir in Kapitel 3, Abschnitt 5 zu sprechen kommen:

- In der Psycholinguistik (vgl. Jackendoff (1990), Pinker (1996)) wird das Lexikon als mentales Konstrukt sowohl eines Sprechers als auch einer Sprachgemeinschaft aufgefasst. Es wird versucht, die Struktur des mentalen Lexikons kognitiv adäquat zu modellieren.
- In der generativen Grammatik fungiert das Lexikon als Modul, in dem Wörter und ihre idiosynkratischen Eigenschaften aufgelistet werden. Diese Wörter werden in eine syntaktische Struktur eingesetzt, sie instantiieren die abstrakten Kategorien der präterminalen Knoten im Syntaxbaum, z.B. $V \rightarrow$ *gehen*. Diese Sichtweise entspricht dem in Chomsky (1965) dargestellten Modell der generativen Grammatik.
- In der Form des gedruckten oder elektronischen Wörterbuchs liefert ein Lexikon dem menschlichen Benutzer Angaben zu den sprachlichen Eigenschaften der verzeichneten lexikalischen Einheiten; im zweisprachigen Wörterbuch findet man vor allem Übersetzungsäquivalente.

- Ein Printwörterbuch kann mit den gleichen Informationssequenzen digitalisiert und in maschinenlesbarer Form zur Verfügung gestellt werden. Maschinenlesbare Wörterbücher werden ebenfalls als Lexika bezeichnet.
- Eine weitere Erscheinungsform des Lexikons ist die Lexikonkomponente in einem Sprachverarbeitungssystem. Die jeweilige sprachverarbeitende Anwendung verlangt von den lexikalischen Ressourcen spezifische lexikalische Information. Diese müssen in einer Form kodiert sein, die vom sprachverarbeitenden System eindeutig interpretiert und verarbeitet werden kann.

Wir werden diese Lesarten im Folgenden voneinander abgrenzen, obgleich sie natürlich Beziehungen zueinander aufweisen.

Von einem Lexikon als Teil oder Modul einer Sprachtheorie, also im Sinne der zweiten Definition, wird man klare Kriterien hinsichtlich des Umfangs der verzeichneten lexikalischen Zeichen sowie hinsichtlich der Angaben zu diesen Zeichen erwarten. Auf der anderen Seite stehen relativ kurzlebige Printwörterbücher wie die neusten Auflagen des Rechtschreibduden, bei denen von Auflage zu Auflage Wörter aufgenommen und wieder entfernt werden. Statt klarer Kriterien bei der Auswahl der linguistischen Beschreibungsebenen haben sich in der praktischen Lexikographie Konventionen ausgebildet darüber, welche Angabetypen in welchen Typen von Wörterbüchern zu finden sind. Dies wird oftmals schon im Namen des Wörterbuchs deutlich (*Rechtschreibwörterbuch, Valenzwörterbuch* etc.). Bei den lexikalischen Ressourcen für sprachverarbeitende Systeme haben sich für die meisten europäischen Staaten als Quasi-Standard formbasierte, vor allem die morphologischen Eigenschaften der sprachlichen Zeichen beschreibende und lexikalisch-semantische Ressourcen im Stile der Wortnetze etabliert.

In allen Fällen ist das Lexikon ein offenes System und somit stärkeren Wandlungen unterworfen als das grammatische System einer Sprache. Lexikalische Neubildungen wie *Topterrorist, Selbstmordattentat* und *Spendensumpf* sind so geläufig, dass sie zumindest zeitweise in das Lexikon einer Sprachgemeinschaft Eingang finden könnten. Ein paar Jahre später sind sie möglicherweise wieder völlig außer Gebrauch. Es ist daher relativ zeitaufwändig, das Lexikon einer Sprache aktuell zu halten, sei es als lexikalische Ressource für sprachverarbeitende Systeme, sei es als Printwörterbuch. Heutzutage werden überwiegend Korpora zeitgenössischer Texte für die Auswahl des zu beschreibenden Wortschatzes und vor allem für die Registrierung neuer Wörter eingesetzt. Wir werden später auf diesen Aspekt der computergestützten Lexikographie eingehen.

Auch wenn wir uns in diesem Buch vornehmlich mit einem anwendungsorientierten Lexikonbegriff befassen, kommen wir nicht umhin, verbindliche

Kriterien dafür festzulegen, was wir unter einem Lexikon im Folgenden verstehen wollen:

- Einem Lexikon sollte ein explizites Lexikonmodell zugrunde liegen, an dem sich zum Beispiel die Erstellung dieses Lexikons ausrichtet. Dieses Modell kann durchaus in Abhängigkeit von einem bestimmten Anwendungszweck entwickelt werden.
- Das Lexikonmodell sollte die elementaren Einheiten im Lexikon festlegen. Die prototypische lexikalische Einheit ist das Wort. Das Lexikon kann aber auch sprachliche Einheiten unterhalb der Wortebene (Morpheme) oder oberhalb der Wortebene (Mehrwortlexeme) zum Gegenstand haben. Wir werden auf beide in separaten Kapiteln eingehen.
- Das Lexikonmodell sollte zu Unterscheidungskriterien für lexikalische Einheiten führen. Diese sind in erster Linie semantischer Natur, können aber auch formbasiert sein. Wir haben diesem wichtigen Gegenstand ein eigenes Kapitel gewidmet.

Im Folgenden werden wir einige Typen digitaler lexikalischer Ressourcen beschreiben, die auch unter dem Sammelbegriff lexikalische Datenbanken zusammengefasst werden.

2 Lexikalische Datenbanken

Wir wollen zunächst festlegen, was wir unter dem Begriff LEXIKALISCHE DATENBANK verstehen.

Lexikalische Datenbanken sind digitale lexikalische Ressourcen, die in einer Form abgespeichert sind, dass die einzelnen Datensätze konsistent im Hinblick auf eine formale Beschreibung ihrer Struktur sind. Ein einzelner Datensatz kann dabei einem Wörterbuchartikel entsprechen oder einem Artikelteil. Er kann aber auch artikelübergreifende Strukturen umfassen. Die formale Beschreibung der Datenstruktur kann in Form eines konzeptuellen Schemas vorliegen, wenn die Daten z.B. in einem relationalen Datenbanksystem abgespeichert sind. Sie kann in Form einer Dokumentgrammatik vorliegen, wenn die Daten als annotierte Dokumente verwaltet werden[1]. Generell gilt, dass relationale Datenbanksysteme und die für die Modellierung der Daten verwendeten konzeptuellen Schemata eine rigidere Strukturierung der Daten erzwingen als Dokumentgrammatiken für XML-annotierte Daten. Man spricht deshalb von relationalen Datenbanken als Verwaltungssystemen für strukturierte Daten und von XML-basierten Datenbanken als Verwaltungssystemen für semi-strukturierte Daten. Die Entscheidung für eine der beiden Alternativen hängt letztlich von der Qualität der zu modellierenden Daten ab. Deshalb kann das Für und Wider beider Alternativen nicht unabhängig von konkreten Projekten diskutiert werden.

Ein LEXIKALISCHES INFORMATIONSSYSTEM ist umfassender als eine LEXIKALISCHE DATENBANK. Es enthält eine lexikalische Datenbank für die Speicherung und Verwaltung der Daten, darüber hinaus aber auch Benutzerschnittstellen für den Zugriff auf diese Daten. Wir werden in Kapitel 5, Abschnitt 1.5 detaillierter auf zwei lexikalische Informationssysteme eingehen.

2.1 Typen lexikalischer Datenbanken

Unter maschinenlesbaren Wörterbüchern (‚machine readable dictionaries', MRDs) verstehen wir die elektronischen Versionen allgemeinsprachlicher Printwörterbücher, die meist in Form von Satzbändern vorliegen, oder auch maschinell hergestellte und genutzte Wörterbücher. Da sie für den menschlichen Benutzer bestimmt sind, sollten möglichst alle Informationen in natürlicher Sprache vorliegen. Insbesondere in Großbritannien sind maschinenlesbare Wörterbücher in der akademischen Forschung intensiv genutzt worden, seit der Longman Verlag, und später auch andere britische Verlage, ihre Daten zur Verfügung stellten[2]. Berühmt ist die maschinenlesbare Version der er-

[1] Vgl. hierzu Kapitel 5, Abschnitt 3 in diesem Buch.

[2] Wir stellen in Kapitel 5, Abschnitt 2.4 ein Projekt vor, in dem Daten aus einem maschinenlesbaren Wörterbuch extrahiert und weiterverwendet wurden.

sten Auflage des *Longman Dictionary of Contemporary English* von 1978[3]. Die Rohdaten maschinenlesbarer Wörterbücher enthalten die für weitgehend manuell erstellte Printwörterbücher charakteristischen Inkonsistenzen. Dies muss bei der Analyse und beim Parsen dieser Daten beachtet werden[4].

Unter maschinenverarbeitbaren Wörterbüchern (‚machine tractable dictionary', MTD) verstehen wir lexikalische Ressourcen, die lexikalisches Wissen in einer Art und Weise kodieren, dass Computersysteme, insbesondere sprachtechnologische Anwendungen, darauf zugreifen können. Die darin enthaltenen Angaben müssen in einem zu spezifizierenden, expliziten Format vorliegen. Maschinenverarbeitbare Wörterbücher können in den unterschiedlichsten sprachtechnologischen Anwendungen eingesetzt werden und dementsprechend auch unterschiedliche Schwerpunkte setzen. Sie können etwa Angaben zur Morphologie oder zur Syntax oder Semantik enthalten oder kombinierte Ressourcen darstellen. Es werden jeweils formal explizite kanonische Angaben bereitgestellt, auf die der Computer zugreifen kann. Ein Beispiel für ein maschinenverarbeitbares Wörterbuch ist das deutsche Wortnetz GermaNet, auf das wir in Kapitel 6 näher eingehen werden. Im Wortnetz sind überwiegend semantische Informationen kodiert. Ein weiteres Beispiel ist das an der Fernuniversität Hagen entwickelte HagenLex[5]. Das Stuttgarter Lexikon IMSLex hingegen enthält überwiegend formbasierte Informationen. Das Informationsprogramm umfasst Flexionsmorphologie, Derivations- und Kompositionsmorphologie und Valenzangaben[6]. MTDs können auf eine spezifische Theorie zugeschnitten sein, etwa auf den HPSG-Formalismus (Head Driven Phrase Structure Grammar) oder die Diskursrepräsentationstheorie (DRT). Ein für die Entwicklung maschinenverarbeitbarer Wörterbücher relevanter Aspekt war und ist die Frage, inwieweit man die lexikalischen Daten theorieneutral modellieren und sie damit für viele sprachtechnologische Anwendungen nutzbar machen kann. GermaNet und IMSLex sind Beispiele für weitgehend theorieneutrale lexikalische Ressourcen.

Wichtig sind in diesem Zusammenhang auch lexikalische Datenbanken, die Fachterminologie enthalten. Oftmals haben große Firmen ihre eigenen Terminologien aufgebaut; und europäische Bemühungen zielten auf die Vereinbarung von Standards, um die Terminologiedaten austauschen bzw. wiederverwerten zu können, z.B. *Interactive Terminology for Europe* (IATE)[7].

---

[3] Vgl. Procter (1978).

[4] Auf das Wörterbuchparsing gehen wir detaillierter in Kapitel 5, Abschnitt 2 ein.

[5] Vgl. Hartrumpf et al. (2003): `http://pi7.fernuni-hagen.de/forschung/hagenlex/hagenlex-de.html`.

[6] Detailliertere Informationen finden sich unter `http://www.ims.uni-stuttgart.de/projekte/IMSLex/`.

[7] Vgl. `http://iate.europa.eu/iatediff/`.

Lexikalische Wissensbanken schließlich sind maschinenverarbeitbare lexikalische Ressourcen, die auch außersprachliches Wissen einbeziehen. Eine klare Unterscheidbarkeit von lexikalischen Datenbanken und lexikalischen Wissensbanken ist nicht immer gegeben. Die Unterscheidung der beiden Ressourcentypen stammt aus der Zeit zu Beginn der Neunziger Jahre, als der Diskurs über Sprach- und Weltwissen einen (vorläufigen) Höhepunkt erfuhr. Eine heute sehr einflussreiche enzyklopädische Ressource, und damit Wissensbank, ist die Wikipedia. Die deutsche Ausgabe dieser Enzyklopädie befindet sich unter `http://de.wikipedia.org`. Die Wikipedia ist der Prototyp einer dynamischen lexikalischen Ressource, mit Hunderten, wenn nicht Tausenden Änderungen täglich. Die Sprachtechnologie beginnt gerade erst, sich den Reichtum dieser Ressource nutzbar zu machen[8].

Das Gegensatzpaar STATISCH und DYNAMISCH zielt auf die Konzeption der Datenbanken schlechthin: eine statische Datenbank ist hinsichtlich der Informationsstruktur, die sie repräsentiert, festgelegt, während eine dynamische Datenbank neue Informationstypen integrieren kann. So weist zum Beispiel der von Petra Ludewig beschriebene Prototyp einer LEXICAL KNOWLEDGE BASE Import- und Exportfunktionen auf, welche die Zusammenführung und Wiederverwendung lexikalischer Information aus externen Ressourcen ermöglicht[9]. Solche dynamischen Systeme benötigen Programme, wie die sog. LEXICON BUILDERS, die automatisch Wörterbücher erstellen können, indem sie Informationen aus bestehenden Wörterbüchern, aus Dokumenten und Korpora akquirieren und zusammenführen.

Hyperlexika sind als Hypertexte realisierte Lexika und Lexikonsysteme, vor allem im World Wide Web (WWW), bei denen ebenfalls zwischen statischen und dynamischen Varianten unterschieden wird: statische Hyperlexika sind vorkompilierte WWW-Versionen eines gedruckten Wörterbuchs als abfragbare Datenbanken. Dynamische Hyperlexika sind nicht vorkompiliert und bieten keine Indexauflösung, sondern eine Suche an, und können daher mit sehr großer Kombinatorik abgefragt werden, ähnlich wie WWW-Suchmaschinen.

Insgesamt gesehen werden dynamische benutzerdefinierte lexikalische Informationssysteme immer wichtiger. Im Kontext der technischen Möglichkeiten einer verbesserten Datengewinnung durch automatische Verfahren der Informationsextraktion aus Dokumenten und Korpora könnte der klassische Lexikonbegriff, der von einem relativ fixen Repertoire lexikalischer Einheiten ausgeht, eine Umbewertung erfahren. Ad-hoc gebildete Lexika für die unterschiedlichsten Zwecke und Anwendungen könnten einerseits in Bezug auf Qualität, Abdeckung und Einsetzbarkeit zu Evaluationsproblemen führen

[8] Vgl. Zesch et al. (2007). Die Autoren haben eine Schnittstelle für die Programmierung (API) entwickelt, mit deren Hilfe man auf die Daten der Wikipedia-Datenbank zugreifen kann.

[9] Vgl. Ludewig (1993).

und für die lange angestrebte Standardisierung kontraproduktiv sein, andererseits für größere Flexibilität und empirisch gesichertes Datenmaterial sorgen.

3 Weiterführende Literatur

Eine lesenswerte Referenz zum Lexikon in der psycholinguistischen Erforschung vor allem des Spracherwerbs ist die Arbeit von Eve Clark (1993). Zum Lexikon in der theoretischen Sprachwissenschaft vor allem der generativen Prägung geben die Arbeiten des Sonderforschungsbereichs „Theorie des Lexikons“ Auskunft[10]. Im Zentrum des Interesses stehen hier aber sicher maschinenlesbare Wörterbücher und Lexika für sprachtechnologische Anwendungen. Für Erstere ist immer noch die Arbeit von Boguraev und Briscoe (1989) die erste Referenz. Kritisch zum Nutzen von maschinenlesbaren Wörterbüchern für die Sprachtechnologie äußern sich Nancy Ide und Jean Véronis (1993). Einen relativ neuen Ansatz präsentiert Daelemans (2000) unter dem Namen „Inductive Lexicon“. Die Standardreferenz zum von Pustejovsky propagierten „Generative Lexicon“ ist sein Aufsatz von 1991, auch wenn es viele neuere, auch in diesem Buch erwähnte Arbeiten aus diesem theoretischen Umfeld gibt. Am Schluss wollen wir mit der Arbeit von Christopher Habel (1985) einen etwas in die Jahre gekommenen, aber zumindest aus historischer Sicht interessanten Artikel zum Platz des Lexikons in der Forschung zur künstlichen Intelligenz empfehlen.

4 Aufgabe

1. Welche Wörterbücher, Lexika und Enzyklopädien kennen Sie bzw. haben Sie schon mal benutzt? Berichten Sie von Ihren Erfahrungen. Was könnte man Ihrer Meinung nach an Wörterbüchern verbessern?

[10] Vgl. `http://www.phil-fak.uni-duesseldorf.de/sfb282/`.

3

Lexikalische Semantik

In diesem Kapitel werden Sie die lexikalisch-semantischen Zusammenhänge kennenlernen, die für die Computerlexikographie von zentraler Bedeutung sind. Sie erfahren insbesondere mehr zur komponentiellen Semantik und zur relationalen Semantik. Wir gehen ausführlich auf das zentrale Konzept der Polysemie ein. Zum Abschluss des Kapitels führen wir die beiden Konzepte der Unterspezifizierung und der Ambiguität ein.

1 Lexikalisches Zeichen und lexikalisches System

Die lexikalische Semantik befasst sich mit den lexikalischen Zeichen sowie dem lexikalischen System oder Lexikon einer Sprache. Lexikalische Zeichen sollten nicht mit Wörtern verwechselt werden. Jede sprachliche Einheit, der eine Bedeutung zugeordnet werden kann, ist ein lexikalisches Zeichen und damit Teil des lexikalischen Systems einer Sprache. Neben einfachen Wörtern sind dies Wortteile, MORPHEME genannt, und wortübergreifende Ausdrücke, vor allem Phraseme, aber auch Kollokationen. Wir unterscheiden also:

- *-bar* – ein Morphem, das als Suffix an verbalen Stämmen Adjektive bildet, z.B. *lernbar*,
- *Sack* – ein Wort bzw. einfaches lexikalisches Zeichen,
- *die Katze im Sack kaufen* – ein Phrasem, dessen Bedeutung nichts mit Katzen und Säcken zu tun hat, sondern mit Dingen, die man unbesehen erwirbt,
- *den Tisch decken* – eine Kollokation.

Wir werden in Kapitel 9 näher auf Morpheme und Wortstrukturen eingehen. Den mehrwortigen Lexemen ist ebenfalls ein eigenes Kapitel gewidmet (Kapitel 11).

Im Folgenden werden wir vor allem auf Wörter und Wortbedeutungen eingehen, möchten aber nochmals betonen, dass der Begriff des lexikalischen Zeichens mehr umfasst als nur Wörter.

Wir werden zunächst auf die Betrachtung des lexikalischen Zeichens in der strukturalistischen Semantik eingehen. Von Saussure ausgehend werden

wir einige semiotische Modelle des sprachlichen Zeichens vorstellen. Diese Modelle behandeln die dichotomische und späterhin trichotomische Struktur von lexikalischer Form, lexikalischer Bedeutung und Referenten von Wörtern. Viele, nicht nur strukturalistische, Semantiker gehen davon aus, dass die Bedeutungsseite des sprachlichen Zeichens in BEDEUTUNGSATOME dekomponierbar ist. Im weiteren Sinn wird die Bedeutung des gesamten Vokabulars einer Sprache als Kombination einer endlichen und zumeist sehr kleinen Menge von Bedeutungsatomen (oder PRIMITIVEN) betrachtet. Diese Annahme geht auf eine Analogie zur Lautform von Wörtern zurück. Die Phoneme einer Sprache können als Kombinationen einer endlichen und sehr kleinen Menge von Lauteigenschaften beschrieben werden. Atomistische Ansätze der Wortsemantik sind sowohl in der allgemeinen Linguistik als auch in der Computerlinguistik sehr beliebt. Wir werden in drei Abschnitten dieses Kapitels einflussreiche klassische und neuere Ansätze der kompositionellen Semantik vorstellen: Katz und Fodors Markertheorie, Wierzbickas semantische Primitive und Pustejovskys Theorie des generativen Lexikons.

Neben dem einzelnen lexikalischen Zeichen ist das lexikalische System oder Subsystem einer Sprache der Untersuchungsgegenstand der lexikalischen Semantik. Ein beliebter Gegenstand war und ist das lexikalische Feld. Ein lexikalisches Feld besteht aus einer Menge lexikalischer Zeichen, deren Bedeutungen über lexikalisch-semantische Relationen verbunden sind. Einige bekannte lexikalische Relationen sind die SYNONYMIE, ANTONYMIE, HYPERONYMIE und HYPONYMIE.

Einige Forscher, vor allem aus dem Bereich der kognitiven Linguistik und künstlichen Intelligenz, postulieren die Existenz konzeptueller Strukturen, von mentalen Strukturen also, die in bestimmter Weise zu den Strukturen im lexikalischen System einer Sprache korrespondieren[1]. Reuland und Ankersmit haben die Beziehungen zwischen konzeptuellen Strukturen und Strukturen von lexikalischen Einträgen genauer untersucht[2].

Lexikalisch-semantische Relationen sind ein wichtiges Strukturierungsmittel in der Computerlexikographie. In den alphabetisch angeordneten Wörterverzeichnissen von Printwörterbüchern werden diese Relationen durch Verweise realisiert. In Spezialwörterbüchern wie zum Beispiel den Wortnetzen, auf die wir später genauer eingehen werden, sind lexikalische Einheiten entsprechend der sie verbindenden lexikalisch-semantischen Relationen gruppiert. In diesem Kapitel werden wir uns noch ausführlicher mit einem Bereich der lexikalischen Semantik beschäftigen, der sich RELATIONALE SEMANTIK nennt.

[1] Vgl. Sowa (1983).
[2] Vgl. Reuland und Ankersmit (1993).

Polysemie von lexikalischen Einheiten ist ein Phänomen, das sich bisher dem vollen Verständnis aller wortsemantischen Theorien entzieht. Mit Polysemie bezeichnet man die Tatsache, dass ein lexikalisches Zeichen in mehr als einer Bedeutung verwendet werden kann (z.B. *Satz* → ‚Einheit der Sprache', ‚großer Sprung', ‚Spielabschnitt beim Tennis' etc.). Die verschiedenen Bedeutungen eines Wortes sind einigen Theorien zufolge miteinander verbunden. Eine Richtung der aktuellen Forschung befasst sich damit, ob Beziehungen zwischen den Bedeutungen von Wörtern sich generalisieren und damit als Regularitäten darstellen lassen (z.B. haben viele Wörter verwandte Bedeutungen, die eine Institution und das Gebäude, das diese Institution beherbergt, bezeichnen, z.B. *Schule*, *Finanzamt*). Man spricht dann von REGULÄRER POLYSEMIE.

In der lexikographischen Praxis stellt sich ständig die Frage, wie viele Bedeutungen oder Lesarten für ein Wort bzw. einen Wörterbucheintrag angesetzt werden sollen – im *Duden Universalwörterbuch* werden 12 Lesarten für das Wort *Satz* unterschieden, in anderen Wörterbüchern sind es weniger oder mehr, die Spannbreite ist gerade bei stark polysemen Wörtern bemerkenswert. Eng mit der Polysemie verbunden ist die Ambiguität von Textwörtern. Eine noch nicht gemeisterte Herausforderung für sprachtechnologische Programme besteht darin, die genaue Bedeutung eines Worts im Kontext eines Textes zu bestimmen. Das Forschungsprogramm, das zur Lösung dieser Frage bzw. zu einem funktionierenden System beitragen möchte, nennt sich WORD SENSE DISAMBIGUATION, was sich in etwa mit ‚Lesartenbestimmung von Textwörtern' übersetzen lässt. Einen Überblick über den Stand der Forschung geben Jean Véronis und Nancy Ide[3]. Ein von Eneko Agirre herausgegebener Sammelband präsentiert die neuesten Forschungsansätze[4].

[3] Vgl. Ide und Véronis (1998).
[4] Vgl. Agirre und Edmonds (2006).

2 Die Struktur des lexikalischen Zeichens

2.1 Die Saussureschen Dichotomien

Um einen Eindruck davon zu bekommen, wie die Form- und Inhaltsseite lexikalischer Zeichen aufeinander bezogen werden können, werden wir uns zunächst die strukturalistische Theorie des lexikalischen Zeichens ansehen. Diese Theorie nahm ihren Ursprung bei Ferdinand de Saussure, der als Wegbereiter der modernen Linguistik gilt. Seine wegweisende Vorlesung ‚Cours de linguistique générale' (Deutsch: Grundfragen der allgemeinen Sprachwissenschaft, de Saussure (2001)) wurde 1916 auf der Basis der Mitschriften von Zuhörern veröffentlicht. Zwei Wortpaare sind grundlegend für Saussures Konzept der Wortbedeutung:

- LANGUE vs. PAROLE (auf Deutsch: Sprachsystem vs. Sprachgebrauch)
- RELATIONS PARADIGMATIQUES vs. RELATIONS SYNTAGMATIQUES (paradigmatische vs. syntagmatische Beziehungen)[5]

Mit der ersten Unterscheidung etabliert Saussure Sprache als System, das von den Verwendungsinstanzen der Sprache, dem Sprachgebrauch, zu unterscheiden ist und einen eigenen Untersuchungsgegenstand der Linguistik darstellt. Entsprechend ist das Lexikon eine Abstraktion aus den zahlreichen Verwendungen lexikalischer Einheiten in Wort und Schrift. Sprache in diesem Sinne ist ein statisches System mit einem sozialen Wert, der durch Konvention festgelegt wird. Das Objekt der linguistischen Forschung ist dieses soziale Produkt, das sich im Gehirn jedes einzelnen Sprechers manifestiert. Dieses Produkt liegt allen konkreten Äußerungen (also dem Sprachgebrauch) zugrunde. Der konkrete Sprachgebrauch wiederum ist geprägt von Varianz in Tonfall, Tonhöhe, dialektaler Einfärbung etc., von welcher auf der Ebene des Sprachsystems abstrahiert wird.

Eine ähnliche Unterscheidung wird im Rahmen der generativen Grammatik durch das Begriffspaar COMPETENCE und PERFORMANCE getroffen. Während die Performanz den aktuellen Sprachgebrauch einer bestimmten Person zu einer bestimmten Zeit bezeichnet, mit allen Idiosynkrasien, individuellen Eigenheiten, Fehlern etc., referiert Chomsky[6] mit dem Begriff der Kompetenz auf das Sprachvermögen als eine kognitive Fähigkeit aller Sprecher. Beide Theoretiker würden sicher der Aussage zustimmen, dass das Abstraktum, LANGUE oder KOMPETENZ genannt, der eigentliche Gegenstand der Linguistik ist. Die generative Grammatik geht hier noch einen Schritt weiter mit der Behauptung, dass Sprachkompetenz ohne Rückgriff auf die Performanz, also einzelne Äußerungen, untersucht werden kann. Viele Lin-

[5] Wir gehen in Abschnitt 4 dieses Kapitels näher auf dieses Gegensatzpaar ein.
[6] Vgl. Chomsky (1969).

guisten folgen dem nicht (mehr). Die Unterscheidung dieser beiden Aspekte von Sprache hat Auswirkungen auf den Begriff der (lexikalischen) Bedeutung. Zum einen spiegelt sich die Unterscheidung wider in dem Begriffspaar der DENOTATIVEN BEDEUTUNG, einem Abstraktum, das sich im Wörterbuch findet, und der REFERENTIELLEN BEDEUTUNG, die eine Eigenschaft der konkreten Äußerung ist. Betrachten wir ein Beispiel:

> The Hitchhiker's Guide to the Galaxy notes that Disaster Area, a plutonium rock band from the Gagrakacka Mind Zones, are generally held to be not only the loudest rock band in the Galaxy, but in fact the loudest noise of any kind at all. (Adams (1980), S. 114)

In diesem Beispiel beziehen sich die Wörter bzw. Wortsequenzen *Disaster Area*, *rock band* und *noise* auf den gleichen Sachverhalt bzw. das gleiche außersprachliche Objekt, obwohl sie verschiedene denotative Bedeutungen haben. In Fällen wie diesen spricht man in der Linguistik übrigens von KOREFERENZ.

Eine weitere wichtige begriffliche Dichotomie, die auf de Saussure zurückgeht, ist die zwischen SUBSTANCE und VALEUR (Substanz vs. Wert). Der Begriff der Substanz bezeichnet die ungeformte Masse der Laute und der Bedeutungen bzw. Begriffe. Das sprachliche Zeichen ist es, das diese Substanz formt und unterteilt. Die Substanz von Laut und Bedeutung existiert unabhängig von einzelnen Sprachen, sie ist universal, wohingegen jede einzelne Sprache diese Substanz anders formt und gliedert. Ein gutes und oft zitiertes Beispiel hierfür ist die Unterteilung des Farbspektrums (der Substanz) in verschiedene lexikalische Felder in verschiedenen Sprachen. Sprecher haben die prinzipielle Fähigkeit, zwischen Farbnuancen zu unterscheiden, haben aber nicht immer Begriffe für diese Unterscheidungen.

Die Form, die der Substanz gegeben wird, bezeichnet de Saussure als Valeur. Der Wert eines sprachlichen Zeichens kann wie folgt formalisiert werden: Sei Z eine endliche Menge von Zeichen $z_1 \ldots z_n$. Der Wert eines bestimmten Zeichens z_i ist nun $Z - (z_1 \ldots z_i - 1 \ldots z_i + 1 \ldots z_n)$. Da N ein endlicher Wert ist, lässt sich diese Formel nur auf endliche Mengen von phonologischen oder semantischen Einheiten anwenden. Diese Auffassung des Wertes eines sprachlichen Zeichens hat sich als besonders fruchtbar erwiesen für die Theorie lexikalischer Felder[7], die auf endlichen Mengen von lexikalischen Einheiten errichtet werden. Für das offene Vokabular lebender Sprachen ist diese Formalisierung aber weniger gut geeignet.

[7] Ein Paradebeispiel ist das lexikalische Feld der Farbbezeichnungen. Das objektiv vorhandene Farbspektrum wird in verschiedenen Sprachen durch unterschiedlich große Mengen von Ausdrücken abgedeckt und unterteilt, so dass ein Ausdruck innerhalb des Vokabulars einer Sprache ein bestimmtes Spektrum bezeichnet: *grün* ist das, was nicht *blau, gelb* etc. ist.

Wenn zum Beispiel eine Sprache nur ein Adjektiv zur Verfügung hat, um auszudrücken, dass etwas groß ist (Z_{de} = groß), dann ist die Valeur dieses Zeichens höher als in einer Sprache, die für diesen Begriff drei Ausdrücke zur Verfügung hat (Z_{en} = big, large, huge).

Ein anderes Beispiel ist die Benennung von Wörterbüchern: Der Ausdruck *Handwörterbuch*, der für sich genommen schwer zu interpretieren ist (ein Wörterbuch, das in eine Hand passt? ein Wörterbuch, das immer zur Hand ist?), erhält eine klare Bedeutung als Teil eines Feldes von Wörterbuchbezeichnern wie *Miniwörterbuch, Taschenwörterbuch, Handwörterbuch, Großwörterbuch*, in dem jeder dieser Bezeichner auf einen Wörterbuchtyp einer gewissen Größe referiert.

Auf die lexikalische Semantik bezogen, nimmt de Saussure an, dass das individuelle lexikalische Zeichen zwei Seiten hat: die Formseite (SIGNIFIANT) und die Inhaltsseite (SIGNIFIÉ). Beide Seiten zusammen bilden das lexikalische Zeichen. Beide Seiten sind dazu geeignet, als Ordnungsaspekt für Wörterbücher zu fungieren.

Saussure verwendet hierfür die Metapher eines Stücks Papier, bei dem die Formseite die Vorderseite und die Inhaltsseite die Rückseite bildet. Wenn die Vorderseite verschwindet, dann verschwindet automatisch auch die Rückseite, und umgekehrt (vgl. de Saussure (2001), S. 101).

Die Beziehung von Form und Inhalt ist allerdings arbiträr, und es wird durch Konvention zwischen den Sprachbenutzern festgelegt, welcher Begriff z.B. mit der Formseite ‚TISCH' verbunden wird (vgl. de Saussure (2001), S. 66ff.). In anderen Sprachen ist dieser Begriff mit einer anderen lexikalischen Form verbunden. Das Wirken der Konvention bei der Ausprägung des Vokabulars einer Sprache kann man anhand der Etablierung neuer Wörter erkennen. So gab es für die Sportschuhe mit Rollen zeitweise zwei Wörter: *Rollerblades* und *Inlineskates*. Die Sprachgemeinschaft hat sich letztendlich für das zweite Wort als die konventionelle lexikalische Form entschieden.

Zusammenfassend kann man sagen, dass Saussures semantische Theorie ATOMISTISCH oder ANALYTISCH ist. Saussure geht davon aus, dass die Inhaltsseite eines lexikalischen Zeichens weiter zerlegt werden kann in individuelle Konzepte. Zugleich ist seine Theorie, wie der Begriff der Valeur zeigt, HOLISTISCH. Sprache ist ein System oder eine Struktur, in der alle Elemente miteinander verbunden sind. Seine Theorie ist UNIVERSAL insofern, als er eine einzelsprachenübergreifende Substanz von Form und Bedeutung annimmt, die in jeder Sprache anders strukturiert wird. Sie ist MENTAL insofern, als Saussure sich auf Lautformen und Bedeutungen als mentale Zustände bzw. Gedanken bezieht.

Seine Annahme einer engen Beziehung zwischen lexikalischer Form und lexikalischer Bedeutung, die in der Papiermetapher zum Ausdruck kommt, setzt seiner Theorie allerdings Grenzen. Im Rahmen dieses Konzepts des le-

xikalischen Zeichens ist es nicht möglich, Phänomene wie Polysemie und Synonymie angemessen darzustellen. Deshalb ist Saussures Theorie unzureichend als Basis für die (Computer)-Lexikographie. Sie wurde in der Folge denn auch modifiziert. Wir werden uns diese Modifikationen in den folgenden Abschnitten ansehen.

2.2 Modifikationen im Rahmen des Strukturalismus

Die strukturalistische Linguistik in der Folge von de Saussure modellierte weitere Aspekte der Form, der Bedeutung und der Funktion lexikalischer Zeichen.

Ogden und Richards (vgl. 1949) entwickelten ein Modell des sprachlichen Zeichens, das sie als Dreieck darstellten. An den Ecken des Dreiecks findet man SYMBOL, also die Formseite des Zeichens, THOUGHT bzw. REFERENCE als die Inhaltsseite des Zeichens und, am rechten unteren Ende, den REFERENT als außersprachlichen Bezugspunkt. Ein Symbol „symbolisiert" ein Gedankenobjekt und steht für ein Referenzobjekt. Das Referenzobjekt ist das außersprachliche Korrelat des sprachlichen Zeichens. Der Akt der Referierens wird als ein kognitiver Prozess betrachtet – durch Gebrauch des sprachlichen Zeichens wird auf etwas Außersprachliches referiert. Während

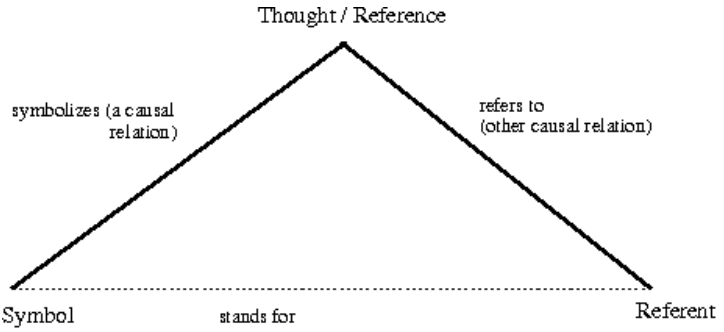


Abbildung 1: Modell des sprachlichen Zeichens nach Ogden und Richards

also de Saussure auf Substanz und Form sprachlicher Zeichen im Verhältnis zu mentalen Zuständen der Sprecher abzielt, erweitern seine Nachfolger das Bild um den außersprachlichen Referenten und die Funktion sprachlicher Zeichen, auf Außersprachliches zu referieren.

Stephen Ullmann (1962) stellt sein Modell des lexikalischen Zeichens ebenfalls als Dreieck dar und projiziert die Formseite des Zeichens (hier NAME genannt) und die Inhaltsseite (SENSE) ebenfalls auf die linke Seite. Auf der rechten Seite finden wir wieder den außersprachlichen Bezugspunkt

(THING). Mit SENSE wird entweder der mentale oder der informationelle Inhalt des Zeichens bezeichnet. Ullmanns Auffassung zufolge ist die Untersuchung der Beziehung zwischen dem Zeicheninhalt und seinem außersprachlichen Bezugsobjekt kein Gegenstand der linguistischen Forschung (vgl. Ullmann (1962)).

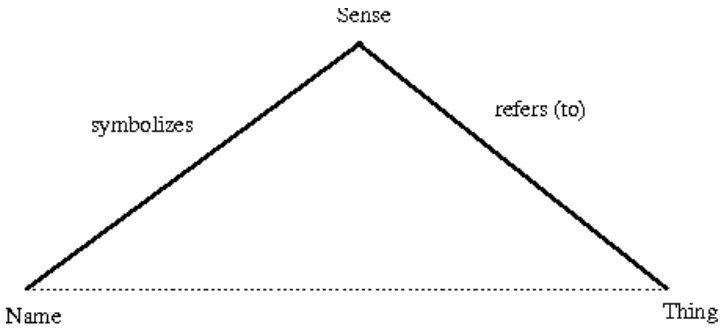


Abbildung 2: Modell des sprachlichen Zeichens nach Ullmann

Heger erweitert das Modell des lexikalischen Zeichens um eine weitere Ebene und entwickelt das Dreieck so zu einem Trapez weiter. Nach wie vor befindet sich die Beziehung zwischen Form- und Inhaltsseite des Zeichens auf der linken Seite und die Beziehung zum außersprachlichen Referenten auf der rechten Seite. Der entscheidende Unterschied ist nun, dass durch das Auffalten der Spitze des Dreiecks die Inhaltsseite des sprachlichen Zeichens als etwas Strukturiertes dargestellt werden kann, nämlich als eine Kombination von Bedeutungselementen, die Heger SEME nennt[8]. Eine Zeichenform kann auf diese Weise mit einem Konglomerat von Bedeutungen verbunden werden. Dies ist das Merkmal der Polysemie (*Satz* → /großer Sprung/, /sprachliche Einheit/ etc.). Mehrere elementare Bedeutungseinheiten formen ein SEMEM, eine komplexe Bedeutungseinheit. Da dieses Modell es also erlaubt, komplexe Bedeutungseinheiten aus einfacheren Elementen zu konstruieren, können damit lexikalisch-semantische Beziehungen definiert werden, für die die älteren Modelle nicht ausgestattet waren. Dazu gehören:

- **Synonymie**: Zwei sprachliche Einheiten verfügen über Inhaltsseiten, die ein Semem gemeinsam haben (*Computer* und *Rechner* haben die Bedeutung ‚elektronische Rechenanlage' gemeinsam. *Rechner* hat darüber ein weiteres Semem, das auf rechnende Menschen referiert). Die Klassen der außersprachlichen Objekte, auf die die beiden sprachlichen Zeichen in

[8] Also ist Semantik die Lehre von den Semen.

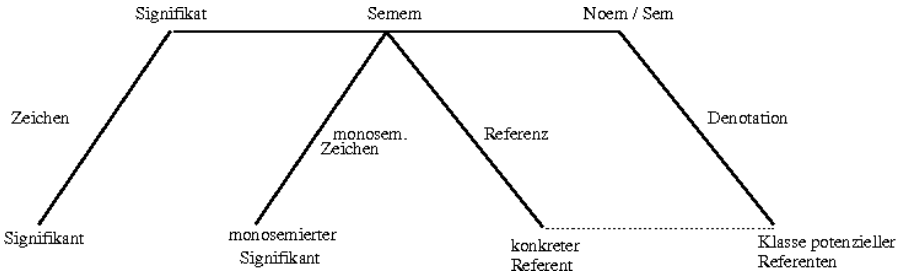


Abbildung 3: Modell des sprachlichen Zeichens nach Heger

dieser durch das gemeinsame Semem repräsentierten Bedeutung referieren, sind extensionsgleich[9].

- **Antonymie**: Die Sememe zweier sprachlicher Zeichen sind so beschaffen, dass alle bis auf ein Sem gleich sind. *Junge* und *Mädchen* teilen die Seme /belebt/, /menschlich/, /jung/, unterscheiden sich aber in dem Sem, das auf das Geschlecht referiert (/männlich/ vs. /weiblich/). Die Klassen der außersprachlichen Objekte, auf der die beiden sprachlichen Zeichen hinsichtlich des gemeinsamen Semems referieren, sind disjunkt.
- **Hyponymie**: Die Sememe zweier lexikalischer Zeichen, die in der Relation der Hyponymie zueinander stehen, teilen sich einen gewissen Bestand an Semen. Die Bedeutungsseite des Hyponyms, also des spezielleren Begriffs, hat darüber hinaus weitere Seme. So teilen sich die lexikalischen Zeichen *Pflanze* und *Rose* einige Seme, z.B. /Ding/, /belebt/, *Rose* hat aber weitere Seme, die die „Rosenhaftigkeit“ ausmachen, z.B. /hat Dornen/. Die Menge der Referenten des Unterbegriffs ist eine Untermenge der Referenten des Oberbegriffs. Mit anderen Worten: jede Rose ist eine Pflanze, aber nicht jede Pflanze ist eine Rose.
- **Kohyponymie**: Die Sememe zweier Kohyponyme haben eine gewisse Menge von Semen gemeinsam, nämlich diejenigen, die sie mit dem gemeinsamen Hyperonym (Oberbegriff) teilen. Darüber hinaus unterscheiden sie sich in mindestens einem Sem (*Schimmel* und *Rappen* teilen sich die Seme, die sie mit ihrem Oberbegriff *Pferd* gemeinsam haben, die Farbe ihres Fells ist ein Bedeutungselement, das die beiden unterscheidet). Die Klassen der Referenten zweier Kohyponyme sind disjunkt.

Zum Abschluss dieses Abschnitts fassen wir die hier dargestellten strukturalistischen Theorien des lexikalischen Zeichens zusammen:

- Diese Theorien sind lokal ATOMISTISCH. Es wird davon ausgegangen, dass die Bedeutung sprachlicher Zeichen sich in Bedeutungselemente zer-

[9] Die Extension eines sprachlichen Zeichens sind die Objekte oder Klassen von Objekten, auf die sich dieses sprachliche Zeichen bezieht.

legen lässt. Die Bedeutungselemente korrespondieren zu mentalen Zuständen und referieren auf Klassen außersprachlicher Dinge und Sachverhalte.

- Der Begriff des Sems als Basiselement der Bedeutung sowie kombinatorische Operationen, die diese Seme zu größeren Einheiten, den Sememen, zusammenbringen, erlauben eine angemessene Darstellung einer Reihe von lexikalisch-semantischen Beziehungen.

Der theoretische Rahmen des Strukturalismus, in welchen diese Modelle der lexikalischen Semantik eingebettet sind, scheint heute überholt. Innerhalb dieses theoretischen Rahmens war und ist es nicht möglich, mehr als einige Bereiche des Vokabulars zu beschreiben, die sich für eine solche Beschreibung besonders gut eignen, z.B. das lexikalische Feld der Verwandtschaftsbeziehungen. Trotzdem haben diese Theorien einen bedeutenden Einfluss auf die Semantik und die (Computer-)Lexikographie gehabt.

3 Komponentielle Semantik

3.1 Der Ansatz von Katz und Fodor

Gerald Katz and Jerry Fodor entwickelten Ende der 60er und Anfang der 70er Jahre des vergangenen Jahrhunderts eine semantische Metatheorie im Rahmen der generativen Semantik. Diese Metatheorie nennt die Kriterien, denen eine semantische Theorie für sprachliche Zeichen natürlicher Sprache genügen muss.

> A semantic metatheory must provide criteria for evaluating individual semantic theories and establish the adequacy of such criteria. (Katz und Fodor (1963), S. 208)

Im Allgemeinen muss eine semantische Theorie die Fähigkeit von Sprechern einer natürlichen Sprache erklären, eine theoretisch unendliche Menge wohlgeformter Äußerungen zu produzieren bzw. zu verstehen, d.h. korrekt zu interpretieren und explizieren zu können. Insbesondere muss eine semantische Theorie erklären können, wie Sprecher einer Sprache

- die unterschiedlichen Lesarten von Sätzen und deren semantischen Inhalt bestimmen können;
- semantische Abweichungen erkennen;
- entscheiden, ob ein Satz die Paraphrase eines anderen Satzes ist oder nicht.

Man erkennt den engen Bezug dieser semantischen Metatheorie zum Programm der generativen Grammatik. So bildet denn auch die Sprecherkompetenz den Bezugspunkt dieser semantischen Metatheorie: zur Kompetenz gehört z.B. die Fähigkeit, semantische Anomalien und bedeutungsgleiche Äußerungen zu erkennen.

Im Bereich der Konstruktion von lexikalischen Einträgen, also Beschreibungen lexikalischer Einheiten, führen Katz und Fodor die Begriffe MARKER und DISTINGUISHER ein.

> The semantic markers and distinguishers are the means by which we can decompose the meaning of one sense of a lexical item into its atomic concepts, and thus exhibit the semantic structure in a dictionary entry and the semantic relations between dictionary entries. That is, the semantic relations among the various senses of different lexical items are represented by formal relations between markers and distinguishers. (Katz und Fodor (1963), S. 185f.)

Die Marker entstammen einem begrenzten Vokabular zu einem gegebenen „konzeptuellen Raum“ (‚conceptual space‘). Sie bilden die primären lexikalischen Deskriptoren. Distinguisher sind sekundäre lexikalische Deskriptoren,

deren Zweck es ist, Wortbedeutungen bis zum notwendigen Detaillierungsgrad zu unterscheiden. In Abbildung 4 ist das Konzept *bachelor* (‚Junggeselle') dargestellt. Marker sind mit runden, Distinguisher mit eckigen Klammern gekennzeichnet.

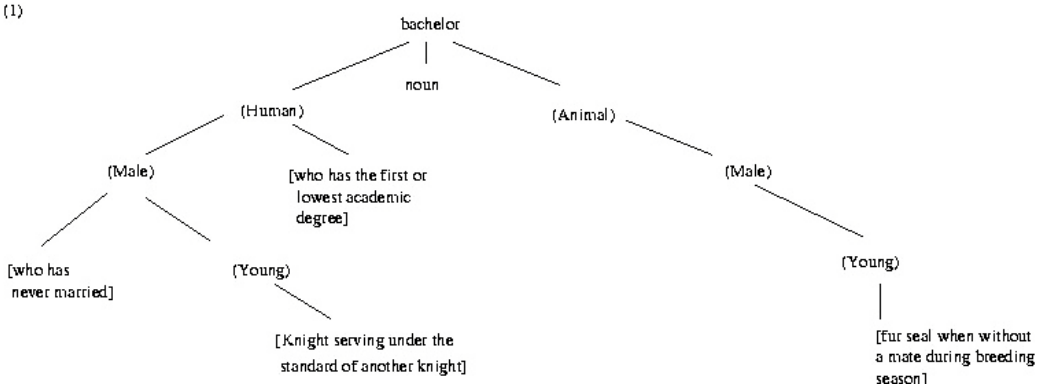


Abbildung 4: Marker und Distinguisher am Beispiel der semantischen Struktur von *bachelor*

Katz und Fodors Annahme, die sie mit den meisten komponentiell orientierten Semantikern teilen, ist, dass es eine Menge von semantischen Basiskomponenten gibt. In diese Basiskomponenten können alle lexikalischen Zeichen – genauer: deren Inhaltsseiten – zerlegt werden. Der Begriff der Wortbedeutung beruht auf diesen Basiselementen und der vollständigen Zerlegbarkeit der Wortbedeutungen in diese Basiselemente, welche ein sprachunabhängiges, UNIVERSALES Vokabular bilden. Die Elemente dieses Vokabulars wiederum repräsentieren KONZEPTE, die die mentalen Korrelate der Wortbedeutungen bilden. Dieses Vokabular von Basiselementen sei hinreichend, um eine unendliche Menge von Äußerungen zu produzieren.

Komponentielle Ansätze der Semantik waren auch im Bereich der künstliche-Intelligenz-Forschung populär. Eine endliche Menge von semantischen Basiseinheiten ist sehr praktisch, wenn man Bedeutungs- und Weltwissen in einer formalen und damit für den Rechner verarbeitbaren Weise modellieren möchte, z.B. für die maschinelle Übersetzung mithilfe einer Interlingua, vgl. Kapitel 6.

Das Prinzip der vollständigen Zerlegbarkeit von Wortbedeutungen in einfachere Basiseinheiten ist aus der Phonologie übernommen. In dieser linguistischen Teildisziplin hat man erfolgreich das Phomeninventar der Sprachen auf eine begrenzte Menge sog. distinktiver Merkmale reduzieren können. Die folgende Äußerung von Chomsky (zitiert bei Pulman) belegt, dass diese Analogie tatsächlich gezogen wurde:

> [T]he very notion ‚lexical entry' presupposes some sort of fixed, universal vocabulary in terms of which these objects are characterized, just as the notion ‚phonetic representation' presupposes some sort of universal phonetic theory. (zit. in Pulman (1983), S. 29)

In der Tat ist das Unterfangen, ein universales Vokabular von semantischen Basiseinheiten zu finden, faszinierend:

- Ein solches Vokabular wäre eine generative Bedeutungskomponente, da prinzipiell eine unendliche Menge von Wortbedeutungen aus dieser endlichen Menge konstruiert werden könnte.
- Aufgrund der möglichen Kombinationen könnten die lexikalischen Lücken von Einzelsprachen als nicht realisierte Kombinationen der Basiseinheiten identifiziert werden (im Deutschen gibt es z.B. kein lexikalisches Zeichen, das den Zustand ‚keinen Durst mehr habend' bezeichnet).
- Lexikalische Zeichen könnten in Felder gruppiert werden, deren Struktur durch Oppositionen der Basiselemente gebildet wird.

Wir haben uns bei dieser Darstellung bewusst für den Irrealis entschieden, da sich dieses Programm im weiteren Verlauf als undurchführbar und die damit verbundenen wissenschaftlichen Perspektiven als unrealistisch erwiesen haben.

Schon bald wurde Kritik an der undifferenzierten Analogie zur Phonologie laut. So schreibt etwa Pulman (vgl. Pulman (1983), S. 30):

> [...] the distinctive features of phonemes are in principle relatable to properties of the human vocal tract, acoustic properties and properties of the perceptual system, and the set of distinctive features is therefore constrained by the observable facts of human physiology. Languages are strictly comparable with respect to these properties. Nothing of this holds for semantic markers or ‚concepts'. The existence of a limited set of basic concepts is mere speculation as is our intuition that ‚the same concept' is expressed by lexical items in different languages.

Es ist unmöglich, ein System von semantischen Markern als Basiselementen der Bedeutung auf die Gegebenheiten eines beobachtbaren, außersprachlichen Systems zu stützen. Wir haben keinen Zugang zu den mentalen Zuständen, die den Konzepten entsprechen könnten.

Wenn man also, wie dies wiederholt gemacht wurde, das Verb *to kill* (‚töten') auf einen Ausdruck CAUSE(X, Y) AND BECOME_NOT_ALIVE(Y) einer semantischen Metasprache abbildet, so verwendet man da-

mit noch lange keine Konzepte, deren Existenz nachgewiesen wäre, sondern lediglich andere Ausdrücke der englischen Sprache[10].

Neben diesen prinzipiellen wissenschaftlichen Problemen der komponentiellen Semantik entstehen die folgenden praktischen Probleme, wenn man dieses Programm ernsthaft verfolgen wollte:

- Es besteht keine Einigkeit über den Inhalt und Umfang des Vokabulars einer semantischen Metasprache. Entsprechend problematisch ist die Abgrenzung zwischen Markern und Distinguishern.
- Es gibt keine Einigkeit über die Methoden, mit denen man semantische Marker entdecken könnte.
- Es gibt keine Einigkeit darüber, welches Vokabular von Basiselementen ausreichend ist, um alle möglichen Wortbedeutungen zu konstruieren, oder zumindest die existierenden Wortbedeutungen einer Sprache. Wenn man das Programm der Ermittlung semantischer Basiseinheiten exhaustiv verfolgt, dann wird dieses Vokabular mit großer Wahrscheinlichkeit den Umfang des natürlichen Vokabulars der untersuchten Sprache erreichen. Damit entfällt natürlich auch jegliche Rechtfertigung für die Bildung eines Metavokabulars.

Trotz dieser prinzipiellen und praktischen Probleme hatte die komponentielle Semantik Auswirkungen auf die praktische Lexikographie.

Die komponentielle Semantik bietet den Rahmen, um die komplexe Bedeutung lexikalischer Zeichen (z.B. *töten*) als eine Kombination von lexikalischen Zeichen mit einfacherer Bedeutung darzustellen (z.B. *bewirken*, *nicht*, *leben*). Die Menge der hierfür benötigten lexikalischen Zeichen mit einfacher Bedeutung könnte das Basisvokabular lexikalischer Bedeutungsbeschreibungen bilden. Basisvokabular und die mit diesem Vokabular beschriebenen sprachlichen Zeichen, also Objektsprache und Metasprache, entstammen dabei derselben natürlichen Sprache, z.B. dem Englischen oder dem Deutschen.

So verwendet z.B. das ‚Longman Dictionary of Contemporary English' eine Liste von Basiseinheiten, das sog. *Defining Vocabulary* (vgl. Quirk (1995), S. B16). Es umfasst ca. 2000 lexikalische Einheiten. In den Bedeutungsbeschreibungen der anderen lexikalischen Einheiten werden, wo immer dies möglich ist, nur diese Basislexeme verwendet. Dahinter steht die Überlegung, dass Lerner, die zunächst die Bedeutungen dieser elementaren lexikalischen Einheiten lernen, mithilfe dieser einfacheren Einheiten die Bedeutungen der schwierigeren lexikalischen Einheiten entschlüsseln können.

[10] Man könnte auch Ausdrücke der polnischen Sprache verwenden, das macht keinen Unterschied. Wichtig ist, dass man keinen Zugang auf die Konzepte hinter den sprachlichen Ausdrücken hat.

3.2 Der Ansatz von Wierzbicka

In seiner kritischen Würdigung der komponentiellen Semantik versucht Pulman, die Idee der semantischen Marker dadurch zu retten, dass er diesen den Status von normalen englischen Wörtern gibt:

> Consider the claim that AND, LIKE and INCHOATIVE are semantic primes in the sense that they are part of a basic sub-vocabulary of English [...] suitable for the partial or total description of many other English words which they can, in combination, paraphrase [...] the enterprise of semantic description on the level of word meaning is the adoption of this sub-vocabulary as a metalanguage. (vgl. Pulman (1983), S. 37)

Dieses Zitat beschreibt ziemlich gut das Forschungsprogramm von Anna Wierzbicka, die versucht, eine Menge von semantischen Primitiven als Untermenge des Vokabulars der Objektsprache festzulegen[11]. Sie schreibt:

- The lexicon of any language can be divided into two parts: a small set of words [...] that can be regarded as indefinable, and a large set of words that [...] in fact can be defined in terms of the words from the set of indefinables.
- For any language, its indefinables can be listed [...]
- Although the set of indefinables is in each case language specific, one can hypothesize that each such set realizes, in its own way, the same universal and innate ‚alphabet of human thought'.

(Wierzbicka (1992), S. 209)

Wierzbicka postuliert zunächst 14 semantische Primitive für das Englische, unter Anderen: I, WANT, KIND, NO[12].

Die Bedeutungen der anderen englischen Wörter seien „Konfigurationen" dieser semantischen Primitive. In ihrem hier zitierten Aufsatz aus dem Jahr 1992 beschreibt sie unter anderem das Wortfeld der (englischen) Verben, die Sprechakte bezeichnen. Sprechakte sind für sie „things that one can do with words".

In den nun folgenden Beispielen aus dieser Arbeit werden wir den definierten Term in Großbuchstaben schreiben. Die Definitionen selbst stehen in einfachen Anführungszeichen.

- ASK und ORDER: ‚(I say:) I want you to do it'
- ORDER impliziert: ‚(I think:) you have to do it'
- ASK impliziert dies nicht: ‚(I think:) you don't have to do it because of this'.

[11] vgl. Wierzbicka (1992), S. 209ff.
[12] Für die vollständige Liste vgl. Wierzbicka (1992), S. 210.

Eine Stärke des Ansatzes, Bedeutungen mithilfe eines kontrollierten Vokabulars zu paraphrasieren, liegt darin, dass man semantische Differenzen, die Unterschieden in den syntaktischen Verwendungsweisen der lexikalischen Einheiten entsprechen, genauer herausarbeiten kann.

So beinhalten die Sprechakte PLEAD, ARGUE und REASON (‚plädieren', ‚streiten', ‚auseinandersetzen') den Austausch von Argumenten. Dementsprechend kann die Rolle des Adressaten syntaktisch realisiert werden: *plead, argue, reason* WITH SOMEBODY.

Diese Verknüpfung von semantischer und syntaktischer Ebene kann allerdings zu einer zirkulären Argumentation führen. Die syntaktischen Verwendungsmuster eines Wortes sind der Beobachtung – z.B. in einem Textkorpus – unmittelbar zugänglich, die Bedeutung eines Wortes aber bestenfalls mittelbar. Man könnte geneigt sein, aus Differenzen in der syntaktischen Verwendungsweise zweier Wörter auf semantische Unterschiede zu schließen und anschließend zu behaupten, dass diese Bedeutungsunterschiede die Differenzen in der syntaktischen Verwendungsweise „bewirken".

Einige der Einwände, die gegen das strukturalistische Konzept von Semen und Sememen und auch gegen den Ansatz von Katz und Fodor vorgebracht wurden, können hier wiederholt werden:

Zunächst wirkt die Auswahl des Basisvokabulars von semantischen Primitiven arbiträr. Es gibt keine außersemantische Argumentation, mit der diese Auswahl gerechtfertigt werden könnte. Wierzbicka erweitert im Laufe ihrer Arbeit das Vokabular der semantischen Primitive von zunächst 14 auf 30.

Ebenso bleibt die Behauptung, dass diese Menge von semantischen Primitiven, seien es nun 14 oder 30, das Basisvokabular des menschlichen Denkens bilde, jedenfalls in seiner englischsprachigen Version, unbewiesen. Diese Behauptung ist für die praktische Arbeit mit diesen semantischen Primitiven allerdings unerheblich.

Die Zerlegung von Bedeutungen in diese semantischen Primitive ist dennoch nützlich, um generische Schemata oder Bedeutungskonfigurationen zu ermitteln sowie Beziehungen zwischen einzelnen Bedeutungen. Damit lassen sich sowohl Polysemiestrukturen einzelner lexikalischer Zeichen als auch lexikalisch-semantische Beziehungen zwischen lexikalischen Zeichen formal als Gemeinsamkeiten und Differenzen in den Bedeutungskomponenten darstellen.

3.3 Das generative Lexikon

Ein neuerer Ansatz der komponentiellen Semantik, der viele Anhänger in der Computerlinguistik gefunden hat, stammt von James Pustejovsky. Seine Beliebtheit bei Computerlinguisten ist allerdings nicht der einzige Grund, diesen

Ansatz hier zu besprechen. Pustejovskys Ansatz hat auch einige interessante Arbeiten in der Computerlexikographie inspiriert.

Sein Begriff der semantischen Primitive weicht stark von der „traditionellen“ Auffassung, wie sie etwa von Wierzbicka vertreten wird, ab. Er sucht stattdessen:

> [...] a new way of viewing primitives, looking more at the generative or compositional aspects of lexical semantics, rather than the decomposition into a specified number of primitives [...] (Pustejovsky (1991), S. 417)

Pustejovsky betrachtet das Verhältnis von logischer und syntaktischer Form sprachlicher Äußerungen. Die syntaktische Struktur sprachlicher Äußerungen ist der Ausgangspunkt seiner Beschreibungen. Ohne deren Untersuchung und Beschreibung sei eine lexikalisch-semantische Theorie zum Scheitern verurteilt[13].

Pustejovskys Ansatz ist es, die logische Form von Äußerungen auf das Lexikon im Allgemeinen und auf generative Mechanismen (GENERATIVE DEVICES) des Lexikons im Besonderen zu stützen.

Durch eine vollständig kompositionelle Semantik natürlicher Sprache versucht Pustejovsky, die generative Kapazität der Sprache zu erklären[14]. Dies umfasst die Fähigkeit der Sprecher, semantisch wohlgeformte von nichtwohlgeformten Äußerungen zu unterscheiden. Wir haben diese Fähigkeit bereits im Ansatz von Katz und Fodor als Kriterium einer semantischen Theorie kennengelernt.

Stärkeren Bezug zur lexikalischen Semantik haben Pustejovskys Versuche, Erklärungen für die sprachlichen Phänomene der METONYMIE[15] und der POLYSEMIE zu finden.

Metonymie

Pustejovskys Begriff der Metonymie geht auf Geoffrey Nunberg (1978) zurück. Danach bedeutet dieser Begriff, dass eine Phrase an Stelle einer anderen Phrase gebraucht wird. Pustejovsky gibt die folgenden Beispiele:

(1) *John began the book* (John begann das Buch). Erläuterung: die Bedeutung kann sein, dass John begann, ein Buch zu lesen oder ein Buch zu schreiben.

[13] Vgl. Pustejovsky (1991), S. 410.

[14] Vgl. Pustejovsky (1991), S. 419.

[15] Die Metonymie ist eine Stilfigur, bei der ein Ausdruck durch einen anderen ersetzt wird, der mit ersterem in sachlichem, aber nicht in semantisch-begrifflichem Zusammenhang steht, z.B. Ersetzung eines Wortes, das ein Getränk bezeichnet, durch ein Wort, das ein Gefäß bezeichnet, das dieses Getränk typischerweise enthält, in *Ich nehme noch ein Glas*.

(2) *John began the cigarette* (John begann die Zigarette). Erläuterung: John begann damit, die Zigarette zu rauchen.

(3) *John began the beer* (John begann das Bier). Erläuterung: John begann damit, das Bier zu trinken.

(4) *Mary enjoyed the book* (Maria genoss das Buch). Erläuterung: Maria genoss es, das Buch zu lesen.

(5) *Mary enjoyed the cigarette* (Maria genoss die Zigarette).

In all diesen Beispielen übernimmt das Objekt des Satzes (*das Buch, das Bier*, etc.) die Rolle der Verbalphrase, die die eigentliche Handlung ausdrückt (*lesen, trinken* etc.). Die Ereignislesart des (Teil-)Satzes, die normalerweise durch das Verb vermittelt wird, das ein Teil des kompletten Arguments des Hauptverbs wäre (*beginnen*), wird nun durch die Nominalphrase getragen. In Pustejovskys Worten wird der Kopf der Objekt-Nominalphrase in die Rolle des Ereignistyps gezwungen.

Reguläre Polysemie

Um das Phänomen der regulären Polysemie sprachlicher Zeichen zu erklären, wählt Pustejovsky die folgenden Beispiele:

(6) *He baked the potato* (Er backte die Kartoffel).

(7) *He baked the cake* (Er backte den Kuchen).

In Beispiel (6) wird eine Zustandsänderung (der Kartoffel) ausgedrückt, wohingegen in Beispiel (7) ein Objekt (der Kuchen) geschaffen wird. Anstatt nun einen Bedeutungswechsel beim Verb anzunehmen, und dieses als polysem zu beschreiben, geht Pustejovksy einen anderen Weg und schreibt den Bedeutungsunterschied in beiden Sätzen allein dem Objekt zu:

> [W]e can derive both word senses of verbs like *bake* by putting some of the semantic weight on the NP. This view suggests that [...] the verb itself is not polysemous. (Pustejovsky (1991), S. 423)

Was an den obigen Beispielen anhand einer Verb-Komplement-Struktur[16] gezeigt wurde, funktioniert auch bei Nomen-Modifikator-Strukturen[17], wie die folgenden Beispiele zeigen:

(8) *She is a fast typist* (Sie ist eine schnelle Tipperin). Erläuterung: Sie ist eine Person, die schnell tippt.

[16] Eine Verb-Komplement-Struktur ist ein Fügung aus einem Verb (z.B. *backen*) und dessen notwendiger Ergänzung (z.B. *Kuchen*).

[17] Eine Nomen-Modifikator-Struktur ist eine Fügung aus einem Substantiv (z.B. *Entscheidung*) und einem modifizierenden Element, meistens einem Adjektiv (z.B. *schnell*).