Universität Zürich
Computerlinguistik

Lexikalisch-semantische Netze: Anwendungsperspektiven für die Computerlinguistik

Lizentiatsarbeit der Philosophischen Fakultät der Universität Zürich

Referent: Prof. Dr. M. Hess

Verfasserin:
Sandra Roth
Langfurren 5
8057 Zürich

31. Mai 2006

Zusammenfassung

Lexikalisch-semantische Netze wie WordNet geniessen in der Computerlinguistik einen hohen Stellenwert, da sie in Anwendungen wie Information Retrieval, Maschineller Übersetzung, Antwortextraktion oder automatischer Rechtschreibekorrektur einen wichtigen Beitrag leisten. Diese Arbeit führt in die lexikalische Semantik ein und stellt gängige maschinenlesbare Wortnetze verschiedener Sprachen vor. Ein weiterer Schwerpunkt liegt auf den Einsatzmöglichkeiten solcher Wortnetze in computerlinguistischen Anwendungen. Lexikalisch-semantische Ressourcen können helfen, die immer wieder auftretenden Probleme wegen ambiger und anaphorischer Konstrukte zu lösen. Ihr Einsatz ermöglicht solide Disambiguierungs- und Anaphernresolutionsverfahren und bietet somit Perspektiven, die Performanz von solchen Systemen zu verbessern.

Abstract

Lexical-semantic nets like WordNet attracted the attention of computational linguists because the nets make an important contribution to systems such as Information Retrieval, Machine Translation, Question Answering or automatic spelling correction. This paper starts with an introduction to lexical semantics and presents different machine-readable wordnets and concentrates on the use of such wordnets for NLP-applications. Lexical-semantic ressources help solving recurring problems due to ambiguities and anaphoras. Therefore, their use enables solid word sense disambiguation and anaphora resolution techniques and provides an outlook to improve the performance of such systems.

Inhaltsverzeichnis

Zusammenfassung i

Inhaltsverzeichnis ii

Abbildungsverzeichnis vi

Tabellenverzeichnis vii

Abkürzungsverzeichnis viii

1 Einleitung 1

1.1 Ausgangslage 1

1.2 Ziel und Aufbau der Arbeit 2

2 Einführung in die lexikalische Semantik 4

2.1 Analyseeinheiten der lexikalischen Semantik 5

2.2 Variabilität von Wortbedeutungen 6

2.3 Lexikalisch-semantische Relationen 7

2.3.1 Synonymie 9

2.3.2 Hyponymie/Hyperonymie 10

2.3.3 Meronymie 11

2.3.4 Antonymie 12

2.4 Frame-Semantik 14

2.5 Einordnung von lexikalischen Einheiten in Wortnetze 15

2.5.1 Taxonomie 16

2.5.2 Thesaurus 16

2.5.3 Ontologie 16

3 Die WordNet-Familie 18

3.1 WordNet 18

3.1.1 Motivation 18

3.1.2 Aufbau und semantische Relationen 19

3.1.2.1 Aufbau der Nomen-Datei 20
3.1.2.2 Aufbau der Verb-Datei 22
3.1.2.3 Aufbau der Adjektiv-Datei 23
3.1.2.4 Aufbau der Adverb-Datei 25
3.2 GermaNet 25
3.2.1 Motivation 25
3.2.2 Aufbau und semantische Relationen 26
3.2.2.1 Aufbau der Nomen-Datei 27
3.2.2.2 Aufbau der Verb-Datei 28
3.2.2.3 Aufbau der Adjektiv-Datei 29
3.2.2.4 Adverbien 30
3.3 GermaNet und WordNet im Vergleich 30
3.4 EuroWordNet 31
3.4.1 Motivation 32
3.4.2 Aufbau und semantische Relationen 32
3.4.3 Integration von GermaNet in EuroWordNet 35
3.5 Weitere Wortnetze der WordNet-Familie 36
3.6 Einordnung 36
3.7 Kritik an der WordNet-Familie 37
3.7.1 Tennis-Problem 37
3.7.2 Unvollständigkeit 38
3.7.3 Sprachbedingte Abhängigkeit von EuroWordNet 38
3.7.4 Vernachlässigung systematischer Polysemie 38

4 Weitere Wortnetze 39
4.1 Die FrameNet-Familie 39
4.1.1 FrameNet 39
4.1.2 SALSA und weitere FrameNets 41
4.2 Roget's Thesaurus 43
4.3 Weitere deutsche Wortnetze 44
4.3.1 OpenThesaurus 45
4.3.2 HaGenLex 46
4.3.3 LexikoNet 47
4.4 Weitere englische Wortnetze 48
4.4.1 Cyc und OpenCyc 48
4.4.2 Omega 50
4.5 Vergleich der Wortnetze 50
4.5.1 Die FrameNet- und WordNet-Familien im Vergleich 50
4.5.2 Roget's Thesaurus und WordNet im Vergleich 52

4.5.3 Weitere deutsche Wortnetze und GermaNet im Vergleich 52
4.5.4 Weitere englische Wortnetze und WordNet im Vergleich 53

5 Grenzen computerlinguistischer Anwendungen 55
5.1 Information-Retrieval-Systeme 55
5.1.1 Problemstellung 56
5.1.2 Lösungsansätze 57
5.2 Antwortextraktionssysteme 58
5.2.1 Problemstellung 59
5.2.2 Lösungsansätze 60
5.3 Maschinelle Übersetzungssysteme 61
5.3.1 Problemstellung 63
5.3.2 Lösungsansätze 63
5.4 Korrekturprogramme 65
5.4.1 Problemstellung 65
5.4.2 Lösungsansätze 66

6 Lesartendisambiguierung mit Wortnetzen 67
6.1 Ansätze zur Lesartendisambiguierung 68
6.1.1 Wörterbuch- und kontextbasierte Ansätze 68
6.1.2 Ähnlichkeitsbasierte Ansätze 69
6.1.2.1 Einbezug von Informationsgehalt 70
6.1.2.2 Einbezug von konzeptueller Dichte 72
6.1.2.3 Einbezug der Klassifikation von Roget's Thesaurus 73
6.1.2.4 Einbezug von topikalischen Relationen 74
6.1.3 Ansätze mit Selektionsbeschränkungen und Wissensbasen . . . 75
6.1.4 Evaluation der Ansätze 76
6.1.4.1 Wörterbuch- und kontextbasierte Ansätze 77
6.1.4.2 Ähnlichkeitsbasierte Ansätze 78
6.1.4.3 Ansätze mit Selektionsbeschränkungen und Wissensbasen . 78
6.2 Einsatz in computerlinguistischen Anwendungen 79
6.2.1 Einsatz im Information Retrieval 79
6.2.1.1 Mögliche Methoden und Ansätze 79
6.2.1.2 Suchmaschinen mit Wortnetzen 80
6.2.2 Einsatz in der Antwortextraktion 83
6.2.2.1 Mögliche Methoden und Ansätze 83
6.2.2.2 Antwortextraktionssysteme mit Wortnetzen 85
6.2.3 Einsatz in der Maschinellen Übersetzung 86
6.2.3.1 Mögliche Methoden und Ansätze 86

6.2.3.2 Maschinelles Übersetzungssystem mit Wortnetzen 87
6.2.4 Einsatz in der Rechtschreibekorrektur 88
6.2.4.1 Mögliche Methoden und Ansätze 88
6.2.4.2 Korrekturprogramm mit Wortnetzen 88
6.2.5 Evaluation und Ausblick 90
6.2.5.1 Information Retrieval 90
6.2.5.2 Antwortextraktion 91
6.2.5.3 Maschinelle Übersetzung 92
6.2.5.4 Korrektursysteme 92

7 Anaphernresolution mit Wortnetzen 94
7.1 Ansätze 94
7.1.1 Einbezug von Assoziationsrelationen 94
7.1.2 Einbezug von Selektionsbeschränkungen 95
7.1.3 Evaluation der Ansätze 95
7.2 Einsatz in computerlinguistischen Anwendungen 96
7.2.1 Einsatz in einem Antwortextraktionssystem 96
7.2.2 Einsatz in Maschinellen Übersetzungssystemen 96
7.2.3 Evaluation und Ausblick 97

8 Rückblick, Fazit und Ausblick 98

Glossar 100

Literaturverzeichnis 105

Lebenslauf 119

Abbildungsverzeichnis

1 Suche nach „Jaguar“ bei Google 1
2 25 Unique Beginners in WordNet 21
3 Adjektiv-Cluster in WordNet 24
4 Künstliche Konzepte in GermaNet 26
5 Kreuzklassifikation in GermaNet 27
6 Architektur von EuroWordNet 34
7 Commerce-Frame visualisiert im FrameGrapher 40
8 SALSA-Annotationsbeispiel des Frame Communication_response . . . 41
9 Das Verb *wonder* in Roget's Thesaurus 43
10 OpenThesaurus-Suchergebnis für die Anfrage „Fahrrad“ 45
11 Ausschnitt aus dem OpenCyc-Eintrag *bicycle* 49
12 Fragment der WordNet-Hierarchie 71
13 Kontextwörter für die Thesauruskategorien 414 und 348 73
14 Auswahl der verschiedenen Lesarten in Clever Search 81
15 Trefferliste auf die Anfrage „Wasserverschmutzung“ in Antonya 83
16 Repräsentation des Getting-Frame für die Antwortextraktion 84
17 Anfrageerweiterung des *Clarification*-Moduls von MySentient Answer . 85
18 Korrekturvorschlag für eine Wortverwechslung 89

Tabellenverzeichnis

1 Sechs Meronymietypen 12
2 Subkategorisierungsrahmen in GermaNet für das Verb „setzen“ 29
3 Thematische Rollenrelationen aus dem italienischen WordNet 33
4 Angaben zum Verb „informieren“ in HaGenLex 47

Abkürzungsverzeichnis

ILI	Interlingualer Index
IR	Information Retrieval
KI	Künstliche Intelligenz
LCS	Lexical Conceptual Structure
NLP	Natural Language Processing
QA	Question Answering
WSD	Word Sense Disambiguation

1 Einleitung

1.1 Ausgangslage

Lexikalisch-semantische Netze sind grosse Wortnetze, die den Wortschatz einer bestimmten Sprache weitgehend abdecken und die Einträge durch Relationen miteinander verknüpfen. So werden beispielsweise Wörter mit gleicher Bedeutung gruppiert, ihren Unter- oder Oberbegriffen zugeordnet oder Bedeutungsgegensätze einander gegenübergestellt. Im Vergleich zu Wörterbüchern ordnen Wortnetze die Einträge nicht alphabetisch, sondern nach ihrer Bedeutung.

Viele Wortnetze sind in einer maschinenlesbaren Form verfügbar. Da die meisten computerlinguistischen Anwendungen nicht nur Angaben zum Wort selbst, sondern auch zur Wortbedeutung erfordern, liegt der Einsatz von Wortnetzen in sprachverarbeitenden Systemen nah.

Besonders interessant ist der Einbezug lexikalisch-semantischer Information für Suchmaschinen. Abbildung 1 zeigt die ersten zwei Treffer einer Suchanfrage in Google. Die Benutzerin plant eine Safari und will sich über das Raubtier Jaguar informieren. Die Ergebnisliste befriedigt ihr Informationsbedürfnis nicht optimal, da Dokumente zu beiden Lesarten durchmischt aufgeführt werden und sie nicht an Treffern zur Automarke interessiert ist.

Abbildung 1: Suche nach „Jaguar“ bei Google

„Jaguar“ ist ein ambiges, d.h. mehrdeutiges Wort. Konventionelle Suchalgorithmen können Wortbedeutungen nicht unterscheiden und geben Treffer zu beiden Lesarten zurück. Eine weitere Problematik entsteht dadurch, dass diese Techniken Anfrage und Dokumente auf der Wortebene und nicht auf der Bedeutungsebene miteinander vergleichen. Die Suche nach „Auto“, findet keine Treffer zum Stichwort „Wagen“, da die Suchmaschine nicht weiss, dass es sich hierbei um synonyme Wörter handelt. Um noch ein letztes Beispiel zu nennen: Man sucht nach „Comicfiguren“ und erwartet eigentlich eine Auflistung von Dokumenten zu Mickey Mouse, Spiderman und Snoopy. Nur wie soll eine Suchmaschine wissen, was alles gängige Comicfiguren sind?

Wortnetze stellen Lösungsansätze in Aussicht, um die Probleme irrelevanter oder fehlender Treffer anzugehen. Eine durch Synonyme erweiterte Suchanfrage ermöglicht neben dem Abgleich auf Wortebene auch das Auffinden von Wörtern mit derselben Bedeutung. Für die Suchanfrage aus der Abbildung muss der Suchraum eingeschränkt werden, um irrelevante Treffer auszufiltern. Dazu sind Methoden der Lesartendisambiguierung (die Bestimmung der beabsichtigten Wortbedeutung) erforderlich. Wortnetze bieten ein Inventar von möglichen Lesarten und können ausserdem bei der Zuweisung der korrekten Lesart eingesetzt werden. Angaben zu möglichen Unterbegriffen oder Instanzen eines Suchbegriffs helfen ebenfalls, die Suche zu verbessern. Das bekannteste maschinenlesbare Wortnetz, WordNet, könnte so rund 70 Comicfiguren ausfindig machen und damit die Suchanfrage erweitern [WordNetBrowser 2006].

Nicht nur Anwendungen des Information Retrieval, auch Antwortextraktions- oder Maschinelle Übersetzungssysteme stossen auf die erwähnten Probleme. Neben Ambiguitäten sind für diese Anwendungsgebiete auch anaphorische Ausdrücke (Einheiten, die auf eine vorangegangene Einheit referieren) problematisch. Das in Wortnetzen enthaltene semantische Wissen kann dazu eingesetzt werden, verschiedene Ambiguitätstypen und Anaphern korrekt aufzulösen.

Lexikalisch-semantische Netze bieten also in verschiedener Hinsicht Perspektiven, die Performanz computerlinguistischer Anwendungen zu erhöhen.

1.2 Ziel und Aufbau der Arbeit

Die vorliegende Arbeit betrachtet Wortnetze hinsichtlich ihres theoretischen Hintergrundes, stellt ihren Aufbau genauer vor und zeigt in einem zweiten Teil Möglichkeiten für den Einsatz von Wortnetzen in computerlinguistischen Systemen auf.

Diese Arbeit beschränkt sich dabei vor allem auf die Präsentation von englischen und deutschen Wortnetzen und Anwendungen.

Die lexikalische Semantik ist dasjenige Teilgebiet der Linguistik, welches sich mit der Bedeutung von sprachlichen Einheiten befasst. Die Arbeit gibt in **Kapitel 2** einen Überblick über verschiedene Theorien der Wortbedeutungen und stellt gängige lexikalisch-semantische Relationen vor. Es folgt eine Präsentation des englischen WordNet, des deutschen Pendants GermaNet und des mehrsprachigen EuroWordNet. **Kapitel 3** geht genauer auf ihren Aufbau und die involvierten Sinnrelationen ein. In einem nächsten Schritt werden in **Kapitel 4** weitere Wortnetze vorgestellt. Der Schwerpunkt liegt auf der FrameNet-Familie. Zudem werden weitere maschinenlesbare Wortnetze für verschiedene Sprachen mit sehr unterschiedlichen Designprinzipien vorgestellt und miteinander verglichen.

Kapitel 5 befasst sich mit den Schwierigkeiten computerlinguistischer Anwendungen, wobei der Fokus auf Systeme des Information Retrieval, der Maschinellen Übersetzung, Antwortextraktion und Rechtschreibekorrektur gelegt wird. Es wird aufgezeigt, auf welche Hindernisse diese Anwendungen treffen. Lösungsansätze werden erarbeitet, welche die erwähnten Problematiken mithilfe von Wortnetzen angehen. Dabei kristallisiert sich die Lesartendisambiguierung als ein wichtiger Zwischenschritt in der Sprachverarbeitung heraus. In **Kapitel 6** werden Ansätze vorgestellt, die durch den Einsatz von Wortnetzen die Unterscheidung von Lesarten ermöglichen. Es folgt eine Aufstellung, wie Lesartendisambiguierung in computerlinguistischen Anwendungen zum Einsatz kommt. Als weitere Einsatzmöglichkeit von Wortnetzen befasst sich **Kapitel 7** mit der Anaphernresolution. Es wird dargelegt, wie sprachverarbeitende Systeme von Anaphernresolutionskomponenten, die auf Informationen aus Wortnetzen zurückgreifen, profitieren können.

Kapitel 8 greift die wichtigsten Erkenntnisse nochmals auf und zeigt zusammenfassend, welche Aussichten der Einsatz von Wortnetzen in der Computerlinguistik hat.

2 Einführung in die lexikalische Semantik

Die lexikalische Semantik ist eine Teildisziplin der Linguistik, die sich mit der Bedeutung von sprachlichen Einheiten, meist Wörtern oder Wortgefügen, befasst. Dieses Kapitel behandelt theoretische Hintergründe und klärt zentrale Begriffe. Verschiedene Relationen zwischen den Wörtern werden vorgestellt und die Aufnahme von Wörtern in Wortnetze thematisiert.

Der **sinnrelationale Ansatz** der lexikalischen Semantik, der vor allem durch [Cruse 1986], [Cruse 2004] und [Lyons 1977] geprägt wurde, betrachtet die semantischen Beziehungen zwischen den Wörtern einer Sprache. Den Wortschatz kann man als hierarchisches Netz abbilden. Die Wörter werden als Knoten dargestellt, die Relationen zwischen ihnen als verzweigende Kanten. Die Bedeutung eines Wortes wird durch die Gesamtheit seiner Relationen zu den anderen Wörtern der Sprache bestimmt.

Neben dem sinnrelationalen Ansatz wird in diesem Kapitel auf die **Frame-Semantik** eingegangen, welche den Anspruch erhebt, zusätzlich Weltwissen einzubeziehen. Als Frame bezeichnet man ein Konstrukt an Kenntnissen über alltägliche Realitätsausschnitte wie Ereignisse, Situationen oder Handlungen.

Ein weiterer Ansatz der lexikalischen Semantik, der an dieser Stelle kurz erwähnt werden soll, ist die **Wortfeldtheorie,** zu der [Trier 1934] grundlegende Arbeiten verfasst hat. Ein Wortfeld (auch semantisches Feld) ist eine Menge gleichartiger und/oder gleichfunktionaler Prozesse oder Gegenstände, welche einen bestimmten begrifflichen Bereich lückenlos abdecken. Das Feld Sitzgelegenheiten umfasst so Stuhl, Sofa, Bank etc. Dieses System befindet sich in ständigem Wandel. Wenn sich die Bedeutung eines Wortes ausweitet, wird diejenige benachbarter Wörter enger. Die Wortfeldtheorie ist nicht unumstritten. Das postulierte Fehlen von Lücken oder Überlappungen von Feldern ist fraglich. Zudem wird auch nicht ganz klar, wie Wortfelder voneinander abgegrenzt werden können. Eine ausführlichere Besprechung semantischer Felder ist bei [Lyons 1977, S. 250–261] zu finden.

Nochmals anders betrachtet wird die lexikalische Semantik aus dem Blickwinkel der **Prototypentheorie,** die ihren Ausgangspunkt in den Arbeiten einer amerikanischen Psychologin fand [Rosch 1973], [Rosch 1975]. Viele Kategorien können nicht über die Angabe notwendiger und hinreichender Bedingungen definiert werden, sondern der Grad der Zugehörigkeit ist zentral. Mitglieder einer bestimmten Kategorie haben also nicht alle denselben Status. Je mehr man sich von den prototypischen (d.h. besten) Mitgliedern distanziert, desto näher kommt man dem Bereich, für den die Zugehörigkeit zur Kategorie fragwürdig wird. Während alle Vertreter der Kategorie „Vogel“ im biologischen Sinne als Vögel gelten, zählen Versuchspersonen in Experimenten Spatzen zu den eindeutigsten Vertretern, Küken weniger und Pinguine noch weniger klar dazu. Kritiker wenden ein, dass die Bestimmung eines Prototypen einer Kategorie nicht universell, sondern von der direkten oder indirekten Erfahrung der Versuchsperson abhängig ist, was nicht unbedingt zu Klarheit und Einfachheit der Theorie führt [Kleiber 1993, S. 98].

2.1 Analyseeinheiten der lexikalischen Semantik

Die lexikalische Semantik befasst sich mit lexikalischen Einheiten und Lexemen. Der Terminus **lexikalische Einheit** geht auf [Cruse 1986] zurück und bezeichnet den kleinsten unabhängigen Form-Bedeutungs-Komplex. Eine lexikalische Einheit ist die Verbindung einer lexikalischen Form mit genau einer Bedeutung [Cruse 1986, S. 77]. Sie muss aus mindestens einer semantischen Konstituente bestehen und mindestens ein Wort sein [Cruse 1986, S. 24]. Eine semantische Konstituente wird dabei als Teil des Satzes definiert, der eine Bedeutung trägt, die mit den Bedeutungen anderer Konstituenten kombiniert dem Satz eine Gesamtbedeutung gibt [Cruse 1986, S. 25]. Obwohl Cruse keine Wortdefinition angibt, wird klar, dass er dabei eine Einheit zwischen zwei Leerschlägen meint. So ist beispielsweise *dis-* in *disobey* eine semantische Konstituente, jedoch keine lexikalische Einheit, da es sich dabei nur um ein Morphem und nicht um ein Wort handelt. Auch *pulled* in der idiomatischen Wendung *Arthur pulled a fast one* ist keine lexikalische Einheit. Es handelt sich zwar um ein Wort, aber nicht um eine semantische Konstituente.

Ein **Lexem** ist laut [Cruse 1986, S. 76] eine Gruppierung von lexikalischen Einheiten. Lexeme sind alphabetisch geordnet die Hauptzugriffseinheiten in einem Wörterbuch.

In seinem neueren Werk spricht [Cruse 2004] nicht mehr von lexikalischen Einheiten (*lexical units*), sondern von *word forms*. In dieser Arbeit soll aber am Begriff lexikalische Einheit festgehalten werden.

2.2 Variabilität von Wortbedeutungen

Ein Lexem mit nur einer Lesart (d.h. einer lexikalischen Einheit) gilt als monosem. Umfasst ein Lexem verschiedene Lesarten, spricht man von Polysemie. Eine lexikalische Form, die lexikalische Einheiten realisiert, die zu verschiedenen Lexemen gehören, gilt als homonym. Im Folgenden werden die beiden Typen lexikalischer Ambiguität genauer vorgestellt.

Polysemie entsteht, wenn sich die Bedeutung eines Lexems im Laufe der Zeit aufspaltet, wie dies zum Beispiel bei „Schloss“ („Schliessvorrichtung“ oder „Burg“) geschehen ist [Linke et al. 1994, S. 141]. Die Lesarten polysemer Lexeme verfügen über dieselben morphosyntaktischen Eigenschaften und beziehen sich semantisch aufeinander.

Systematische Polysemie kann für Gruppen von Nomen festgestellt werden, die sich vorhersagbar polysem verhalten. Solche Alternationen werden bei [Pustejovsky 1995, S. 31ff.] unter dem Begriff komplementäre Polysemie abgehandelt und treten in verschiedenen Bereichen auf. Zur Veranschaulichung dient Beispiel 1, in dem sich „Fenster“ sowohl auf die Öffnung im ersten Satz wie auch auf das Objekt selbst im zweiten Satz beziehen kann.

(1) a. Die Schlange kam durch das Fenster herein.
    b. Das Fenster war dermassen beschädigt, dass wir es ersetzen mussten.

Wenn die verschiedenen Lesarten keinerlei Bezug zueinander haben, spricht man von **Homonymie.** Sprachhistorisch ist dieses Phänomen so zu verstehen, dass die Lautgestalten zweier ursprünglich verschieden lautender Morpheme im Verlauf der Geschichte zur Deckung gekommen sind, so geschehen mit „Kiefer“ („Baum“ oder „Gebiss“) [Linke et al. 1994, S. 141]. Homonyme weisen oft ein unterschiedliches morphosyntaktisches Verhalten auf, im Beispiel von „Kiefer“ zeigt sich dies durch das ungleiche Genus. Weiter kann man zwischen Homographie und Homophonie unterscheiden. Homographe werden gleich geschrieben. Die unterschiedliche Bedeutung kann beispielsweise durch die Betonung ausgedrückt werden, wie dies beim Paar „modern“ („zeitgemäss“) und „modern“ („verschimmeln“) der Fall ist. Homophonie bezeichnet das Phänomen gleich lautender lexikalischer Einheiten, beispielsweise „mehr“ und „Meer“.

Die Zuweisung einer Lesart zu einer lexikalischen Einheit wird nicht nur durch polyseme und homonyme Lexeme erschwert, sondern auch durch die Verwendung von **Metaphern.** Mit Metaphorik wird die bildhafte Verwendung eines Ausdrucks

bezeichnet, der gegenüber seiner Basisbedeutung eine Sinnveränderung erfährt. Es muss eine bestimmte Ähnlichkeit zwischen dem verwendeten Ausdruck und der intendierten Bedeutung bestehen [Schiehlen und Klabunde 2004, S. 323]. Im folgenden Beispiel 2 wird das Nomen „Fuchs“ metaphorisch verwendet und damit die Schlauheit des Vorgesetzten betont.

(2) Mein Vorgesetzter ist ein Fuchs.

Eine Metapher ist von einer Metonymie nicht immer klar abzugrenzen. **Metonymie** bezeichnet eine Verschiebung in der begrifflichen Interpretation, indem eine nichtwörtliche Bedeutung verwendet wird. Diese Bedeutung ist mit der wörtlichen Bedeutung auf eine bestimmte Weise verknüpft [Schiehlen und Klabunde 2004, S. 322]. Häufige Verschiebungen sind beispielsweise im Bereich Gefäss–Inhalt (e.g. „ein Glas trinken“) oder Ort–Institution (e.g. „Bern hat die Steuern erhöht“) zu finden.

Obwohl Metaphorik und Metonymie erst auftreten, wenn nicht einzelne Lexeme, sondern ganze Phrasen betrachtet werden, stellen sie Schwierigkeiten dar. Im Lexikon darf „Fuchs“ nicht ohne weitere Erklärung mit der Bedeutung „schlauer, raffinierter Mensch“ und „Glas“ nicht als „Getränk“ aufgenommen werden.

2.3 Lexikalisch-semantische Relationen

Eine lexikalisch-semantische Relation (auch Sinnrelation, semantische Relation) ist eine Beziehung, die zwischen zwei oder mehreren lexikalischen Einheiten aufgrund ihrer lexikalisch-semantischen Eigenschaften gestiftet wird.

Man unterscheidet zwischen paradigmatischen und syntagmatischen Sinnrelationen. **Paradigmatische Sinnrelationen** sind Beziehungen zwischen lexikalischen Einheiten, deren Bedeutungen in systematischer Weise zusammenhängen, wie dies zum Beispiel zwischen den Adjektiven „warm“ und „kalt“ der Fall ist. Typischerweise involvieren paradigmatische Relationen lexikalische Einheiten, die derselben Wortart angehören [Cruse 2004, S. 146]. Die bekanntesten Relationen sind Synonymie, Antonymie, Hyponymie/Hyperonymie und Meronymie. In der sinnrelationalen lexikalischen Semantik ergibt sich die Bedeutung eines Wortes aus der Gesamtheit der paradigmatischen Beziehungen zu anderen Wörtern im Sprachsystem.

Syntagmatische Sinnrelationen befassen sich mit dem Auftreten lexikalischer Einheiten im Satz [Cruse 2004, S. 146f.] und beschreiben Bindungskräfte auf lexikalischer Ebene [Lutzeier 1995, S. 88f.]. So folgt auf das Adjektiv „blond“ in

den meisten Fällen das Nomen „Haar“, man spricht von einer Kollokation. Syntagmatische Relationen befassen sich auch mit semantischen Beschränkungen, die ein Prädikat seinen Argumenten auferlegt, auch Selektionsbeschränkungen genannt. Sätze wie in Beispiel 3 [Wagner 2005, S. 1] sind zwar syntaktisch korrekt, aber semantisch anomal[1].

(3) * Der Stein denkt.

Das Verb „denken“ verlangt ein belebtes Subjekt. Unbelebte Subjekte werden demnach durch diese Beschränkung zurückgewiesen.

Ein weiterer Relationstyp ist die **derivationelle Sinnrelation** [Cruse 2004, S. 147]. Sie manifestiert sich zwischen Elementen einer Wortfamilie, d.h. lexikalischen Einheiten, die vom selben Wortstamm abgeleitet sind, beispielsweise zwischen „kochen“ und „Koch“.

Im Folgenden befasst sich dieses Kapitel mit einer genauen Analyse der wichtigsten paradigmatischen Relationen. Es besteht kein Konsens darüber, wie viele solche Relationen es gibt. Wohl am meisten Relationen aufgelistet haben [Melcuk und Zholkovsky 1988, S. 55–63], die für den Aufbau des *Explanatory Combinatorial Dictionary* für die russische Sprache über 50 so genannte lexikalische Funktionen ausfindig gemacht haben. Diese Zahl umfasst neben paradigmatischen jedoch auch derivationelle Relationen. [Cruse 2004, S. 143f.] schlägt Kriterien für die Bestimmung signifikanter Relationen vor. Wichtig ist erstens wiederholtes Auftreten (engl. *recurrence*) einer Relation, d.h., sie muss zwischen vielen lexikalischen Einheiten bestehen. Zweitens führt er an, dass eine Relation nicht zwischen allen lexikalischen Einheiten bestehen darf (engl. *discrimination*). An dritter Stelle steht schliesslich das Kriterium, dass das hinter der Relation stehende Konzept lexikalisiert sein soll (engl. *lexicalizability*), wie das beispielsweise im Deutschen bei der Antonym-Relation durch das Alltagswort „Gegensatz“ der Fall ist.

[Cruse 2004, S. 141ff.] unterscheidet paradigmatische Relationen der Identität oder Inklusion (Synonymie, Hyponymie, Meronymie) und solche des Gegensatzes oder der Exklusion (Inkompatibilität, Komplementarität, Antonymie, Reversivität und Konversion). Auf die einzelnen Relationen wird im Folgenden eingegangen. Die Aufteilung von [Cruse 2004] deckt sich grösstenteils mit derjenigen von [Lutzeier 1995, S. 73–87], der zwischen vertikalen und horizontalen[2] Relationen unterscheidet.

[1] Syntaktisch oder semantisch ungültigen Beispielsätzen wird ein Asterisk vorangestellt, fragwürdigen Beispielen ein Fragezeichen.

[2] [Lutzeier 1995] betrachtet hier, ob die Relationen zu einer vertikalen oder horizontalen Strukturierung des Wortschatzes führen. Die Synonymie zählt er dabei zu keiner der beiden Gruppen.

2.3.1 Synonymie

Synonymie ist ein Terminus zur Kennzeichnung der Bedeutungsgleichheit von Wörtern und wird auch Äquivalenzrelation genannt.

[Cruse 2004, S. 154ff.] unterscheidet zwischen absoluter, propositionaler und Beinahe-Synonymie. **Absolute Synonymie** liegt dann vor, wenn die betreffenden Ausdrücke in allen Kontexten uneingeschränkt austauschbar sind. Dabei handelt es sich um eine strikte Einschränkung und es fällt schwer, unkontroverse Beispiele zu finden. Absolut synonym ist zweifelsohne das Paar „beginnen“ - „anfangen“. Ansonsten ist absolute Synonymie ein rares Phänomen, denn zwei bedeutungsgleiche Wörter im mentalen Lexikon zu speichern, ist ein Luxus und widerspricht dem Ökonomie-Prinzip [Linke et al. 1994, S. 154]. Dieses besagt, dass Sprachen dazu tendieren, den kommunikativen Zweck mit möglichst einfachen, spärlichen Mitteln zu erfüllen. Häufig differenzieren sich zwei Ausdrücke für dieselbe Bedeutung im Verlauf der Zeit, indem sie sich mit Bedeutungsnuancen aufladen und unterschiedlich verwendet werden.

Propositionale Synonymie betitelte [Cruse 1986, S. 270f.] früher mit kognitiver Synonymie. Diese Art von Synonymie besteht dann, wenn lexikalische Einheiten ausgetauscht werden können, ohne dass dies Auswirkung auf den Wahrheitswert hat [Cruse 2004, S. 155]. So kann man im englischen Beispiel 4 *violin* durch *fiddle* ersetzen, ohne dass sich die Aussage verfälscht. Das Resultat des Ersatzes ist jedoch ein etwas seltsam klingender Satz, hervorgerufen durch einen Unterschied im Stil [Cruse 2004, S. 155].

(4) a. She's going to play a violin concerto.
b. ? She's going to play a fiddle concerto.

Generell haben propositionale Synonympaare die gleiche Denotation[3], aber eine unterschiedliche Konnotation[4]. Die konnotative Bewertung kann neben stilistischen auch aufgrund von fachsprachlichen, regionalen, soziolektalen oder politischen Besonderheiten erfolgen.

Plesionymie wird häufig auch unter dem Begriff Beinahe-Synonymie (engl. *near-synonymy*) aufgeführt [Cruse 2004, S. 156]. Der Austausch eines Wortes durch sein Plesionym hat Auswirkung auf den Wahrheitswert einer Aussage. Das Paar *foggy* - *misty* sind beispielsweise Plesionyme [Cruse 1986, S. 285]. Man kann zugleich

[3]Denotation bezeichnet den Bedeutungskern.

[4]Konnotation steht für das Phänomen, dass der Bedeutungskern durch zusätzliche Bedeutungsaspekte anders bewertet wird.

das eine assertieren und das andere verneinen, ohne dass dies zu einer paradoxen Aussage führt (vgl. *It wasn't foggy last Friday – just misty.).*

2.3.2 Hyponymie/Hyperonymie

Die Sinnrelation, welche den Wortschatz vertikal ordnet, also eine hierarchische Struktur aufbaut, wird als Hyponymie/Hyperonymie bezeichnet. Hyponyme sind Unterbegriffe, Hyperonyme Oberbegriffe. X ist nach [Lyons 1977, S. 292] ein Hyponym von Y, wenn gilt: „X is a kind of Y".

Die verschiedenen Hyponyme desselben Hyperonyms werden Kohyponyme genannt, so sind „Dackel" und „Spaniel" beides Kohyponyme von „Hund". Kohyponyme unterscheiden sich in mindestens einem semantischen Merkmal und sind inkompatibel, d.h., sie schliessen sich gegenseitig aus. Daher ist auch Beispiel 5 semantisch ungültig.

(5) * Dieser Dackel ist ein Spaniel.

[Wierzbicka 1984] identifiziert fünf Typen von Hyponymie, welche sich primär nach der Art der Ähnlichkeit der Mitglieder eines Typs unterscheiden. Mitglieder des **taxonomischen Typs** sind motiviert durch eine physische Ähnlichkeit, die beispielsweise zwischen „Spatz" und „Vogel" besteht. Ein Paar der **funktionalen Kategorie** erfordert ein gemeinsames Einsatzgebiet, wie dies bei „Rassel" und „Spielzeug" der Fall ist. Einen weiteren Typ nennt die Autorin **Singular-Kollektiv** und meint damit Paare mit einem Singularetantum wie „Obst" als Oberbegriff. Mitglieder dieses Typs teilen sich ebenfalls eine Funktion und oft auch ihre Geschichte, die den gemeinsamen Einsatz erklärt. Der vierte Typ wird **Plural-Kollektiv** genannt und verweist auf Mitglieder einer Kategorie, die zwar keine gemeinsame Funktion, aber dieselbe Herkunft oder Geschichte teilen. Oberbegriffe dieses Typs sind zum Beispiel „Zutaten" oder „Lebensmittel". Bei der letzten Kategorie handelt es sich um eine **heterogene Klasse,** die Paare wie „Medizin" - „Kräuter" umfasst, die zwar eine Funktion teilen, aber nicht am selben Ort auftreten.

[Chaffin 1992, S. 280] kritisiert, dass der Testrahmen für Hyponymie („X is a kind of Y") nicht auf die Plural-Kollektiv-Kategorie anwendbar ist und die den anderen Kategorien gemeinsame physische oder funktionale Ähnlichkeit der Mitglieder nicht vorhanden zu sein scheint. Er ist deshalb der Ansicht, dass der Typ Plural-Kollektiv nicht als Hyponymie-, sondern als Meronymie-Relation angesehen werden sollte (vgl. 2.3.2). Allgemein ist die Abgrenzung zwischen Hyponymie und Meronymie nicht immer klar. Auf die Beziehung zwischen den beiden Relationen geht [Miller 1990, S. 255] genauer ein.

Die Hyponymie-Relation ist transitiv, irreflexiv und asymmetrisch[5]. Es sind jedoch einige Fälle bekannt, in denen die Transitivität fehlschlägt. Das wohl am häufigsten zitierte Beispiel 6 stammt von [Hampton 1982, S. 160].

(6) a. A car seat is a type of seat.
b. A seat is a type of furniture.
c. * A car seat is a type of furniture.

Damit die Hyponymie-Relation transitiv ist, muss nach [Hampton 1982] und [Chaffin 1992, S. 279] in beiden Prämissen derselbe Typ von Hyponymie bestehen. Dies ist im obigen Beispiel nicht der Fall, denn a) ist taxonomisch und b) Singular-Kollektiv. Die Erklärung von [Cruse 2004, S. 149] geht in eine andere Richtung. Seiner Meinung nach schlägt die Transitivität in diesem Beispiel fehl, da der Autositz kein prototypischer Vertreter von Möbeln ist. Der erste Erklärungsansatz scheint verständlicher. Der Einbezug von Prototypentheorie zieht weitere Probleme mit sich, insbesondere wenn man sich damit beschäftigt, wann es sich um einen prototypischen Vertreter handelt und wann nicht.

2.3.3 Meronymie

Eine weitere Relation der Inklusion ist die Meronymie, worunter man die semantische Relation zwischen sprachlichen Ausdrücken zur Bezeichnung der Relation des Teils zum Ganzen bzw. zur Bezeichnung von Besitzverhältnissen versteht. Häufig wird Meronymie auch Partonymie oder Teil-Ganzes-Beziehung genannt. Beim Paar „Finger“ - „Hand“ bezeichnet man den „Finger“ als Meronym, die „Hand“ als Holonym.

Tabelle 1 zeigt die von [Winston et al. 1987, S. 420f.] aufgelisteten sechs Typen von Meronymie, die sich mit der Auffassung von [Chaffin 1992, S. 262f.] decken. Eine Ausnahme ist die Prozess-Phase-Relation („Heranwachsen“ - „Kindheit“) von [Chaffin 1992], die [Winston et al. 1987] unter der Relation Merkmal-Ereignis subsumieren.

Die Meronymie-Relation ist transitiv, irreflexiv und asymmetrisch. Wie auch die Hyponymie ist die Meronymie nicht durchwegs transitiv, wie das an [Chaffin 1992, S. 275] angelehnte Beispiel 7 zeigt:

[5] Transitiv, da gilt: Wenn A ein Hyponym von B ist und B ein Hyponym von C, dann ist A ein Hyponym von C. Irreflexiv, weil kein Begriff ein Unterbegriff von sich selber sein kann. Asymmetrisch, weil gilt: Wenn A ein Hyponym von B ist, kann B nicht ein Hyponym von A sein.

Relationstyp	**Beispiel**
Komponente–Objekt	„Henkel“ - „Tasse“
Mitglied–Menge	„Schiff“ - „Flotte“
Portion–Masse	„Brotscheibe“ - „Brot“
Objekt–Material	„Bierglas“ - „Glas“
Merkmal–Ereignis	„zahlen“ - „einkaufen“
Ort–Gebiet	„Oase“ - „Wüste“

Tabelle 1: Sechs Meronymietypen nach [Winston et al. 1987, S. 420f.]

(7) a. Der Kopf ist Teil der Statue.
b. Die Statue ist Teil der etruskischen Sammlung.
c. * Der Kopf ist Teil der etruskischen Sammlung.

Auch hier wird das Fehlschlagen der Transitivität durch unterschiedliche Meronymietypen in den Prämissen erklärt. In a) ist die Meronymie durch den Komponenten-Objekt-Typ motiviert, in b) hingegen durch den Mitglied-Mengen-Typ, was zu der ungültigen Schlussfolgerung in c) führt.

2.3.4 Antonymie

Die semantischen Relationen, welche Exklusion und Gegensätzlichkeit ausdrücken, werden unter dem Begriff Antonymie zusammengefasst. Auf die problematische Vielfalt der Gegensatzrelation wurde oft hingewiesen (beispielsweise bei [Cruse 1986, S. 197ff.]). In dieser Arbeit wird im Gegensatz zum Verständnis von [Lyons 1977], [Cruse 1986] und [Lutzeier 1995] unter Antonymie ganz generell die Gegensatzrelation zwischen zwei lexikalischen Einheiten verstanden.

Die allgemeinste Relation des Bedeutungsgegensatzes ist die **Inkompatibilität.** Zwei Ausdrücke sind genau dann inkompatibel, wenn sie grössere semantische Gemeinsamkeiten aufweisen, sich aber bezüglich einer inhaltlichen Dimension unterscheiden. Bei der Annahme des gemeinsamen Bestehens tritt ein Widerspruch auf [Lutzeier 1995, S. 80f.] (vgl. dazu Beispiel 5).

Von **Komplementarität** spricht man, wenn der Sachverhaltsbereich durch ein Antonympaar in zwei disjunkte Abschnitte ohne Zwischenbereich unterteilt wird. Durch die so aufgebaute „Entweder-oder-Situation“ kann man nicht nur aus dem Bestehen des einen Elementes auf das Nichtbestehen des anderen Elementes schliessen, sondern auch vom Nichtbestehen des einen Elementes auf das Bestehen des

anderen Elementes. Für [Lyons 1977, S. 279] handelt es sich dabei immer um Gegensatzpaare, die nicht gesteigert werden können. Häufig wird diese Gegensatzrelation auch Kontradiktion genannt ([Lang 1994], [Linke et al. 1994]). Viel zitierte Beispiele komplementärer lexikalischer Einheiten sind „tot“ - „lebendig“ oder „Hengst“ - „Stute“.

Was [Lyons 1977], [Cruse 1986] und [Lutzeier 1995] als Antonymie bezeichnen, wird hier in Anlehnung an [Linke et al. 1994] und [Lang 1994] **Kontrarität** benannt. Im Gegensatz zu Komplementarität ist der Bereich nicht dichotomisch aufgeteilt, sondern es gibt Zwischenbereiche. Es ist zwar nicht möglich, dass zwei zueinander in der Kontraritäts-Relation stehende lexikalische Einheiten zugleich zutreffen, doch dass beide nicht zutreffen, ist möglich. Ein Beispiel für ein konträres Paar ist „gut“ - „schlecht“. [Lutzeier 1995, S. 81] weist darauf hin, dass konträre Einheiten graduierbar sein müssen. Seiner Meinung nach beschränkt sich die Kontrarität auf die Wortart Adjektive, andere Autoren wie [Lang 1994] zählen auch Verbpaare wie „lieben“ - „hassen“ in ihren Beispielen auf.

In der **Konversions-Relation** stehen Einheiten, die denselben Sachverhalt aus zwei entgegengesetzten Blickwinkeln darstellen [Linke et al. 1994]. Konverse Paare haben gemeinsam, dass sie im selben semantischen Feld vorkommen und auf dieselbe Aktivität referieren, aber die assoziierten thematischen Rollen auf der Oberflächenstruktur anders verteilt sind [Fellbaum 1998a, S. 81]. Im Beispiel „kaufen“ - „verkaufen“ sind Quelle (engl. *source*) und Ziel (engl. *goal*) vertauscht („der X verkauft dem Y das Z“ vs. „der Y kauft vom X das Z“). Ein Beispiel für verteilte Rollenverhältnisse bei Nomen ist das Paar „Nachfolger“ - „Vorgänger“.

Zum Schluss wird noch auf die **Reversivität** eingegangen. Hier muss der Anfangszustand des ersten Geschehens den Endzustand des zweiten Geschehens darstellen und umgekehrt [Lutzeier 1995, S. 84]. [Cruse 2004, S. 166f.] ordnet nur Verben dieser Bedeutungsrelation zu, so beispielsweise „aufsperren“ - „zusperren“. Interessanterweise spielt die Art des Vorgangs zwischen den beiden Zuständen keine signifikante Rolle. So unterscheidet sich beispielsweise die Handlung, eine Krawatte umzubinden, stark von derjenigen, eine Krawatte wieder loszubinden, und trotzdem sind *tie* und *untie* Reversive.

2.4 Frame-Semantik

Im Vergleich zum soeben eingeführten sinnrelationalen Ansatz ist die Frame-Semantik eine Theorie, die neben lexikalischer Information auch Alltagswissen über zentrale Konzepte speichert. Sie kann als Weiterentwicklung der Kasusgrammatik von [Fillmore 1968] betrachtet werden und beruht auf weiteren Arbeiten von ihm (vgl. [Fillmore 1977a], [Fillmore 1977b], [Fillmore 1985]). Nach Ansicht dieses Ansatzes kann Wortbedeutung am besten verstanden werden, wenn man die ihr zugrunde liegenden konzeptuellen Strukturen betrachtet. Lexikalische Einheiten werden beschrieben, indem eine Verknüpfung mit einer schematischen Darstellung von solchen konzeptuellen Strukturen, so genannten **Frames,** hergestellt wird. Frames stellen die Konzepte von Situationen, Handlungen und Zuständen dar und sind zusammen mit den lexikalischen Einheiten, die analog zur Definition von [Cruse 1986] eingeführt werden, primäre Analyseeinheiten der Frame-Semantik [Fillmore et al. 2003, S. 235]. Unterschiedliche Lesarten eines polysemen Lexems entsprechen verschiedenen semantischen Frames, an welchen sich das Lexem beteiligen kann. So kann das englische Adjektiv *hot* entweder den Temperatur-Skala-Frame oder den Geschmacks-Empfindungs-Frame hervorrufen [Fillmore et al. 2003, S. 236].

In der Satzstruktur wird Verben eine spezielle Rolle zuteil, da sie die Unterstützung anderer Phrasen im Satz (Argumente) verlangen. Die Anforderungen, dass ein Wort mit einer Anzahl bestimmter Satzglieder kombiniert werden muss, um wohlgeformte Sätze zu bilden, nennt man Valenz. Valenz kann auf syntaktischer und auf semantischer Ebene spezifiziert werden. Die syntaktische Ebene spezifiziert die Anzahl der Argumente und umfasst Informationen zu grammatikalischen Eigenschaften[6] wie Kasus oder Phrasentyp (Nominalphrase, Verbalphrase) oder zu grammatikalischen Funktionen wie Subjekt oder Objekt. Subkategorisierungsrahmen legen die Anzahl und den Typ der Argumente fest und ordnen Verben Kategorien wie transitiv, intransitiv etc. zu.

Auf der semantischen Ebene werden **Frame-Elemente** identifiziert. Sie charakterisieren semantische Grundverhältnisse und Beziehungen innerhalb eines Satzes und legen fest, welche Strukturen an einem Frame teilnehmen können, also welche Selektionsbeschränkungen bestehen. Frame-Elemente basieren auf semantischen Rollen (auch thematische Rollen, Theta-Rollen, Tiefenkasus), die Relationen zwischen dem Prädikat und seinen Argumenten beschreiben. Typische semantische Rollen sind Agens (handlungsausführendes Argument) oder Patiens (betroffenes Argument). Es wurde nicht festgelegt, wie viele und welche Relationen in die Frame-

[6] In der Kasusgrammatik spricht man von Oberflächenkasus.

Semantik einbezogen werden sollten, wodurch sich Frame-Elemente eher auf Frame-spezifische situationale Rollen als auf die Tiefenkasus der Kasusgrammatik beziehen [Fillmore et al. 2003, S. 240].

Das folgende Beispiel 8 von [Fillmore et al. 2003, S. 237] zeigt den Frame Transfer mit den Frame-Elementen donor, theme, recipient. Die Verben „geben“ und „bekommen“ rufen beide diesen Frame hervor.

(8) Transfer-Frame:
Der donor ist in Besitz von etwas (theme), und verursacht, dass jemand anders (recipient) in Besitz des theme kommt, zum Beispiel durch Veranlassung einer Bewegung des theme zum recipient.

Wenn man weiss, dass der Lehrer dem Studenten ein Buch „gegeben“ hat, kann man daraus schliessen, dass der Student vom Lehrer ein Buch „bekommen“ hat. Die Verben unterscheiden sich in der Perspektive, die sie dem Transfer-Event auferlegen.

Relationen bestehen nicht nur innerhalb eines Frame, sondern auch zwischen den Frames. Manchmal ist ein Frame Teil eines komplexeren Frame, wie zum Beispiel der Transfer-Frame eine Rolle im komplexeren Commercial Transaction-Frame spielt, welcher die Frame-Elemente buyer, seller, goods, und money umfasst [Fillmore et al. 2003, S. 239].

2.5 Einordnung von lexikalischen Einheiten in Wortnetze

Das Urmodell von Wortnetzen geht auf [Quillian 1968] zurück, der im Forschungskontext der Künstlichen Intelligenz untersuchte, wie semantische Information im Gehirn gespeichert ist. Das so genannte „semantische Gedächtnis“ besteht analog zur Auffassung der sinnrelationalen Semantik aus Wörtern (Knoten) und Relationen (Verzweigungen zwischen den Knoten) und wurde auf dem Computer modelliert.

Später wurden Wortnetze als Netze von Konzepten (Bedeutungen) basierend auf den lexikalischen Einheiten einer Sprache definiert [Vossen 2004a, S. 162]. In dieser Definition kennen Wortnetze keine Konzepte, für die es kein Wort oder keinen Ausdruck in einer Sprache gibt. Die Bedeutung einer lexikalischen Einheit wird hauptsächlich durch die Relationen zu anderen Einheiten explizit. In einem Wortnetz werden also einerseits lexikalische Einheiten mit ihrer Bedeutung verknüpft und andererseits Bedeutungen untereinander in Beziehung gebracht.

Je nach Ausdrucksstärke und Abdeckung des Wortnetzes spricht man von einer Taxonomie, einem Thesaurus oder einer Ontologie.

2.5.1 Taxonomie

Taxonomien (von griechisch *taxon:* Gruppe von Lebewesen) kommen aus den Naturwissenschaften, wo sie zur systematischen Einteilung von Organismen in Gruppen verwendet werden. Die relevanten Begriffe sind in einer hierarchischen Beziehung angeordnet, basierend auf asymmetrischen Sinnrelationen wie Hyponymie und Meronymie. Taxonomien sind auf andere systematisch strukturierbare Bereiche wie lexikalische Hierarchien übertragen worden und umfassen meist nur Nomen.

2.5.2 Thesaurus

Der Ausdruck Thesaurus wurde vom griechischen Wort *thesauros* abgeleitet und bezeichnete ursprünglich „Schatz“. Nun versteht man unter einem Thesaurus ein nach der Bedeutungsähnlichkeit der aufgenommenen Wörtern gegliedertes Vokabular, dessen Begriffe durch Relationen miteinander verbunden sind. Da Wörter nicht nach ihrer Lautgestalt oder Orthografie, sondern nach ihrer Bedeutung klassifiziert werden, spricht man von einer onomasiologischen Anordnung, d.h., es wird betrachtet, wie Konzepte bezeichnet werden können. Im Kontrast dazu stehen semasiologische (nach Wörtern und ihrer Bedeutung) geordnete Wörterbücher, die dem Anwender helfen, das Konzept hinter einem ihm unbekannten Wort zu verstehen [Fellbaum 1998b, S. 7]. In vielen Thesauri sind die Relationen zwischen den Konzepten nicht explizit benannt. Neben Synonymie und Hyponymie/Hyperonymie werden teils auch Antonyme oder Meronyme aufgeführt. Ein Thesaurus enthält Nomen und oft Adjektive und Verben.

2.5.3 Ontologie

Ontologien werden in der Informatik als „formale Spezifikation einer Konzeptualisierung“ definiert [Gruber 1993, S. 200]. Es handelt sich dabei um Konstrukte, mit denen Wissen über ein bestimmtes Sachgebiet semantisch komplex repräsentiert wird. Daneben sollten sie auch für den Datenaustausch verwendet werden können. Im Gegensatz zu Taxonomien und Thesauri, welche Wörter einer Sprache in Verbindung stellen, erheben Ontologien zusätzlich den Anspruch, (möglicherweise nicht lexikalisierte) Konzepte zu repräsentieren. Die Grundbausteine einer Ontologie sind

Klassen und Beziehungen. Mit einer Klasse ist eine Menge von individuellen Objekten der Welt gemeint, die gewisse gemeinsame Eigenschaften aufweisen. Klassen sind in einer Klassenhierarchie geordnet und durch Beziehungen untereinander in Zusammenhang gebracht. Gängig ist die Unterscheidung zwischen Typ und Rolle. So ist ein „Schäferhund“ sowohl als ein „Wirbeltier“ (Typ) als auch als ein „Haustier“ (Rolle) einzuordnen. Weitere Bestandteile sind Eigenschaften (genauere Beschreibungen von Klassen), ihre Werte und Nebenbedingungen (Einschränkungen des Wertebereiches) und Inferenzregeln. Welche semantischen Relationen ausgedrückt werden können, hängt von der Ausdrucksmächtigkeit einer Ontologie ab.

3 Die WordNet-Familie

Das wohl bekannteste maschinenlesbare Wortnetz ist das englische WordNet. Dieses und darauf aufbauende Wortnetze für andere Sprachen fassen die häufigsten lexikalischen Einheiten zu Konzepten zusammen und verbinden Konzepte durch verschiedene lexikalisch-semantische Relationen miteinander.

In diesem Kapitel werden das Princeton WordNet, das deutsche Pendant GermaNet und das multilinguale EuroWordNet vorgestellt und miteinander verglichen.

3.1 WordNet

Das Princeton WordNet (im Folgenden nur WordNet) wird seit 1985 vom Cognitive Science Laboratory der Universität Princeton unter der Leitung von George A. Miller entwickelt [WordNet 2006]. Es handelt sich um eine nach psycholinguistischen Theorien über das menschliche Gehirn aufgebaute lexikalische Datenbank für die englische Sprache. WordNet existiert bereits in der Version 2.1, ist kostenlos verfügbar und darf sowohl für akademische als auch für kommerzielle Anwendungen eingesetzt werden. Eine Online-Abfrage ist ebenfalls möglich unter [WordNetBrowser 2006][7].

3.1.1 Motivation

WordNet war anfänglich ein Experiment, ein bestimmtes Modell der Struktur des menschlichen lexikalischen Gedächtnisses zu implementieren [Fellbaum 1998b, S. 4]. Dazu stützte man sich auf Experimente von [Collins und Quillian 1969] über Prozesse des Zugriffs auf lexikalische Information im Gehirn. WordNet sollte die wichtigsten lexikalischen Knoten identifizieren und die semantischen Relationen zwischen ihnen modellieren.

[7] Die in diesem Kapitel aufgeführten englischen Beispiele stammen alle aus der Online-Abfrage der WordNet-Version 2.1.

Obwohl WordNet auf einer psycholinguistischen Motivation beruht, wurde es von Kognitionswissenschaftlern weitgehend ignoriert [Miller 1998a, S. 43]. Dafür ist es aber auf breites Interesse in Kreisen der Computerlinguistik gestossen.

3.1.2 Aufbau und semantische Relationen

WordNet ist aufgeteilt in vier Dateien, je eine für Nomen, Verben, Adjektive und Adverbien, und umfasst rund 155 000 lexikalische Einheiten [WordNetStatistik 2006][8]. Als primäre Repräsentationseinheit dienen Synsets (Synonymgruppen), welche eine Menge von Wörtern derselben Wortart, die in einem bestimmten Kontext austauschbar sind, zu einem Konzept zusammenfassen. In WordNet sind Synsets wie {*violin, fiddle*} oder {*brumous, foggy, hazy, misty*} zu finden, was zeigt, dass neben absoluten auch proportionale Synonyme und Plesionyme zu Synsets zusammengefasst werden.

Polyseme und Homonyme gehören zu verschiedenen Synsets. So lässt sich beispielsweise das systematisch polyseme Wort *door* in mehreren Synsets finden. Die beiden Lesarten (Objekt selbst bzw. Öffnung) sind im Beispiel 9 dargestellt.

(9) a. {door} (a swinging or sliding barrier that will close the entrance to a room or building or vehicle)

b. {doorway, door, room access, threshold} (the entrance [the space in a wall] through which you enter or leave a room or building; the space that a door can close)

Ursprünglich sollte WordNet nur aus Synsets und Pointers (d.h. lexikalisch-semantischen Relationen) zwischen Synsets bestehen. Im Verlauf der Entwicklung erkannte man, dass Wortbedeutungen durch Betrachtung von Synonymen und Relationen nicht immer klar wurden. Um ähnliche Synsets zu kontrastieren, wurden die Synsets in der Folge durch Glossen, bestehend aus Wörterbuchdefinitionen und teils auch Beispielsätzen, ergänzt [Miller 1990, S. 18f.].

Semantische Relationen sind entweder zwischen Synsets als konzeptuelle Relationen oder zwischen einzelnen lexikalischen Einheiten (Synonymen aus den Synsets, also auf Wortebene) als lexikalische Relationen kodiert [Kunze 2005b, S. 165]. Zentrale konzeptuelle Relationen in WordNet sind Hyponymie/Hyperonymie, Troponymie (Unter-/Oberbegriffsrelation bei Verben) und Meronymie. Bei den peripheren Relationen handelt es sich um die Implikations-, die Kausations-, die Instanz- und

[8] Alle Angaben zum Umfang von WordNet stammen aus dieser Quelle.

die Domänen-Relation. Die wichtigsten lexikalischen Relationen in WordNet sind die synsetinterne Synonymie und die Antonymie, daneben sind morphosemantische Relationen wie Derivations-, Attribut-, Partizip- und *See-Also*-Relationen kodiert. Diese Relationen decken sich zum grössten Teil mit den in Abschnitt 2.3 vorgestellten Beziehungen. Nicht aufgeführte, WordNet-spezifische Relationen werden später genauer vorgestellt.

Wie im Kapitel 2 erwähnt, gibt es detaillierte Analysen semantischer Relationen. [Fellbaum 1998b, S. 10] erscheint die Anzahl solcher Relationen im menschlichen Gedächtnis aber relativ klein, weswegen sich WordNet auf etwa ein Dutzend Relationen beschränkt.

Lexikalisch-semantische Relationen, die in mehreren Wortartendateien vorkommen, sind die Derivations-, die Domänen- und die *See-Also*-Relation.

Die **Derivations-Relation** ist eine wortartenübergreifende lexikalische Relation, welche Einträge desselben Wortstamms aber mit verschiedener Wortart miteinander verknüpft. Sie besteht zum Beispiel zwischen Paaren wie *succeeder* - *succeed* oder *black* - *blackness*.

Zur **Domänen-Relation** sind keine genaueren Dokumentationen erhältlich. Es handelt sich wahrscheinlich um eine kürzlich eingeführte Relation. Im [WordNet-Glossar 2006] wird sie als topikalische Klassifikation beschrieben. Eine Recherche hat ergeben, dass es sich um eine konzeptuelle, kategorienübergreifende Relation handelt. Die Beispiele *cloudy* - *meteorology*, *fry* - *cooking*, *bullshit* - *vulgarism* decken drei der vier Wortarten ab.

Auch die ***See-Also*-Relation** ist eher dürftig dokumentiert. Die Adjektiv-Synsets *quiet* und *peaceful* sind beispielsweise durch diese konzeptuelle Relation miteinander verknüpft.

3.1.2.1 Aufbau der Nomen-Datei

WordNet umfasst rund 81 000 Nomen-Synsets. Die Synsets liegen in einer hierarchischen Struktur vor, wobei es prinzipiell möglich wäre, dass alle Nomen genau einen Baum bildeten. Dazu müsste es ein leeres Synset an oberster Stelle geben. In WordNet entschied man sich gegen diese Lösung, da dieses oberste Synset keine semantische Information tragen würde [Miller 1998a, S. 28]. Die Nomen-Datei ist in 25 semantische Felder aufgeteilt, welche je einen *Unique Beginner* (Synset ohne Hyperonym) haben. Abbildung 2 zeigt die 25 Unique Beginners, die durch Betrachtung möglicher Adjektiv-Nomen-Kombinationen motiviert sind.

{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}
{motive}	

Abbildung 2: 25 Unique Beginners in WordNet nach [Miller 1998a, S. 29]

In den 25 Dateien werden durch die **Hyponymie-Relation** 10–12 Hierarchie-Ebenen erreicht. WordNet macht nur limitierten Gebrauch von multipler Vererbung, d.h., wenige Synsets weisen mehr als ein Hyperonym auf. Eine spezielle Form von Hyponymie besteht, wenn ein Eigenname auf einen bestimmten, einzigartigen Referenten verweist. WordNet kennzeichnet Eigennamen wie *Snoopy* als **Instanz** vom Synset *fictional character*.

Meronymie ist vor allem in den semantischen Feldern body, artifact und quantity zu finden [Miller 1998a, S. 37]. In WordNet sind die drei Meronymietypen Komponente-Objekt, Mitglied-Menge und Material-Objekt kodiert. Der erstgenannte Typ kommt dabei laut [Miller 1998a, S. 39] bei weitem am häufigsten vor, beispielsweise zwischen *toe* und *foot*.

Antonymie tritt vorwiegend zwischen Nomen auf, die aus Adjektiven abgleitet sind und dem semantischen Feld attribute angehören (e.g. *happiness* - *unhappiness*) [Miller 1998a, S. 40]. Zudem findet man Antonyme in Wortpaaren wie *man* - *woman*. Es lassen sich nur wenige komplementäre Antonyme finden. So fehlen beispielsweise Oppositionen in der Tierwelt wie „Hengst“ - „Stute“ (bzw. *stallion* - *mare*). Das Paar *man* - *woman* ist ein seltener Vertreter für konträre Nomen. Für konverse Antonyme wie „Nachfolger“ - „Vorgänger“ (bzw. *successor* - *predecessor*) konnte in WordNet keine Verknüpfung festgestellt werden. [Miller 1998a, S. 40] weist darauf hin, dass die wohl interessanteste Eigenschaft antonymer Nomen diejenige sei, dass sie meistens dasselbe, häufig sogar dasselbe unmittelbare Hyperonym haben.

Der Assoziation zwischen Nomen wie *size* und den Adjektiven *small* und *big* wurde in WordNet mit Einführung der **Attribut-Relation** Folge geleistet. Aufgenommen wurden jedoch nur semantisch wenig komplexe Adjektiv-Attribut-Verbindungen [Miller 1998a, S. 41].

3.1.2.2 Aufbau der Verb-Datei

Wie die Nomen sind auch die rund 14 000 Verb-Synsets hierarchisch strukturiert und aufgrund semantischer Kriterien in 15 Felder aufgeteilt worden, welche durch die entsprechenden Unique Beginners benannt werden. 14 Dateien beschreiben je ein semantisches Feld, beispielsweise Kommunikations- oder Wetterverben. Die 15. Datei umfasst Verben, die Zustände ausdrücken, so genannte stative Verben. Ebenfalls in dieser Gruppe sind Hilfs- und Kontrollverben (e.g. *succeed*) zu finden [Fellbaum 1998a, S. 7f.]. In den entsprechenden Synsets werden auch Phrasal Verbs wie beispielsweise *let down,* idiomatische Wendungen wie *kick the bucket* und metaphorische Verwendungen wie *die* im Synset *break down* aufgeführt.

Im Gegensatz zur Nomen-Datei spricht [Fellbaum 1998a, S. 79] bei in der Verb-Datei nicht von Hyponymie, sondern von **Troponymie**. Der Grund liegt darin, dass sie den Testrahmen für Hyponymie bei Nomen („X is a kind of Y“) für Verben als nicht geeignet betrachtet. Die Troponymie-Relation besteht, wenn man von zwei Verben V_1 und V_2 sagen kann: „To V_1 is to V_2 in some particular manner“ [Fellbaum 1998a, S. 79]. In seltenen Fällen konnte multiple Vererbung festgestellt werden. Das Verb *to eat (take in solid food)* hat beispielsweise zwei Hyperonyme, nämlich *to eat (eat a meal, take a meal)* und *consume, ingest, take in, take, have (serve oneself to, or consume regularly).*

Die **Antonymie-Relation** umfasst komplementäre Verbpaare wie *live - die,* konträre Verben wie *love - hate* und konverse Paare wie *sell - buy.* [Fellbaum 1998a, S. 82] weist darauf hin, dass die Gegensatzrelation vieler verbaler Paare auf morphologischer Markiertheit eines Verbs basiert, wie dies in Paaren wie *appear - disappear* der Fall ist[9].

Die **Implikations-Relation** (engl. *entailment*) gilt dann, wenn ein Verb ein anderes notwendigerweise voraussetzt [Kunze 2005b, S. 167]. Sie gleicht der Meronymie bei Nomen und wurde deshalb nicht als solche bezeichnet, da die Beziehung zwischen Teilen einer Aktivität, Ereignis oder Status sich von der Beziehung zwischen Teilen eines nominalen Referenten unterscheidet [Fellbaum 1998a, S. 78]. Eine Implikation besteht, wenn der Satz *someone* V_1 logisch den Satz *someone* V_2 bedingt, was beispielsweise zwischen *sleep* und *snore* der Fall ist. Ein Verb kann gleichzeitig durch Troponymie und Implikation mit einem anderen Verb verbunden sein, was beim Paar *walk - march* beobachtet werden kann.

Die **Kausations-Relation** (engl. *cause*) trifft auf kausativ-resultative Paare wie *kill*

[9] *Disappear* ist im Vergleich zu *appear* durch das Präfix *dis-* markiert.

- *die* zu oder auf Wörter, die sowohl eine transitiv-kausative als auch eine intransitiv-antikausative Lesart haben (e.g. *melt*)[10].

Weiter werden für Verben Subkategorisierungsrahmen aufgeführt (in WordNet Satz-Frames, engl. *Sentence Frames*, genannt), welche die möglichen Verwendungen in einem Satz benennen [Tengi 1998, S. 108]. Sie bestehen aus paraphrastischen Sätzen mit pronominalen Variablen für die Komplemente, illustriert am Beispiel 10 für das Verb *accuse*.

(10) Satz-Frames für *accuse*

a. Somebody - - - - s somebody.

b. Somebody - - - - s somebody of something.

Die Satz-Frames beschreiben *accuse* als transitives Verb, da es nur mit einem Objekt zusammen auftritt (*somebody*). Optional kann *accuse* durch eine durch die Präposition *of* eingeleitete Präpositionalphrase ergänzt werden. Satz-Frames sind in WordNet allerdings nicht systematisch erfasst worden.

3.1.2.3 Aufbau der Adjektiv-Datei

Die 19 000 Synsets der Adjektiv-Datei beinhalten neben Adjektiven auch einige Nomen und Präpositionalphrasen, die häufig als Modifikatoren auftreten (e.g. *home* in *home office* oder *at that place*). Adjektiv-Synsets werden in zwei Klassen aufgeteilt. Man unterscheidet zwischen deskriptiven Adjektiven wie *big* oder *interesting* und den seltener vorkommenden relationalen Adjektiven wie *electrical* [Miller 1998b, S. 47].

Ein deskriptives Adjektiv schreibt einem Nomen einen Attributwert zu. Bereits erwähnt wurde die **Attribut-Relation,** welche Nomen mit den passenden modifizierenden deskriptiven Adjektiven verknüpft.

Die wichtigste semantische Relation zwischen deskriptiven Adjektiven ist die **Antonymie.** In der Adjektiv-Datei sind sowohl komplementäre Antonyme wie *dead - alive* als auch konträre wie *fat - thin* berücksichtigt. Obwohl die Antonymie-Relation bei Adjektiven zentral ist, haben viele kein Antonym (e.g. *moist*). Für solche Fälle haben die WordNet-Entwickler so genannte **Ähnlichkeits-Pointer** eingeführt; ein solcher Pointer führt im Beispielsfall zu *wet*. Da *wet* das Antonym von *dry* ist,

[10] Ein Verb ist kausativ, wenn das Agens Nummer eins Agens Nummer zwei dazu veranlasst, eine Handlung auszuführen (e.g. „Die Sonne schmilzt die Schokolade.“). Bei der antikausativen Form spielt das Agens keine Rolle (e.g. „Die Schokolade schmilzt.“).

besteht eine konzeptuelle Opposition[11] zwischen *moist* und *dry* [Miller 1998b, S. 50]. [Gross et al. 1989, S. 95] unterscheiden dann zwischen direkten Antonymen (e.g. *wet* - *dry*) und indirekten Antonymen (e.g. *moist* - *dry*). Durch Einbezug des Ähnlichkeitsmasses verfügen, abgesehen von wenigen Ausnahmen, alle deskriptiven Adjektive über ein direktes oder indirektes Antonym.

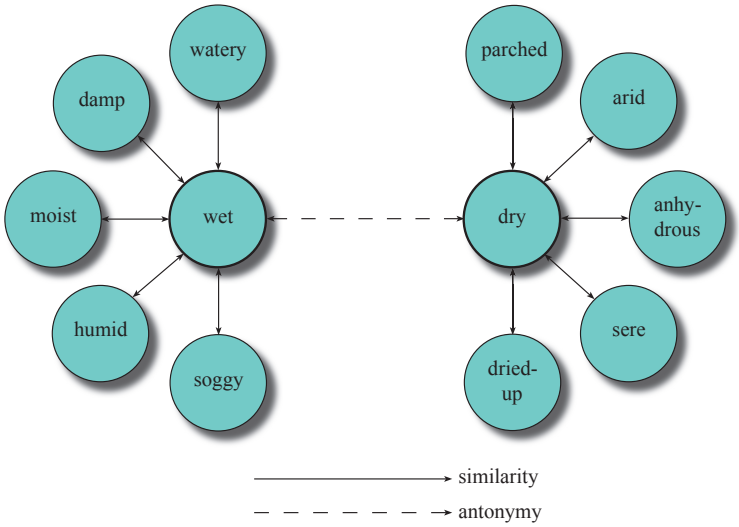

Abbildung 3: Adjektiv-Cluster in WordNet basierend auf [Miller 1998b, S. 51]

Die Abbildung 3 zeigt die Beziehung zwischen der Ähnlichkeits- und der Antonymie-Relation schematisch auf. Mehrere Adjektiv-Synsets sind durch die Ähnlichkeitsbeziehung mit den antonymen Adjektiven *wet* und *dry* verknüpft. Ein Adjektiv-Cluster enthält also ein Antonympaar und mehrere so genannte Satelliten-Synsets.

Eine Unterklasse der deskriptiven Adjektive sind partizipiale Adjektive wie *elapsed*. Diejenigen ohne direkte oder indirekte Antonyme wurden in einer separaten Datei erfasst und durch die **Partizip-Relation** mit den entsprechenden Verben (*elapse*) verknüpft.

Relationale Adjektive sind eine grosse, offene Klasse, die sich semantisch oder morphologisch auf Nomen beziehen, wie dies bei *musical* und *music* der Fall ist [Miller 1998b, S. 59ff.]. Die Derivations-Relation zwischen *musical* und *music* wird häufig **Pertonymie**[12] genannt. Typischerweise sind relationale Adjektive nicht graduierbar oder steigerbar und haben selten direkte Antonyme [Miller 1998b, S. 59f.].

[11] Damit ist eine Gegensätzlichkeit auf Synset-Ebene und nicht auf Wortebene gemeint.

[12] Der Begriff Pertonymie wurde aus dem englischen *pertain* („zu etw. gehören“) und dem griechischen *onoma* („Name“) gebildet und bezeichnet die Derivations-Relation bei Adjektiven.

3.1.2.4 Aufbau der Adverb-Datei

Unter den 3600 Adverb-Synsets in WordNet sind viele Adverbien durch Suffigierung von Adjektiven abgeleitet worden (e.g. *beautifully*). Daneben umfasst die Adverb-Datei viele mit der Adjektivform homomorphe Adverbien (e.g. *fast*) und adverbial verwendete Phrasen wie *home* (beschrieben als *at or to or in the direction of one's home or family*). Zentrale Relationen sind **Synonymie** und **Antonymie.** Abgeleitete Adverbien sind zusätzlich durch **Pertonymie** mit den entsprechenden Adjektiven verlinkt und erben von ihnen Eigenschaften wie Antonymie, wie dies zwischen dem Adjektivpaar *specific* - *general* bzw. dem Adverbpaar *specifically* - *generally* beobachtet werden kann.

Die Anzahl Synsets in WordNet beträgt insgesamt rund 118 000. Während Nomen und Verben in einer Baumstruktur vorliegen, bilden Adjektive eine cluster-ähnliche Struktur. Neben Synonymie sind je nach Wortart andere lexikalisch-semantische Relationen zentral. Von den insgesamt rund ein Dutzend Relationen ist dies für Nomen und Verben Hyponymie, für Adjektive und Adverbien hingegen die Antonymie.

3.2 GermaNet

Mit GermaNet ist ein lexikalisch-semantisches Wortnetz für das Deutsche aufgebaut worden, das die wichtigsten und häufigsten Nomen, Verben und Adjektive sowie ihre elementaren semantischen Relationen zueinander modelliert.

3.2.1 Motivation

Der Aufbau von GermaNet wurde durch fehlende maschinenlesbare lexikalisch-semantische Netze für die deutsche Sprache angeregt. Ursprüngliche Zielsetzung war es, eine Grundlagenressource für die Lesartendisambiguierung zu schaffen [Kunze 2005b, S. 163]. Von 1996–2001 wurde an der Universität Tübingen daran gearbeitet, vier Jahre ruhte die Weiterentwicklung, seit 2005 wird GermaNet wieder erweitert [Kunze und Lemnitzer 2005, S. 345].

GermaNet ist im Gegensatz zu WordNet kostenpflichtig. Auf [canoonet 2006] kann man sich die zentralen Relationen (Synonymie, Hyponymie/Hyperonymie, Antonymie) durch Eingabe des Konzepts im Suchfeld anzeigen lassen[13].

[13] Die in diesem Kapitel aufgeführten deutschen Beispiele stammen aus dieser Abfrage, welche Daten aus dem Jahr 2001 als Grundlage hat.

3.2.2 Aufbau und semantische Relationen

GermaNet ist keine Übersetzung von WordNet, sondern wurde von Grund auf neu aufgebaut. Der Grundwortschatz wurde anhand von Häufigkeitslisten aus Korpora und lexikalischen Ressourcen wie zum Beispiel dem *Duden 8* ermittelt [Kunze 2005b, S. 163].

GermaNet-Daten liegen in je einer Nomen-, Verb- und Adjektiv-Datei vor. Die Erfassung von Adverbien ist geplant [Kunze und Lemnitzer 2005, S. 346]. Eigennamen wie geografische Bezeichnungen (e.g. „Holland“), stilistische (e.g. „schiffen“ für „urinieren“) und orthografische Varianten (e.g. „Xylofon“) werden explizit markiert[14].

Im Gegensatz zu WordNet, welches nur tatsächlich auftretende Wörter einer Sprache abbildet, macht GermaNet Gebrauch von **künstlichen Konzepten**[15]. Künstliche Konzepte werden einerseits für lexikalische Lücken (e.g. fehlendes Antonym von „durstig“) verwendet, andererseits können sie auf rein konzeptuelle Konstrukte wie „?Charakterbeschaffener“ als Oberbegriff von „Feigling“ verweisen [Kunze und Wagner 1999, S. 8]. Die Verwendung künstlicher Konzepte soll die Hierarchie besser strukturieren und wird durch die Forderung von [Cruse 1986, S. 150] begründet, dass Kohyponyme eines dominierenden Knotens möglichst inkompatibel zueinander sein sollen. Beispiele für künstliche Konzepte sind in der Abbildung 4 dargestellt.

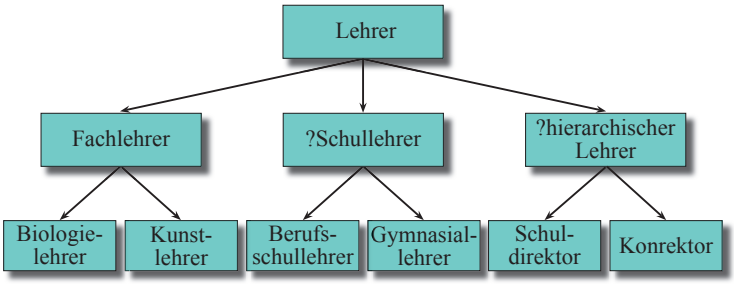

Abbildung 4: Künstliche Konzepte in GermaNet nach [Kunze 2005b, S. 172]

„?Schullehrer“ und „?hierarchischer Lehrer“ helfen, das Teilnetz „Lehrer“ symmetrischer zu strukturieren [Kunze 2005b, S. 171]. Würde man auf diese beiden künstlichen Konzepte verzichten, wären die terminalen Knoten im Beispielnetz nicht inkompatibel genug. Zusätzlich befänden sich „Fachlehrer“, „Gymnasiallehrer“ und „Schuldirektor“ völlig unmotiviert auf derselben Hierarchiestufe. Die Hinzunahme

[14] Eigennamen werden mit einem Dollarzeichen vor dem Wort gekennzeichnet, stilistische Varianten mit *s, orthografische mit *o nach dem Wort.

[15] Künstlichen Konzepten wird ein Fragezeichen vorangestellt.

der nicht lexikalisierten Knoten führt so zu einer besseren Strukturierung der Konzepthierarchie. Der Begriff künstliches Konzept kann insofern als ungünstig gewählt beurteilt werden, als er suggeriert, dass das Konzept nicht existiert. Besser verständlich wäre es wohl, solche Konzepte durchgängig als lexikalische Lücken oder nicht lexikalisierte Konzepte zu bezeichnen.

Lexikalisch-semantische Relationen, die in allen Dateien auftreten, sind Synonymie, Hyponymie/Hyperonymie, Antonymie, *See-Also*-Relation und Derivations-Relation. Bei der ***See-Also*-Relation** handelt es sich laut [Kunze 2005b, S. 167] um eine unterspezifizierte Assoziationsbeziehung, welche sie weder zu den lexikalischen noch zu den konzeptuellen Relationen zählt. Die Relation wurde in allen drei Dateien kodiert, operiert jedoch nicht über die Wortartengrenzen. So wird zum Beispiel das Paar „Brauerei“ - „Bier“ durch diese Relation verknüpft.

Die **Derivations-Relation** ist eine kategorienübergreifende Relation [Kunze 2005b, S. 166], beispielsweise für Nominalisierungen (e.g. „beobachten“ - „Beobachtung“) oder Verbalisierungen (e.g. „Luft“ - „entlüften“).

3.2.2.1 Aufbau der Nomen-Datei

Die Unterteilung der 36 000 Nomen-Synsets[16] in semantische Felder und die entsprechenden Unique Beginners stimmen in GermaNet mit wenigen Ausnahmen mit WordNet überein. **Hyponymie** ist die wichtigste konzeptuelle Relation für Nomen und bildet eine hierarchische Struktur. Diese Hierarchie ist wegen multipler Vererbung jedoch kein Baum. Konzepte, welche unterschiedlichen Hierarchien angehören, werden kreuzklassifiziert, d.h. sind mehr als einem Hyperonym untergeordnet.

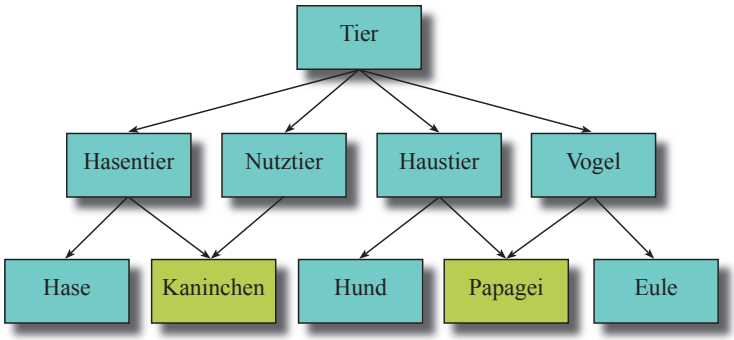

Abbildung 5: Kreuzklassifikation in GermaNet nach [Kunze und Wagner 2001, S. 233]

[16] Die Daten zum Umfang von GermaNet beruhen auf einer Präsentation von [Kunze 2005a].

Wie Abbildung 5 zeigt, ist beispielsweise „Papagei“ sowohl Unterbegriff von „Vogel“ als auch von „Haustier“, „Eule“ hingegen nur von „Vogel“. In WordNet gibt es zwar einige Mehrfachvererbungen, die Kreuzklassifikation ist jedoch nicht systematisch und ordnet beispielsweise *parrot* nur unter *bird* ein. Die Untersuchung kreuzklassifizierter Synsets kann helfen, Muster regulärer Polysemie herauszufinden, die für die Strukturierung des Lexikons nützlich sind [Kunze und Wagner 2001, S. 233].

Eine weitere prominente Sinnrelation ist die **Meronymie-Relation,** die nur für konkrete Nomen spezifiziert wird. In GermaNet wurde nur der Typ Komponente-Objekt realisiert. Die Teil-Ganzes-Beziehung besteht so beispielsweise zwischen „Chromosom“ - „DNA“.

Antonymie kommt in der Nomen-Datei zwischen komplementären Paaren wie „Hengst“ - „Stute“ und konversen Paaren wie „Vorgänger“ - „Nachfolger“ vor. GermaNet verknüpft auch so genannte fakultative Gegensatzpaare [Kunze 2005b, S. 176f.]. Dieser Begriff stammt von [Agricola und Agricola 1982, S. 20] und bezieht sich auf „nicht eigentliche Bedeutungsgegensätze“ wie beispielsweise „Planwirtschaft“ - „Marktwirtschaft“ oder „römisch“ - „arabisch“ (in Bezug auf das Zahlensystem).

3.2.2.2 Aufbau der Verb-Datei

Die Verbklassifikation der rund 9000 Synsets folgt nur in den obersten Stufen derjenigen von WordNet. Sie orientiert sich vor allem an der Klassifikation semantischer Felder von [Levin 1993]. GermaNet hat ebenfalls 15 Klassen, wobei in der 15. Klasse stative Verben (dort als „allgemeine Verben“ bezeichnet) eingeordnet sind.

Neben **Hyponymie** wird auch **Antonymie** kodiert. Komplementäre Paare wie „leben“ - „sterben“, konträre Paare wie „lieben“ - „hassen“ und auch konverse Paare wie „kaufen“ - „verkaufen“ lassen sich finden.

Die **Kausations-Relation** verknüpft Ursache und Wirkung und besteht etwa zwischen „zeigen“ und „sehen“. Auch im Deutschen gibt es Verben wie „schmelzen“, die sowohl eine transitive als auch eine intransitive Lesart haben und bei denen zwischen den beiden Lesarten eine Kausations-Relation besteht. Diese Relation operiert jedoch nicht nur zwischen Verbpaaren, sondern auch zwischen Verb-Adjektiv-Paaren (e.g. „zerstören“- „kaputt“). Innerhalb der Kausations-Relation werden systematisch verschiedene Alternationstypen von Verben[17] berücksichtigt. So sind beispielsweise die kausative und die antikausative Lesart von „kochen“ („Er kocht die Suppe“ vs. „Die Suppe kocht“) miteinander verbunden.

[17] Verbalternation beschreibt einen Wechsel der Realisierung der Argumentstruktur eines Verbs.

Die **Implikations-Relation** verbindet Verben, wovon eines das andere voraussetzt (e.g. „gelingen“ - „versuchen“). In GermaNet kommt diese Relation nicht häufig vor, da es laut [Kunze 2005b, S. 167] nur wenige echte Implikationen gibt. Selbst Paare wie „suchen“ - „finden“ sind nicht unproblematisch. Man kann durchaus 5000 Euro finden, ohne danach gesucht zu haben.

Um lexikalisch-semantische und syntaktische Information zu verknüpfen, werden Subkategorisierungsrahmen angegeben. Verbeinträge werden mit Informationen zu ihrem syntaktischen Komplementierverhalten ergänzt. Jede Lesart wird mit dem entsprechenden Rahmen verbunden, welche auf den *Complementation Codes* der CELEX-Datenbank[18] basieren. Die Tabelle 2 zeigt die acht Lesarten vom Verb „setzen“ mit den dazugehörenden Subkategorisierungsrahmen[19].

	Code	**Beispiel**
setzen**1**	NN.AN	Er setzt die Fahnen.
setzen**2**	NN.AR	Sie setzt sich.
setzen**3**	NN.AN.BL	Er setzt den Schüler ans Fenster.
setzen**4**	NE.AN	Es setzt Prügel.
setzen**5**	NN.AN.PP	Er setzt seine Hoffnungen auf sie.
setzen**6**	NN.An	Die Häsin hat (Junge) gesetzt. (Jägersprache)
setzen**7**	NN.BL	Das Pferd setzt über die Hürde.
setzen**8**	NN.AN.Dn	Das Protokoll setzt (ihnen) Schranken.

Tabelle 2: Subkategorisierungsrahmen in GermaNet für das Verb „setzen“ nach [Kunze und Wagner 2001, S. 235]

3.2.2.3 Aufbau der Adjektiv-Datei

Die ca. 5500 Adjektiv-Synsets umfassende Datei wird in GermaNet in 14 semantische Felder eingeteilt, welche auf der Klassifikation von [Hundsnurscher und Splett 1982] basieren. Eine solche Klasse ist die wetterbezogene, welcher beispielsweise Adjektive wie „schwül“ zugeordnet werden. Zusätzlich gibt es eine 15. Klasse für Pertonyme (e.g. „finanziell“ - „Finanzen“). Somit werden Adjektive wie Nomen und Verben hierarchisch eingeteilt und nicht wie in WordNet aufgrund von Ähnlichkeit. Daher ist die **Hyponymie-Relation**

[18] Die *Complementation Codes* wurden von der CELEX German Database Release 2.5 übernommen. Die webbasierte Version kann unter [CELEX 2006] aufgerufen werden.

[19] Wofür die Abkürzungen in der zweiten Spalte stehen, ist bei [GermaNet 2006], Verbs -> Verb Frames dokumentiert.

zentral, die beispielsweise die Adjektive „fröhlich“ (Hyperonym) und „lustig“ (Hyponym) verbindet.

Mit der Wahl der hierarchischen Anordnung entfällt die Unterscheidung zwischen direkten und indirekten Antonymen. Die **Antonymie-Relation** besteht sowohl zwischen komplementären Paaren wie „tot“ - „lebendig“ als auch zwischen konträren Paaren wie „dick“ - „dünn“.

Eine Sinnrelation, die nur bei Adjektiven auftritt, ist die **Partizip-Relation.** Sie besteht zwischen Adjektiven, die auf eine verbale Partizipform zurückgehen, und ihrem Basisverb [Kunze 2005b, S. 166], beispielsweise zwischen „beschlossen“ und „beschliessen“. Interessant sind Partizipien eines nicht mehr produktiven Basisverbs, welche auch durch die Partizip-Relation verknüpft werden, so beobachtet zwischen „verschossen“ (im Sinne von „verliebt sein“) und dem ungebräuchlichen „(sich) verschiessen“.

3.2.2.4 Adverbien

Es ist geplant, eine Adverb-Datei aufzubauen, wobei nur Adverbien berücksichtigt werden sollen, die keine identische Adjektivform haben, also nur „echte“ Adverbien wie „gestern“ oder „dort“. Zudem ist vorgesehen, die Datei stärker zu strukturieren als diejenige von WordNet. Die hierarchische Gliederung der Synsets soll beibehalten werden. Als dominierende Relationen werden Antonymie und Synonymie erwartet [Kunze und Lemnitzer 2005, S. 346f.].

Neben dem Aufbau einer Adverbien-Datei planen die Entwickler von GermaNet, Selektionsbeschränkungen zu erfassen. Dazu sollen Prädikate und ihre typischerweise selegierten Argumente aus Korpora extrahiert werden [Lemnitzer 2005, persönliche Kommunikation]. Hierbei soll auf das Inventar entsprechender Relationen zurückgegriffen werden, das im Rahmen des EuroWordNet-Projekts (vgl. 3.4.2) festgelegt wurde [Kunze und Wagner 2001, S. 244].

3.3 GermaNet und WordNet im Vergleich

Obwohl WordNet und GermaNet ähnlich beschaffen sind, gibt es grundlegende Unterschiede. Das wichtigste Kriterium ist die Grösse der Datenbanken. WordNet hat mit rund 118 000 Synsets mehr als doppelt so viele wie GermaNet. Dann unterscheiden sich die Wortnetze in ihrer Motivation. Während WordNet sich an psycholinguistischen Prinzipien orientiert und als Experiment einer Implementation eines

Lexikonmodells gedacht war, beruht GermaNet auf linguistischen Prinzipien und wurde als Ressource zur Lesartendisambiguierung geschaffen.

Der Aufbau der Wortnetze weicht vor allem in der Strukturierung der Adjektiv-Datei ab. WordNet setzt auf Ähnlichkeitsverbünde, oft auch Satelliten-Ansatz [Kunze und Wagner 2001, S. 235] genannt. GermaNet hingegen gliedert Adjektive analog zu Verben und Nomen hierarchisch. GermaNet hat im Gegensatz zu WordNet bis anhin erst Pläne für eine Adverb-Datei.

Weitere Abweichungen sind die künstlichen Konzepte für nicht lexikalisierte Konzepte und rein konzeptuelle Konstrukte, welche nur GermaNet verwendet. GermaNet greift viel häufiger als WordNet auf multiple Vererbung zurück, welche Kreuzklassifikation genannt wird.

Die lexikalisch-semantischen Relationen sind ähnlich, jedoch haben die Pointers unterschiedliche Gültigkeitsbereiche. So können in GermaNet vermehrt Synsets verschiedener Wortarten miteinander in Beziehung gebracht werden. Die Kausations-Relation, die in WordNet nur zwischen Verben vorgesehen ist, kann beispielsweise in GermaNet zwischen allen Wortarten kodiert werden. Zudem verknüpft die Meronymie-Relation in GermaNet nur Komponenten und ihre Objekte. WordNet kodiert hingegen drei verschiedene Typen von Teil-Ganzes-Beziehungen. Ein letztes zentrales Unterscheidungskriterium in der Struktur ist, dass GermaNet andere und mehr Subkategorisierungsrahmen aufführt als WordNet.

Schliesslich soll auf die Tatsache hingewiesen werden, dass es sich bei WordNet um ein Open-Source-Projekt handelt, GermaNet hingegen Lizenzgebühren erhebt.

3.4 EuroWordNet

EuroWordNet war von 1996–1999 ein Projekt der Europäischen Gemeinschaft, das WordNet-ähnliche Wortnetze weiterer europäischer Sprachen erstellt, in einer multilingualen Datenbank gespeichert und miteinander verknüpft hat [Vossen 2004a].

Bei Projektschluss umfasste EuroWordNet Wortnetze von acht europäischen Sprachen (Holländisch, Italienisch, Spanisch, Englisch, Tschechisch, Estnisch, Deutsch und Französisch). EuroWordNet beinhaltet für jede der vier erstgenannten Sprachen je 30 000 Konzepte, für die vier letztgenannten je 15 000 Konzepte [Kunze und Wagner 1999, S. 10]. Später stiessen weitere Sprachen wie Schwedisch und Russisch hinzu.

3.4.1 Motivation

Als wichtigstes Einsatzgebiet für diese multilinguale Datenbank wurde anfänglich die sprachübergreifende Informationserschliessung angesehen. Der Vergleich verschiedener Wortnetze ermöglichte aber auch Forschungen zu Lexikalisierungsmustern einzelner Sprachen und konnte zudem zu einer Überprüfung der Konsistenz einzelner Wortnetze beitragen. Allgemein kann EuroWordNet als ein mächtiges Werkzeug für die Untersuchung lexikalisch-semantischer Ressourcen betrachtet werden [Vossen 2004a, S. 74].

3.4.2 Aufbau und semantische Relationen

Drei Ebenen formen EuroWordNet. Erstens ist dies die sprachabhängige Ebene, die monolinguale WordNets[20] umfasst, zweitens die sprachunabhängige Ebene mit dem Interlingualen Index, und drittens die ebenfalls sprachunabhängige Ontologie-Ebene.

Laut [Wagner 2005, S. 11] wurden die einzelnen WordNets entweder wie GermaNet von Grund auf neu aufgebaut (*„merge approach“*), oder man übersetzte das Princeton WordNet, so erfolgt für das französische WordNet (*„expand approach“*). Beide Ansätze orientieren sich im Aufbau und den semantischen Relationen stark am Princeton WordNet. Ein wichtiger Unterschied ist, dass in EuroWordNet auf eine Unterteilung der Dateien nach Wortarten verzichtet und eine an [Lyons 1977] angelehnte Einordnung nach semantischen Stufen vorgenommen wird [Alonge et al. 1998, S. 98]. Einheiten erster Stufe sind demnach konkrete, physikalische Dinge (meist konkrete Nomina), die zweite Stufe umfasst unter anderem Eigenschaften, Prozesse, Ereignisse (Nomen, Verben und auch Adjektive). Einheiten der dritten Stufe sind beispielsweise abstrakte Nomen.

Somit werden semantische Relationen zwischen Konzepten verschiedener Wortarten möglich. So wird das Nomen *arrival* mit dem Verb *to leave* durch die wortartenübergreifende **Antonymie-Relation** verknüpft [Alonge et al. 1998, S. 97f.]. **Meronymie** wird im Gegensatz zu WordNet, das drei Unterscheidungen macht, und GermaNet mit nur einem generellen Pointer in EuroWordNet durch einen generellen und fünf spezialisierte Pointers erfasst [Kunze und Wagner 1999, S. 11f.]. Alle kodierten Relationen haben [Alonge et al. 1998, S. 111–113] aufgelistet.

Die sprachspezifischen WordNets wurden miteinander verknüpft, so dass eine Verbindung zwischen den Synset einer Sprache und ihren Übersetzungsäquivalenten in

[20] Mit dem Begriff WordNets werden die verschiedenen Wortnetze der WordNet-Familie bezeichnet.

der anderen Sprache besteht. Die Verknüpfung erfolgt mittels des **Interlingualen Index** (kurz ILI), der aus einer unstrukturierten Liste von Einträgen besteht, die durch einen eindeutigen Code gekennzeichnet sind [Vossen 2004a, S. 164]. Die ILI-Einträge wurden zum grössten Teil aus Synsets von WordNet 1.5 übernommen. Sie werden nicht untereinander, sondern durch **einfache Äquivalenz-Relationen** mit dem entsprechenden Synset eines WordNet verknüpft. Ein Vorteil dabei ist, dass neue Sprachen einfach zu EuroWordNet hinzugefügt werden können, indem sie zum ILI verlinkt werden. So wird sowohl das deutsche Konzept „fahren“ als auch das italienische *guidare* mit dem ILI-Eintrag *drive* in Bezug gebracht.

Da Konzepte in verschiedenen Sprachen unterschiedlich lexikalisiert sind, ist das Mapping der beteiligten Wortnetz-Strukturen oft problematisch. In solchen Situationen sind die so genannten **komplexen Äquivalenz-Relationen** nützlich. Sie kommen zum Einsatz, wenn es in einer Sprache eine lexikalische Lücke gibt oder die Bedeutungen nicht genau übereinstimmen. Zur Illustration erwähnt [Vossen 2004a, S. 165f.] das holländische *schoonmaken* („sauber machen“), welches nur eine Lesart hat, wohingegen die Lesarten des englischen *to clean* mit mindestens vier ILI-Einträgen übereinstimmen. Das holländische Synset *schoonmaken* wird daher durch die **Beinahe-Synonymie-Relation** mit allen Lesarten von *to clean* verknüpft. Ein weiteres Beispiel für ein deutsches Konzept ohne englische Äquivalente ist der „Buss- und Bettag“. Das Konzept wird mit dem ILI-Eintrag *Christian holy day* durch **Hyperonymie** und zum Eintrag *penitence* durch **Meronymie** verknüpft [Wagner 2005, S. 12].

Konzept A	thematische Rolle	Konzept B
abbeverarsi <drink>	*Involved_agent*	animale <animal>
lessare <boil>	*Involved_patient*	cibo <food>
lavare <wash>	*Involved_instrument*	acqua <water>

Tabelle 3: Thematische Rollenrelationen aus dem italienischen WordNet nach [Wagner 2005, S. 13]

Einige WordNets in EuroWordNet umfassen Relationen, welche Selektionsbeschränkungen ausdrücken. Diese Relationen kodieren thematische Rollen zwischen Verb- und Nomenkonzepten. Im Grunde genommen spezifizieren sie, welche Art von Nomenkonzepten in den Ereignissen der Verbkonzepte involviert sind. Diese Relationen sind bidirektional, das heisst, der Pointer vom Verb zum Nomen heisst ***Involved,*** die umgekehrte Richtung ***Role.*** Tabelle 3 zeigt einige Beispiele von thematischen Rol-

lenrelationen des italienischen WordNet. Die ILI-Einträge, welche den italienischen Konzepten entsprechen, sind in spitzen Klammern aufgeführt.

Diese thematischen Rollenrelationen sind aber in EuroWordNet nicht systematisch kodiert, und es besteht auch keine Information dazu, wie die semantischen Argumente syntaktisch realisiert werden.

Die ILI-Einträge wurden mit zwei verschiedenen Ontologien verknüpft. Zum einen ist dies die so genannte **Top-Ontologie,** welche aus einer Hierarchie von sprachunabhängigen Konzepten besteht und wichtige semantische Unterscheidungen (e.g. Objekt vs. Substanz, dynamisch vs. statisch) reflektiert. Der ILI-Eintrag *car* wird durch die Top-Konzepte *artifact, object* und *vehicle* klassifiziert. Da das deutsche Konzept „Auto“ mit *car* verknüpft ist, erbt „Auto“ die Klassifizierung von *car*.

Zum anderen besteht eine Verknüpfung zur **Domänenontologie,** welche aus Wissensstrukturen besteht, die Begriffe nach Themengebiet oder semantischen Feldern (e.g. Verkehr, Sport) gruppieren.

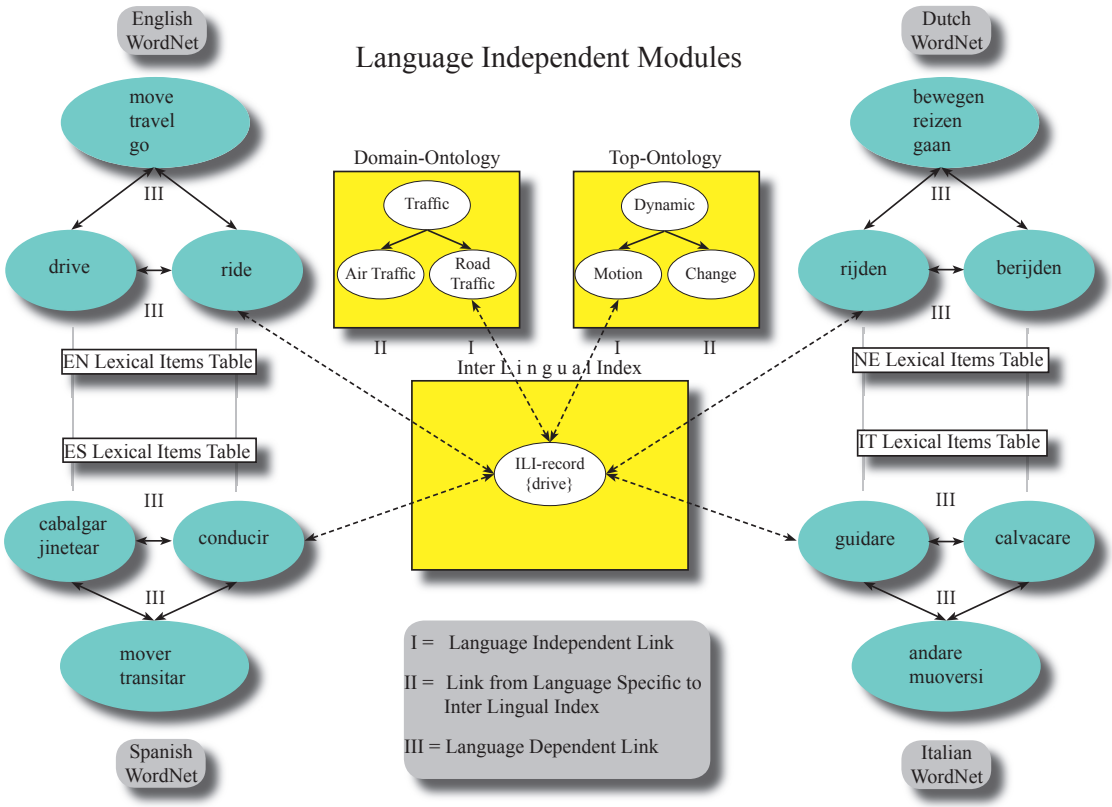

Abbildung 6: Architektur von EuroWordNet nach [Vossen 1998, S. 80]

Abbildung 6 zeigt eine vereinfachte Übersicht über die Architektur von EuroWordNet. Die verschiedenen Module und ihre Beziehungen zueinander werden schema-

tisch dargestellt. In der Mitte befindet sich das sprachunabhängige Modul ILI, welches einen bestimmten Eintrag einerseits mit der Domänenontologie und der Top-Ontologie in Beziehung setzt, andererseits mit den sprachspezifischen Entsprechungen verknüpft wurde. Das Konzept *Motion* der Top-Ontologie ist direkt mit dem ILI-Eintrag *drive* verbunden und bezieht sich auf alle sprachspezifischen Konzepte, die mit diesem ILI-Eintrag verknüpft sind. Rund um die sprachunabhängigen Module sind die einzelnen sprachspezifischen WordNets mit ihren lexikalisch-semantischen Relationen zu sehen.

3.4.3 Integration von GermaNet in EuroWordNet

1998 wurden 15 000 GermaNet-Konzepte in EuroWordNet aufgenommen. In einem ersten Schritt sollten die definierten 1400 Basiskonzepte mit dem ILI verknüpft werden [Kunze und Wagner 1999, S. 13]. Zu diesem Zeitpunkt wurde schon erkannt, dass in GermaNet nicht alle diese Konzepte enthalten waren. Die Lücken wurden geschlossen und die sprachinternen Relationen auf die relevanten Pointers von EuroWordNet abgebildet. In einem zweiten Schritt wurden Hyponyme der Basiskonzepte kodiert.

Problematisch waren laut [Kunze und Wagner 1999, S. 15] verschiedene Lexikalisierungsmuster, so hat beispielsweise das Konzept „gemütlich“ keine exakte Übersetzungsäquivalente im Englischen. Weiter sind kulturelle Verschiedenheiten aufzuführen, die Übersetzungen von kulturkreisspezifischen Konzepten wie „Gymnasiallehrer“ erschwerten.

Insgesamt schätzen [Kunze und Wagner 1999, S. 16] die Integration von GermaNet in EuroWordNet als sehr nützlich ein. Die Grundwortschatzabdeckung von GermaNet konnte auf Inkonsistenzen überprüft, Lücken entdeckt und behoben werden. [Kunze und Wagner 2001, S. 238] nennen den Bereich der Ereignisnominalisierung, der zuvor noch nicht optimal abgedeckt war. Wie schon erwähnt, sollen die Rollenrelationen von EuroWordNet in GermaNet als Basis für Selektionsbeschränkungen dienen [Kunze und Wagner 2001, S. 239]. Die Integration und die daraus resultierende Datenüberprüfung hat sich also positiv auf die Weiterentwicklung von GermaNet ausgewirkt.

3.5 Weitere Wortnetze der WordNet-Familie

Die internationale Zusammenarbeit im EuroWordNet-Projekt hat mit dem ILI als grundlegendes Lesarten-Inventar einen Quasi-Standard für den Aufbau von Wortnetzen entwickelt. Zur Entwicklung von Richtlinien und Standards für Wortnetze und zum Austausch der in Wortnetzen enthaltenen Daten und Strukturen ist die *Global WordNet Association* (kurz GWA) gegründet worden. Dabei handelt es sich um eine Plattform für Kooperationen, die auch die Integration sprachspezifischer WordNets in EuroWordNet fördern will [Wagner 2005, S. 10].

Laut Angaben der [GWA 2006] gibt es heute WordNets für 37 Sprachen, worunter sich auch viele Vertreter nichtindoeuropäischer Sprachen wie Baskisch oder Thai finden.

3.6 Einordnung

Handelt es sich bei den Wortnetzen der WordNet-Familie um Taxonomien, Thesauri oder Ontologien? [Fellbaum 2002] erläutert, dass WordNet die Kriterien für eine **Ontologie** nicht erfüllt, da nicht strikt zwischen Typen und Rollen unterschieden wird. Einerseits führt das Prinzip der Kreuzklassifikation teilweise zu einer solchen Differenzierung, doch sind in WordNet und in GermaNet beispielsweise als Hyponyme des Synsets „Hund“ sowohl Typen wie „Pudel“ als auch Rollen wie „Jagdhund“ aufgeführt. Zudem werden zumindest in WordNet keine nicht lexikalisierten Konzepte berücksichtigt. Der sprachspezifische Aufbau widerspricht der intendierten Allgemeingültigkeit einer Ontologie. Ontologien sind meist für ein bestimmtes, eingeschränktes Fachgebiet (e.g. Medizin) konzipiert und erheben nicht wie die WordNets den Anspruch, den ganzen Wortschatz einer Sprache abzudecken. Als einziger Vertreter der WordNet-Familie verfügt EuroWordNet über eine Ontologie-Komponente. Die Top-Ontologie und die Domänenontologie bestehen aus sprachunabhängigen Konzepten und Wissensstrukturen.

Wortnetze sind ausdrucksstärker als **Taxonomien,** zumal die Einträge nicht nur durch Hierarchiebeziehung, sondern auch durch Relationen wie Antonymie, Meronymie, Implikation oder Kausation miteinander verknüpft werden.

Bleibt also eine mögliche Zuordnung zu den **Thesauri.** Die lexikalischen Einheiten der WordNet-Familie sind innerhalb der syntaktischen Kategorien eher nach ihrer Bedeutung als nach ihrer Wortform geordnet. Diese onomasiologische Anordnung rückt die Wortnetze in der Tat in die Nähe eines Thesaurus. Um die Maschinen-

lesbarkeit zu untermauern, bezeichnet [Kunze 2005b, S. 161] GermaNet als einen Online-Thesaurus.

3.7 Kritik an der WordNet-Familie

3.7.1 Tennis-Problem

In den WordNets sind Konzepte durch verschiedene lexikalisch-semantische Relationen miteinander verknüpft. Dabei wurde die Diskurs-Semantik und somit mögliche syntagmatische Umgebungen vernachlässigt. Fragt jemand nach dem spezialisierten Vokabular für Tennis, hilft ihm WordNet nicht weiter, da es an topikalischer Information fehlt. Solche Angaben sind jedoch in einem Thesaurus zu finden. Motiviert durch dieses Beispiel spricht [Fellbaum 1998b, S. 10] vom so genannten Tennis-Problem. Semantisch nahe stehende Wörter wie *racquet*, *ball* und *net* sind alle Teil des Konzepts *court game,* stehen in WordNet aber nicht in Beziehung zueinander.

WordNet hat mit der kürzlich erfolgten Einführung der Domänen-Relation einen Schritt Richtung Kodierung topikalischer Information gemacht. Das Synset *tennis* ist beispielsweise durch diese Relation mit *break* oder *match point* verknüpft. Die Domänen-Relation ist aber zu wenig systematisch erfasst worden, so dass beispielsweise *racquet* immer noch fehlt.

In den Dokumentationen zu GermaNet wird nie explizit auf das Tennis-Problem eingegangen. Bewusst oder unbewusst wird jedoch in Ansätzen eine Lösung des Problems vorgeschlagen. Unter dem künstlichen Konzept *?Computerterm* werden durch Hyponymie computerbezogene Konzepte wie *Datenstruktur, Zeichenkette, ?Computersprache* etc. verknüpft. In einer solchen Struktur liegt aber nur dieses eine semantische Feld vor. Laut [Kunze und Wagner 1999, S. 14] entstand die Computerterm-Klassifikation im Kontext von Anforderungen einer spezifischen sprachverarbeitenden Anwendung.

EuroWordNet behandelt das Tennis-Problem durch Hinzufügen von Domänenlabels, geht aber systematischer vor als WordNet. Konzepte wie *Tennisschuh* und *Tennisschläger* werden durch ein solches Label assoziiert [Gilarranz et al. 1997, S. 51].

3.7.2 Unvollständigkeit

Zu Recht bemängelt wird die Unvollständigkeit der WordNets. Einerseits umfasst GermaNet erst rund 50 000 Synsets, andererseits sind viele Synsets nur ungenügend durch lexikalisch-semantische Relationen mit anderen verknüpft. Das Synset „Computer“ hat nur gerade ein Meronym, nämlich „Speicher“.

WordNet und GermaNet verfügen über wenige wortartenübergreifende Relationen. Allerdings handelt es sich dabei nur um periphere Relationen wie beispielsweise die Kausations-Relation (GermaNet) oder die Domänen-Relation (WordNet). EuroWordNet hingegen enthält keine wortartenseparierte Dateien und kodiert somit verschiedene wortartenübergreifende Relationen.

3.7.3 Sprachbedingte Abhängigkeit von EuroWordNet

[Kunze und Wagner 1999, S. 16] machen darauf aufmerksam, dass der Interlinguale Index von EuroWordNet auf WordNet-Konzepten aufgebaut ist. Dadurch ist der ILI prinzipiell durch die englische Sprache beeinflusst, obwohl er eigentlich für alle Sprachen anwendbar sein soll. Aus diesem Grund ist vorgesehen, den ILI durch Konzepte zu erweitern, die in mindestens zwei EuroWordNet-Sprachen grundlegend sind.

3.7.4 Vernachlässigung systematischer Polysemie

[Pustejovsky et al. 1997] kritisieren, dass die WordNets Regularitäten zwischen Bedeutungen ignorieren. [Lemnitzer und Kunze 2003, S. 152] schlagen vor, systematisch polyseme Wortfelder wie *Baum-Holzart* genauer zu analysieren. Es soll geprüft werden, ob generische Regeln formuliert werden können, auf welcher hierarchischen Ebene sie angewendet werden und wie mit Ausnahmen umgegangen werden könnte.

4 Weitere Wortnetze

In diesem Kapitel werden weitere lexikalisch-semantische Netze vorgestellt. Die getroffene Auswahl berücksichtigt die grossen, maschinenlesbaren Wortnetze der FrameNet-Familie, als Klassiker unter den Thesauri den Roget's Thesaurus und weitere deutsche Wortnetze. Daneben bietet die Präsentation zweier englischer Wortnetze, welche nicht nur Wortinformation, sondern auch Weltwissen erfasst haben, eine etwas erweiterte Sicht auf das Themengebiet. Für jede vorgestellte Ressource wird am Schluss des Kapitels ein Vergleich mit dem entsprechenden Vertreter der WordNet-Familie gezogen.

4.1 Die FrameNet-Familie

4.1.1 FrameNet

Das Berkeley FrameNet-Projekt (im Folgenden nur FrameNet) unter der Leitung von Charles Fillmore befasst sich seit 1997 mit dem Aufbau einer maschinenlesbaren lexikalischen Ressource für die englische Sprache [FrameNet 2006]. FrameNet, das es schon in der dritten Version gibt, ist für die Forschergemeinschaft frei verfügbar.

Basierend auf der Frame-Semantik werden Frames, die Informationen über eine Alltagssituation mit allen beteiligten Teilnehmern und Ereignissen enthalten, die beteiligten Frame-Elemente und die Relationen zwischen den Frames definiert. In einem zweiten Schritt erfolgt die Extraktion von Beispielsätzen aus grossen Korpora wie aus dem *British National Corpus*. Spezifisch amerikanische Ausdrücke und weitere Korpusevidenzen werden durch Hinzunahme amerikanischer Korpora ermittelt [Ruppenhofer et al. 2005, S. 7]. Mit der Annotierung von Beispielsätzen sollen typische Beispiele aller möglichen Kombinationen von Frame-Elementen erfasst werden. Die Wörterbuchdefinitionen wurden entweder von Wörterbüchern übernommen oder von Hand durch das FrameNet-Team erstellt [Ruppenhofer et al. 2005, S. 7].

Auf der Projektwebseite wird die aktuelle Abdeckung von FrameNet mit 625 Frames und über 8900 lexikalischen Einheiten angegeben [FrameNet 2006]. Rund

135 000 Sätze wurden seit Beginn des Projektes annotiert. Alle mit Wörterbuchdefinitionen versehenen lexikalischen Einheiten, Frames, Frame-Elemente und die Beziehungen zwischen Frames werden in einer lexikalischen Datenbank gespeichert [Fillmore et al. 2003, S. 240].

[Ruppenhofer et al. 2005, S. 93–96] listen die verschiedenen Relationen zwischen Frames auf. Abbildung 7 zeigt eine Visualisierung des commerce-Frame im FrameGrapher-Tool des FrameNet-Projekts [FrameGrapher 2006]. Die ***Subframe*-Relation** (in der Abbildung durch blaue Pfeile dargestellt) verknüpft einen komplexen Frame mit einem untergeordneten Frame, wie dies in der Abbildung zwischen Commercial_transaction und Commerce_goods-transfer der Fall ist. Frame-Elemente des übergeordneten Frame können mit denjenigen des Subframe übereinstimmen, müssen aber nicht.

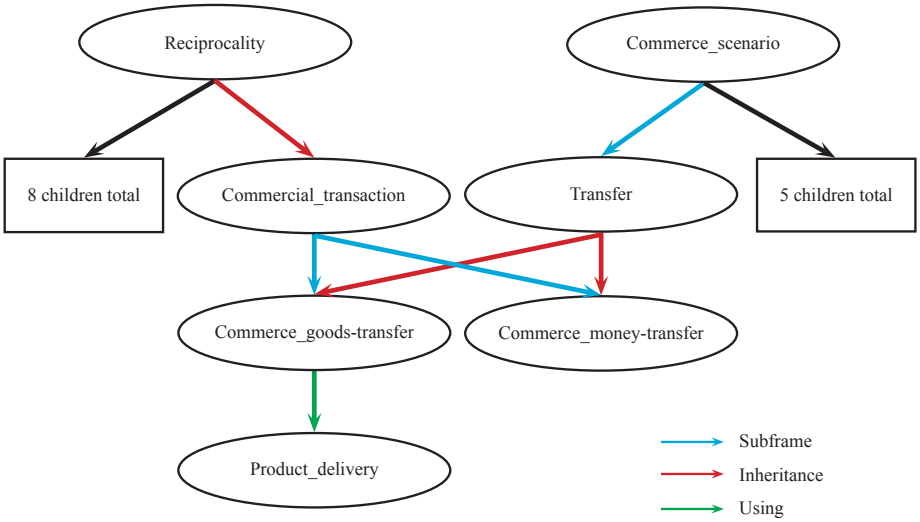

Abbildung 7: Commerce-Frame in der Visualisierung im [FrameGrapher 2006]

Eine ähnliche Relation ist die ***Inheritance*-Relation.** Es handelt sich um eine Beziehung zwischen zwei Frames, wovon der erste eine spezifische Ausprägung des übergeordneten zweiten Frame ist. Alle Frame-Elemente und Subframes des übergeordneten Frame haben gleiche oder spezifischere Entsprechungen im untergeordneten Frame. In der Abbildung 7 wird durch die roten Pfeile dargestellt, dass der Commerce_goods-transfer-Frame vom Transfer-Frame erbt.

Weitere Relationen umfassen unter anderem die ***See-Also*-Relation** zwischen ähnlichen Frames und die ***Using*-Relation,** bei der ein spezifischer Frame eine spezielle Betrachtungsweise eines anderen liefert. Die letztgenannte Relation verbindet die Frames Commerce_goods-transfer und Product_delivery durch einen grünen Pfeil.

4.1.2 SALSA und weitere FrameNets

SALSA steht für *SAarbrücken Lexical Semantics Annotation and Acquisition Project* und wird an der Universität Saarbrücken unter der Leitung von Manfred Pinkal und Annette Frank entwickelt [SALSA 2006]. Ein erster Release ist Mitte dieses Jahres (2006) geplant. Die Daten sollen für den akademischen Gebrauch kostenlos erhältlich sein.

In Anlehnung an FrameNet basiert auch SALSA auf der Frame-Semantik und hat zum Ziel, das syntaktisch annotierte, 1,5 Millionen Token umfassende TIGER-Korpus mit Frames zu annotieren und somit den Mangel an semantisch annotierten Korpora für nichtenglische Sprachen zu beheben. Laut [Burchardt et al. 2006, S. 2] soll SALSA über 500 deutsche Verbprädikate in insgesamt 20 000 annotierten Sätzen abdecken. Im Gegensatz zu FrameNet verfolgt SALSA also einen korpusbasierten Ansatz. Die SALSA-Annotierer betrachteten ein Prädikat nach dem anderen, und nicht wie im FrameNet-Projekt einen Frame nach dem anderen [Burchardt et al. 2006, S. 3].

Abbildung 8 zeigt einen einfachen annotierten Beispielsatz. Das Verb „antworten“ ruft den Frame Communication_response mit den Frame-Elementen Message und Speaker hervor.

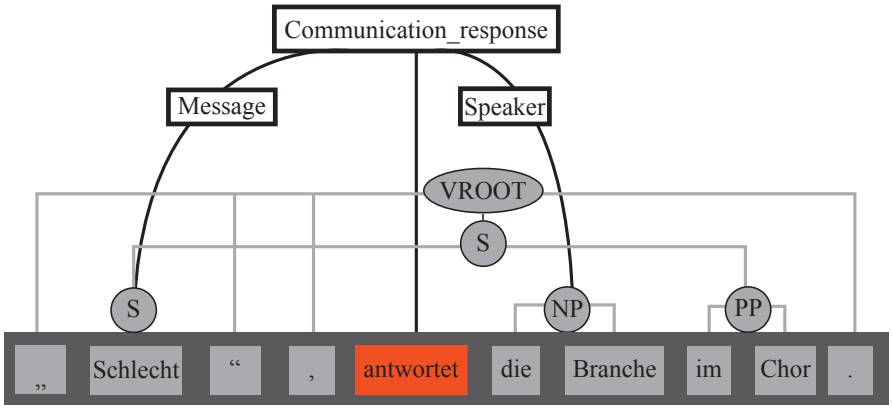

Abbildung 8: SALSA-Annotationsbeispiel des Frame Communication_response nach [Burchardt et al. 2006, S. 2]

Frames und Frame-Elemente wurden im Rahmen des FrameNet-Projekts in erster Linie für das Englische identifiziert. Der Frage, ob sie auch auf andere Sprachen anwendbar sind, konnte mit dem SALSA-Projekt erstmals nachgegangen werden. [Burchardt et al. 2006, S. 3] halten fest, dass ungefähr ein Drittel der im TIGER-Korpus angetroffenen Prädikatslesarten durch FrameNet nicht abgedeckt waren, so

dass fehlende Frames oder Frame-Elemente ergänzt werden mussten. Ein weiteres Problem werfen verschiedene Lexikalisierungsmuster der beiden Sprachen auf. Ein Beispiel ist das deutsche Verb „fahren“, das im Gegensatz zum Englischen nicht zwischen der Nutzung eines Fahrzeuges als Fahrer oder als Passagier unterscheidet. Im Englischen gibt es zwei Wörter, nämlich *ride* im Frame Ride_vehicle und *drive* im Frame Operate_vehicle [Burchardt et al. 2006, S. 3].

Partner von SALSA ist das Projekt **German FrameNet** der Universität Texas unter der Leitung von Hans C. Boas [German-FrameNet 2006]. Im Unterschied zu SALSA plant das German FrameNet, ein ca. 100-mal grösseres Korpus zu parsen, um eine bessere Abdeckung von Beispielen zu erreichen [Boas 2006, persönliche Kommunikation].

Neben dem Berkeley FrameNet und der deutschen Version bestehen auch FrameNets für weitere Sprachen. Das **Spanish FrameNet** der Autonomen Universität Barcelona orientiert sich im Aufbau am englischen FrameNet und soll laut Angaben auf der Projekt-Webseite wie SALSA ab Mitte 2006 verfügbar sein [Spanish-FrameNet 2006]. Es wird mindestens 1000 lexikalische Einheiten der Wortarten Verb, Nomen und Adjektiv umfassen.

Ein weiteres FrameNet-Projekt entsteht an der Keio Universität in Tokio für die japanische Sprache [Japanese-FrameNet 2006]. Entwickler des **Japanese FrameNet** befassen sich momentan mit der Analyse japanischer Bewegungsverben anhand eines morphologisch und syntaktisch annotierten Korpus [Ohara et al. 2004, S. 9]. Im Verlauf dieses Projekts wird auch ersichtlich, inwiefern die typologisch gesehen mit dem Englischen nicht verwandte japanische Sprache mit dem Frame-Semantik-Ansatz in der gleichen Art wie Englisch, Deutsch oder Spanisch beschrieben werden kann [Ohara et al. 2004, S. 10].

Nach Angaben der FrameNet-Webseite wird zudem der Aufbau eines **sprachübergreifenden FrameNet** diskutiert [FrameNet 2006]. Der Plan von [Boas 2005] strebt an, eine multilinguale lexikalische Datenbank mit semantischen Frames zu erstellen. Dazu werden von der FrameNet-Datenbank alle sprachspezifischen Komponenten (lexikalische Einheiten, Beispielsätze etc.) beseitigt, so dass nur noch Informationen zu Frames, Frame-Relationen und Frame-Elementen zurückbleiben [Boas 2005, S. 454f.]. Danach kann die Datenbank mit lexikalischen Angaben nichtenglischer Sprachen ergänzt werden. Mithilfe von Wörterbüchern und parallelen Korpora wird nach lexikalischen Einheiten gesucht, die denselben Frame hervorrufen. Dazu werden die Resultate der FrameNet-Projekte anderer Sprachen einbezogen. [Boas 2005, S. 446] geht auch auf die auftretenden Probleme im Bereich von abweichenden Polysemiestrukturen, Valenzen und Lexikalisierungsmustern ein, die sich

bei der Verlinkung paralleler FrameNet-Segmente unweigerlich ergeben und durch die das ganze Unterfangen sehr ressourcenintensiv wird.

4.2 Roget's Thesaurus

Der wohl bekannteste und älteste Thesaurus für die englische Sprache ist der Roget's Thesaurus [Rogets 1852]. Rogets Inspiration für den Aufbau eines Thesaurus war die Pflanzentaxonomie [Jarmasz 2003, S. 8], bald aber klassifizierte er weitere Wörter und Phrasen methodologisch und zwar geordnet nach den Konzepten, die sie ausdrücken.

Alle Einträge werden einer von sechs Klassen zugeordnet, welche die semantischen Felder Abstract relations, Space, Matter, Intellect, Volition, und Emotion, Religion and Morality abdecken. Die Klassen werden in 39 Sektionen unterteilt, weiter in Subsektionen, Kopfgruppen und schliesslich in rund 1000 Köpfe. Jeder Kopf ist nach Wortart in Paragraphen aufgetrennt. Semantisch verwandte Wörter werden in durch Semikolon separierten Gruppen aufgelistet. Das Klassifikationssystem von Roget's Thesaurus scheint sehr robust zu sein, denn in den 150 Jahren seit der Erstveröffentlichung wurde selten etwas daran geändert.

Die Abbildung 9 zeigt, wie der Kopf *wonder* in der Klassifikation von Roget's Thesaurus eingeordnet wird und listet für die Wortarten Nomen (N.) und Verb (Vb.) jeweils den ersten Paragraphen der Semikolon-Gruppen auf.

Class six: Emotion, religion and morality
Section two: Personal emotion
Sub-section: Contemplative
Head Group: 864 Wonder — 865 Lack of wonder
Head: 864 Wonder
N. wonder, state of wonder, wonderment, raptness; admiration, hero worship, 887 *love*; awe, fascination; cry of wonder, gasp of admiration, whistle, wolf wolf, exclamation, exclamation mark; shocked silence, 399 *silence*; open mouth, popping eyes, eyes on stalks; shock, surprise, surprisal, 508 *lack of expectation*; astonishment, astoundment, amazement; stupor, stupefaction; bewilderment, bafflement, 474 *uncertainty*; consternation, 854 *fear*.
...
Vb. wonder, marvel, admire, whistle; hold one's breath, gasp, gasp with admiration; hero-worship, 887 *love*; stare, gaze and gaze, goggle at, gawk, open one's eyes wide, rub one's eyes, not believe one's eyes; gape, gawp, open one's mouth, stand in amazement, look aghast, 508 *not expect*; be awestruck, be overwhelmed, 854 *fear*; have no words to express, not know what to say, be reduced to silence, be struck dumb, 399 *be silent*.
...

Abbildung 9: Das Verb *wonder* in Roget's Thesaurus nach [Jarmasz 2003, S. 19]

Der Roget's Thesaurus umfasst je nach Ausgabe zwischen 200 000 und 300 000 Wörter in rund 65 000 Semikolon-Gruppen. Nur zwei semantische Relationen zwischen Köpfen oder Semikolon-Gruppen sind explizit kodiert. Die ***Cross-Reference*-Relation** verweist in Abbildung 9 beispielsweise vom letzten Wort der zweiten Semikolon-Gruppe (*love*) durch Angabe der Nummer 887 auf den entsprechenden Kopf. Die ***See*-Relation** weist auf einen anderen Paragraphen im selben Kopf hin. So können allgemeine Paragraphen auf solche spezifischer Ausprägung verweisen [Jarmasz 2003, S. 25]. Weitere Relationen sind in Roget's Thesaurus implizit enthalten. **Antonymie** kann beispielsweise zwischen Kopfpaaren auftreten, die den negativen oder positiven Aspekt einer Idee verkörpern, wie es in Abbildung 9 zwischen den Kopfgruppen *Wonder* und *Lack of wonder* der Fall ist.

Der Roget's Thesaurus wurde in mehr als hundert verschiedenen Ausgaben publiziert und in den 1950er Jahren erstmals in ein maschinenlesbares Format gebracht [Jarmasz 2003, S. 10]. Diese elektronischen Versionen von Roget's Thesaurus sind laut [Hirst und St-Onge 1998, S. 330] für computerlinguistische Applikationen nicht brauchbar, da der für Algorithmen unerlässliche Index nicht enthalten war. Weitere Bestrebungen, Roget's Thesaurus maschinenlesbar zu machen, scheiterten daran, dass ältere verfügbare Versionen den Nachteil hatten, dass der notwendige Wortschatz zum Verarbeiten zeitgemässer Texte fehlte, und neuere Versionen von Roget's Thesaurus nicht lizenziert werden konnten.

Diesen Mangel wollten [Jarmasz und Szpakowicz 2001] beseitigen. Sie beschreiben den Aufbau einer elektronischen lexikalischen Datenbank, die Rogets Klassifikationssystem und andere Aspekte der ursprünglichen Ressource in eine elektronische Version bringen sollte. Das Vorgehen wurde später bei [Jarmasz 2003] detailliert beschrieben. Die wichtigste Aufgabe dabei war die Ausfindung und Benennung der impliziten semantischen Relationen [Jarmasz und Szpakowicz 2001, S. 7f.].

4.3 Weitere deutsche Wortnetze

Verschiedene deutsche Universitäten und Forschungsinstitute arbeiten am Aufbau lexikalisch-semantischer Ressourcen. Im Rahmen dieser Arbeit werden drei weitere Projekte vorgestellt. Es handelt sich dabei um den OpenThesaurus von Daniel Naber, HaGenLex der FernUniversität Hagen und LexikoNet der Berlin-Brandenburger Akademie der Wissenschaften.

4.3.1 OpenThesaurus

OpenThesaurus ist ein frei verfügbares deutsches Wortnetz, das 2003 mit der Motivation entstand, es in OpenOffice als Thesaurus verwenden zu können [Naber 2005]. Das Wortnetz darf kostenlos heruntergeladen und in akademischen oder kommerziellen Anwendungen eingesetzt werden. OpenThesaurus kann auf der Projekt-Webseite [OpenThesaurus 2006] über ein Webinterface betrachtet oder editiert werden.

Anfänglich wurden frei zugängliche elektronische Wörterbücher eingelesen und in eine Datenbank importiert [Naber 2005, S. 423]. An der Weiterarbeit sind nicht nur Linguisten beteiligt, sondern OpenThesaurus steht allen zur Mitarbeit offen, so dass [OpenThesaurus 2006] durch die Community auf aktuell 14 645 Synonymgruppen mit insgesamt 36 800 lexikalischen Einheiten angewachsen ist[21].

OpenThesaurus erfasst Synonymgruppen, die einem von 25 Fachgebieten (Medizin, Botanik etc.) zugeordnet werden können. Ebenfalls optional ergänzen Kennzeichnungen wie *umgangssprachlich, derb, vulgär, fachsprachlich* oder regionale Informationen (e.g. *österr.*) die Synonymgruppen. Seit neustem steht der OpenThesaurus auch in einer Schweizer Version zum Download bereit.

Abbildung 10: Suchergebnis des [OpenThesaurus 2006] für die Anfrage „Fahrrad“

Wie OpenThesaurus Suchresultate präsentiert und zum Editieren bereitstellt, stellt Abbildung 10 dar. Aus der Synonymgruppe, die „Bike“, „Drahtesel“, „Fahrrad“ etc. enthält, können Einträge gelöscht oder neue hinzugefügt werden. Der unterste graue

[21] Stand der Daten: 30.4.2006.

Kasten beinhaltet Information zur Verknüpfung mit anderen Synonymgruppen. Hier betrifft es die Synonymgruppe („Fahrzeug, Gefährt, Vehikel, ...“), der die ursprüngliche Synonymgruppe untergeordnet wird.

Neben **Synonymie** erfasst OpenThesaurus also auch **Ober- und Unterbegriffs-Relationen.** Jede Synonymgruppe kann genau einer Synonymgruppe als Hyponym zugeordnet werden. Sieben von WordNet übernommene Unique Beginners der Nomen-Datei wurden unter der künstlichen Synonymgruppe „Irgendetwas“ eingeordnet. Die Vermeidung von multipler Vererbung und Verwendung nur eines Unique Beginners führt zu einer echten Hierarchie. [Naber 2005, S. 432] begründet die im Vergleich zu anderen maschinenlesbaren Wortnetzen geringe Anzahl berücksichtigter Relationen damit, dass ein offenes Wortnetz einfach gehalten werden muss, um die Benutzer und Mitarbeiter nicht zu überfordern und um die Fehlerrate von Nutzerbeiträgen gering zu halten. Zudem sei Synonymie zentral für die wichtigste Anwendung, den Einsatz in OpenOffice.

Die Erfassung oder Abänderung von Beiträgen erfordert eine Anmeldung. Neue Einträge und Änderungen werden regelmässig von einem Administrator kontrolliert. Die Community wird ebenfalls aufgefordert, sich an der Qualitätskontrolle zu beteiligen, indem auf der Startseite eine selten aufgerufene Synonymgruppe verlinkt wird.

Zusätzlich verfügt OpenThesaurus auch über ein Vollformenlexikon mit 380 000 Einträgen [Naber 2005, S. 430]. Die Suche nach der Pluralform „Fahrräder“ führt zwar nicht gleich zur gewünschten Synonymgruppe, „Fahrrad“ wird aber als grammatisch verwandtes Wort aufgeführt und mit einem Link hinterlegt, der die Benutzerin schnell zum gewünschten Eintrag führt.

Bisher existieren OpenThesaurus-Installationen für Spanisch, Polnisch und Slowakisch. Die einzelnen Projekte sind allerdings unabhängig voneinander. Im Gegensatz zu GermaNet/EuroWordNet besteht also zwischen den Synonymgruppen der verschiedenen Sprachen keine Verknüpfung [Naber 2005, S. 433].

4.3.2 HaGenLex

HaGenLex ist ein Akronym für *HAgen GErmaN LEXicon*. Es wird auf der Projektseite als domänenunabhängiges Computerlexikon fürs Deutsche beschrieben, das seit 1996 am Lehrgebiet für Intelligente Informations- und Kommunikationssysteme in Hagen entwickelt wird [Osswald 2006].

HaGenLex wurde vorwiegend auf der Basis von Frequenzlisten und Wörterbüchern von Hand erstellt. Im Juli 2005 umfasste HaGenLex rund 13 000 Nomen, 7000 Ver-

ben, 3200 Adjektive und 600 Adverbien, alle mit detaillierter morphosyntaktischer und semantischer Information versehen [Osswald 2006].

HaGenLex basiert auf dem Multinet-Paradigma von [Helbig 2001], einem Formalismus zur Darstellung der Semantik natürlicher Sprache mittels mehrschichtiger, erweiterter semantischer Netze. Multinet umfasst eine Hierarchie von 45 ontologischen Sorten wie zum Beispiel Objekt, Eigenschaft und mehr als 100 semantische Relationen und Funktionen wie beispielsweise **Synonymie.** Dazu werden 16 binäre semantische Merkmale wie „belebt“ oder „menschlich“ verwendet, um Selektionsbeschränkungen angeben zu können. Verben verfügen über Angaben zu Subkategorisierung in Form von Kasusrahmen [Osswald 2006].

Thematische Rolle	AGT	OBJ	MCONT
Selektionsbeschränkung	LEGPER+	LEGPER+	
Kasusrahmen	NP[nom]	NP[acc]	PP[*über*, acc]
Realisierung	obligatorisch	optional	optional

Tabelle 4: Angaben zum Verb „informieren“ in HaGenLex basierend auf [Osswald 2006]

Tabelle 4 skizziert die Angaben zum Verb „informieren“. Es denotiert eine nichtmentale Handlung mit den drei Partizipanten Agens (AGT), Objekt (OBJ) und mentalem Inhalt (MCONT)[22]. Die zweite Zeile legt fest, dass als Agens und Objekt nur natürliche oder juristische Personen (LEGPER) zulässig sind. In der dritten Zeile werden die Kasusrahmen aufgeführt und in der vierten wird durch den Eintrag „obligatorisch“ klar, dass von den drei Partizipanten nur der erste syntaktisch realisiert sein muss.

4.3.3 LexikoNet

LexikoNet ist ein Wortnetz deutscher Nomen, an dem die Berlin-Brandenburger Akademie der Wissenschaften seit 2000 arbeitet [LexikoNet 2006]. [Geyken und Schrader 2006, S. 1] bezeichnen LexikoNet als Nomen-Ontologie, dessen Daten zurzeit vor allem als Filter für lexikografische Extraktionsaufgaben im DWDS-Projekt (*Digitales Wörterbuch der deutschen Sprache*) eingesetzt werden. Die LexikoNet-Daten können online abgefragt [LexikoNet 2006][23] aber nicht kostenlos heruntergeladen werden.

[22] Mit mentalem Inhalt wird bezeichnet, *worüber* man informiert.

[23] `Wortinformation` anklicken und den gewünschten Eintrag ins Suchfeld eingeben. Die Trefferanzeige beinhaltet Belegstellen aus Korpora und Informationen zu Kollokationen.

Die Einträge stammen aus dem *Wörterbuch der Gegenwartssprache* und wurden mit weiteren Text- und Wörterbuchressourcen verknüpft [Geyken und Schrader 2006, S. 1]. Insgesamt sind rund 90 000 lexikalische Einheiten erfasst, wobei es sich bei rund der Hälfte davon um Berufsbezeichnungen handelt.

LexikoNet basiert auf einer Konzepthierarchie von mehr als 1200 Konzeptknoten [Geyken und Schrader 2006, S. 1]. Die Hierarchie beginnt mit den Konzepten *Konkrete Nomen* und *Abstrakte Nomen,* die dann weiter unterteilt werden. Zentrale Relation ist somit die **Hyponymie/Hyperonymie,** wobei Mehrfachvererbung möglich ist. In LexikoNet werden verschiedene Typen von Hyponymie kodiert. So erfolgt eine Trennung von Typ und Rolle („Esel“ entweder als Hyponym von „Huftier“ oder „Arbeitstier“). Auch Instanzen und generische Begriffe werden unterschieden. „Moschee“ als generischer Begriff wird folglich anders behandelt als die Instanz „Al-Aqsa-Moschee“.

In der Online-Abfrage ist die Darstellung der semantischen Relationen etwas unbefriedigend gelöst worden. Verschiedene Lesarten eines polysemen Lexems werden zwar aufgeführt, doch Synonyme beider Lesarten durchmischt aufgelistet.

4.4 Weitere englische Wortnetze

4.4.1 Cyc und OpenCyc

Das Cyc-Projekt[24] befasst sich seit 1984 mit der Erfassung eines breiten Spektrums von Weltwissen. Die Motivation der Initianten [Lenat und Guha 1990] war, dass Menschen beim Lesen von Texten und beim Verstehen von Gesprächen auf Wissen zurückgreifen, das in den Äusserungen selbst gar nicht vorhanden ist. Kommunikation ist erst durch die Kombination von explizit in gesprochenen oder geschriebenen Texten vorhandenem Wissen und den impliziten Weltwissensbeständen möglich. Viele sprachverstehende Systeme scheitern denn auch an ihrer zu kleinen Wissensbasis.

Die Einträge und Aussagen von Cyc werden mit der Sprache CycL beschrieben, mit der Aussagen über die Welt in Form von prädikatenlogischen Ausdrücken erster und höherer Stufe gemacht werden können. Die Cyc-Wissensbasis umfasst mehr als 2,2 Millionen Aussagen (Fakten und Regeln), die über 250 000 Begriffe beschreiben [Matuszek et al. 2006, S. 1].

[24] Der Name „Cyc“ ist von en/cyc/lopedia abgeleitet.

Da Cyc zwischen Klassen und Instanzen und zwischen Typen und Rollen unterscheidet, handelt es sich um eine Ontologie, die aus drei Teilen besteht. Die obere Ontologie **Upper Cyc** ist die allgemeinste Ebene. Sie umfasst breite, abstrakte Konzepte und beschreibt ihre Inhalte und Beschaffenheit. Upper-Cyc-Konzepte sind durch verschiedene Relationen miteinander verbunden und Mehrfachvererbung ist erlaubt. Ein Ausschnitt der Komponenten von Upper Cyc ist in der Online-Cyc-Dokumentation zu finden [Upper-Ontology 2006]. Sowohl die mittlere als auch die untere Ontologie sind weniger abstrakt und decken spezifische Domänen (e.g. Chemie) ab [Matuszek et al. 2006, S. 3]. Selektionsbeschränkungen können beschrieben werden, indem Argumente für bestimmte Klassen eingeschränkt werden.

1994 wurde die Firma Cycorp gegründet, welche die Entwicklung von Cyc übernahm. Die Cyc-Daten sind nur zu einem kleinen Teil frei erhältlich. Daneben wird auf das Open-Source-Projekt **OpenCyc** hingewiesen, das eine Ontologie mit 47 000 Konzepten und 306 000 Aussagen über diese Konzepte enthält und in erster Linie eine taxonomische Projektion von Cyc ist [Matuszek et al. 2006, S. 1]. Klassen werden in einer Hierarchie eingeordnet. Im [OpenCyc-Browser 2006] können die Wissensbestände betrachtet, neue Einträge erfasst oder Aussagen über bestehende Einträge assertiert werden.

Collection : Bicycle

GAF Arg : 1

Mt : UniversalVocabularyMt
isa : ArtifactTypeByGenericCategory ExistingObjectType

Mt : ProductGMt
isa : ManufacturedGoodsType

Mt : UniversalVocabularyMt
quotedIsa : CycLConstant

Mt : BookkeepingMt
quotedIsa : CycLConstant

Mt : UniversalVocabularyMt
genls : Device-UserPowered Device-UserControlled MountableTransporter WheeledTransportationDevice
disjointWith : Tricycle Unicycle
comment : "Bicycle is a specialization of WheeledTransportationDevice, MountableTransporter, Device-UserPowered, and Device-UserControlled. Each instance of Bicycle is a personal wheeled transport device powered by human peddling. Most bicycles have two main wheels arranged in-line, but some tandem bicycles have more than two wheels."

Mt : ProductUsageMt
deviceTypesCommonlyUsedTogether : BicycleHelmet

Abbildung 11: Ausschnitt aus dem OpenCyc-Eintrag *bicycle,* visualisiert im [OpenCyc-Browser 2006]

Abbildung 11 zeigt einen Teil des Eintrages für *bicycle*. Der zweitunterste Abschnitt *„UniversalVocabularyMt“* ist besonders interessant, denn hier ist beispielsweise festgehalten, dass Fahrräder Transportmittel sind, die aus Körperkraft angetrieben wer-

den. Im untersten Abschnitt *„ProductUsageMt“* wird noch darauf hingewiesen, dass man Fahrräder gewöhnlicherweise zusammen mit Fahrradhelmen benutzen sollte.

4.4.2 Omega

Wie ihre Vorgängerin Sensus ist auch Omega eine Ontologie, die halbautomatisch aus WordNet, Mikrokosmos (eine weitere lexikalische Ressource) und FrameNet generiert wurde [Philpot et al. 2005, S. 65].

Der untere Teil der WordNet-Hierarchie wird durch viele lexikalische Einheiten abgedeckt, der obere Teil hingegen nicht [Philpot et al. 2005, S. 59]. Mikrokosmos ist ein viel kleineres Netz mit weniger Fokus auf die Lexikalisierung der Konzepte. Die Stärken von Mikrokosmos liegen in der oberen Struktur und der Ausdruckskraft und nicht in der Abdeckung. Omega vereint nun die Vorteile der beiden Ressourcen und gliedert die rund 120 000 Konzepte in einer klar verständlichen Organisation.

Das so genannte *New Upper Model* wurde erstellt, um eine gemeinsame Basis für die beiden Ressourcen zu schaffen [Philpot et al. 2005, S. 60]. Es beginnt mit einem generellen Konzept und ist durch Verzweigungen mit rund 200 Konzepten verknüpft. Omega hat von WordNet die Relationen **Hyponymie/Hyperonymie,** alle drei **Meronymietypen, Pertonymie** und von Mikrokosmos zwei nicht weiter spezifizierte Relationen übernommen, um Konzepte miteinander zu verknüpfen [Philpot et al. 2005, S. 60]. Lexikalische Einheiten sind mit morphologischer und orthografischer Information versehen und werden nicht wie in WordNet einem Synset zugeordnet, sondern es wird ein Bedeutungsobjekt erstellt, das zwischen dem Konzept und der lexikalischen Einheit steht [Philpot et al. 2005, S. 60f.].

Omega kann über das Internet abgefragt werden [Omega-Browser 2006]. Die Darstellung ist jedoch nicht so gelungen wie die WordNet-Onlineabfrage und die Verknüpfung durch semantische Relationen fehlt weitgehend.

4.5 Vergleich der Wortnetze

4.5.1 Die FrameNet- und WordNet-Familien im Vergleich

Wortnetze beider Familien sind lexikalisch-semantische Ressourcen, die den Wortschatz einer bestimmten Sprache erfassen und in einer maschinenlesbaren Form bereitstellen.

Die primäre Repräsentationseinheit der WordNet-Familie ist das Synset, in welchem synonyme lexikalische Einheiten gruppiert werden. WordNet-Relationen verknüpfen entweder lexikalische Einheiten oder Synsets untereinander. Die FrameNet-Familie geht von Frames aus, die das notwendige Hintergrundwissen festhalten, das zum Verständnis der Wortbedeutung beiträgt. Ihnen werden lexikalische Einheiten und Frame-Elemente zugewiesen. Frames werden ebenfalls untereinander verknüpft, allerdings ist das Inventar an möglichen Relationen um einiges kleiner als in WordNet. Gemeinsam sind die zentralen Relationen, um Synsets bzw. Frames in eine Hierarchie einzureihen.

Wie bei den WordNets werden auch die lexikalischen Einträge von FrameNet mit Definitionen versehen, die im Vergleich zu WordNet zum grössten Teil nicht von Lexikografen, sondern aus einem englischen Wörterbuch stammen. Die annotierten Beispielsätze von FrameNet erheben den Anspruch, *alle* kombinatorischen Möglichkeiten einer lexikalischen Einheit abzudecken, während WordNet nur einzelne Beispielsätze erwähnt und somit auch nur über unvollständige syntaktische Informationen verfügt [Boas 2005, S. 468]. Analog zu den Definitionen wurden auch die Beispielsätze aus grossen Korpora extrahiert und nicht durch Linguisten oder Lexikografen erstellt.

Das Vorgehen der FrameNet-Entwickler stellt die möglichst genaue Beschreibung eines Frame in den Mittelpunkt. Häufig vorkommende lexikalische Einheiten werden somit zuerst erfasst. Fehlende Lexikoneinträge oder fehlende Lesarten homonymer oder polysemer Wörter sind dabei die Folge [Burchardt et al. 2006, S. 2]. SALSA wählte ein korpusbasiertes Vorgehen und betrachtet ein Prädikat nach dem anderen. Auch die WordNets werden ergänzt, indem in Korpora Belegstellen für noch nicht in der Datenbank enthaltene Wörter gesucht werden.

WordNet umfasst rund 155 000 lexikalische Einheiten. FrameNet deckt mit knapp 9000 einen viel kleineren Bereich der englischen Sprache ab. Die Betrachtung der deutschen Sprache ergibt ein ähnliches Bild. SALSA beschränkt sich momentan auf 500 Verben, während die Verb-Datei von GermaNet mehr als 9000 Verb-Synsets umfasst. [Ruppenhofer et al. 2005, S. 8] merken an, dass WordNet über eine grössere Abdeckung von hierarchischen Relationen in bestimmten Gebieten wie beispielsweise der Pflanzen- oder Tierwelt verfügt. WordNet kann daher besser für Anwendungen in diesen Gebieten eingesetzt werden.

Die meisten FrameNet- und WordNet-Projekte sind in erster Linie lexikografisch ausgerichtet, obwohl für Definitionen und Beispiele Korpora verwendet werden. Eine realistische Wiedergabe der tatsächlichen sprachlichen Situation könnte mit Projekten wie SALSA erfolgen, da Korpora annotiert werden. So besteht die Möglichkeit,

die Datenbankstruktur um Informationen zu Auftretenshäufigkeiten von lexikalischen Einheiten, Frames oder Frame-Elementen zu erfassen.

Die Frage, ob es sich bei den Vertretern der FrameNet-Familie um Thesauri oder Ontologien handelt, ist nicht abschliessend zu beantworten. Einerseits haben sie Eigenschaften von Thesauri, indem sie einen grossen Teil des Grundwortschatzes nach der Bedeutung ordnen und die Einträge durch Relationen verknüpfen. Zum anderen erheben sie den Anspruch, durch Frame-Konstrukte Weltwissen zu erfassen, was sie in die Nähe von Ontologien rückt.

4.5.2 Roget's Thesaurus und WordNet im Vergleich

WordNet und Roget's Thesaurus gehören zu den grössten lexikalisch-semantischen Ressourcen für die englische Sprache. Roget's Thesaurus ist mit 250 000 Einträgen sogar noch umfassender als WordNet.

WordNet war von Anfang an als elektronische Version vorhanden und wurde dementsprechend konzipiert. Die weitaus ältere Ressource des Roget's Thesaurus war lange Zeit ausschliesslich in gedruckter Form erhältlich und wurde erst vor kurzem für die weitere computerlinguistische Nutzung maschinenlesbar gemacht.

Der Aufbau der beiden Thesauri ist in einigen Bereichen vergleichbar. Die Klassen des Roget's Thesaurus können als Äquivalente zu den Unique Beginners der WordNet-Familie angesehen werden. Semikolon-Gruppen fassen wie die WordNet-Synsets Synonyme und Plesionyme zusammen. Lexikalisch-semantische Relationen behandeln die beiden Wortnetze jedoch verschieden. Während in Roget's Thesaurus nur zwei Relationen explizit gekennzeichnet sind, unterscheidet WordNet rund 15 Relationen. Im Gegensatz zu WordNet verfügt Roget's Thesaurus über viel mehr wortartenübergreifende Relationen, indem Wörter verschiedener Wortarten im selben Kopf aufgeführt werden. Durch diese topikalische Gruppierung fällt das Tennis-Problem (vgl. 3.7.1) nicht mehr ins Gewicht.

4.5.3 Weitere deutsche Wortnetze und GermaNet im Vergleich

Bezüglich der Abdeckung kommt HaGenLex den im Abschnitt 4.3 vorgestellten deutschen Wortnetzen GermaNet am nächsten, weist jedoch nur rund halb so viele Einträge auf. OpenThesaurus ist mit rund 14 000 Synonymgruppen mehr als dreimal kleiner als GermaNet. Viele gebräuchliche Wörter sind noch nicht erfasst. LexikoNet beindruckt mit einer hohen Abdeckung von Nomen, enthält aber keine anderen

Wortarten. Adverbien sind zurzeit nur in HaGenLex enthalten.

Während GermaNet und OpenThesaurus gleichbedeutende lexikalische Einheiten durch Synsets bzw. Synonymgruppen repräsentieren, steht in HaGenLex und LexikoNet eine Repräsentationseinheit für jedes lexikalisierte Konzept. HaGenLex macht im Gegensatz zu GermaNet bislang nahezu keinen Gebrauch von künstlichen Konzepten [Osswald 2003, S. 43]. Für LexikoNet und OpenThesaurus konnten hinsichtlich dieses Punktes keine Belege ausfindig gemacht werden.

Das Inventar an lexikalisch-semantischen Relationen wird in den verschiedenen Wortnetzen sehr unterschiedlich gehalten. OpenThesaurus und LexikoNet berücksichtigen mit Synonymie und Hyponymie/Hyperonymie nur einen kleinen Teil des Relationenbestandes von GermaNet. HaGenLex hingegen kennt viel mehr Relationen als GermaNet. Bislang verfügt nur HaGenLex über Angaben zu Selektionsbeschränkungen. Angaben zu Subkategorisierungsrahmen hingegen kennt auch GermaNet.

Während GermaNet mittels Korpora erweitert wird, zählt OpenThesaurus auf eine aktive Nutzergemeinschaft. Der Download der Daten steht damit auch allen offen, was OpenThesaurus zum einzigen frei verfügbaren deutschen Wortnetz macht [Naber 2005, S. 423]. LexikoNet wird durch halbautomatische Extraktion lexikalisch-semantischer Relationen aus einem elektronischen Wörterbuch ergänzt [Geyken und Schrader 2006, S. 2].

[Geyken und Schrader 2006, S. 2] machen darauf aufmerksam, dass WordNet im Vergleich zu LexikoNet keine Unterscheidung zwischen Rollen- und Typenlesarten vornimmt. In WordNet und auch in GermaNet hat „Esel“ beispielsweise nur die Typenlesart. Ausserdem trennt LexikoNet Instanzen und generische Begriffe. Für [Geyken und Schrader 2006, S. 1] rechtfertigen diese Punkte die Zuordnung von LexikoNet zu den Ontologien. HaGenLex mit seiner Vielfalt an semantischen Informationen und die erwähnten Charakteristika von LexikoNet rücken die beiden Wortnetze in die Nähe von Ontologien. Trotzdem fällt eine unumstrittene Zuordnung entweder zu Ontologien oder zu Thesauri abgesehen von OpenThesaurus, dessen Name schon für die Zuordnung zu den Thesauri spricht, nicht einfach.

4.5.4 Weitere englische Wortnetze und WordNet im Vergleich

Cyc deckt mit 250 000 Begriffen einen rund doppelt so grossen Wortschatz wie WordNet und Omega ab. Alle drei Ressourcen kodieren mehrere semantische Relationen. Während Cyc eigene Relationen definiert hat, sind diejenigen von Omega zu einem

grossen Teil von WordNet übernommen worden.

Der grösste Unterschied zwischen WordNet und Cyc besteht darin, dass Cyc über die Einträge in Form von Regeln und Fakten aussagt und dadurch Weltwissen speichert. Dieser Eigenschaft ist es zuzuschreiben, dass Cyc klar zu den Ontologien zählt. Omega wird ebenfalls als Ontologie beschrieben [Philpot et al. 2005, S. 59], reicht aber nicht an die Ausdrucksmächtigkeit von Cyc heran.

5 Grenzen computerlinguistischer Anwendungen

Dieses Kapitel führt in die wichtigen computerlinguistischen Anwendungsgebiete Information Retrieval, Antwortextraktion, Maschinelle Übersetzung und Rechtschreibekorrektur ein. Bei der Verarbeitung natürlicher Sprache stossen diese Systeme vielfach auf Probleme. Die prominenteste Schwierigkeit stellen Ambiguitäten dar. Nicht nur auf Wortebene, sondern auch auf Satzebene führen ambige Strukturen zu mehreren Lesarten, von denen ein System die beabsichtigte auswählen muss. Die Wahl falscher Lesarten hat beispielsweise im Information Retrieval unbefriedigende Suchresultate oder in der Maschinellen Übersetzung unzureichende Übersetzungsqualität zur Folge. Neben Ambiguitäten werden auch die Problematiken mit anaphorischen Ausdrücken skizziert und Lösungsansätze dafür vorgestellt, die auf Wortnetzen basieren.

5.1 Information-Retrieval-Systeme

Die Verbreitung elektronischer Medien wie dem Internet führte im letzten Jahrzehnt zu einem starken Anstieg der erhältlichen Information, die dank zunehmender Vernetzung auch immer mehr Personen zugänglich wurde. Über die richtige Information zu verfügen ist häufig für die Entscheidungsfindung unerlässlich. Da man unmöglich alles wissen kann, ist das schnelle Auffinden relevanter Information erforderlich und daher das Bedürfnis nach effizienten Information-Retrieval-Systemen gross.

Im Information Retrieval (kurz IR) geht es darum, aus einem umfangreichen Inventar diejenigen natürlichsprachlichen Dokumente aufzufinden, deren Inhalt dem Informationsbedarf eines Benutzers genügen [Salton und McGill 1983, S. 2]. Der Benutzer beschreibt sein Informationsbedürfnis in einer Suchanfrage, die aus einem oder mehreren Wörtern besteht. Das IR-System vergleicht die Anfrage mit dem Index der Dokumentensammlung und liefert dem Benutzer eine Auswahl von relevanten Dokumenten zurück. Die bekannteste Anwendung sind Internet-Suchmaschinen.

Die Performanz von IR-Systemen wird anhand einer Sammlung von Dokumenten getestet. Die Testsammlung, die von der Text REtrieval Conference (kurz TREC) zur Verfügung gestellt wird, besteht aus einer Anzahl Dokumenten, einer Anzahl Anfragen und aus einer Liste mit relevanten Dokumenten hinsichtlich jeder Anfrage [Voorhees 1998, S. 287]. Die üblichen Vergleiche betrachten einerseits die Ausbeute (engl. *recall*) und die Treffergenauigkeit (engl. *precision*) eines IR-Systems. Die beiden Werte messen den Anteil relevanter zurückgelieferter Dokumente an allen relevanten Dokumenten (Ausbeute) und den Anteil zurückgelieferter relevanter Dokumente an allen zurückgelieferten Dokumenten (Treffergenauigkeit). Um die beiden Messwerte zu mitteln, wurde das F-Mass (engl. *f-measure*) eingeführt, das häufig für die Angabe der Performanz von IR-Systemen verwendet wird.

5.1.1 Problemstellung

Das Retrieval-Modell eines IR-Systems legt fest, wie Dokumente und Anfragen repräsentiert und miteinander verglichen werden sollen. In herkömmlichen IR-Systemen, die nach dem **booleschen Modell** funktionieren, werden die Dokumente der Testsammlung erst computerlinguistisch verarbeitet[25] und indexiert [Salton und McGill 1983, S. 25f.]. Im zweiten Schritt werden Anfragen gegen den Index abgeglichen, Dokumente in relevante und irrelevante eingeteilt und nach einem bestimmten Algorithmus gewichtet.

[Smeaton 1999, S. 103f.] spricht von *matching of bags* und führt auch gleich die Problematik an, die dieser Ansatz mit sich bringt: Wie in Kapitel 1 dargelegt, führt eine Suche zu keinen Treffern, wenn im Index lediglich ein Synonym des Suchbegriffs enthalten ist. Nur aufgrund eines Zeichenkettenvergleiches fehlen viele relevante Dokumente. Andererseits sind viele gefundene Dokumente völlig irrelevant, da beispielsweise bei polysemen oder homonymen Anfragetermen nicht nur die gewünschte Lesart, sondern auch die anderen Bedeutungen einbezogen werden. [Kunze und Wagner 2001, S. 239] merken an, dass häufig verwendete Wörter eher mehrdeutig sind als seltene, was dieses Problem noch zentraler erscheinen lässt. Die zwei grossen Probleme, für die IR-Systeme eine Lösung finden müssen, sind also

a) Vermeidung von **fehlenden** Treffern und
b) Vermeidung von **irrelevanten** Treffern.

Die dritte Problematik beschränkt sich auf die Suche in mehrsprachigen Dokumentensammlungen. In bestimmten Fällen ist es sinnvoll, auf Anfragen in der einen Spra-

[25] Darunter versteht man Tokenisierung, morphologische Analyse, Ausfiltern von Stoppwörtern etc.

che auch Dokumente zurückzuliefern, die in einer anderen Sprache verfasst wurden. [Clough und Stevenson 2004, S. 97] illustrieren dies am Anwendungsbeispiel des elektronischen Handels. Wenn der Benutzer Dokumente in einer anderen Sprache zwar verstehen kann, aber nicht derart gewandt ist, eine konkrete Anfrage in der Fremdsprache zu formulieren, könnten Suchresultate in der anderen Sprache dem Benutzerbedarf entgegenkommen.

Wie auf eine Anfrage in verschiedenen Sprachen verfasste Dokumente gefunden werden, ist das zentrale Problem im Gebiet des **sprachübergreifenden IR.** Wenn Anfrageterme mithilfe von Wörterbüchern in die Zielsprache übersetzt und die Entsprechungen dem System übergeben werden, verringern mehrdeutige Anfrageterme in zweierlei Hinsicht die Treffergenauigkeit. Zum einen wird die Anfrage durch irrelevante Entsprechungen erweitert („Kiefer“, ursprünglich in der Baum-Lesart, wird nicht nur durch *pine,* sondern auch durch *jaw* erweitert), zum anderen sind die Entsprechungen selber wieder mehrdeutig und führen zu vielen irrelevanten Dokumenten in der Zielsprache (*pine* steht auch für das Verb „sich sehnen“).

Die grösste Schwierigkeit im IR entsteht durch **lexikalische Ambiguitäten.** Deshalb müssen IR-Systeme in der Lage sein, nicht Wörter, sondern die Bedeutungen der Anfrage- und Indexterme miteinander abzugleichen.

5.1.2 Lösungsansätze

Das Problem fehlender Treffer lösen IR-Systeme, indem semantisch ähnliche Wörter der Anfrage in einem Wortnetz nachgeschlagen und der Suche mitgegeben werden. Für eine solche **Suchraumerweiterung** (engl. *Query Expansion*) muss die Anfrage erst disambiguiert, d.h. die beabsichtigte Lesart ermittelt werden. Ohne Disambiguierung leidet die Ausbeute, da Begriffe der nicht intendierten Lesart im Suchprozess nicht ausgeschlossen werden.

Um irrelevante Treffer zu vermeiden, muss der **Suchraum eingeschränkt** werden. Nach der Disambiguierung weiss zwar das System, dass der Benutzer mit der Suchanfrage „Jaguar“ nach Autos, und nicht nach Tieren sucht, doch werden irrelevante Treffer noch nicht ausgeschlossen, da die Indexterme der Dokumente nicht disambiguiert sind. Statt Wörter erfasst man daher Konzepte im Index, und gleicht die Anfragen dann mit diesem Index ab [Gonzalo et al. 1998a, S. 40]. Im Index eines Textes zu Autos würde nicht mehr einfach „Jaguar“, sondern zusätzlich ein Vermerk auf die richtige Lesart (realisiert beispielsweise durch WordNet-Lesartennummern) stehen. So vergleicht man nicht mehr *Wörter,* sondern *Bedeutungen* miteinander und spricht von **konzeptueller** oder **semantischer Indexierung.**

Um im sprachübergreifenden IR das Problem von lexikalischen Ambiguitäten und somit von fehlenden oder irrelevanten Treffern zu verringern, können **multilinguale lexikalisch-semantische Ressourcen** eingesetzt werden. Wenn das Wort „Kiefer“ in einer Beispielanfrage disambiguiert ist, verbessert sich der Suchprozess gleich in zweifacher Hinsicht. So findet das System zum einen die korrekte Entsprechung in der englischen Sprache (e.g. *pine*) und weiss zudem auch, dass nach „Kiefern“ und nicht nach dem Verb „sich sehnen“ (e.g. *to pine*) gesucht werden soll. Konzeptuelle Indexierung mit sprachunabhängigen Strukturen wie Frames könnten helfen, sprachübergreifende Systeme zu erstellen.

Eine weitere Option bietet ein **alternatives IR-Modell.** Das klassische boolesche Modell wird abgelöst durch ein Modell, welches die Relevanz eines Dokumentes aufgrund der semantischen Ähnlichkeit zwischen den Anfragetermen und den Indextermen des Dokumentes berechnet [Smeaton 1999, S. 106f.]. Je näher in einem Wortnetz sich die Terme stehen, desto ähnlicher sind sie sich. [Gonzalo et al. 1998b, S. 189f.] gehen derselben Idee nach. Sie sprechen vom Mass der konzeptuellen Distanz, welches Aufschluss über die Ähnlichkeit von Anfrage und Dokument gibt. Der Vergleich aufgrund von Ähnlichkeit bzw. Distanz verspricht aber auch nur dann Erfolg, wenn die Wörter in der Anfrage und auch in den Dokumenten disambiguiert sind, da sonst fälschlicherweise eine Ähnlichkeitsbeziehung zwischen „Jaguar“ in der Automarkenlesart und „Gepard“ hergestellt werden könnte.

Grob unterscheidet man also zwei Möglichkeiten, wie Wortnetze im IR eingesetzt werden können. Einerseits bieten sie ein **Inventar an semantischen Konzepten,** mit denen Wortbedeutungen repräsentiert werden können. Sie helfen so, in Dokumenten und Anfragen Lesarten zu unterscheiden oder sie gleich mit Konzepten zu indexieren. Andererseits geben sie durch die **Relationen zwischen den Einträgen** Aufschluss darüber, wie nah sich Konzepte sind. Dies hilft, Anfrage und Dokumente hinsichtlich ihrer Ähnlichkeit abzugleichen.

Alle Lösungsansätze haben gemeinsam, dass sie zuverlässige Disambiguierungsmethoden erfordern. Im IR ist in erster Linie die Auflösung von lexikalischen Ambiguitäten zentral.

5.2 Antwortextraktionssysteme

Bei der Antwortextraktion (engl. *Question Answering,* kurz QA) geht es darum, in einer grossen Dokumentensammlung einen Textausschnitt zu finden, der die Antwort auf eine in natürlicher Sprache formulierte Frage enthält [Harabagiu et al. 2001,

S. 282]. IR-Systeme können diesem Bedürfnis nicht nachkommen, da sie nur eine gewichtete Liste von Dokumenten und keine Antworten zurückliefern.

Ein QA-System muss die Frage analysieren, eine oder mehrere Antworten in einer Textsammlung finden und dem Anwender die Antwort in einem geeigneten Format darlegen. Die Antworten gängiger Systeme liegen dabei in Textausschnitten einer bestimmten Länge vor.

Um Antwortextraktionssysteme vergleichen zu können, wurde an der TREC im Jahre 1999 eine QA-Aufgabe eingeführt [Voorhees und Tice 2000]. Es müssen Faktenfragen beantwortet werden, indem Dokumente des TREC-Korpus konsultiert werden. Die Performanz wird gemessen, indem die Relevanz der ersten fünf Antworten betrachtet wird.

5.2.1 Problemstellung

Wie auch im IR stösst in QA-Systemen die *Bag-matching*-Methode an ihre Grenzen. Wenn sich die Schlüsselwörter der Antwort nicht mit den gleichbedeutenden der Anfrage decken, kann keine zufrieden stellende Antwortextraktion erfolgen und ambige Suchanfragen führen zu einer grossen, ungenauen Treffermenge.

Zusätzlich besteht die Schwierigkeit darin, dass die Anfrage nicht aus Stichwörtern, sondern aus einer natürlichsprachlichen Anfrage besteht. Somit ist nicht nur die Auflösung von lexikalischen, sondern auch von **anderen Ambiguitätstypen** ein Thema. Im Beispiel 11 ist es für den menschlichen Leser eindeutig, dass hier nach jemandem gefragt wird, der einen Hut *auf dem Kopf* trug.

(11) Wer sah den Mann mit dem Hut?

Hier handelt es sich aber um einen mehrdeutigen Satz. Die Ambiguität resultiert aus den verschiedenen Anschlussmöglichkeiten der Präpositionalphrase. Da Computer meist nicht über das Wissen verfügen, dass man *mittels* eines Hutes nicht sehen kann, können sie diese Ambiguität nicht auflösen.

Ein weiteres Problem bei der Antwortextraktion sind **Anaphern**[26]. Anaphorische Ausdrücke können verschieden realisiert sein. Im folgenden Beispiel 12 von [Meyer und Dale 2002, S. 3] handelt es sich bei der fett markierten Anapher um

[26] Eine Anapher ist eine sprachliche Einheit, die für eine im Text vorangegangene sprachliche Einheit steht. Die bereits erwähnte sprachliche Einheit, auf die sich eine Anapher bezieht, wird als Antezedens bezeichnet.

eine so genannte assoziative Anapher, die als definite Nominalphrase realisiert ist und sich auf das Antezedens *tuna* bezieht.

(12) The tuna has no respiratory mechanism to ensure the flow of water over **the gills.**

Assoziative Anaphern sind problematisch, da definite Ausdrücke gebraucht werden, um auf eine Einheit zu referieren, die nicht vorgängig erwähnt wurde, deren Existenz aber mithilfe einer vorgängig erwähnten Einheit hergeleitet werden kann [Meyer und Dale 2002, S. 1].

Auch pronominale anaphorische Ausdrücke wie das fett markierte Pronomen in 13b) von [Vicedo und Ferrández 2000, S. 560] sind problematisch.

(13) a. Question: Who is the village head man of Digha?
b. Answer: **He** is the sarpanch, or village head man of Digha, a hamlet or mud-and-straw huts 10 miles from ...

Wenn ein System die Anapher nicht vorgängig aufgelöst hat, das heisst, für *he* das richtige Antezedens noch nicht gefunden hat, kann die Frage 13a) nicht beantwortet werden.

Eine weitere Art von Pronominalanaphern, die QA-Systeme an ihre Grenzen stossen lässt, wird in Beispiel 14 von [Cristea 2001, S. 16] gezeigt.

(14) If the baby does not thrive on the raw milk, boil it.

Dem menschlichen Leser ist klar, dass man üblicherweise Milch kocht, und nicht Kinder. In der maschinellen Verarbeitung ist die automatische Auflösung solcher Pronominalanaphern keine triviale Aufgabe. Anaphernresolution benötigt in vielen Fällen eine substanzielle Menge an **Weltwissen.** Da ein QA-System den Inhalt einer Anfrage und der Dokumente „verstehen“ muss, ist eine Resolutionskomponente erforderlich, die dieses Problem angeht.

5.2.2 Lösungsansätze

Analog zu den Lösungsansätzen von Abschnitt 5.1.2 erfordern auch QA-Systeme Verfahren zur Suchraumerweiterung, Suchraumeinschränkung durch konzeptuelles Indexieren oder Abgleich aufgrund semantischer Ähnlichkeit zwischen der Anfrage und den Dokumenten.

Anaphern und Antezedenten, die durch Nominalphrasen realisiert sind, stehen häufig durch lexikalisch-semantische Relationen miteinander in Beziehung. In Beispiel 12 werden *tuna* und *the gills* durch die **Meronymie-Relation** verknüpft. Information aus Wortnetzen kann also helfen, das korrekte Antezedens einer NP-Anapher aufzufinden und so in der weiteren Verarbeitung im QA-System Probleme zu vermeiden.

Anaphern wie in Beispiel 13 können durch Einbezug syntaktischen Wissens behandelt werden, sofern zuvor Eigennamenerkennung erfolgte. Nur so kann *he* mit dem richtigen Antezedens in Beziehung gebracht werden.

Da QA-Systeme in vielen Fällen die Anfrage und die Dokumente verstehen müssen, um korrekte Antworten zu liefern, ist nicht nur die Auflösung von lexikalischen Ambiguitäten, sondern von verschiedenen Ambiguitätstypen zentral. **Selektionsbeschränkungen** helfen bei der Resolution pronominaler Anaphern wie in Beispiel 14. Angaben zum Verb *boil* sollten darüber informieren, dass es als Objekt ein Lebensmittel selegiert. So wird erkannt, dass sich das Pronomen *it* auf *milk* bezieht und nicht auf *baby*.

Auch im Beispiel 11 kann die Anschlussambiguität durch semantisches Wissen aufgelöst werden. Selektionsbeschränkungen vom Verb „sehen“ legen fest, dass es durch ein Instrument ergänzt werden kann, welches ein optisches Gerät sein muss. Da „Hut“ kein optisches Gerät ist, wird klar, dass die Präpositionalphrase nicht an das Verb, sondern an das Akkusativobjekt angehängt werden muss. Einige Wortnetze enthalten Angaben zu Selektionsbeschränkungen von Verben, die eingesetzt werden können, um Ambiguitäten und Anaphern aufzulösen.

Wie IR-Systeme verlangen QA-Systeme eine Disambiguierung der Frage und der Dokumentensammlung, um die Textausschnitte mit der korrekten Antwort zu finden. Im Gegensatz zu IR-Systemen sind aber nicht nur lexikalische Ambiguitäten, sondern auch weitere Typen von Ambiguität sowie Anaphern problematisch und erfordern zuverlässige Resolutionsmethoden.

5.3 Maschinelle Übersetzungssysteme

Mit der fortschreitenden Globalisierung wächst der Bedarf an Übersetzungen ständig, denn ohne Übersetzung treten Verständigungsprobleme zwischen verschiedensprachigen Personen auf. Um in internationalen Firmen und Organisationen die grosse Arbeitslast der Übersetzer zu reduzieren, werden maschinelle Übersetzungssysteme eingesetzt.

Die Geschichte der maschinellen Übersetzung begann mit der Idee, man könne übersetzen, indem man einfach die Wörter der Quellsprache durch Wörter der Zielsprache ersetzt [Locke und Booth 1955]. Um die simple **Wort-für-Wort-Übersetzung** zu verbessern, ordnete man ganze Wortkategorien um. Während im Deutschen beispielsweise die Reihenfolge der Nominalphrasen Artikel, Adjektiv und dann Nomen ist, hält sich das Französische weitgehend an die Folge Artikel, Nomen, Adjektiv. Dies Methode der so genannten **direkten Übersetzung** hat neben der unbefriedigenden Übersetzungsqualität auch die vielen notwendigen Regeln zum Nachteil [Hess 2005, S. 63].

Später folgte der **Transfer-Ansatz,** wobei nicht Wortkategorien, sondern ganze Wortfolgen (d.h. verschiedene Varianten von Nominalphrasen) betrachtet wurden. Die Übersetzungsarbeit wird dabei in drei Schritte zerlegt: Als erster Schritt wird die Ausgangssprache analysiert und in eine Syntaxstruktur überführt. Danach folgt der Transfer der Syntaxstruktur und der Lexikon-Lookup, woraus die Syntaxstruktur der Zielsprache resultiert. Schliesslich erfolgt die Synthese der Zielsprache und die Ausgabeaufbereitung [Hess 2005, S. 66–78].

Maschinelle Übersetzungssysteme werden evaluiert, indem die entstandene Übersetzung mit einer Übersetzung eines professionellen Übersetzers verglichen wird. [Papineni et al. 2002] schlagen die automatische Evaluationsmethode BLEU (Bi-Lingual Evaluation Understudy) vor, die auf diesem Konzept aufbaut. Mit BLEU kann man nicht die Qualität jedes beliebigen maschinell übersetzten Textes messen, sondern es dient zur Evaluation, wenn in einem Maschinellen Übersetzungssystem neue Komponenten (beispielsweise Anaphernresolution) eingeführt werden. Als Input dient immer derselbe Text, und der übersetzte Output wird in der Folge mit der professionellen Übersetzung verglichen.

Der Grundstein des Masses von [Papineni et al. 2002] ist die Treffergenauigkeit. Man zählt die Anzahl der Wörter der automatischen Übersetzung, die ebenfalls in der professionellen Übersetzung vorkommen und teilt sie durch die Gesamtanzahl Wörter der automatischen Übersetzung. Die Berücksichtigung von n-Grammen[27] verbessert dieses Mass gegenüber der Verwendung von Unigrammen. Das Mass scheint durchgängig zu funktionieren, denn in mehreren Evaluationsdurchgängen schlossen verschiedene Übersetzungen von professionellen Übersetzern und Übersetzungssystemen immer in der gleichen Reihenfolge ab [Papineni et al. 2002, S. 313f.].

[27] Ein n-Gramm ist eine geordnete Menge mit n Zeichen. Ein Unigramm besteht aus einem, ein Bigramm aus zwei Zeichen. Hier geht es darum, dass nicht einzelne Wörter, sondern Wortfolgen miteinander verglichen werden.

5.3.1 Problemstellung

Obwohl transferbasierte Übersetzungssysteme immer noch das dominierende Paradigma der Maschinellen Übersetzung sind, werden oft fehlerhafte Übersetzungen erzielt. Eine Fehlerquelle sind typologische Unterschiede der Sprachen, wenn beispielsweise die Quellsprache keine Artikel kennt, die Zielsprache hingegen schon (vgl. [Hess 2005, S. 82]).

Die bedeutendste Schwierigkeit Maschineller Übersetzung stellen aber ambige Strukturen dar. Häufig ist aufgrund **lexikalischer Ambiguitäten** nicht klar, welche Lesart beabsichtigt ist, und die falsche wird in die Zielsprache übertragen. Die Übersetzungsqualität nimmt jedoch nicht nur wegen homonymer oder polysemer Lexeme ab. Auch **weitere Ambiguitätstypen** wie Anschlussambiguität erschweren die automatische Übersetzung. Es ist zwar unproblematisch, Sätze wie das Beispiel 11 ins Englische zu übersetzen, da im übersetzten Satz (*He saw the man with the hat.*) derselbe Typ von Ambiguität erhalten bleibt. Die Übersetzung in Sprachen wie das Japanische wäre aber in diesem Fall nicht trivial, da instrumental verwendete Komplemente eine andere grammatikalische Markierung haben als Adjunkte.

Auch **Anaphernresolution** ist in der Maschinellen Übersetzung ein grosses Thema. Für einen Satz wie im Beispiel 15a) von [Hess 2005, S. 84] schlägt ein Übersetzungssystem mindestens zwei Übersetzungsvarianten vor, nämlich b) und c).

(15) a. When John drank the wine in the glass he spilled **it.**
b. Als Hans den Wein im Glas trank, verschüttete er ihn.
c. Als Hans den Wein im Glas trank, verschüttete er es.

Ohne Berücksichtigung von semantischer Information ist die Auflösung dieser fett gedruckten Pronominalanapher nicht möglich und Maschinelle Übersetzungssysteme wählen eventuell die falsche Übersetzungsvariante c) aus. Auch die Übersetzung des Beispiels 14 ins Deutsche kann ein unbeabsichtigtes Ergebnis zur Folge haben.

5.3.2 Lösungsansätze

Unbefriedigende Übersetzungsqualität ist in den meisten Fällen auf fehlende oder ungenügende Ambiguitäts- und Anaphernresolutionsverfahren zurückzuführen.

In erster Linie sollten lexikalische Ambiguitäten aufgelöst werden. Hierzu kann der **Kontext des ambigen Wortes** in der Quellsprache einbezogen werden. Kontextwörter stehen oft in lexikalisch-semantischer Beziehung mit dem ambigen Wort und

können so möglicherweise Aufschluss über die korrekte Lesart des Wortes geben. Die Auflösung von Anschlussambiguitäten erfordert den Einbezug semantischer Information. In vielen Fällen reichen Angaben zu **Selektionsbeschränkungen** von Verben aus, um die korrekte Lesart herauszufinden.

Ebenfalls mit semantischer Information können gewisse Anaphern aufgelöst werden. In den Beispielen 14 bzw. 15 kann durch Berücksichtigung von Typenwissen (Milch ist ein Lebensmittel, bzw. Wein ist ein Getränk) und Selektionsbeschränkungen (kochen tut man üblicherweise nur Lebensmittel, bzw. verschütten kann man nur etwas Flüssiges) das Pronomen *it* mit dem richtigen Antezedens verbunden werden und so korrekt in die Zielsprache übersetzt werden.

Durch das Einführen einer **alternativen Übersetzungsmethode** werden die problematischen Punkte von einer ganz anderen Seite angegangen. Im so genannten Interlingua-Modell wird der disambiguierte Text der Quellsprache in eine sprachunabhängige Repräsentation übersetzt. Aus dieser Darstellung wird der Text der Zielsprache direkt generiert. Bei der Interlingua handelt es sich um eine Art Logiksprache, die von der Ausgangs- und Zielsprache unabhängig ist [Dorr et al. 2004]. Das Interlingua-Modell ist insbesondere attraktiv für die Übersetzung zwischen Sprachen, die in ihrem Vokabular und der Struktur sehr stark divergieren, wie es beispielsweise für Japanisch-Deutsch der Fall ist.

Ein weiterer Vorteil dieses Verfahrens ist, dass es nicht für jedes Sprachpaar ein neues Übersetzungsprogramm braucht, sondern dass es für jede Sprache nur je ein Analyse- und ein Syntheseprogramm benötigt und neue Sprachen nur die Äquivalenzrelationen zu der Interlingua in Betracht ziehen müssen [Vossen 2004a, S. 165]. Obwohl der **Einsatz einer Interlingua** in der Maschinellen Übersetzung vielversprechend ist, ist es schwierig, eine Interlingua zu erstellen, die alle Sprachen abdeckt. Momentan gibt es noch keine allgemein akzeptierte Interlingua-Repräsentation [Reeder et al. 2004, S. 1]. Auch [Hess 2005, S. 89] macht darauf aufmerksam, dass die Interlingua-Methode bisher wegen verlorener Oberflächeninformation oder fehlender Ontologie einer logischen Repräsentation erfolglos blieb.

Um erfolgreiche Interlingua-Systeme zu schaffen, müssen sprachliche Ausdrücke in eine semantische Repräsentation dekomponiert werden. Wie im folgenden Kapitel erläutert wird, bieten sich der Interlinguale Index von EuroWordNet oder die sprachunabhängigen FrameNet-Strukturen an, diese Funktion zu übernehmen.

5.4 Korrekturprogramme

Zur Erkennung von Fehlern in Texten wurden schon früh Programme entwickelt, die Fehler markieren und Korrekturvorschläge machen. Korrekturprogramme befassen sich

a) mit der Korrektur von Zeichenketten, welche in der Sprache nicht vorkommen (so genannte Nichtwörter),
b) mit der Korrektur von Schreibfehlern, die zufällig gültige Wörter ergeben haben, und
c) mit der Korrektur von Grammatikfehlern wie Kongruenz- oder Wortstellungsfehler.

Alle drei Aufgaben erfordern ein grosses Lexikon, in dem die Wörter einer Sprache aufgeführt sind. Es soll sich dabei um ein Vollformenlexikon oder ein Grundformenlexikon mit den entsprechenden Generierungsregeln handeln. Für stark flektierende Sprachen wie Deutsch können durch die Wortbildungsmechanismen Komposition und Derivation ständig neue Wörter gebildet werden, die ebenfalls als korrekt erkannt werden sollen.

5.4.1 Problemstellung

Für Nichtwörter, die vielfach durch Tippfehler entstehen, wird in Korrekturprogrammen nach denjenigen Wörtern gesucht, die sich aus einer minimalen Anzahl von Änderungen ergeben. Diesen Algorithmus bezeichnet man als **Levenshtein-Distanz** [Fliedner 2004, S. 465]. Die Korrekturvorschläge werden der Benutzerin zur Auswahl dargelegt. Leider führt der Einbezug der Levenshtein-Distanz in den meisten Fällen zu vielen teils unsinnigen Vorschlägen, die nichts mit dem beabsichtigten Wort gemeinsam haben.

Aufgabe b) soll Wortverwechslungen erkennen. Ähnlich geschriebene oder ausgesprochene Wörter (e.g. *there*, *their* und *they're*) werden am häufigsten verwechselt. Traditionellerweise werden solche Wörter in so genannten **Verwechslungsmengen** zusammengefasst [Fliedner 2004, S. 467]. Sobald eines dieser Wörter gebraucht wird, werden mithilfe von Heuristiken mögliche Fehler entdeckt und eine Warnung herausgegeben. Dieses Verfahren kann allerdings nur Fehler korrigieren, bei denen genau ein Wort aus einer Verwechslungsmenge vorkommt, andere Fehler bleiben unbemerkt. Es müssen für jedes Wort passende Heuristiken entwickelt werden, was sehr ressourcenintensiv ist.

Aufgabe c) verlangt zusätzlich nach einer syntaktischen Analyse der Eingabe, worauf hier nicht näher eingegangen wird.

5.4.2 Lösungsansätze

Laut [Agirre et al. 1994] kann die Anzahl Korrekturvorschläge bei Nichtwörtern markant reduziert werden, wenn dem Benutzer nur **in den Kontext passende Korrekturvorschläge** unterbreitet werden.

[Budanitsky und Hirst 2001] schlagen eine vergleichbare Methode vor, um Wortverwechslungen zu korrigieren. Sie wollen nach Wörtern suchen, deren Lesarten nicht semantisch ähnlich sind mit allen anderen Wörtern. Solche Wörter könnten potenzielle Fehler sein und sollten markiert werden.

Computerlinguistische Anwendungen stossen meist an ihre Grenzen, wenn ambige Strukturen verarbeitet werden müssen. Wortnetze helfen bei der Disambiguierung. Sie enthalten ein Inventar an Konzepten und ihre lexikalisch-semantischen Relationen geben Aufschluss über die Ähnlichkeit zwischen dem ambigen Wort und den Kontextwörtern. Auch für die Auflösung von Anaphern sollen die in Wortnetzen gespeicherten Informationen einbezogen werden.

6 Lesartendisambiguierung mit Wortnetzen

Nachdem Kapitel 5 eine Übersicht über die Probleme computerlinguistischer Anwendungen bot, sollen nun in Kapitel 6 Disambiguierungs- und in Kapitel 7 Anaphernresolutionsansätze präsentiert werden, welche die Schwierigkeiten der Anwendungen durch Einbezug von Wortnetzen angehen und dabei die zuvor erwähnten Lösungsansätze aufgreifen.

Lesartendisambiguierung (engl. *Word Sense Disambiguation,* kurz WSD) hat zum Ziel, einem Wort in einem bestimmten Kontext eine Lesart zuzuweisen, welche sich klar von möglichen anderen Lesarten dieses Wortes unterscheidet [Ide und Véronis 1998, S. 3]. In einem ersten Schritt müssen die möglichen Lesarten für alle Wörter in einem Textabschnitt bestimmt werden. In einem zweiten Schritt geht es um die Zuweisung der korrekten Lesarten.

Da zum Einsatz von Wortnetzen für die Lesartendisambiguierung sehr viel geforscht und publiziert wurde, wird dem Thema ein eigenes Kapitel gewidmet. Es handelt sich um ein sehr zentrales Problem, da etwa jedes fünfte Wort mehrere Lesarten (durchschnittlich 3) aufweist [WordNetStatistik 2006]. Die Zahl von [Orasan et al. 2005, S. 206] ist noch eindrücklicher: In WordNet haben die häufigsten 1000 englischen Wörter durchschnittlich 7,83 Lesarten. Dabei wurden wohlgemerkt nur lexikalische Ambiguitäten berücksichtigt! Es wundert daher nicht, dass viele Systeme an ambigen Strukturen scheitern.

Um Lesartendisambiguierungsverfahren zu evaluieren, wurde eine mit WordNet-Lesarten annotierte Version des *Brown Corpus,* genannt *Semantic Concordance* (kurz SemCor) erstellt [Miller et al. 1993]. Auch hier verwendet man die IR-Masse Treffergenauigkeit, Ausbeute und F-Mass, um die Performanz von WSD-Systemen zu ermitteln.

6.1 Ansätze zur Lesartendisambiguierung

[Ide und Véronis 1998, S. 4–18] unterscheiden zwischen Ansätzen, die auf Künstlicher Intelligenz (KI), Wissen oder Korpora basieren. KI-Ansätze stellen sich dabei als beschränkt nützlich heraus, da sie meistens nur für eine limitierte Domäne einsetzbar sind. In den 1980er Jahren begann man, wissensbasierte Ansätze zu verfolgen, die grosse lexikalische Ressourcen wie Wörterbücher oder Thesauri benützen. Korpusbasierte Ansätze machen von statistischen Informationen wie Auftretenshäufigkeit einer Lesart oder von grossen Korpora Gebrauch.

Hier werden schwerpunktmässig **wissensbasierte Ansätze** behandelt. Dabei unterscheidet man drei Vorgehen: Beim wörterbuch- und kontextbasierten Ansatz werden Wörterbuchdefinitionen und Kontextwörter berücksichtigt, ohne die lexikalisch-semantischen Relationen zwischen den Kontextwörtern und dem ambigen Wort als wichtig zu erachten. Die beiden anderen Ansätze verfolgen ein relationales Vorgehen und ziehen bei der Disambiguierung solche Beziehungen mit in Betracht. Zum einen werden semantische Ähnlichkeiten zwischen den Wörtern, zum anderen Angaben zu Selektionsbeschränkungen einbezogen.

Im Folgenden geht es hauptsächlich darum, wie man mit den verschiedenen Vorgehen lexikalische Ambiguitäten auflösen kann. Am Rande wird auch auf eine Methode zur Resolution weiterer Ambiguitätstypen hingewiesen.

6.1.1 Wörterbuch- und kontextbasierte Ansätze

Hinter wörterbuch- und kontextbasierten Ansätzen steckt die Hypothese, dass eine Anzahl von gemeinsam vorkommenden Wörtern die richtige Lesart füreinander bestimmen. Viel zitiert wurden die Nomen *base, bat, glove* und *hit,* die alle mehrere Lesarten haben, sich aber zusammen vorkommend jeweils nur auf das Baseballspiel beziehen können [Krovetz und Croft 1992, S. 120].

Genauer gesagt ist die Annahme solcher Ansätze diejenige, dass in Wörterbuchdefinitionen semantisch ähnlicher Wörter viele **Wortüberlappungen** zu finden sind. Ambige Wörter weisen im besten Fall nur in einer Lesart Überlappungen mit Wörterbuchdefinitionen von Kontextwörtern auf. Mithilfe dieser Methoden lässt sich dann auf die beabsichtigte Lesart schliessen.

Erstmals Wörterbuchdefinitionen eingesetzt hat [Lesk 1986], indem er selbst eine Wissensbasis gebildet hat, die jede Lesart in einem Wörterbuch mit einer Liste von Wörtern assoziiert, die in der Definition dieser Lesart vorkommen. Für jede

Lesart und für alle Wörter im Kontext des polysemen Wortes wird ein so genannter *Bag of Words* erstellt. Disambiguierung erfolgt dann durch die Auswahl derjenigen Lesart des Zielwortes, deren Wörterbuchdefinition die grösstmögliche Anzahl von Überlappungen mit der Definition der Kontextwörter aufweist.

Beispielsweise hat das Wort *pine* zwei Lesarten, nämlich „Baum“ und „sich sehnen, schmachten“, und das Wort *cone* kann entweder „Kegel“, „etwas Kegelförmiges“ oder „(Tannen-)Zapfen“ bedeuten. Lesks Algorithmus berechnet die Überlappung der Lesarten von *pine* und *cone*. Die Wörterbuchdefinition von *pine* im Sinne von „Baum“ (*kinds of evergreen tree with needle-shaped leaves ...*) und diejenige von *cone* als „(Tannen-)zapfen“ (*fruit of certain evergreen trees*) haben die meisten gemeinsamen Wörter, was zur Auswahl der entsprechenden Lesarten führt. Diese Methode führt zu 50–70% korrekter Disambiguierung und dient vielen folgenden Experimenten als Grundlage.

[Fragos et al. 2003] wenden die Methode von Lesk an, um Nomen aus SemCor zu disambiguieren. Da sie so nur schlechte Resultate (36%) erzielten, schlugen die Autoren zur Verbesserung den *Weighted Overlapping*-Ansatz vor, bei dem WordNet-Glossen und ihre Hyperonyme betrachtet werden. Nach computerlinguistischer Vorverarbeitung der Texte wurden zwei verschiedene Arten von Bags erstellt, einmal ein Bag für jede Lesart mit entsprechend gewichteten Hyperonymen der Glossen, andererseits ein Bag für alle Kontextwörter mit Hyperonymen. Die Gewichtung innerhalb eines Bag hängt von der Position eines Hyperonyms in der Hierarchie ab. Dieser Algorithmus erreicht in SemCor knapp 50% Performanz.

[Erk 2005] verfolgt den *Bag of Words*-Ansatz mit FrameNet-Ressourcen. Der Satz mit dem ambigen Wort dient dabei als Kontext. Dazu werden Informationen zur Wortart, zu Komplementen und Adjunkten des Zielwortes einbezogen. Zur Evaluation des Algorithmus wurde nicht FrameNet, sondern ein Ausschnitt von SALSA gewählt, da der Aufbau von SALSA nach Lemma und nicht nach Frame erfolgt und daher das Problem von fehlenden Lesarten für ein Wort nicht derart gravierend ist wie bei FrameNet. Der Algorithmus führte zu einer Performanz von rund 75%.

6.1.2 Ähnlichkeitsbasierte Ansätze

Die Idee ähnlichkeitsbasierter Ansätze ist, dass sich Wörter wie „Auto“ und „Fahrrad“ näher stehen als Wörter wie „Auto“ und zum Beispiel „Gabel“. Die Berechnung der Ähnlichkeit basiert auf der **Position der Wörter in einem Wortnetz.** Je kürzer der Pfad zwischen ihnen, desto ähnlicher sind sie sich. Das Ähnlichkeitsmass wird in der Lesartendisambiguierung mit der Motivation eingesetzt, dass damit bestimmt

werden kann, welche Lesart eines ambigen Wortes Kontextwörtern näher steht und diese dann als die beabsichtigte bestimmt wird. Anders als bei den wörterbuch- und kontextbasierten Ansätzen werden aber hier lexikalisch-semantische Relationen einbezogen.

Semantische Ähnlichkeit beschreiben [Budanitsky und Hirst 2001] als Mass, wie nahe sich zwei Einträge in einem Wortnetz stehen, wenn man nur Hyponymie/Hyperonymie-Verzweigungen betrachtet. Obwohl viele Autoren die Begriffe **semantische Verwandtschaft** und semantische Ähnlichkeit synonym verwenden, bestehen [Budanitsky und Hirst 2001] auf einer Unterscheidung. Semantische Verwandtschaft berücksichtigt weitere lexikalisch-semantische Relationen wie Meronymie oder Antonymie. WSD-Algorithmen, die auf der Berechnung semantischer Verwandtschaft beruhen, sind im Vergleich zu solchen semantischer Ähnlichkeit selten. Nur einmal wurde ein Mass aufgestellt, welches die semantische Verwandtschaft mithilfe des kürzesten und „nicht oft die Richtung wechselnden“ Pfades berechnet [Hirst und St-Onge 1998, S. 309]. Ihr Algorithmus wurde aber nicht für Lesartendisambiguierung, sondern für die Erkennung von Rechtschreibefehlern verwendet.

Oft wird zur Messung der Performanz ähnlichkeitsbasierter Disambiguierungsmethoden das Experiment von [Miller und Charles 1991] verwendet. Sie liessen Testpersonen beurteilen, wie ähnlich einander zwei Konzepte sind. Nun kann untersucht werden, wie das Resultat der ähnlichkeitsberechnenden Algorithmen mit den Beurteilungen der Testpersonen korreliert. Einfache Berechnungen des kürzesten Pfades ergeben eine Korrelation von 66% [Resnik 1995b, S. 450]. Um dieses Resultat zu verbessern, gibt es verschiedene Ansätze, welche auf der Berechnung des kürzesten Pfades beruhen, aber zusätzliche Ideen berücksichtigen. Einige der hier vorgestellten Ansätze verfolgen nicht ein rein wissensbasiertes, sondern teilweise auch ein korpusbasiertes Vorgehen.

6.1.2.1 Einbezug von Informationsgehalt

[Resnik 1995a] berechnet die Ähnlichkeit, indem er anfänglich die Distanz zwischen den Knoten der betrachteten Konzepte in WordNet berechnet. Zusätzlich zieht er den Informationsgehalt in Betracht, denn je mehr Information sich zwei Konzepte teilen, desto ähnlicher sind sie sich. Die gemeinsame Information wird hierbei durch den Informationsgehalt der gemeinsamen Hyperonyme festgelegt. Im Wesentlichen wird der Informationsgehalt eines Synsets jeweils berechnet, indem die **Auftretenswahrscheinlichkeit** aller in diesem Synset vorkommenden Nomen und die Auftretenswahrscheinlichkeit aller Nomen der untergeordneten Synsets innerhalb eines

grossen Korpus betrachtet werden [Smeaton 1999, S. 106]. Je häufiger ein Nomen auftritt, desto weniger informativ ist es.

Abbildung 12 illustriert dieses Vorgehen. Angenommen, man wollte die Ähnlichkeiten zwischen *car* und *bicycle* und zwischen *car* und *fork* berechnen. In der WordNet-Hierarchie sieht man, dass der gemeinsame Informationsgehalt von *car* und *bicycle* 8,30 ist, nämlich derjenige des gemeinsamen Hyperonyms *vehicle*. *Car* und *fork* teilen sich nur einen Informationsgehalt von 3,53 (Synset *artifact*). Dieses Resultat geht mit der zuvor erwähnten Intuition überein, dass sich *car* und *bicycle* ähnlicher sind als *car* und *fork*.

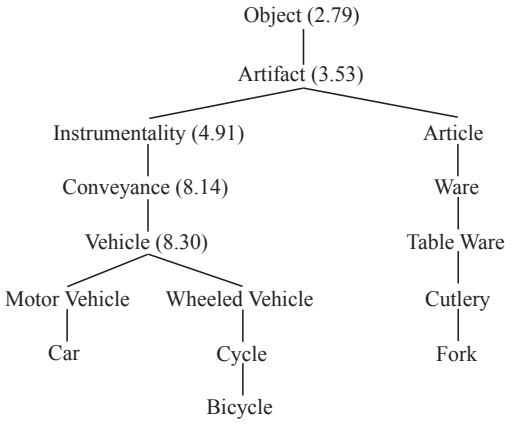

Abbildung 12: Fragment der WordNet-Hierarchie nach [Jiang und Conrath 1997, S. 114]

Wie kann diese Methode nun helfen, Lesarten zu disambiguieren? Als Beispiel gibt [Resnik 1995a, S. 57f.] das Paar *doctor* und *nurse*. Er sucht nun in der WordNet-Hierarchie nach gemeinsamen Hyperonymen. Für diese Synsets berechnet er den Informationsgehalt und findet so als informativsten Oberbegriff das Synset *health professional*. Daher wird die Lesart von *doctor* im Sinn von „Mediziner“ gegenüber derjenigen als „Doktortitel“ bevorzugt, ebenso wird die Lesart von *nurse* im Sinn von „Krankenschwester“ und nicht diejenige von „Kindermädchen“ ausgewählt.

[Resnik 1995b] evaluiert dieses Ähnlichkeitsmass mithilfe des Testsets von [Miller und Charles 1991]. Während mit einfachen Berechnungen nach dem kürzesten Pfad wie erwähnt 66% Korrelation erzielt wird, erreicht seine Methode 79% Übereinstimmung [Resnik 1995b, S. 450].

[Jiang und Conrath 1997] folgen in den Grundsätzen dem Ansatz von Resnik, machen aber darauf aufmerksam, dass bei der Berechnung des Informationsgehaltes aufgrund von Auftretenshäufigkeit polyseme Nomen einen übertrieben hohen Wert erhalten. Dazu kommt es, da die Auftretenshäufigkeit des Wortes und nicht die

der Lesart benutzt wird. Die Autoren schlagen die Verwendung von SemCor vor. Da dieses Korpus mit Synsets indexiert ist, kann die Auftretenswahrscheinlichkeit von Lesarten berechnet werden. Zudem berücksichtigen sie weitere Faktoren wie die lokale Synsetdichte oder die Position in der Hierarchie. Durch ihr *Combined Distance Model* erreichen sie eine Korrelation von 82% [Jiang und Conrath 1997, S. 121].

Einen weiteren informationsgehaltsbasierten Ansatz beschreibt [Lin 1998]. Er definiert semantische Ähnlichkeit mit einer Formel aus der Informationstheorie, welche anwendungs-, domänen- und ressourcenunabhängig sein soll. Dazu zieht er eine Anzahl von Annahmen hinzu, die bei [Lin 1998, S. 297] detailliert beschrieben sind. So schliesst das Mass im Vergleich zu demjenigen von [Jiang und Conrath 1997] einen Prozentpunkt besser ab.

6.1.2.2 Einbezug von konzeptueller Dichte

Der Ansatz zur Disambiguierung von [Agirre und Rigau 1995] basiert auf einer Formel der **konzeptuellen Distanz zwischen Konzepten,** der konzeptuellen Dichte. Zusätzlich zur Berechnung des kürzesten Pfades wird die Position und die Dichte der Konzepte (e.g. Anzahl Hyponyme) in einem Ausschnitt der Hierarchie betrachtet. Konzepte, die durch einen kurzen Pfad verbunden sind, in einem tiefer gelegenen und in einem dichteren Teil der Hierarchie vorkommen, stehen sich semantisch näher als andere.

Die Autoren evaluieren ihr Ähnlichkeitsmass mithilfe von SemCor. Für jedes zu disambiguierende Nomen werden im Text benachbarte Nomen und ihre Hyperonyme betrachtet und diejenige Lesart mit der höchsten konzeptuellen Dichte als die korrekte ausgewählt. Dieser Ansatz erreicht eine Treffergenauigkeit von 66% und eine Ausbeute von 59% [Agirre und Rigau 1995, S. 6]. Einen ähnlichen Ansatz verfolgen [Leacock und Chodorow 1998]. Im Gegensatz zu [Agirre und Rigau 1995] lassen die Autoren die Hierarchiedichte ausser Acht.

Um die Ähnlichkeit zwischen zwei Wörtern bestimmen zu können, stellen [Pedersen und Michelizzi 2002] ein frei zugängliches Webinterface bereit, das auf Basis von WordNet (Version 2.0) die Ähnlichkeit zwischen zwei Eingaben berechnet und sie nach WordNet-Lesarten disambiguiert. Die Abfrage erlaubt es unter anderem, die Masse von [Leacock und Chodorow 1998], [Resnik 1995b], [Jiang und Conrath 1997], [Lin 1998] und [Hirst und St-Onge 1998] zu testen. Dabei werden die Resultate sehr übersichtlich dargestellt.

6.1.2.3 Einbezug der Klassifikation von Roget's Thesaurus

Einen ähnlichkeitsbasierten Ansatz verfolgt auch [Yarowsky 1992]. Er verwendet nicht WordNet, sondern die rund 1000 Köpfe[28] des Roget's Thesaurus, um Nomen zu disambiguieren. Für jedes Wort in einer Kategorie werden aus einem Korpus **Kontextwörter extrahiert,** gewichtet und der Kategorie zugewiesen. Eine Kategorie umfasst typischerweise über 3000 Kontextwörter, also mehr, als von Wörterbuchdefinitionen hergeleitet werden könnten.

Abbildung 13 zeigt zwei Kategorien mit den zugewiesenen Kontextwörtern. Es handelt sich dabei um Wörter, die häufig mit den Mitgliedern dieser Kategorie zusammen auftreten. Die Zahlen in den Klammern geben an, wie stark ein Kontextwort gewichtet wird. Die Berechnung basiert unter anderem auf der Auftretenshäufigkeit des Wortes.

ANIMAL/INSECT (Category 414):
species (2.3), family (1.7), bird (2.6), fish (2.4), breed (2.2), cm (2.2), animal (1.7), tail (2.7), egg (2.2), wild (2.6), common (1.3), coat (2.5), female (2.0), inhabit (2.2), eat (2.4), nest (2.5), ...
TOOLS/MACHINERY (Category 348):
tool (3.1), machine (2.7), engine (2.6), blade (3.8), cut (2.6), saw (5.1), lever (4.1), pump (3.5), device (2.2), gear (3.5), wheel (2.8), shaft (3.3), wood (2.0), tooth (2.5) piston (3.6), ...

Abbildung 13: Kontextwörter für die Thesauruskategorien 414 und 348 nach [Yarowsky 1992, S. 456]

Wenn ein Wort in einem neuen Text disambiguiert werden soll, werden 50 Wörter links und rechts davon betrachtet und gezählt, wie viele Übereinstimmungen sie mit den Kontextwörtern der möglichen Kategorien haben. Dem Lexem *crane* konnte die intendierte Baumaschinen-Lesart zugewiesen werden, da [Yarowsky 1992, S. 457] in den 100 Umgebungswörtern viele Übereinstimmungen mit den Kontextwörtern der Thesaurusklasse 348 fand.

[Jarmasz und Szpakowicz 2003] verwenden eine elektronische Version des Roget's Thesaurus, um zu überprüfen, in welchem Ausmass daraus errechnete semantische Ähnlichkeit mit der menschlichen Beurteilung des Experiments von [Miller und Charles 1991] übereinstimmt. Ihr Ähnlichkeitsmass berechnet den

[28] [Yarowsky 1992] spricht von Kategorien.

Pfad zwischen zwei Nomen in Roget's Thesaurus. Der kürzeste Pfad besteht zwischen einem Nomenpaar in derselben Semikolon-Gruppe, der längste, wenn der Pfad über eine der sechs Klassen (Unique Beginners) führt.

[Jarmasz und Szpakowicz 2003] evaluieren die vorgängig vorgestellten Ähnlichkeitsmasse von [Resnik 1995b], [Jiang und Conrath 1997] und der simplen Berechnung des kürzesten Pfades mit einer neueren WordNet-Version (1.7) am Experiment von [Miller und Charles 1991]. Dabei schliesst das Mass von [Jarmasz und Szpakowicz 2003, S. 113] mit 87% Korrelation mit der menschlichen Beurteilung am besten ab. Das Mass von [Resnik 1995b] erreicht 77%, [Jiang und Conrath 1997] 69%, [Lin 1998] und auch [Leacock und Chodorow 1998] 82% und die Berechnung des kürzesten Pfades kommt auf immerhin 73% Übereinstimmung [Jarmasz und Szpakowicz 2003, S. 113]. Dass neben den eigenen Algorithmen auch andere Methoden besser abschliessen, führen [Jarmasz und Szpakowicz 2003, S. 114] auf den Einsatz der neuen WordNet-Version zurück, in der die Taxonomie ausführlicher ausgearbeitet ist.

[Gurevych und Niederlich 2005] evaluieren die bis anhin aufgeführten Ansätze zur Lesartendisambiguierung mit einem deutschen Datensatz unter Zuhilfenahme von GermaNet. Den Informationsgehalt berechnen sie mithilfe einer Häufigkeitsliste deutscher Wortstämme. [Gurevych und Niederlich 2005, S. 471] führen auf, dass der Korrelations-Koeffizient bei [Lesk 1986] 69%, bei [Resnik 1995b] 71% und bei [Lin 1998] 73% beträgt. Verglichen mit den Resultaten von [Jarmasz und Szpakowicz 2003] schneiden die Ähnlichkeitsmasse an deutschen Datensätzen also schlechter ab.

6.1.2.4 Einbezug von topikalischen Relationen

Die Berücksichtigung topikalischer Relationen ist dadurch motiviert, dass man durch Betrachtung bestehender WordNet-Relationen im Beispiel 16 von [Moldovan und Novischi 2002, S. 674] den Zusammenhang der beiden Sätze, genauer gesagt der Nomen *food* und *refrigerator,* nicht ermitteln kann.

(16) a. Jim was hungry.
b. He opened the refrigerator.

Zieht man aber die **in Glossen enthaltene Information** mit in Betracht, kann eine Verbindung hergestellt werden. In den Glossen der Synsets *hungry: feeling a need or desire to eat* ***food*** und *refrigerator: a kitchen appliance in which* ***food*** *can be stored*

at low temperatures wird der Zusammenhang durch das gemeinsame Konzept *food* klar. Zwischen *hungry* und *refrigerator* kann eine so genannte **lexikalische Kette** (engl. *lexical chain*) hergestellt werden, also eine Abfolge semantisch verwandter Wörter, durch welche zwei Konzepte verknüpft sind.

Mit der Extraktion topikalischer Relationen befasst sich das Extended WordNet-Projekt von [Harabagiu und Moldovan 1999]. Dort werden erst alle WordNet-Glossen disambiguiert, die Inhaltswörter der Glossen und durch verschiedene Relationen verbundene Glossen berücksichtigt. Hierbei werden verschiedene Stufen topikalischer Relationen, Heuristiken (beispielsweise werden kürzere Pfade zwischen Synsets höher gewichtet als längere Pfade), konzeptuelle Dichte und Statistiken grösserer Korpora einbezogen. Durch diese Vorgehensweise erreicht man, dass Relationen über die Wortartengrenzen einbezogen werden können.

6.1.3 Ansätze mit Selektionsbeschränkungen und Wissensbasen

[Wagner 2005] schlägt vor, bei der Auflösung verschiedener Ambiguitätstypen Angaben zu Selektionsbeschränkungen zu verwenden. Informationen zu Argumentstypen eines Prädikates können helfen, die fett gedruckten, mehrdeutigen Wörter im Beispiel 17 von [Wagner 2005, S. 12] bzw. [Kunze und Wagner 2001, S. 241] aufzulösen.

(17) a. The **dishwasher** read the newspaper.
b. Der Mann tritt gegen den **Ball.**

Das polyseme Lexem *dishwasher* in a) kann auf ein Gerät oder einen Menschen verweisen. Da aber die Selektionsbeschränkungen vom Verb *read* ein menschliches Agens verlangen, kann die Geräte-Lesart ausgeschlossen werden. In b) wird das homonyme Nomen „Ball“ aufgrund der Selektionsbeschränkungen von „treten“ im Sinne von „Spielgerät“ und nicht als „Tanzveranstaltung“ disambiguiert.

Der Satz in Beispiel 18 von [Kunze und Wagner 2001, S. 241] weist keine lexikalische, sondern eine syntaktische Ambiguität auf.

(18) Das Brot schneidet die Mutter.

Nach rein morphologischen und syntaktischen Kriterien ist es nicht möglich, herauszufinden, welche Nominalphrase das Subjekt und welche das Objekt ist. Da das

Verb „schneiden“ in den meisten Fällen ein menschliches Agens und ein unbelebtes Patiens verlangt, kann die Nominalphrase „die Mutter“ als Subjekt identifiziert werden.

Auch einige **Anschlussambiguitäten**[29] können mithilfe von Selektionsbeschränkungen aufgelöst werden, was am Beispiel 19a) von [Wagner 2005, S. 2] gezeigt wird.

(19) a. Peter watched the woman with the umbrella.
b. Peter watched the man with the telescope.

Die Selektionsbeschränkungen von *watch* legen fest, dass ein Instrument hinzukommen kann, welches ein optisches Gerät sein muss. Dies schliesst die Möglichkeit aus, die Präpositionalphrase *with the umbrella* dem Verb anzuhängen. Wie das häufig zitierte Beispiel 19b) zeigt, führt diese Methode aber nicht in jedem Fall zu einer eindeutigen Entscheidung, da es sich beim „Teleskop“ um ein optisches Gerät handelt, also keine der beiden Lesarten ausgeschlossen werden kann.

Angaben zu Selektionsbeschränkungen reichen auch im Fall von Beispiel 20 von [Cycorp 2006] nicht aus, um herauszufinden, welcher Konstituente die Satzergänzung *flying over Zurich* angehängt werden soll.

(20) a. Fred saw the plane flying over Zurich.
b. Fred saw the mountains flying over Zurich.

Der menschliche Leser weiss, dass Flugzeuge über Zürich fliegen können, fliegende Berge aber eine aussergewöhnliche Erscheinung wären. Grosse Wissensbasen wie Cyc speichern genügend Weltwissen, um zu wissen, dass Flugzeuge fliegen, Menschen in Flugzeugen fliegen, Berge nicht fliegen können und dass Zürich eine Stadt ist.

6.1.4 Evaluation der Ansätze

Algorithmen zur Lesartendisambiguierung können an verschiedenen Textsammlungen oder im Vergleich mit der menschlichen Beurteilung evaluiert werden. Da in jedem Versuch wieder andere Wörter (meistens nur Nomen) disambiguiert werden, ist ein aussagekräftiger Vergleich oft schwierig zu vollziehen.

Die Genauigkeit der hier vorgestellten Ansätze pendelt zwischen 60 und 80%. Die Evaluation an SemCor schliesst dabei vielfach schlechter ab als diejenige mit dem

[29] [Wagner 2005, S. 12] spricht von strukturellen Ambiguitäten.

Experiment von [Miller und Charles 1991]. Der Einsatz von SemCor ist aber grundsätzlich vorzuziehen. Schliesslich werden zuverlässige Disambiguierungsmethoden verlangt, die nicht nur von Hand zusammengesuchte ambige Wortpaare, sondern ein Korpus disambiguieren können.

Ein generelles Problem ist, dass die meisten Wörterbücher **zu feinkörnige Lesarten** unterscheiden, um für die Disambiguierung effizient eingesetzt werden zu können. [Ide und Véronis 1998, S. 22] rechnen vor, dass der Satz in Beispiel 21 bemerkenswerte 284 592 verschiedene Lesartenkombinationen hat[30].

(21) There is a huge envelope of air around the surface of the earth.

Obwohl die Berechnung nicht auf einem der hier vorgestellten Wortnetze beruht, besteht dieses Problem auch beim Einsatz von WordNet.

6.1.4.1 Wörterbuch- und kontextbasierte Ansätze

Die grosse Schwierigkeit bei den wörterbuch- und kontextbasierten Ansätzen ist die Bestimmung der **angemessenen Kontextberücksichtigung.** Wie gross soll das Fenster sein, in dem nach Kontextwörtern gesucht wird? Wann ist die Grenze erreicht, an der die Berücksichtigung zu vieler Kontextwörter eher schadet als nützt? [Leacock und Chodorow 1998, S. 266] haben herausgefunden, dass der Einbezug von Kontext bei der Disambiguierung von homonymen Wörtern nützlicher ist als bei polysemen. Diese Feststellung erstaunt nicht. Homonyme haben im Vergleich zu Polysemen keinen gemeinsamen Bedeutungskern und weisen daher auch weniger gemeinsame ähnliche Wörter auf.

[Vossen 2004b] erkannte als weiteres Problem, dass in gewissen Fällen die beabsichtigte Lesart „überschrieben“ wird. Bei der Disambiguierung des Wortes *Java* in Beispiel 22 von [Vossen 2004b, S. 20], das drei Lesarten hat (Insel, Kaffeemarke oder Programmiersprache), leitet der Kontext fehl, da die Kontextwörter *internet services* die Auswahl der Programmiersprachen-Lesart von *Java* begünstigen.

(22) internet services on Java

Die beabsichtigte Insel-Lesart von *Java* konnte mithilfe der Kontextwörter nicht erkannt werden. Statt Aufschluss über die korrekte Lesart zu geben, führt ihr Einbezug in diesem Beispiel ins Abseits.

[30] Grundlage der Berechnung war der *Longman Dictionary of Contemporary English.*

6.1.4.2 Ähnlichkeitsbasierte Ansätze

Während bei den wörterbuch- und kontextbasierten Ansätzen die effiziente Kontextberücksichtigung zentral ist, sind es bei den ähnlichkeitsbasierten Ansätzen die lexikalisch-semantischen Relationen, die betrachtet werden. Welche Relationen soll man einbeziehen? Wie wirken sich fehlende Relationen auf die Performanz aus?

Die meisten ähnlichkeitsbasierten Ansätze setzen WordNet ein. Neben dem eingangs erwähnten Problem der zu feinen Lesartenunterscheidungen machen [Rila et al. 1998, S. 32] auf weitere Nachteile aufmerksam. Da zur Berechnung des Ähnlichkeitsmasses lexikalisch-semantische Relationen eingesetzt werden, sind **fehlende Relationen** problematisch. Insbesondere die Meronymie-Relation ist in WordNet bis anhin immer noch nicht sehr ausführlich erfasst. Zudem haben viele ähnliche Begriffe nicht dieselbe Wortart und sind daher durch keine Relation miteinander verbunden.

Auch mithilfe ähnlichkeitsbasierter Ansätze ist die Auflösung der Ambiguität in Beispiel 22 schwierig. *Java* wird in dieser Phrase mit hoher Wahrscheinlichkeit ebenfalls als „Programmiersprache“ und nicht als „Insel“ disambiguiert, da die unbeabsichtigte Lesart eher durch Relationen mit dem ambigen Wort verbunden ist.

6.1.4.3 Ansätze mit Selektionsbeschränkungen und Wissensbasen

Das zentrale Problem beim Einsatz von Selektionsbeschränkungen ist die Frage nach dem **Inventar der Beschränkungen.** [Wagner 2005, S. 6] stellt fest, dass immer noch unklar ist, welche Rollen überhaupt bestehen. Die Frage nach der angemessenen Generalisierung und nach dem Inventar an semantischen Typen ist ebenfalls noch nicht geklärt [Kunze und Wagner 2001, S. 242].

Obwohl der Einsatz von Selektionsbeschränkungen und Wissensbasen für die Lesartendisambiguierung vielversprechend scheint, darf man nicht ausser Acht lassen, dass **Verletzungen der Selektionsbeschränkungen** nicht selten sind. Obwohl die Verben *drink* oder *agree* normalerweise nur belebte Agenten selegieren, sind Sätze wie im Beispiel 23 von [Wagner 2005, S. 3f.] alles andere als aussergewöhnlich.

(23) a. The car drinks gasoline.

b. Paris and Berlin have agreed on matters for the French presidency of the EU.

Metaphorische (a) oder metonymische (b) Konstrukte heben Selektionsbeschränkungen auf, weswegen [Wagner 2005, S. 4] vorschlägt, nicht Selektionsbeschränkungen, sondern **Selektionspräferenzen** zu betrachten. Diejenige Lesart eines Satz, welche am wenigsten Präferenzen verletzt, wird ausgewählt.

Man sollte nicht vergessen, dass bisher die wenigsten lexikalisch-semantischen Ressourcen über ausführliche Angaben zu Selektionsbeschränkungen verfügen. In erster Linie ist also erforderlich, Selektionsinformation aus Korpora zu akquirieren. Dazu schlagen [Kunze und Wagner 2001, S. 242] ein Vorgehen vor, das die relative Häufigkeit von Prädikats-Arguments-Kookkurrenzen[30] in Korpora berücksichtigt und so Selektionsbeschränkungen extrahiert.

Trotz der erwähnten Nachteile können durch die Berücksichtigung von Kontextwörtern, Ähnlichkeitsmassen, topikalischen Relationen und Selektionsbeschränkungen viel mehr lexikalische Ambiguitäten aufgelöst werden als ohne den Einbezug von Wortnetzen. Als Alternative können nur statistische Methoden genannt werden, welche die Lesart nach ihrer Auftretenshäufigkeit selegieren. Es bleibt zu untersuchen, welche Vorhergehensweise mit den bestehenden Ressourcen besser abschneidet.

6.2 Einsatz in computerlinguistischen Anwendungen

Im Folgenden wird aufgezeigt, wie die oben erwähnten Ansätze zur Lesartendisambiguierung im Information Retrieval, in der Antwortextraktion, Maschinellen Übersetzung und Rechtschreibekorrektur eingesetzt werden können und wie sich dies auf die Performanz dieser Systeme auswirkt.

6.2.1 Einsatz im Information Retrieval

6.2.1.1 Mögliche Methoden und Ansätze

Erste Versuche, Lesartendisambiguierung in IR-Systemen einzusetzen, waren eher entmutigend. [Voorhees 1994] hat von Hand Anfragen aus der TREC erweitert, indem sie Synonyme, Hyperonyme und Hyponyme einbezog. Die Erweiterung war nur bei kurzen Anfragen nützlich. Die Autorin vermutet, dass sich durch eine automatische Erweiterung zudem die Forderung nach automatischer Lesartendisambiguierung stellen würde, was insbesondere bei kurzen Anfragen schwierig wäre.

[30] Eine solche Kookkurrenz wäre beispielsweise, dass das Prädikat „kochen“ oft mit dem Argument „Wasser“ vorkommt.

Statt die Anfrage mit einem gewöhnlichen Index abzugleichen, ermöglicht die **konzeptuelle Indexierung** der Dokumente einen Schritt in Richtung verbesserter Suchalgorithmen. [Smeaton und Quigley 1996] erreichen gute Resultate in der Auffindung von Bildüberschriften. Die Nomen in den Überschriften und Anfragen werden manuell mit WordNet-Synsets indexiert. Bei kurzen Dokumenten können mit einem Ähnlichkeitsmass gute Ergebnisse erzielt werden. Anfragen wie *children running on the beach* finden so Bildüberschriften wie *boys playing in the sand*.

Auch [Gonzalo et al. 1998a] schlagen vor, die Dokumentensammlung mit WordNet-Synsets zu indexieren. In einer Evaluation an SemCor führt die manuelle Synset-Indexierung eines Korpus verglichen mit der Indexierung mit Wörtern zu einer beinahe 30% besseren Performanz eines IR-Systems [Gonzalo et al. 1998a, S. 38].

[Philpot et al. 2005] erwähnen Forschungsprojekte, die sich mit der konzeptuellen Indexierung basierend auf Omega befassen. Interessant ist der Einbezug von Omega insbesondere, weil die Ontologie nicht nur für die Indexierung englischer Dokumente, sondern sprachunabhängig einsetzbar ist. [Gilarranz et al. 1997] wollen zur multilingualen Indexierung den Interlingualen Index von EuroWordNet einsetzen. Dazu werden erst die Dokumente mit dem ILI indexiert, was die Treffergenauigkeit erhöhen soll. Zusätzlich erkennt die Suchmaschine äquivalente Terme, da sie mit demselben ILI-Eintrag verknüpft sind. So kann die Ausbeute signifikant erhöht werden.

6.2.1.2 Suchmaschinen mit Wortnetzen

In diesem Abschnitt werden vier Suchsysteme vorgestellt, die Lesartendisambiguierung durchführen und alle andere lexikalisch-semantische Ressourcen einsetzen.

Die IR-Methode von [Gurevych 2005] beruht auf **Information aus GermaNet** und kommt in einer Anwendung der elektronischen Berufsberatung zum Einsatz. Die Interessensprofile von Personen werden gegen eine Dokumentensammlung mit natürlichsprachlichen Beschreibungen der Ausbildungsberufe abgeglichen. Da dieses System in einer ganz spezifischen Domäne eingesetzt wird, reicht es nicht aus, zur Evaluation generelle Textsammlungen wie TREC einzusetzen. Als Gold Standard für die beiden durchgeführten Experimente ist daher extra ein geeignetes wissensbasiertes System erstellt worden [Gurevych 2005, S. 289].

Im ersten Experiment geht es darum, die Anfragen an dieses System durch GermaNet-Angaben zu ergänzen. Die Erweiterung durch Hyponyme führt dabei zu

den besten Resultaten [Gurevych 2005, S. 293]. Gibt jemand „Computer“ ein, wird in der Anfrage beispielsweise der Unterbegriff „Laptop“ eingefügt. Erst das zweite Experiment führt Lesartendisambiguierung durch. [Gurevych 2005] berücksichtigt die semantische Ähnlichkeit zwischen Wörtern und setzt dazu die Methode von [Lin 1998] ein, da sie in der Evaluation mit deutschen Wortpaaren gut abgeschnitten hat (vgl. 6.1.2.3). Um Anfrage und Dokument hinsichtlich ihrer Ähnlichkeit zu vergleichen, wird die Dokumentensammlung in eine Zwischenrepräsentation gebracht, in der alle Nomen mit GermaNet-Synsets konzeptuell indexiert werden. Die relevantesten Dokumente sind diejenigen, die der Anfrage semantisch am ähnlichsten sind. [Gurevych 2005, S. 297] will das Suchsystem noch verbessern, indem die Masse semantischer Verwandtschaft und die computerlinguistische Vorverarbeitung der Dokumente verfeinert werden.

Eine andere Vorgehensweise schlagen [Kruse et al. 2005] vor, die eine zusätzliche Komponente für Suchmaschinen vorstellen. In einem ersten Schritt wird der Suchraum durch Lesartendisambiguierung eingeschränkt und in einem zweiten die beabsichtigte Lesart durch ähnliche Terme ergänzt. Das zusätzliche Modul ihres Suchalgorithmus Clever Search wird als *Wrapper* bezeichnet und kann in Kombination mit verschiedenen Suchmaschinen wie Google oder Yahoo eingesetzt werden. Bei der Anfrage wird die Anwenderin aufgefordert, die Anfrage zu konkretisieren, indem sie die beabsichtigte WordNet-Lesart des Anfrageterms auswählt. Dazu werden alle **WordNet-Wörterbuchdefinitionen und Beispielsätze** der möglichen Lesarten angezeigt.

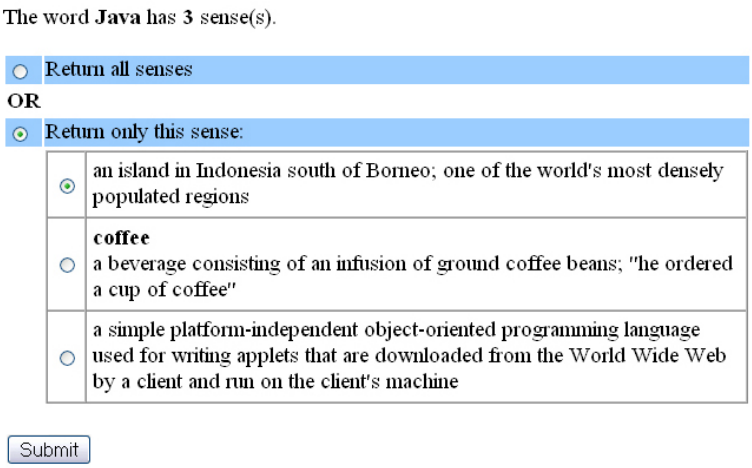

Abbildung 14: Auswahl der verschiedenen Lesarten in Clever Search nach [Kruse und Naujoks 2006]

Clever Search kann bei [Kruse und Naujoks 2006] getestet werden. Abbildung 14 zeigt eine Anfrage mit dem Suchbegriff *Java*. Durch Anklicken der Option `Return`

`all senses` wird die Benutzeranfrage ohne Erweiterung an die ausgewählte Suchmaschine weitergegeben. Die Option `Return only this sense` fordert dazu auf, die beabsichtigte Lesart auszuwählen. Nun werden die konzeptuellen Relationen von WordNet einbezogen und das gewählte Synset durch Wörter der Hyponym-Synsets erweitert. Gibt es keine Hyponyme, werden Hyperonyme in Betracht gezogen [Kruse et al. 2005, S. 370]. Die erweiterte Anfrage wird so der Suchmaschine übergeben. Wählt der Benutzer die Programmiersprachen-Lesart, sieht die Google-Anfrage so aus: „Java“ „object-oriented programming language“ OR „object-oriented programming language“.

Suchraumerweiterung ist auch das Paradigma der Suchtechnologie iFinder der Firma Intrafind[31], die laut Angaben von [Naber 2006, persönliche Kommunikation] die Daten des **OpenThesaurus** verwendet, um in ca. 230 000 XML-Dokumenten der deutschen Presseagentur zu suchen. Die Demoapplikation listet neben der Trefferliste eine Anzahl ähnlicher Wörter auf, die der Benutzer in eine erweiterte Suchanfrage einbeziehen kann [iFinder 2006]. Für die Suchanfrage „Bank“ sind als mögliche Erweiterungen die Nomen „Bankhaus“, „Geldinstitut“, „Kreditanstalt“, „Kreditinstitut“, „Sitzbank“ oder „Sparkasse“ zu verzeichnen. Durch die Suchraumerweiterung kann einerseits die Ausbeute erhöht werden, andererseits können dadurch, dass die Anfrage nur nach Rückfrage beim Benutzer erweitert wird, nicht beabsichtigte Lesarten (e.g. „Bank“ im Sinne von „Sitzbank“) ausgeschlossen und so relevante Treffer höher gewichtet werden.

Eine mehrsprachige Suchmaschine ist [Antonya 2006]. Zurzeit stehen die Sprachen Deutsch, Englisch, Französisch, Italienisch, Niederländisch, Spanisch zur Verfügung. Laut Informationen von [Joop, Firma Irion[32] 2006, persönliche Kommunikation] ist Antonya eine Demonstrationsapplikation einer kommerziellen Suchmaschine TwentyOne Search. Sie ist besonders stark im Gebiet Umwelt und hat circa 3000 Webseiten indexiert [Vossen 2004b, S. 26]. Im System werden ebenfalls die Synonyme der im Sucheintrag enthaltenen Wörter berücksichtigt [Antonya 2006]. Dabei greift Antonya auf ein lexikalisch-semantisches Netz namens **SemNet** zurück, das laut Angaben von [Joop 2006, persönliche Kommunikation] auf Lexika eines Verlages basiert, die umfangreicher sind als WordNet.

Abbildung 15 zeigt die Suchresultate der Anfrage „Wasserverschmutzung“. Unter den drei ersten Treffern befinden sich Dokumente dreier verschiedener Sprachen. Fährt man mit der Maus über kursiv gesetzte Wörter, werden zusätzlich Synonyme in der Suchsprache angezeigt.

[31] `http://www.intrafind.de`.

[32] `http://www.irion.nl`.

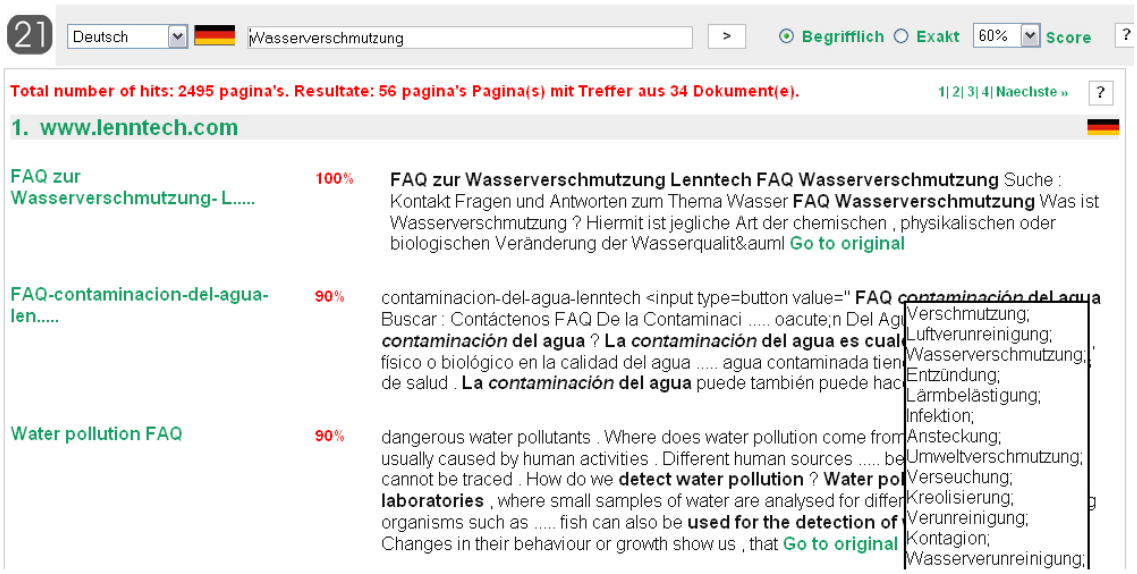

Abbildung 15: Auflistung der Treffer auf die Anfrage „Wasserverschmutzung“ in [Antonya 2006]

Durch Ausweitung auf weitere Fachgebiete hätte Antonya das Potenzial, eine der ersten sprachübergreifenden Suchmaschinen zu werden.

6.2.2 Einsatz in der Antwortextraktion

6.2.2.1 Mögliche Methoden und Ansätze

[Harabagiu et al. 2001] setzen lexikalisch-semantische Ressourcen und Heuristiken für verschiedene Teilaufgaben der Antwortextraktion ein. Zentral ist hierbei der Einsatz von WordNet in der **Anfrageerweiterung** durch semantisch ähnliche Wörter und Information in den Glossen. Die TREC-Frage *Who killed Martin Luther King?* konnte beispielsweise nur durch Erweiterung durch das Synonym *assassin* richtig beantwortet werden [Harabagiu et al. 2001, S. 286]. Auch eine vollständige Umformulierung der Anfrage mithilfe von WordNet kann helfen, die korrekten Antworten zu finden [Harabagiu et al. 2001, S. 287f.].

[Moldovan und Novischi 2002] verwenden für das QA **lexikalische Ketten,** die aus topikalischen Relationen aus WordNet extrahiert werden (vgl. 6.1.2.4). Damit kann einerseits die Ausbeute erhöht werden, indem die Anfrage mit topikalisch verwandten Wörtern erweitert wird. Andererseits kann die Antwortextraktion verbessert werden, da lexikalische Ketten bis zu einem gewissen Grad **Weltwissen** gespeichert haben. Die Beispielanfrage 24a) von [Moldovan und Novischi 2002, S. 674] kann mithilfe der beiden Sätze aus Beispiel 16, der WordNet-Glosse von *want: feel or*

have a desire und der Hinzunahme lexikalischer Ketten richtig beantwortet werden (vgl. 24b).

(24) a. Qestion: What did Jim want?
b. Answer: Jim wanted to eat. (or Jim wanted food.)

Topikalische Relationen verbinden *want* mit *desire, hungry* und schliesslich *food,* und verfügen so beispielsweise über das Weltwissen, dass man im Normalfall essen will, wenn man hungrig ist.

[Fliedner 2004] arbeitet mit FrameNet im Bereich Antwortextraktion für das Deutsche. Dazu werden die Sätze erst computerlinguistisch vorverarbeitet, unter anderem werden Eigennamen erkannt. Er schlägt vor, Textsammlungen mit Frames konzeptuell zu indexieren und die Anfragen ebenfalls in diese bedeutungsorientierte Repräsentation zu übersetzen. Durch den **Abgleich der Frame-Strukturen** kann die Antwort gefunden werden. Dieses Vorgehen wird am Satz aus Beispiel 25 von [Fliedner 2004, S. 63] illustriert.

(25) Lockheed hat von Grossbritannien den Auftrag für 25 Transportflugzeuge erhalten.

Aus diesem Satz wurde der Frame Getting extrahiert, den die Abbildung 16 zeigt.

Getting
donor: Grossbritannien
recipient: Lockheed [organisation]
theme: Auftrag für 25 Transportflugzeuge

Abbildung 16: Repräsentation des Getting-Frame für die Antwortextraktion basierend auf [Fliedner 2004, S. 63]

Die Anfrage „Von wem hat Lockheed einen Auftrag erhalten?“ wird nun in eine ganz ähnliche Repräsentation übersetzt. Hier bleibt aber das Frame-Element donor leer, „Grossbritannien“ wird durch eine Lücke oder lediglich mit Informationen zum erwarteten Antworttyp (Person oder Organisation) ersetzt. Durch den Frame-Abgleich kann die Lücke gefüllt und somit die Antwort „Grossbritannien“ gefunden werden. Durch verbesserte Eigennamenerkennung und Selektionsbeschränkungen von GermaNet (in Planung) sollen Verbesserungen für diesen Prototypen erzielt werden. [Fliedner 2004, S. 65] vermutet, dass durch die sprachunabhängige Repräsentation es ausserdem ein Einfaches wäre, so ein sprachübergreifendes QA-System zu erstellen.

6.2.2.2 Antwortextraktionssysteme mit Wortnetzen

[Curtis et al. 2005] beschreiben das kommerzielle QA-System MySentient Answer, das von der Firma MySentient[33] und Cycorp entwickelt wird. Cyc wird in MySentient Answer für mehrere Aufgaben verwendet. Unter anderem werden die Einträge der Ontologie dazu verwendet, um Schlüsselwörter in der Anfrage zu erweitern und um Anaphern aufzulösen (mehr dazu später unter 7.2.1).

Hier wird die erste Aufgabe von MySentient Answer, die Anfrageerweiterung, genauer vorgestellt. Das *Clarification*-Modul identifiziert die Schlüsselwörter der Anfrage und formuliert sie neu. Abbildung 17 zeigt, wie die Anfrage *Who is offering the best APR for an auto loan?* erweitert wird.

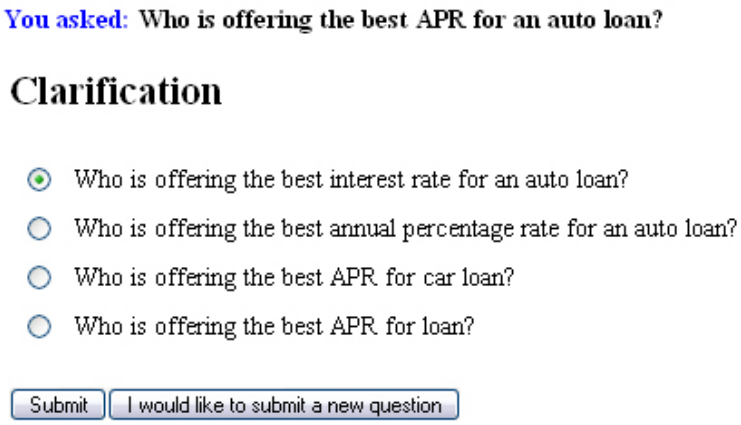

Abbildung 17: Anfrageerweiterung des *Clarification*- Moduls von MySentient Answer nach [Curtis et al. 2005, S. 4]

Mögliche Erweiterungen werden dem Benutzer zur Auswahl dargelegt. In der ersten Auswahl wird *annual percentage rate* durch das semantisch ähnliche *interest rate* ersetzt. Das Akronym *APR* in der zweiten Auswahl wird durch *annual percentage rate* ersetzt. Cyc enthält also in den Synonymgruppen auch Abkürzungen. Auswahl drei ersetzt *auto* durch sein Synonym *car* und die vierte Auswahl formuliert die Frage etwas generischer, indem *auto* weggelassen wird.

[Curtis et al. 2005, S. 9] planen, das voll funktionsfähige QA-System noch zu erweitern, indem bessere Disambiguierungsverfahren oder Anaphernresolutionsmethoden eingesetzt werden.

Ein Antwortextraktionssystem für die deutsche Sprache ist InSicht, bei dem unter anderem HaGenLex, Namenslexika und grosse Textkorpora zum Einsatz kommen

[33] `http://www.mysentient.com`.

[Hartrumpf 2006]. HaGenLex wird dazu verwendet, die in Anfragen und Dokumenten explizit und implizit enthaltene Information in so genannte semantische Netze abzufüllen, woraus ebenfalls **konzeptuelle Indexierung** resultiert. Die Anfrage wird mit ähnlichen Netzen erweitert und dann mit der Dokumentensammlung abgeglichen [Hartrumpf 2005]. InSicht hat am Cross-Language Evaluation Forum[34] teilgenommen und war 2004 das beste System für die deutsche Sprache [Hartrumpf 2006]. [Hartrumpf 2005, S. 7] weisen gleichwohl auf Mängel hin. So führt beispielsweise unzureichende linguistische Vorverarbeitung und fehlerhafte Anaphernresolution zu einer niedrigen Ausbeute.

6.2.3 Einsatz in der Maschinellen Übersetzung

6.2.3.1 Mögliche Methoden und Ansätze

[Dorr et al. 1998] schlagen ein Übersetzungsverfahren vor, das EuroWordNet und lexikalisch-konzeptuelle Strukturen (LCS) einsetzt. LCS sind semantische Repräsentationen von Verben basierend auf den Klassen von [Levin 1993], welche zur **konzeptuellen Indexierung** zur Hand gezogen werden. Der Interlinguale Index von EuroWordNet übernimmt dabei die Verknüpfungen zu anderen Sprachen. Beide Ressourcen liefern im Übersetzungsprozess Informationen, die richtige Lesart eines mehrdeutigen Wortes zu finden.

Einen Schritt weiter geht der Ansatz, zur Maschinellen Übersetzung das **Interlingua-Modell** zu verwenden und so das Problem der Übersetzungsambiguitäten zu lösen. Da es sich bei einer Interlingua um eine sprachunabhängige Repräsentation handelt, können mehrsprachige Wortnetze diese Funktion übernehmen. [Boas 2005] schlägt vor, parallele FrameNet-Fragmente von verschiedenen Sprachen zu verknüpfen. So entsteht ein **multilinguales FrameNet,** bei dem es sich um eine Kombination einer Interlingua mit einem transferbasierten System handelt. [Boas 2005, S. 465] spricht dann von einer interlingualen Repräsentation, die für die Maschinelle Übersetzung eingesetzt werden kann.

Auch **EuroWordNet** kann als Interlingua eingesetzt werden. [Reeder et al. 2004] beschreiben den Aufbau einer Interlingua und von semantisch annotierten Korpora sechs verschiedener Sprachen. Die erste der drei Ebenen der Interlingua erfasst syntaktische Informationen wie Wortarten oder grammatikalische Funktionen. Die zweite verknüpft die lexikalischen Einheiten mit Konzepten und ersetzt syntaktische

[34] Das Cross-Language Evaluation Forum, kurz CLEF, hat eine Testinfrastruktur für multilinguale IR- und QA-Systeme entwickelt [CLEF 2006].

Relationen durch semantische Rollen. Die Konzepte und semantischen Relationen sind dabei zu einem grossen Teil von der Omega-Ontologie übernommen worden. Die dritte Ebene ist die eigentliche Interlingua. Sie ist sprachunabhängig und befasst sich beispielsweise mit verschiedenen Lexikalisierungsmustern in den Sprachen. Auch die parallelen Korpora werden nach dieser Vorgehensweise annotiert. Korpora und Interlingua können in der Folge nicht nur für Maschinelle Übersetzungssysteme, sondern für weitere mehrsprachige Anwendungen eingesetzt werden.

6.2.3.2 Maschinelles Übersetzungssystem mit Wortnetzen

[Orasan et al. 2005] beschreiben eine Komponente des kommerziellen Übersetzungssystems Translution Central der Firma Translution[35], das E-Mails übersetzt. Zurzeit sind Übersetzungen zwischen den fünf Sprachen Englisch, Französisch, Deutsch, Spanisch und Italienisch möglich. Das Modul des Translation Checkers unterstützt den Übersetzungsprozess, indem der Benutzer beim Abschicken der Mails spezifiziert, welche Lesart eines ambigen Wortes er in seinem E-Mail intendiert. Die Auflistung in Beispiel 26 von [Orasan et al. 2005, S. 206] zeigt die Auswahl, wenn ein Benutzer das Verb *to address* verwendet.

(26) a. to speak to someone formally
b. to put an address on something
c. to deal with something particular

Aus den drei möglichen Definitionen wählt er die beabsichtigte Lesart aus. Das disambiguierte Verb kann nun korrekt in die Zielsprache übersetzt werden.

Das Wörterbuch des Translation Checkers basiert auf dem Interlingualen Index von **EuroWordNet.** Würde man alle Lesarten für ambige Wörter anzeigen, wäre der Benutzer überfordert. Um den Polysemiegrad zu reduzieren, wurden zu spezifische oder seltene Lesarten des ILI entfernt, sehr nahe liegende Lesarten zusammengeführt und fehlende Lesarten ergänzt [Orasan et al. 2005, S. 206]. Zusätzlich zu den anfallenden Arbeiten am ILI mussten für einige WordNets Wörterbuchdefinitionen ausgearbeitet oder erweitert werden.

Um die Anzahl ambiger Wörter weiter zu reduzieren, mussten verschiedene Filter eingesetzt werden [Orasan et al. 2005, S. 207f.]. Durch Eigennamen-Erkennung können Eigennamen wie *Bill Gates* identifiziert und so 13 Lesarten von *bill* und 7

[35] `http://www.translution.com`.

Lesarten von *gate* eliminiert werden. Ein weiterer Filter erkennt diejenigen Ambiguitäten, die eins zu eins in die Zielsprache übersetzt werden können. Beispielsweise besteht zwischen dem englischen *opponent* und dem französischen *adversaire* eine solche Beziehung. Durch diese Filter konnte die durchschnittliche Lesartenanzahl pro Wort von rund 3,5 auf 2,2 reduziert werden [Orasan et al. 2005, S. 210].

[Orasan et al. 2005, S. 211] planen, ihr System durch noch bessere Wörterbücher zu ergänzen und den Einfluss des Translation Checkers auf den ganzen Übersetzungsprozess zu evaluieren.

6.2.4 Einsatz in der Rechtschreibekorrektur

6.2.4.1 Mögliche Methoden und Ansätze

Die Korrektur von Nichtwörtern oder Wortverwechslungen wird durch Berücksichtigung von Ähnlichkeitsmassen ermöglicht. Dahinter verbirgt sich die Idee, dass es eher unwahrscheinlich ist, dass Fehler mit anderen Wörtern des Textes in einer semantischen Ähnlichkeitsbeziehung stehen.

6.2.4.2 Korrekturprogramm mit Wortnetzen

Der Korrekturalgorithmus von [Agirre et al. 1994] für Baskisch und Französisch korrigiert primär **Nichtwörter.** Er ergänzt ein bestehendes Korrekturprogramm, indem die Korrekturvorschläge gefiltert werden. Das Auswahlverfahren basiert auf dem **Mass konzeptueller Dichte** (vgl. 6.1.2.2), wobei für die Berechnung ein Wortnetz zum Einsatz kommt, das in dieser Arbeit nicht vorgestellt wurde. Für den fett markierten Tippfehler im französischen Beispielsatz 27 von [Agirre et al. 1994, S. 6] wurden die Wörter *chef, cher, chez, chié, chieé, chéri* und *chic* vorgeschlagen.

(27) le **cheé** de la police reunit vingt hommes sur la place du village

Der Algorithmus der Autoren identifiziert nun *chef* als den wahrscheinlichsten Kandidaten, da dieser den anderen Wörtern semantisch am nächsten steht.

[Hirst und Budanitsky 2005] schlagen vor, **Wortverwechslungen**[36] zu korrigieren, indem WordNet einbezogen wird. Erst werden alle Wörter in einem Textfenster disambiguiert. Ihre Methode sucht nun nach Wörtern, die nur einmal vorkommen und deren Lesarten nicht semantisch ähnlich sind mit allen anderen Wörtern. Dazu

[36] Die Autoren sprechen durchgängig von Malapropismen.

evaluierten die Autoren verschiedene Ähnlichkeitsmasse und entschlossen sich dann zum Einsatz des Masses von [Jiang und Conrath 1997]. Findet das System einen potenziellen Fehler, also ein Wort, das den Kontextwörtern nicht ähnlich ist, sucht der Algorithmus nach Wörtern, die zum möglicherweise fehlerhaften Wort eine geringe Levenshtein-Distanz und zu den Kontextwörtern eine **semantische Ähnlichkeit** aufweisen. Gibt es ein solches Wort, wird es als Korrekturvorschlag dargelegt.

Ihr System soll also Fehler wie in Beispiel 28 von [Hirst und Budanitsky 2005, S. 88] erkennen und den Benutzer auf die richtige Schreibweise (in eckigen Klammern) aufmerksam machen.

(28) It is my sincere **hole** [hope] that you will recover swiftly.

Da es kein Korpus gibt, das Wortverwechslungen enthält, haben die Autoren selber eines erstellt und in Zeitungsartikeln Wörter durch Varianten ersetzt. Zur Evaluation verwendeten sie die üblichen IR-Masse. Treffergenauigkeit ist beispielsweise die Anzahl richtig erkannter Fehler über der Anzahl vermuteter Fehler. Die Autoren schätzen ihr System mit einer Ausbeute von 50% und Treffergenauigkeit von 20% als praktisch einsetzbar ein. Sie halten fest, dass auch eine tiefe Treffergenauigkeit vertretbar ist, denn ein Benutzer akzeptiere es, drei von vier Korrekturvorschlägen abzulehnen.

Beim Korrekturprogramm von [Hirst und Budanitsky 2005] handelt es sich um einen Prototypen. Die Autoren merken an, dass neben dem Korrekturalgorithmus eine benutzerfreundliche Darstellung des Korrekturvorschlages entscheidend sei. Wie eine solche Mitteilung aussehen könnte, ist in Abbildung 18 dargestellt.

Wrong word?

Ontologist means someone who studies the nature of existence. Did you mean *oncologist*, someone who studies or treats cancer?

Abbildung 18: Korrekturvorschlag für eine Wortverwechslung basierend auf [Hirst und Budanitsky 2005, S. 107]

6.2.5 Evaluation und Ausblick

6.2.5.1 Information Retrieval

Bringt der Einsatz von Lesartendisambiguierung überhaupt etwas in IR-Systemen? [Sanderson 1994] zeigt, dass Disambiguierung nur eine geringe Verbesserung der Performanz zur Folge hat. Wenn die Lesartendisambiguierung zu weniger als 90% genau ist, sind die Resultate schlechter als ohne jegliche Disambiguierung. [Gonzalo et al. 1998a, S. 42] relativieren, dass sich weniger als 10% Disambiguierungsfehler nicht negativ auf die Performanz auswirken. Bei 30–60% Disambiguierungsfehlern nimmt die Performanz zwar drastisch ab, führt aber nicht zu schlechteren Resultaten als herkömmliche IR-Systeme.

Am meisten wurde WordNet im IR eingesetzt. Obwohl dieses Wortnetz eine Vielzahl an Informationen kodiert, fehlen beispielsweise häufig für das IR zentrale Beziehungen über Wortartengrenzen [Gonzalo et al. 1998b, S. 190]. **Fehlende topikalische Relationen,** die in der Beschreibung des Tennis-Problems angesprochen wurden (vgl. 3.7.1), verringern insbesondere die Effizienz ähnlichkeitsbasierter Ansätze. Ein weiterer Nachteil von WordNet ist, dass pro Wort **zu viele Lesarten** unterschieden werden. Dies erhöht die Schwierigkeit unnötig, die korrekte WordNet-Lesart ausfindig zu machen.

Dieses Problem wollten [Mihalcea und Moldovan 2001] lösen, indem sie automatisch eine **grobkörnigere Version von WordNet** erstellten. Der durchschnittliche Polysemie-Grad der Wörter konnte so um rund einen Viertel reduziert werden. Der Einsatz des neuen WordNet in sprachverarbeitenden Systemen scheint vielversprechend, insbesondere wenn man nebenbei den Effekt der Lesartenreduktion beim Maschinellen Übersetzungssystem von [Orasan et al. 2005] vor Augen hält (vgl. 6.2.3.2).

Auch die Entwickler von EuroWordNet gingen das Problem der Feinkörnigkeit von WordNet an, indem sie die **Polysemie** für Nomen und Verben **reduzierten.** Ähnliche Lesarten werden in einem so genannten *sense cluster* aufgenommen [Peters et al. 1998]. So steigt die Treffermenge einer Suchanfrage nicht mehr exponentiell, wenn die Anfrage mit Synonymen erweitert wird. EuroWordNet geht auch das Problem fehlender Relationen über die Wortartengrenzen an, indem die verschiedenen Wortarten nicht in getrennten Subnetzen erfasst werden und wortartenübergreifende Relationen erlaubt sind.

Topikalische Relationen können nicht nur aus EuroWordNet durch Domänenlabels extrahiert werden. Auch WordNet bietet neben den Domänen-Relationen eine Mög-

lichkeit, solche Angaben zu erhalten. Der **Einsatz von lexikalischen Ketten** kann helfen, die Ausbeute im Dokumentenretrieval zu erhöhen, indem die Anfrage mit topikalisch verwandten Wörtern erweitert wird. Sie helfen auch, die Präzision zu verbessern, indem Weltwissen, das aus lexikalischen Ketten abgeleitet wird, einbezogen wird.

[Gurevych 2005, S. 296] hält fest, dass ihr auf semantischer Ähnlichkeit basierender IR-Ansatz etwa gleich gut abschneidet wie ein konventionelles System. Weiter bemerkt sie, dass nur die Berücksichtigung zu einem hohen Grad ähnlicher Wörter wie Synonyme oder direkte Hyponyme positive Auswirkungen auf die Performanz hat. Bei solchen Vergleichen darf nie vergessen werden, dass der Lösungsansatz tatsächlich erfolgversprechend ist, die computerlinguistische Vorverarbeitung der natürlichen Sprache wie beispielsweise das Stemming aber meist fehlerhaft ist und daher Erfolge verhindert.

Die mit FrameNet-Strukturen annotierten deutschen Texte plant [Fliedner 2004] für das QA zu verwenden. Der Einsatz im IR ist aber auch denkbar, insbesondere, wenn sein System noch durch die hoffentlich bald verfügbaren Selektionsbeschränkungen von GermaNet erweitert wird.

IR-Systeme im Kontext des Internets können auch ohne den Einsatz von lexikalisch-semantischen Ressourcen gute Resultate liefern. Die Ausbeute ist zwar sehr gering, doch scheint dies nicht problematisch zu sein, zumal viel Information auf dem Internet redundant ist. Was auch immer man im Suchfeld eingibt, man bekommt meist viele Treffer. Wer nicht nur Wert auf eine hohe Ausbeute legt, sondern Treffergenauigkeit wünscht, darf nicht auf Lesartendisambiguierung verzichten. Wenn automatische Disambiguierungsmethoden noch nicht die erwünschten Resultate erzielen, müsssen Anwender die Ambiguität auflösen. Entwickler sind also gefordert, bei Suchmaschinen die Möglichkeit einzubauen, den **Benutzer einzubeziehen.** Eine Variante hat die Suchmaschine Clever Search gezeigt. Nur so bietet sich die Möglichkeit, auf die Suchanfrage nach „Jaguar“ eine hohe Treffergenauigkeit zu erzielen und diejenigen Dokumente zuoberst aufzufinden, die sich tatsächlich mit der beabsichtigten Lesart befassen.

6.2.5.2 Antwortextraktion

Die vorgestellten Überlegungen und Systeme für die Antwortextraktion schlagen in der Regel zwei Lösungsansätze vor. Einerseits soll die Anfrage durch semantisch oder topikalisch verwandte Begriffe erweitert werden, andererseits legen viele Autoren Wert auf konzeptuelle Indexierung.

Konzeptuelle Indexierung wird häufig durch die **fehlende Abdeckung lexikalisch-semantischer Ressourcen** erschwert. Der Ansatz von [Fliedner 2004] verspricht erst Erfolg, wenn einerseits FrameNet mehr Frames und lexikalische Einheiten abdeckt und andererseits Angaben zu Selektionsbeschränkungen von GermaNet-Verben erhältlich sind.

Da automatische Lesartendisambiguierung nicht immer zuverlässige Resultate liefert, vertrauen auch Systeme wie MySentient Answer die Disambiguierung dem Benutzer an.

6.2.5.3 Maschinelle Übersetzung

Der Einsatz von mehrsprachigen Wortnetzen wie EuroWordNet oder das geplante multilinguale FrameNet für die Übersetzung ist häufig diskutiert worden. Der Interlinguale Index bzw. die sprachunabhängigen Frame-Strukturen können entweder als Übersetzungshilfsmittel oder als **Interlingua** eingesetzt werden. Das Interlingua-Modell hat zum Vorteil, dass in einem Übersetzungssystem neue Sprachen inkrementell hinzugefügt werden können und nicht nach Äquivalenzrelationen zu jedem bereits erfassten Sprachpaar verlangen.

Auch die Maschinelle Übersetzung ist stark auf zuverlässige Disambiguierungsalgorithmen angewiesen. Solange diese nicht präzis genug sind, schafft der **Einbezug des Benutzers** in den Disambiguierungsprozess Abhilfe. [Orasan et al. 2005] machen einen Vorschlag, das Problem lexikalischer Ambiguitäten in einem kommerziellen System so zu lösen.

Laut [Dorr 2006, persönliche Information] ist in den letzten Jahren nicht viel zum Einsatz von Wortnetzen in der Maschinellen Übersetzung geforscht worden. Es bleibt zu hoffen, dass der Einsatz des auf FrameNet-Strukturen basierenden Prototypen von [Boas 2005] neue Inputs und Anreize für die Forschung liefert.

6.2.5.4 Korrektursysteme

Mit dem Ähnlichkeitsmass kann bereits ein Teil der Nichtwörter und Wortverwechslungen erkannt werden. Um diese Algorithmen noch zu verbessern, sollten laut [Hirst und Budanitsky 2005, S. 110] die eingesetzten lexikalisch-semantischen Ressourcen vermehrt Verknüpfungen zwischen verschiedenen Wortarten enthalten. Auch im Bereich der Rechtschreibekorrektur wird die Feinkörnigkeit von WordNet kritisiert. Beide angesprochenen Mängel verlangen nach einer **Anpassung der Wortnetze.** Das neue WordNet von [Mihalcea und Moldovan 2001] könnte das

Problem der Feinkörnigkeit lösen. Der Einsatz von Ressourcen wie EuroWordNet bietet eine Lösung für das erstgenannte Problem, indem wortartenübergreifende Relationen kodiert werden.

Eine weitere Perspektive bei der Korrektur von Tippfehlern wäre, statt das Ähnlichkeitsmass die semantische Verwandtschaft zu betrachten. Zusätzlich zu den Unter- und Oberbegriffsrelationen gibt der **Einbezug weiterer Sinnrelationen** wie Meronymie oder Antonymie Aufschluss über mögliche Fehler.

Auch wenn das Ziel vieler Forschungsbestrebungen automatische Lesartendisambiguierung ist, können die vorgeschlagenen Lösungsansätze jetzt schon für die Disambiguierung verwendet werden. Viele der hier vorgestellten Ansätze und Systeme überlassen die Entscheidung über die beabsichtigte Lesart den Anwendern und gehen so die erwähnten Probleme an. Grundlage in beiden Fällen ist das **Wortbedeutungsinventar** von Wortnetzen, mit dem die möglichen Lesarten benannt werden können. Für die Disambiguierung durch den Anwender helfen zudem **Beispielsätze und Wörterbuchdefinitionen** von Wortnetzen bei der Auswahl. Automatische Disambiguierung verwendet zudem die **Relationen zwischen den Konzepten,** um beispielsweise für die Disambiguierung wichtige Wortähnlichkeit zu berechnen. In jeder Hinsicht ist der Einsatz von Wortnetzen unerlässlich und hat gerade mit der geplanten Verbesserung der bestehenden lexikalisch-semantischen Ressourcen Zukunft.

7 Anaphernresolution mit Wortnetzen

Neben Lesartendisambiguierung ist die Auflösung von Anaphern eine nicht zu unterschätzende Aufgabe in computerlinguistischen Anwendungen. Wie in Kapitel 5 dargelegt, bedürfen insbesondere Antwortextraktions- und Maschinelle Übersetzungssysteme einer Anaphernresolutionskomponente.

In diesem Kapitel wird gezeigt, wie Anaphern mithilfe von lexikalisch-semantischen Ressourcen aufgelöst werden können und damit die Performanz computerlinguistischer Anwendungen verbessert werden kann.

7.1 Ansätze

In der Anaphernresolution kann Information aus Wortnetzen in zweierlei Hinsicht nützlich sein. Einerseits wird semantische Information über lexikalisch-semantische Relationen, andererseits über Selektionsbeschränkungen einbezogen.

7.1.1 Einbezug von Assoziationsrelationen

[Vieira und Poesio 2000, S. 558] konzentrieren sich auf die Auslösung von assoziativen Anaphern, die mit ihrem Antezedens in einer lexikalisch-semantischen Beziehung wie Synonymie, Hyperonymie oder wie in Beispiel 29 Meronymie stehen.

(29) the flat ... the living room.

Eine assoziative Relation zu finden, erfordert laut [Vieira und Poesio 2000, S. 559] oft komplexe Suchvorgänge in der WordNet-Hierarchie. Besonders unbefriedigend fanden die Autoren die **Berücksichtigung der Meronymie-Relation.** [Poesio et al. 2002, S. 1221] erklären dies dadurch, dass in WordNet viele Einträge und Relationen fehlen. So besteht beispielsweise eine Meronymie-Relation zwi-

schen *floor* und *house,* aber keine zwischen *room* und *house.* Dass die Meronymie-Information in WordNet nur ungenügend erfasst ist, ist insofern bedauerlich, dass assoziative Anaphern und ihre Antezedenten häufig genau in dieser Beziehung stehen.

Um diesen Mangel aufzuheben und Informationen zu Meronymie-Relationen anderweitig zu erhalten, schlagen [Meyer und Dale 2002] vor, **assoziative Konstruktionen** aus Korpora zu extrahieren und dann mithilfe von WordNet zu generalisieren. Ein Beispiel einer solchen Struktur ist *the head of the stingray.* Aus WordNet folgt, dass Köpfe Körperteile sind und Stachelrochen Fische. Aus den obigen Belegen wird generalisiert, dass Fische Körperteile haben. Solche Informationen können helfen, dass der Ausdruck *the tuna* aus Beispiel 12 ein mögliches Antezedens für die assoziative Anapher *the gills* ist [Meyer und Dale 2002, S. 3].

Auch [Poesio et al. 2002] wollen aus Korpora extrahierte Konstruktionen[37] extrahieren. In einer Evaluation konnte der Einbezug so erlangter Teil-Ganzes-Beziehungen die Performanz im Vergleich zum Einbezug von WordNet-Relationen um mehr als das Doppelte steigern [Poesio et al. 2002, S. 1223].

7.1.2 Einbezug von Selektionsbeschränkungen

[Wagner 2005] will Selektionsbeschränkungen aus Wortnetzen nicht nur für die Auflösung von Ambiguitäten, sondern auch für die Auflösung von Anaphern einsetzen. In Beispiel 30 von [Wagner 2005, S. 2] kann mithilfe von Selektionsbeschränkungen das richtige Antezedens für das Relativpronomen *which* gefunden werden.

(30) Claire enjoyed the view of the cows on top of the hills, which were mooing gently.

Die Selektionsbeschränkungen des Verbs im Relativsatz legen fest, dass *moo* ein belebtes Agens verlangt. Das Antezedens *the hills,* welches aufgrund syntaktischer Kriterien ebenfalls zur Auswahl stünde, kann somit ausgeschlossen werden.

7.1.3 Evaluation der Ansätze

Die Extraktion assoziativer Konstruktionen aus Korpora bietet dem Problem fehlender Meronymie-Relationen in WordNet Abhilfe. Ungeklärt bleibt jedoch die Frage

[37] [Poesio et al. 2002] sprechen nicht von assoziativen Konstruktionen, sondern von syntaktischen Mustern.

nach dem **angemessenen Generalisierungsgrad,** denn die Generalisierung ist nicht immer für alle Subtypen gültig. Wenn man beispielsweise weiss, dass Katzen Beine haben und Katzen Tiere sind, folgt daraus nicht, dass alle Tiere Beine haben. Auf Schlangen trifft die Generalisierung beispielsweise nicht zu.

Der Einsatz von Selektionsbeschränkungen ist vielversprechend. Die praktische Umsetzung steckt allerdings noch in den Kinderschuhen, da die wenigsten Wortnetze solche Angaben erfasst haben.

7.2 Einsatz in computerlinguistischen Anwendungen

Anaphernresolutionsverfahren, welche Informationen aus Wortnetzen benützen, können in Anwendungen der Antwortextraktion und der Maschinellen Übersetzung eingesetzt werden.

7.2.1 Einsatz in einem Antwortextraktionssystem

[Vicedo und Ferrández 2000, S. 556] machen darauf aufmerksam, dass die wenigsten QA-Systeme über Anaphernresolutionskomponenten verfügen. Es konnte nur ein solches System gefunden werden, dass Anaphernresolution mithilfe semantischen Wissens durchführt. Dabei handelt es sich um das kommerzielle QA-System My-Sentient Answer (vgl. 6.2.2.2), welches Pronominal- und NP-Anaphern behandelt [Curtis et al. 2005, S. 9]. Die Auflösung geht so vor, dass das wahrscheinlichste Antezedens ausgewählt wird, indem **linguistisches und semantisches Wissen** eingesetzt wird. Unter linguistischem Wissen versteht man die Eliminierung von Kandidaten aufgrund von Kasus- oder Numerusinkongruenz. Auf der semantischen Seite wird vermieden, dass Anaphern wie *he* auf ein Antezedens wie *my mother* verweist, da Letzteres im internen Diskursmodell als Frau repräsentiert ist.

7.2.2 Einsatz in Maschinellen Übersetzungssystemen

Auch im Bereich der Maschinellen Übersetzung ist der Einsatz von Anaphernresolution entscheidend. [Mitkov 1999, S. 159] führt einige aussichtsreiche Arbeiten vor. Bei der Vorstellung ihres Maschinellen Übersetzungssystems sprechen [Orasan et al. 2005] das Problem der Auflösung anaphorischer Ausdrücke an keiner Stelle an. So können hier leider keine Maschinellen Übersetzungssysteme aufgeführt werden, die zur Auflösung von Anaphern Informationen von Wortnetzen

miteinbeziehen.

7.2.3 Evaluation und Ausblick

Jeder Typ von Anapher verlangt eine eigene Resolutionsmethode. Der Einsatz von Wortnetzen ist insbesondere für die Auflösung von assoziativen und pronominalen Anaphern vielversprechend. Bei assoziativen anaphorischen Ausdrücken sind Anapher und Antezedens häufig durch lexikalisch-semantische Relationen miteinander verbunden und können daher mithilfe von Informationen aus Wortnetzen aufgelöst werden. Pronominalanaphern können zu einem grossen Teil durch syntaktisches Wissen aufgelöst werden. In Fällen, in denen dies nicht ausreicht, können Informationen über Selektionsbeschränkungen zum richtigen Antezedens führen.

[Vicedo und Ferrández 2000] befassen sich mit der Frage, ob die Auflösung von pronominalen Anaphern einen Mehrwert bringt und experimentieren mit einem QA-System. Im ersten Testlauf werden Pronominalanaphern in den Dokumenten und Anfragen nicht mit ihren Antezedenten verbunden, im zweiten hingegen erfolgt die Verknüpfung mithilfe eines Resolutionsmoduls. Der Prozentsatz korrekt beantworteter Fragen steigt dadurch um 12 Prozentpunkte auf 54%.

Wenn bei der Auflösung von Pronominalanaphern wie in Beispiel 14 **semantisches Wissen** erforderlich ist, kann der Einsatz von Wortnetzen die Performanz von QA-Systemen erhöhen. Angaben zu Selektionsbeschränkungen helfen, dass die Resolutionsverfahren über das notwendige semantische Wissen verfügen.

Ähnliches gilt für die Maschinelle Übersetzung. Informationen aus Wortnetzen können helfen, Anaphern aufzulösen und so zu einer besseren Übersetzungsqualität beizutragen.

Für den Einsatz in beiden computerlinguistischen Anwendungsgebieten sind Anaphernresolutionsmodule, die Informationen aus Wortnetzen berücksichtigen, noch nicht ausgereift. Die fortschreitende Erfassung semantischer Information stimmt zuversichtlich und lässt hoffen, dass zuverlässige Resolutionsverfahren in einigen Jahren vorhanden sein werden.

8 Rückblick, Fazit und Ausblick

Wortnetze modellieren den Grundwortschatz einer Sprache auf konzeptueller Ebene und verbinden Wörter durch elementare lexikalisch-semantische Relationen miteinander. In dieser Arbeit wurden verschiedene Wortnetze vorgestellt, die sich in ihrem Aufbau, den enthaltenen Sinnrelationen und weiteren semantischen Informationen teilweise stark unterscheiden.

Verschiedene sprachverarbeitende Systeme stossen auf ähnliche Probleme. Vor allem ist dies auf ambige Strukturen zurückzuführen. IR-Systeme, Maschinelle Übersetzungs- oder Antwortextraktionssysteme liefern wegen mehrdeutiger Wörter schlechte Resultate und die Wahl der falschen Lesart führt zu ungenügenden Resultaten im Verlauf der weiteren Verarbeitung. Auch das falsche oder unzureichende Auflösen von Anaphern ist häufig Ursache von mangelhafter Performanz. Ausgereifte Lesartendisambiguierungs- und Anaphernresolutionsverfahren sind daher nicht zu vernachlässigende Komponenten computerlinguistischer Anwendungen.

Der Einsatz lexikalisch-semantischer Ressourcen im IR bietet neue Such- und Indexiermöglichkeiten. Vor allem das Erstellen eines konzeptuellen Index scheint aussichtsreich, da damit Anfrage und Dokument nicht auf Wortebene, sondern auf der Bedeutungsebene verglichen werden können. Methoden der Suchraumerweiterung und -einschränkung helfen, fehlende und ungenaue Treffer zu vermeiden.

In der Antwortextraktion und der Maschinellen Übersetzung werden ähnliche Techniken wie im IR eingesetzt. Neben lexikalischen Ambiguitäten müssen hier auch weitere Ambiguitätstypen und Anaphern aufgelöst werden, um gute Resultate zu erhalten. Dazu werden semantische Informationen aus Wortnetzen in Form von Selektionsbeschränkungen verwendet.

Schliesslich thematisierte diese Arbeit auch das Anwendungsgebiet der Rechtschreibekorrektur, wo Nichtwörter und Wortverwechslungen durch den Einbezug von Angaben aus Wortnetzen erkannt und Korrekturvorschläge dargelegt werden können.

Methoden zur automatischen Lesartendisambiguierung stellen zurzeit noch nicht in jeder Hinsicht zufrieden. Das Problem liegt vor allem bei den verwendeten Ressour-

cen. Viele Wortnetze haben zum Nachteil, dass sowohl Einträge als auch Relationen zwischen den Einträgen fehlen oder die unterschiedenen Lesarten zu feinkörnig sind, um effizient eingesetzt werden zu können. Solange die automatische Lesartendisambiguierung noch nicht besser ausgereift ist, stellt der Einbezug des Anwenders eine sinnvolle Alternative dar, dieses Problem zu umgehen. Hier wird die in Wortnetzen vorhandene Information dazu verwendet, die Auswahlmöglichkeiten zu benennen und dem Benutzer die Auswahl durch Wörterbuchdefinitionen zu erleichtern. In vielen der vorgestellten Systeme führt die Disambiguierung durch Benutzer zu guten Ergebnissen.

Systematische Angaben von Selektionsbeschränkungen und Subkategorisierungsrahmen erleichern die Auflösung von Anaphern und verschiedenen Ambiguitätstypen. Solche Informationen zu erfassen steht zuoberst auf der Prioritätenliste von GermaNet. Man kann hoffen, dass die Daten bald verfügbar sind und sprachverarbeitende Systeme davon profitieren können.

Jedes Wortnetz ist typischerweise für die Lösung eines Teilproblems prädestiniert. Häufig ermöglicht die Kombination verschiedener lexikalisch-semantischer Ressourcen, Synergien zu nutzen und so computerlinguistische Anwendungen zu verbessern. Für das konzeptuelle Indexieren schlagen [Shi und Mihalcea 2005] beispielsweise eine Kombination dreier Ressourcen vor, darunter die grossen Bestände von WordNet und FrameNet.

Der Einsatz von Wortnetzen ist nicht nur für die in dieser Arbeit vorgestellten Anwendungen interessant. Auch im Bereich der Textverarbeitung und -zusammenfassung, E-Learning oder im Semantic Web können lexikalisch-semantische Ressourcen eingesetzt werden. Zur Anwendung solcher Ressourcen in diesen Gebieten gibt es zwar interessante Ansätze, doch noch sehr wenig kommerzielle Systeme. Hoffentlich fühlen sich viele Computerlinguisten angesprochen, Wortnetze in diesen Fachgebieten anzuwenden.

Da momentan intensiv an der Erstellung lexikalisch-semantischer Ressourcen, an verbesserter Abdeckung und an der Neuerfassung weiterer semantischer Information gearbeitet wird (beispielsweise bei GermaNet), darf man in den nächsten Jahren mit weiteren Forschungsresultaten und funktionstüchtigen Systemen rechnen. Der Einsatz lexikalisch-semantischer Ressourcen für sprachverarbeitende Anwendungen bietet also in vielerlei Hinsicht Perspektiven für die Zukunft und wird weiterhin ein wichtiges Thema der Computerlinguistik bleiben.

Glossar

Algorithmus
Vorgang mit Handlungsanweisungen zur Lösung eines Problems.

Ambiguität
Mehrdeutigkeit von sprachlichen Ausdrücken. Ambige Strukturen können lexikalischer, syntaktischer oder semantischer Natur sein. Lexikalische Ambiguitäten entstehen aufgrund von Polysemie oder Homonymie.

Anapher
Sprachliche Einheit, die für eine im Text vorangegangene sprachliche Einheit (Antezedens) steht.

Anaphernresolution
Verfahren zur Auflösung einer Anapher, d.h. Bestimmung des Antezedens einer Anapher.

Antezedens
Bereits erwähnte sprachliche Einheit, auf die sich eine Anapher bezieht.

Antonymie, Antonym
Lexikalisch-semantische Relation der Gegensätzlichkeit. Ein Antonym eines Begriffs ist das Gegenteil dieses Begriffs.

Antwortextraktion
Engl. *Question Answering* (QA). Fachgebiet, das auf eine natürlichsprachliche Abfrage Dokumente findet. Fragen werden mit denjenigen Textausschnitten beantwortet, welche die Antwort enthalten.

Ausbeute
Engl. *recall*. Evaluationsmesswert im Information Retrieval, der den Anteil relevanter zurückgelieferter Dokumente an allen relevanten Dokumenten misst.

Denotation
Begriff der Semantik, der den Bedeutungskern eines sprachlichen Ausdruckes bezeichnet.

Frame
Analyseeinheit der Frame-Semantik. Konstrukt an Kenntnissen über alltägliche Realitätsausschnitte wie Ereignisse, Situationen, Handlungen und Zustände.

F-Mass
Engl. *f-measure*. Evaluationsmesswert im Information Retrieval, der die Werte Ausbeute und Treffergenauigkeit mittelt. Wird oft verwendet, um die Performanz von IR-Systemen anzugeben.

Heuristik
Lehre zur methodischen Gewinnung neuer Erkenntnisse mithilfe der Erfahrung, oft auf Faustregeln beruhend.

Homonymie
Die Lesarten eines lexikalisch mehrdeutigen Begriffs haben keinen Bezug zueinander und haben selten gemeinsame morphosyntaktische Eigenschaften. Beispiel: „Kiefer“ (mit den Lesarten „Gebiss“/ „Baum“). Bei gleich lautenden Ausdrücken spricht man von Homophonie, bei gleich geschriebenen von Homographie.

Hyperonymie, Hyperonym
Lexikalisch-semantische Relation der Überordnung. Ein Hyperonym eines Begriffs ist ein Oberbegriff dieses Begriffs.

Hyponymie, Hyponym
Lexikalisch-semantische Relation der Unterordnung. Ein Hyponym eines Begriffs ist ein Unterbegriff dieses Begriffs.

Information Retrieval (IR)
Fachgebiet, das sich mit dem Auffinden von natürlichsprachlichen Dokumenten befasst. Beschäftigt sich mit der Indexierung, Suche und Wiedergewinnung von Information mittels verschiedener Methoden.

Konnotation
Damit bezeichnet man das Phänomen, dass der Bedeutungskern eines sprachlichen Ausdrucks durch zusätzliche Bedeutungsaspekte anders bewertet wird.

Korpus
Umfassende Textsammlung, die nach bestimmten Kriterien zusammengestellt worden ist.

Lesartendisambiguierung
Verfahren zur Auflösung von lexikalischen Ambiguitäten, d.h. Zuweisung der beabsichtigten Lesart.

Lexem
Gruppierung von lexikalischen Einheiten.

Lexikalische Einheit
Verbindung einer lexikalischen Form mit genau einer Bedeutung. Besteht aus mindestens einer semantischen Konstituente und mindestens einem Wort.

Lexikalisch-semantische Relation
Auch Sinnrelation, semantische Relation. Beziehung, die zwischen zwei oder mehreren lexikalischen Einheiten aufgrund ihrer lexikalisch-semantischer Eigenschaften gestiftet wird. Gängige Relationen sind: Synonymie, Hyponymie, Hyperonymie, Meronymie und Antonymie.

Lexikalisch-semantisches Netz
Auch Wortnetz. Datenstruktur, in der Wörter durch lexikalisch-semantische Relationen miteinander verknüpft sind. Bekanntestes maschinenlesbares Wortnetz ist WordNet.

Maschinelle Übersetzung
Automatische Übersetzung zwischen zwei Sprachen durch den Computer. Verschiedene Ansätze: Wort-für-Wort-Übersetzung, direkte Übersetzung, transferbasierte Übersetzung, Interlingua-basierte Übersetzung.

Meronymie, Meronym
Lexikalisch-semantische Relation, auch Partonymie, Teil-Ganzes-Beziehung. Ein Meronym eines Begriffs ist Teil dieses Begriffs.

Metapher
Eine Metapher bezeichnet eine bildhafte Verwendung eines Ausdruckes, welcher gegenüber seiner Basisbedeutung eine Sinnveränderung erfährt.

Ontologie
Ausdrucksmächtiges lexikalisch-semantisches Netz, kann zur Wissensrepräsentation verwendet werden.

Pertonymie, Pertonym
Lexikalisch-semantische Derivations-Relation, die ein Adjektiv mit einem vom gleichen Wortstamm hergeleiteten Wort verknüpft. Ein Pertonym eines Begriffs ist eine Derivationsform dieses Begriffs.

Plesionymie, Plesionym
Siehe Synonymie, Synonym.

Polysemie
Die Lesarten eines lexikalisch mehrdeutigen Begriffs haben einen Bezug zueinander und teilen sich häufig morphosyntaktische Eigenschaften. Polysemie entsteht, wenn sich die Bedeutung eines Lexems im Laufe der Zeit aufspaltet. Beispiel: „Schloss“ (mit den Lesarten „Schliessvorrichtung“ und „Gebäude“).

Selektionsbeschränkungen
Beschreiben das semantische Verhalten einer lexikalischen Einheit, meist von Verben. Beschränkungen, die ein Prädikat seinen Argumenten auferlegt.

Semantisches Feld
Auch Wortfeld, Menge gleichartiger und/oder gleichfunktionaler Prozesse oder Sachen, die einen bestimmten begrifflichen Bereich lückenlos abdecken.

Subkategorisierungsrahmen
Beschreiben das syntaktische Verhalten einer lexikalischen Einheit und legen die Anzahl und den Typ der Argumente fest.

Synonymie, Synonym
Bedeutungsgleichheit. Ein Synonym eines Begriffs hat dieselbe Bedeutung wie dieser Begriff. Weitere Unterteilung in absolute Synonymie (e.g. „beginnen“ - „anfangen“), propositionale Synonymie (e.g. „Geige“ - „Violine“) und Plesionymie (e.g. „neblig“ - „dunstig“).

Synset
Synonymgruppe in Wortnetzen der WordNet-Familie. Absolute, proportionale Synonyme und Plesionyme werden in Synsets zusammengefasst.

Taxonomie
Lexikalisch-semantisches Netz, in dem die Einträge in einer hierarchischen Struktur vorliegen.

Thematische Rolle
Auch semantische Rolle, Theta-Rolle. Beschreiben die Relationen zwischen dem Prädikat und seinen Argumenten. Gängigste Rollen: Agens (handlungsausführendes Argument), Patiens (von der Handlung betroffenes Argument).

Thesaurus
Lexikalisch-semantisches Netz, in dem die Einträge durch verschiedene lexikalisch-semantische Relationen miteinander verbunden sind.

Treffergenauigkeit
Engl. *precision*. Evaluationsmesswert im Information Retrieval, der den Anteil zurückgelieferter relevanter Dokumente an allen zurückgelieferten Dokumenten misst.

Valenz
Die Valenz legt fest, mit welchen Satzgliedern ein Verb ergänzt werden muss, um wohlgeformte Sätze zu bilden.

Verbalternation
Beschreibt einen Wechsel der Realisierung der Argumentstruktur eines Verbs.

Wortnetz
Siehe lexikalisch-semantisches Netz.

Literaturverzeichnis

[Agirre et al. 1994] Agirre, Eneko, X. Arregi, A. Diaz de Ilarraza und K. Sarasola (1994). *Conceptual Distance and Automatic Spelling Correction*. In: *Proceedings of the Workshop on Computational Linguistics for Speech and Handwriting Recognition*, S. 1–8, Leeds, England.

[Agirre und Rigau 1995] Agirre, Eneko und G. Rigau (1995). *A Proposal for Word Sense Disambiguation Using Conceptual Distance*. In: *International Conference on Recent Advances in Natural Language Processing*, Tsigov Chark, Bulgaria.

[Agricola und Agricola 1982] Agricola, Christiane und E. Agricola (1982). *Wörter und Gegenwörter*. Bibliographisches Institut, Leipzig, Vierte Aufl.

[Alonge et al. 1998] Alonge, Antonietta, N. Calzolari, P. Vossen, L. Bloksma, I. Castellon, M. A. Marti und W. Peters (1998). *The Linguistic Design of the EuroWordNet Database*. In: Vossen, Piek, Hrsg.: *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*, S. 81–112. Kluwer Academic Publishers, Dordrecht. Reprinted from Computers and the Humanities, 32(2–3).

[Antonya 2006] Antonya (2006). *Webseite der sprachübergreifenden Suchmaschine Antonya*. `http://antonya.irion.nl/antonya/web/main.do` [Stand: 1.5.2006].

[Boas 2005] Boas, Hans C. (2005). *Semantic Frames as Interlingual Representations for Multilingual Lexical Databases*. International Journal of Lexicography, 18(4):445–478.

[Budanitsky und Hirst 2001] Budanitsky, Alexander und G. Hirst (2001). *Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures*. In: *Proceedings of the 2nd meeting of the North American*

Chapter of the Association for Computational Linguistics (NAACL-2001). Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA.

[Burchardt et al. 2006] Burchardt, Aljoscha, K. Erk, A. Frank, A. Kowalski, S. Padó und M. Pinkal (2006). *The SALSA Corpus: a German Corpus Resource for Lexical Semantics*. In: *Proceedings of LREC*, Genoa. To appear.

[canoonet 2006] canoonet (2006). *Webseite der Online-Grammatik und Wörterbücher von canoo.net*. `http://www.canoo.net` [Stand: 1.5.2006].

[CELEX 2006] CELEX (2006). *Webseite des CELEX-Projekts*. `http://www.mpi.nl/world/celex/` [Stand: 1.5.2006].

[Chaffin 1992] Chaffin, Roger (1992). *The Concept of a Semantic Relation*. In: Lehrer, Adrienne und E. Kittay, Hrsg.: *Frames, Fields, and Contrasts. New Essays in Semantic and Lexical Organization*, S. 253–288. Lawrence Erlbaum, Hillsdale.

[CLEF 2006] CLEF (2006). *Webseite des Cross-Language Evaluation Forums*. `http://www.clef-campaign.org/` [Stand: 1.5.2006].

[Clough und Stevenson 2004] Clough, Paul und M. Stevenson (2004). *Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval*. In: Sojka, Petr, K. Pala, P. Srmz, C. Fellbaum und P. Vossen, Hrsg.: *Proceedings of the Second Global WordNet Conference*, S. 97–105, Brno, Czech Republic.

[Collins und Quillian 1969] Collins, A. M. und M. R. Quillian (1969). *Retrieval Time from Semantic Memory*. Journal of Verbal Learning and Verbal Behaviour, 8:240–247.

[Cristea 2001] Cristea, Dan (2001). *Anaphora*. In: *EUROLAN-2001. Creation and Exploitation of Annotated Language Resources*, Iasi, Romania. `http://www.racai.ro/EUROLAN-2001/page/resources/profs/Cristea/Anaphora.ppt` [Stand: 1.5.2006].

[Cruse 1986] Cruse, Alan (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

[Cruse 2004] Cruse, Alan (2004). *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford, London, Zweite Aufl.

[Curtis et al. 2005] Curtis, Jon, G. Matthews und D. Baxter (2005). *On the Effective Use of Cyc in a Question Answering System*. In: *Papers from the IJCAI Workshop on Knowledge and Reasoning for Answering Questions*, S. 1–10, Edinburgh, Scotland.

[Cycorp 2006] Cycorp (2006). *The Cyc Knowledge Base*. `http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc` [Stand: 1.5.2006].

[Dorr et al. 2004] Dorr, Bonnie, E. Hovy und L. Levin (2004). *Machine Translation: Interlingual Methods*. In: Brown, Keith, Hrsg.: *Encyclopedia of Language and Linguistics, Band 7*, S. 383–394. Elsevier, Oxford, Zweite Aufl.

[Dorr et al. 1998] Dorr, Bonnie, M. Martí und I. Castellón (1998). *Evaluation of EuroWordNet- and LCS-Based Lexical Resources for Machine Translation*. In: *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

[Erk 2005] Erk, Katrin (2005). *Frame Assignment as Word Sense Disambiguation*. In: *Proceedings of IWCS-6*, Tilburg University, Tilburg, The Netherlands.

[Fellbaum 1998a] Fellbaum, Christiane (1998a). *A Semantic Network of English Verbs*. In: Fellbaum, Christiane, Hrsg.: *WordNet: an Electronic Lexical Database*, S. 69–104. MIT Press, Cambridge.

[Fellbaum 1998b] Fellbaum, Christiane (1998b). *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge.

[Fellbaum 2002] Fellbaum, Christiane (2002). *Parallel Hierarchies in the Verb Lexicon*. In: *Proceedings of the OntoLex Workshop, LREC*, Las Palmas, Spain.

[Fillmore 1968] Fillmore, Charles J. (1968). *The Case for Case*. In: Bach, Emmon und R. Harms, Hrsg.: *Universals in Linguistic Theory*, S. 1–88. Holt, Reinhardt and Winston, New York.

[Fillmore 1977a] Fillmore, Charles J. (1977a). *The Need for a Frame Semantics in Linguistics*. Statistical Methods in Linguistics, 12:5–29.

[Fillmore 1977b] Fillmore, Charles J. (1977b). *Scenes-and-frames Semantics*. In: Zampolli, Antonio, Hrsg.: *Linguistic Structures Processing*, S. 55–81. North-Holland Publishing Company, Amsterdam.

[Fillmore 1985] Fillmore, Charles J. (1985). *Frames and the Semantics of Understanding*. Quaderni di Semantica, 6(2):222–254.

[Fillmore et al. 2003] Fillmore, Charles J., C. Johnson und M. Petruck (2003). *Background to FrameNet*. International Journal of Lexicography, 16:235–251.

[Fliedner 2004] Fliedner, Gerhard (2004). *Towards Using FrameNet for Question Answering*. In: *Proceedings of the LREC 2004 Workshop „Building Lexical Resources from Semantically Annotated Corpora“*, S. 61–65, Lissabon.

[Fragos et al. 2003] Fragos, Kostas, Y. Maistros und C. Skourlas (2003). *Word Sense Disambiguation Using WordNet Relations*. In: *Proceeding of the 1st Balkan Conference in Informatics*, Thessaloniki, Greek.

[FrameGrapher 2006] FrameGrapher (2006). *Webseite des FrameGrapher Tools des FrameNet-Projekts*. `http://framenet.icsi.berkeley.edu/FrameGrapher/` [Stand: 1.5.2006].

[FrameNet 2006] FrameNet (2006). *Webseite des Projekts Berkeley FrameNet*. `http://framenet.icsi.berkeley.edu/` [Stand: 1.5.2006].

[German-FrameNet 2006] German-FrameNet (2006). *Webseite des Projekts German FrameNet*. `http://gframenet.gmc.utexas.edu/` [Stand: 1.5.2006].

[GermaNet 2006] GermaNet (2006). *Webseite des GermaNet-Projekts*. `http://www.sfs.nphil.uni-tuebingen.de/lsd` [Stand: 1.5.2006].

[Geyken und Schrader 2006] Geyken, Alexander und N. Schrader (2006). *LexikoNet - a Lexical Database Based on Type and Role Hierarchies*. In: *Proceedings of LREC*, Genoa. To appear.

[Gilarranz et al. 1997] Gilarranz, Julio, J. Gonzalo und F. Verdejo (1997). *An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database*. In: *Proceedings of AAAI-97 Spring Symposium Cross-Language Text and Speech Retrieval*, S. 49–55, Stanford.

[Gonzalo et al. 1998a] Gonzalo, Julio, F. Verdejo, I. Chugur und J. Cigarrán (1998a). *Indexing with WordNet Synsets can Improve Text Retrieval*. In: *Proceedings of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing*, S. 38–44, Montreal.

[Gonzalo et al. 1998b] Gonzalo, Julio, F. Verdejo, C. Peters und N. Calzolari (1998b). *Applying EuroWordNet to Cross-Language Text Retrieval*. Computers and the Humanities, Double Special Issue on EuroWordNet, S. 185–207.

[Gross et al. 1989] Gross, Derek, U. Fischer und G. A. Miller (1989). *The Organization of Adjectival Meanings*. Journal of Memory and Language, 28:92–106.

[Gruber 1993] Gruber, Thomas (1993). *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5:199–220.

[Gurevych 2005] Gurevych, Irina (2005). *Anwendungen des semantischen Wissens über Konzepte im Information Retrieval*. In: *Proceedings of „Knowledge eXtended: Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten“*, S. 287–300, Jülich, Deutschland.

[Gurevych und Niederlich 2005] Gurevych, Irina und H. Niederlich (2005). *Computing Semantic Relatedness of GermaNet Concepts*. In: Fisseni, Bernard, H.-C. Schmitz, B. Schröder und P. Wagner, Hrsg.: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, S. 462–474. Peter Lang, Frankfurt am Main.

[GWA 2006] GWA (2006). *Webseite der Global WordNet Association*. `http://www.globalwordnet.org/` [Stand: 1.5.2006].

[Hampton 1982] Hampton, James A. (1982). *A Demonstration of Intransitivity in Natural Categories*. Cognition, 12:151–164.

[Harabagiu und Moldovan 1999] Harabagiu, Sandra und D. Moldovan (1999). *WordNet 2 – A Morphologically and Semantically Enhanced Resource*. In: *Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources*, S. 1–8, Maryland.

[Harabagiu et al. 2001] Harabagiu, Sandra, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, B. R., R. Girju, V. Rus und P. Morarescu (2001). *The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering*. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, S. 282–289.

[Hartrumpf 2005] Hartrumpf, Sven (2005). *University of Hagen at QA@CLEF 2005: Extending Knowledge and Deepening Linguistic Processing for Question Answering*. In: Peters, Carol, Hrsg.: *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*, Wien, Österreich. Centromedia.

[Hartrumpf 2006] Hartrumpf, Sven (2006). *Das deutschsprachige Frage-Antwort-System InSicht*. `http://pi7.fernuni-hagen.de/Frage-Antwort-System/Frage-Antwort-System.html` [Stand: 1.5.2006].

[Helbig 2001] Helbig, Hermann (2001). *Die semantische Struktur natürlicher Sprache. Wissensrepräsentation mit MultiNet*. Springer, Berlin.

[Hess 2005] Hess, Michael (2005). *Einführung in die Computerlinguistik I*. Offizielles Skript zur Überblicksvorlesung ECL I an der Universität Zürich, Wintersemester 2005/2006. `http://www.ifi.unizh.ch/CL/hess/classes/ecl1/ecl1.0.pdf` [Stand: 1.5.2006].

[Hirst und Budanitsky 2005] Hirst, Graeme und A. Budanitsky (2005). *Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion*. Natural Language Engineering, 11(1):87–111.

[Hirst und St-Onge 1998] Hirst, Graeme und D. St-Onge (1998). *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 305–332. MIT Press, Cambridge.

[Hundsnurscher und Splett 1982] Hundsnurscher, Franz und J. Splett (1982). *Semantik der Adjektive des Deutschen: Analyse der semantischen Relationen*. Westdeutscher Verlag, Opladen.

[Ide und Véronis 1998] Ide, Nancy und J. Véronis (1998). *Introduction to the Special Issue on Word Sense Disambiguation: the State of the Art*. Computational Linguistics, 24(1):1–40.

[iFinder 2006] iFinder (2006). *Webseite der Online-Demo der iFinder Suchtechnologie*. `http://demo.intrafind.org/iFinder/www/?indexID=4` [Stand: 1.5.2006].

[Japanese-FrameNet 2006] Japanese-FrameNet (2006). *Webseite des Projekts Japanese FrameNet*. `http://jfn.st.hc.keio.ac.jp/` [Stand: 1.5.2006].

[Jarmasz 2003] Jarmasz, Mario (2003). *Roget's Thesaurus as a Lexical Resource for Natural Language Processing*. Diplomarbeit, University of Ottawa, Ottawa, Canada. `http://www.site.uottawa.ca/~mjarmasz/thesis/Roget_for_NLP.pdf` [Stand: 1.5.2006].

[Jarmasz und Szpakowicz 2001] Jarmasz, Mario und S. Szpakowicz (2001). *The Design and Implementation of an Electronic Lexical Knowledge Base*. In: *Canadian Conference on AI*, S. 325–334.

[Jarmasz und Szpakowicz 2003] Jarmasz, Mario und S. Szpakowicz (2003). *Roget's Thesaurus and Semantic Similarity*. In: *RANLP*, S. 111–120.

[Jiang und Conrath 1997] Jiang, Jay J. und D. W. Conrath (1997). *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. In: *Proceedings of International Conference on Research in Computational Linguistics*, S. 111–120, Taiwan.

[Kleiber 1993] Kleiber, Georges (1993). *Prototypensemantik: eine Einführung*. Narr, Tübingen.

[Krovetz und Croft 1992] Krovetz, Robert und W. B. Croft (1992). *Lexical Ambiguity and Information Retrieval*. ACM Transactions on Information Systems, 10(2):115–141.

[Kruse und Naujoks 2006] Kruse, Peter und A. Naujoks (2006). *Clever Search Suchmaschine*. `http://oslo.cs.uni-magdeburg.de/~cegeie/` [Stand: 1.5.2006].

[Kruse et al. 2005] Kruse, Peter M., A. Naujoks, D. Rösner und M. Kunze (2005). *Clever Search: A WordNet Based Wrapper for Internet Search Engines*. In: Fisseni, Bernard, H.-C. Schmitz, B. Schröder und P. Wagner, Hrsg.: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn.*, S. 367–380. Peter Lang, Frankfurt am Main.

[Kunze 2005a] Kunze, Claudia (2005a). *Präsentation am Ontologie-Workshop in Dortmund*. `http://www.hytex.uni-dortmund.de/hytex/Ontologie-Workshop-2005-11-25/Onto_Kunze.pdf` [Stand: 1.5.2006].

[Kunze 2005b] Kunze, Claudia (2005b). *Semantische Relationstypen in GermaNet*. In: Langer, Stefan und D. Schnorbusch, Hrsg.: *Semantik im Lexikon*, S. 163–180. Gunter Narr, Tübingen.

[Kunze und Lemnitzer 2005] Kunze, Claudia und L. Lemnitzer (2005). *Die Zukunft der Wortnetze oder die Wortnetze der Zukunft. Ein Roadmap-Beitrag*. In: Fisseni, Bernard, H.-C. Schmitz, B. Schröder und P. Wagner, Hrsg.: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn.*, S. 337–352. Peter Lang, Frankfurt am Main.

[Kunze und Wagner 1999] Kunze, Claudia und A. Wagner (1999). *Integrating GermaNet into EuroWordNet, a Multi-Lingual Lexical-Semantic Database*. Sprache und Datenverarbeitung, 23(3):5–20.

[Kunze und Wagner 2001] Kunze, Claudia und A. Wagner (2001). *Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das*

Deutsche. In: Lemberg, Ingrid, B. Schröder und A. Storrer, Hrsg.: *Chancen und Perspektiven computergestützter Lexikographie*, S. 229–246. Niemeyer, Tübingen.

[Lang 1994] Lang, Ewald (1994). *Das Spektrum der Antonymie. Semantische und konzeptuelle Strukturen im Lexikon und ihre Darstellung im Wörterbuch*. In: Harras, Gisela, Hrsg.: *Die Ordnung der Wörter*. de Gruyter, Berlin.

[Leacock und Chodorow 1998] Leacock, Claudia und M. Chodorow (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 265–284. MIT Press, Cambridge.

[Lemnitzer und Kunze 2003] Lemnitzer, Lothar und C. Kunze (2003). *Using Wordnets in Teaching Virtual Courses of Computational Linguistics*. In: Sojka, Petr, K. Pala, P. Srmz, C. Fellbaum und P. Vossen, Hrsg.: *Proceedings of the Second Global WordNet Conference*, S. 150–156, Brno, Czech Republic.

[Lenat und Guha 1990] Lenat, Douglas B. und R. V. Guha (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Addison-Wesley, Reading, Massachusetts.

[Lesk 1986] Lesk, Michael (1986). *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone*. In: *SIGDOC '86: Proceedings of the 5th Annual International Conference on Systems Documentation*, S. 24–26, New York. ACM Press.

[Levin 1993] Levin, Beth (1993). *English Verb Classes and Alternations*. University of Chicago Press, Chicago.

[LexikoNet 2006] LexikoNet (2006). *Webseite des LexikoNet-Projekts*. `http://www.dwds.de/erschliessung/LexikoNet` [Stand: 1.5.2006].

[Lin 1998] Lin, Dekang (1998). *An Information-Theoretic Definition of Similarity*. In: *Proceedings of the 15th International Conference on Machine Learning*, S. 296–304, Madison, WI.

[Linke et al. 1994] Linke, Angelika, M. Nussbaumer und P. R. Portmann (1994). *Studienbuch Linguistik*. Niemeyer, Tübingen, Zweite Aufl.

[Locke und Booth 1955] Locke, William N. und A. D. Booth (1955). *Machine translation of languages: fourteen essays*. Technology Press of the MIT and Wiley, New York.

[Lutzeier 1995] Lutzeier, Rolf (1995). *Lexikologie*. Stauffenburg Verlag, Tübingen.

[Lyons 1977] Lyons, John (1977). *Semantics*, Bd. 1. Cambridge University Press, Cambridge.

[Matuszek et al. 2006] Matuszek, Cynthia, J. Cabral, M. Witbrock und J. DeOliveira (2006). *An Introduction to the Syntax and Content of Cyc*. In: *2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering*, Stanford University. To appear.

[Melcuk und Zholkovsky 1988] Melcuk, I. und A. Zholkovsky (1988). *The Explanatory Combinatorial Dictionary*. In: Evens, Martha W., Hrsg.: *Relational Models of the Lexicon*. Cambridge University Press, Cambridge.

[Meyer und Dale 2002] Meyer, Josef und R. Dale (2002). *Using the WordNet Hierarchy for Associative Anaphora Resolution*. In: *Proceedings of SemaNet'02: Building and Using Semantic Networks*, S. 1–7, Taipei, Taiwan.

[Mihalcea und Moldovan 2001] Mihalcea, Rada und D. Moldovan (2001). *Automatic Generation of a Coarse Grained WordNet*. In: *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, S. 35–41, Pittsburgh, PA.

[Miller 1990] Miller, George A. (1990). *Nouns in WordNet: a Lexical Inheritance System*. International Journal of Lexicography, 3(4):245–264.

[Miller 1998a] Miller, George A. (1998a). *Nouns in WordNet*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 23–46. MIT Press, Cambridge.

[Miller und Charles 1991] Miller, George A. und W. Charles (1991). *Contextual Correlates of Semantic Similarity*. Language and Cognitive Processes, 6(1):1–28.

[Miller et al. 1993] Miller, George A., C. Leacock, T. Randee und R. Bunker (1993). *A Semantic Concordance*. In: *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, S. 303–308, Plainsboro, NJ.

[Miller 1998b] Miller, Katherine J. (1998b). *Modifiers in WordNet*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 47–68. MIT Press, Cambridge.

[Mitkov 1999] Mitkov, Ruslan (1999). *Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP*. Machine Translation, 14(3–4):159—-161.

[Moldovan und Novischi 2002] Moldovan, Dan und A. Novischi (2002). *Lexical Chains for Question Answering*. In: *Proceedings of COLING 2002*, S. 674–680, Geneva, Switzerland.

[Naber 2005] Naber, Daniel (2005). *OpenThesaurus: ein offenes deutsches Wortnetz*. In: Fisseni, Bernard, Schmitz Hans-Christian Schröder Bernhard und Petra Wagner, Hrsg.: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, S. 422–433. Peter Lang, Frankfurt am Main.

[Ohara et al. 2004] Ohara, Kyoko, S. Fujii, T. Ohori, R. Suzuki, H. Saito und S. Ishizaki (2004). *The Japanese FrameNet Project: An Introduction*. In: *Proceedings of LREC*, S. 9–11, Genoa.

[Omega-Browser 2006] Omega-Browser (2006). *Webseite mit der Auflistung verschiedener Omega Ontologie-Browser*. `http://omega.isi.edu/doc/browsers.html` [Stand: 1.5.2006].

[OpenCyc-Browser 2006] OpenCyc-Browser (2006). *Webseite des Cycorp OpenCyc Ontologie-Browsers (opencyc1)*. `http://opencyc1.cyc.com:3602/cgi-bin/cyccgi/cg?cb-start` [Stand: 1.5.2006].

[OpenThesaurus 2006] OpenThesaurus (2006). *Webseite und Online-Abfrage des OpenThesaurus-Projekts*. `http://www.openthesaurus.de` [Stand: 13.4.2006].

[Orasan et al. 2005] Orasan, Constantin, T. Marshall, R. Clark, L. A. Ha und R. Mitkov (2005). *Building a WSD Module Within an MT system to Enable Interactive Resolution in the User's Source Language*. In: *10th EAMT conference „Practical Applications of Machine Translation“*, S. 205–211, Budapest.

[Osswald 2003] Osswald, Rainer (2003). *Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons HaGenLex*. In: Kunze, Claudia, L. Lemnitzer und A. Wagner, Hrsg.: *GermaNet Workshop „Anwendungen des deutschen Wortnetzes in Theorie und Praxis“*, S. 43–51. Tagungsband.

[Osswald 2006] Osswald, Rainer (2006). *Webseite des HaGenLex-Projekts*. `http://pi7.fernuni-hagen.de/forschung/hagenlex/hagenlex-de.html` [Stand: 1.5.2006].

[Papineni et al. 2002] Papineni, Kishore, S. Roukos, T. Ward und W.-J. Zhu (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. In: *Proceedings of the 40th ACL*, S. 311–318.

[Pedersen und Michelizzi 2002] Pedersen, Ted und J. Michelizzi (2002). *WordNet::Similarity Web Interface*. `http://marimba.d.umn.edu/cgi-bin/similarity.cgi` [Stand: 1.5.2006].

[Peters et al. 1998] Peters, Wim, I. Peters und P. Vossen (1998). *The Reduction of Semantic Ambiguity in Linguistic Resources*. In: Rubio, Antonio, N. Gallardo, R. Catro und A. Tejada, Hrsg.: *Proceedings of First International Conference on Language Resources and Evaluation*, S. 59–66, Granada.

[Philpot et al. 2005] Philpot, Andrew, E. Hovy und P. Pantel (2005). *The Omega Ontology*. In: *Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP)*, S. 59–66, Korea. Refereed one-day workshop paper.

[Poesio et al. 2002] Poesio, Massimo, T. Ishikawa, S. Schulte im Walde und R. Vieira (2002). *Acquiring Lexical Knowledge for Anaphora Resolution*. In: *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*, S. 1220–1224, Las Palmas de Gran Canaria, Spain.

[Pustejovsky 1995] Pustejovsky, James (1995). *The Generative Lexicon*. MIT Press, Cambridge.

[Pustejovsky et al. 1997] Pustejovsky, James, B. Boguraev, M. Verhagen, P. Buitelaar und J. M. (1997). *Semantic Indexing and Typed Hyperlinking*. In: *Proceedings of the American Association for Artifical Intelligence Conference*, S. 120–128, Stanford University, CA.

[Quillian 1968] Quillian, M. R. (1968). *Semantic Memory*. In: Minsky, Marvin, Hrsg.: *Semantic Information Processing*, S. 227–270. MIT Press, Cambridge, MA.

[Reeder et al. 2004] Reeder, Florence, B. Dorr, D. Farwell, N. Habash, S. Helmreich, E. Hovy, L. Levin, T. Mitamura, K. Miller, O. Rambow und A. Siddhartan (2004). *Interlingual Annotation for MT Development*. In: *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, S. 1–9, Washington DC.

[Resnik 1995a] Resnik, Philip (1995a). *Disambiguating Noun Groupings with Respect to WordNet Senses*. In: *Proceedings of the Third Workshop on Very Large Corpora*, S. 54–68.

[Resnik 1995b] Resnik, Philip (1995b). *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, S. 448–453, Montreal, Canada.

[Rila et al. 1998] Rila, Mandala, T. Tokunaga und H. Tanaka (1998). *The Use of WordNet in Information Retrieval*. In: *Proceedings of ACL Workshop on the Usage of WordNet in Natural Language Processing Systems*, S. 31–37, Somerset, NJ.

[Rogets 1852] Rogets, Peter Mark (1852). *Roget's Thesaurus of English Words and Phrases*. Longman Group Ltd, Harlow, Essex.

[Rosch 1973] Rosch, Eleanor H. (1973). *Natural Categories*. Cognitive Psychology, 7:573–605.

[Rosch 1975] Rosch, Eleanor H. (1975). *Cognitive Reference Points*. Cognitive Psychology, 7:532–547.

[Ruppenhofer et al. 2005] Ruppenhofer, Josef, M. Ellsworth, M. Petruck und C. Johnson (2005). *FrameNet: Theory and Practice*. `http://framenet.icsi.berkeley.edu/book/book.pdf` [Stand: 1.5.2006].

[SALSA 2006] SALSA (2006). *Webseite des SALSA-Projekts*. `http://www.coli.uni-saarland.de/projects/salsa/` [Stand: 1.5.2006].

[Salton und McGill 1983] Salton, Gerard und M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

[Sanderson 1994] Sanderson, Mark (1994). *Word Sense Disambiguation and Information Retrieval*. In: *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, S. 142–151, Dublin, Ireland.

[Schiehlen und Klabunde 2004] Schiehlen, Michael und R. Klabunde (2004). *Semantik*. In: Carstensen, Kai-Uwe, C. Eber, C. Endriss, S. Jekat, R. Klabunde und H. Langer, Hrsg.: *Computerlinguistik und Sprachtechnologie: eine Einführung*, S. 276–332. Elsevier, München, Zweite Aufl.

[Shi und Mihalcea 2005] Shi, Lei und R. Mihalcea (2005). *Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing*. In: *CICLing 2005*, S. 100—-111, Mexico.

[Smeaton 1999] Smeaton, Alan F. (1999). *Using NLP or NLP Resources for Information Retrieval Tasks*. In: Strzalkowski, Tomek, Hrsg.: *Natural Language Information Retrieval*, S. 99–111. Kluwer Academic Publishers, Dordrecht.

[Smeaton und Quigley 1996] Smeaton, Alan F. und I. Quigley (1996). *Experiments on Using Semantic Distances Between Words in Image Caption Retrieval*. In: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, S. 174–180, Zurich.

[Spanish-FrameNet 2006] Spanish-FrameNet (2006). *Webseite des Projekts Spanish FrameNet*. `http://gemini.uab.es/SFN` [Stand: 1.5.2006].

[Tengi 1998] Tengi, Randee I. (1998). *Design and Implementation of the WordNet Lexical Database and Searching Software*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 105–128. MIT Press, Cambridge.

[Trier 1934] Trier, Jost (1934). *Das sprachliche Feld. Eine Auseinandersetzung*. Neue Jahrbücher für Wissenschaft und Jugendbildung, 10:428–449.

[Upper-Ontology 2006] Upper-Ontology (2006). *Webseite von Cycorp OpenCyc Selected Vocabulary and Upper Ontology*. `http://www.cyc.com/cycdoc/vocab/upperont-diagram.html` [Stand: 1.5.2006].

[Vicedo und Ferrández 2000] Vicedo, José L. und A. Ferrández (2000). *Importance of Pronominal Anaphora Resolution in Question Answering Systems*. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, S. 555–562, Hong Kong, China.

[Vieira und Poesio 2000] Vieira, Renata und M. Poesio (2000). *An Empirically Based System for Processing Definite Descriptions*. Computational Linguistics, 26(4):539–594.

[Voorhees und Tice 2000] Voorhees, Ellen und D. Tice (2000). *Building a Question Answering Test Collection*. In: *Proceedings of SIGIR-2000*, S. 200–207, Athens, Greek.

[Voorhees 1994] Voorhees, Ellen M. (1994). *Query Expansion Using Lexical-Semantic Relations*. In: *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, S. 61–69, Dublin, Ireland.

[Voorhees 1998] Voorhees, Ellen M. (1998). *Using WordNet for Text Retrieval*. In: Fellbaum, Christiane, Hrsg.: *WordNet: An Electronic Lexical Database*, S. 285–304. MIT Press, Cambridge.

[Vossen 1998] Vossen, Piek (1998). *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht. Reprinted from Computers and the Humanities, 32(2–3).

[Vossen 2004a] Vossen, Piek (2004a). *EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual-Index*. International Journal of Lexicography, 17(2):161–173.

[Vossen 2004b] Vossen, Piek (2004b). *Wordnets for Information Retrieval: a Hole in One*. OmniPaper seminar. `www.vossen.info/docs/2004/Omnipaper/Wordnets%20for%20information%20retrieval.ppt` [Stand: 1.5.2006].

[Wagner 2005] Wagner, Andreas (2005). *Learning Thematic Role Relations for Lexical Semantic Nets*. Doktorarbeit, Universität Tübingen, Tübingen. `http://w210.ub.uni-tuebingen.de/dbt/volltexte/2005/2005/pdf/AW_diss.pdf` [Stand: 1.5.2006].

[Wierzbicka 1984] Wierzbicka, Anna (1984). *Apples are not a „Kind of Fruit“: The Semantics of Human Categorization*. American Ethnologist, 11:313–328.

[Winston et al. 1987] Winston, Morton E., R. Chaffin und D. Herrmann (1987). *A Taxonomy of Part-Whole Relations*. Cognitive Science, 11(4):417–444.

[WordNet 2006] WordNet (2006). *Webseite des Projekts Princeton WordNet*. `http://wordnet.princeton.edu/` [Stand: 1.5.2006].

[WordNetBrowser 2006] WordNetBrowser (2006). *Webseite des WordNet-Browsers*. `http://wordnet.princeton.edu/perl/webwn` [Stand: 1.5.2006].

[WordNet-Glossar 2006] WordNet-Glossar (2006). *Webseite des WordNet-Glossars*. `http://wordnet.princeton.edu/man/wngloss.7WN` [Stand: 1.5.2006].

[WordNetStatistik 2006] WordNetStatistik (2006). *Webseite der WordNet-Statistik*. `http://wordnet.princeton.edu/man/wnstats.7WN` [Stand: 1.5.2006].

[Yarowsky 1992] Yarowsky, David (1992). *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora*. In: *Proceedings of COLING-92*, S. 454–460, Nantes, France.

Lebenslauf

Persönliche Angaben

Sandra Roth
Langfurren 5
8057 Zürich
sandraroth@gmail.com
Geboren am 26. Juni 1980

Schulbildung

1987–1992	Primarschule Erlinsbach AG
1992–1996	Bezirksschule Aarau
1996–2000	Kantonsschule Aarau, Typus D
1998	Austauschjahr in Japan
2001–2006	Studium der Allgemeinen Sprachwissenschaft, Computerlinguistik und Informatik an der Universität Zürich

Berufliche und nebenberufliche Tätigkeiten

2000–2001	Flight Attendant bei Swissair
2002–2005	Teilzeitanstellung bei Swiss Re im Records Management
2005	Praktikum und anschliessende Teilzeitanstellung bei Canoo Engineering
seit 2005	Hilfsassistenz am Institut für Informatik, Computerlinguistik
seit 2006	Teilzeitanstellung bei der E-Learning Koordination der Phil. Fakultät
seit 2004	Vorstandsmitglied des Fachvereins für Computerlinguistik, seit 2005 Präsidentin

Hobbies

Japanische Sprache und Küche
Sport im Freien
Asienreisen