Claudia Kunze, Andreas Wagner

Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche

1 GermaNet, eine lexikalische Ressource für die Informationserschließung

Elektronische Wörterbücher, welche über die Modellierung kleinerer Sprachdomänen hinausgehend den strukturierten Zugriff auf lexikalische Einheiten des Grundwortschatzes gestatten, sind ein zentrales Desiderat der Informationsgesellschaft. Lexikalische Wissensbasen stehen im Mittelpunkt des Interesses, da die Verfügbarkeit elektronischer Bedeutungswörterbücher für zahlreiche Anwendungen innerhalb der Maschinellen Sprachverarbeitung die unabdingbare Voraussetzung darstellt.

Semantische Netze im Stil des Princeton WordNet (vgl. Miller et al. 1990, Fellbaum 1998), die eine Vielzahl lexikalischer Einheiten in ihren grundlegenden semantischen Relationen abbilden, stellen geeignete Grundlagen-Ressourcen für effiziente computerlinguistische Verfahren zur Bedeutungsdisambiguierung bereit.

Das in solchen Ressourcen repräsentierte Wissen fungiert in computerlinguistischen Anwendungen als Referenzwissen und wird über statistische Verfahren mit den im zu verarbeitenden Text vorkommenden Wörtern abgeglichen (vgl. Yarowsky 1992; Harley 1994). WordNet kann so wie ein klassischer Thesaurus oder ein Wörterbuch sinn- und sachverwandter Wörter eingesetzt werden.

Folgende Anwendungen bedürfen der lexikalisch-semantischen Disambiguierung (vgl. Kapitel 4):

- die Maschinelle Übersetzung;
- die Informationserschließung;
- die semantische Annotierung von Korpora;

- die Entwicklung von Sprachlernwerkzeugen, Übersetzungswerkzeugen und Werkzeugen zum Informationserwerb;
- die Entwicklung automatischer Summarizer,
- die Realisierung von Sprachgenerierungswerkzeugen.

Mit der Entwicklung des deutschen Wortnetzes GermaNet ist die Lücke, die es in bezug auf deutschsprachige semantische Lexika zu verzeichnen gab, gefüllt worden.

In Kapitel 2 beschreiben wir die Aufbauprinzipien und die Merkmale des GermaNet. Kapitel 3 zeigt, wie GermaNet in einen multilingualen Kontext gestellt worden ist. Die Rolle, die GermaNet in computerlinguistischen Anwendungen spielen kann, insbesondere bei der Akquisition von Selektionspräferenzen, wird im 4. Kapitel erörtert. Das Schlußwort faßt zusammen, welche Perspektiven und Modifikationen sich für GermaNet ergeben.

2 Aufbau des GermaNet

Der festgestellte Mangel an computertechnisch verfügbaren lexikalisch-semantischen Ressourcen für das Deutsche hat die Entwicklung eines deutschen semantischen Wortnetzes motiviert, das im wesentlichen an den Strukturierungsprinzipien des Princeton WordNet 1.5 orientiert ist.[1]

Diese Anlehnung am Datenbankmodell und an den Aufbauprinzipien des WordNet bedeutet allerdings nicht, daß GermaNet aus einer Übersetzung der WordNet-Konzepte hervorgegangen ist. Vielmehr ist GermaNet aus verschiedenen lexikographischen Quellen (u.a. aus dem DEUTSCHEN WORTSCHATZ und dem DUDEN 8 der sinn- und sachverwandten Wörter) unter Berücksichtigung von Korpusfrequenzen von Hand aufgebaut worden. Darüber hinaus setzt GermaNet eigene Schwerpunkte sowohl auf der strukturellen als auch auf der konzeptuellen Ebene (zu den Unterschieden zwischen GermaNet und WordNet s. u.).

GermaNet modelliert den Grundwortschatz des Deutschen auf konzeptueller Ebene und verbindet Nomen, Verben und Adjektive durch elementare semantische Relationen und leistet somit einen wichtigen Beitrag zur Schaffung einer geeigneten Dateninfrastruktur für deutschsprachige computerlinguistische Anwendungen.

2.1 Abdeckung

Ausgangsziel des SLD-Projekts war, mit GermaNet einen on-line Thesaurus, der den deutschen Grundwortschatz abdeckt, zu erstellen (vgl. Hamp/Feldweg 1997). Zentrales Konzept der lexikalischen Kodierung sind die sogenannten *synsets*, die als abstrakte Bedeutungseinheiten zu gegebenen Konzepten eine Synonymenmenge bereitstellen. Es gibt semantische Relationen zwischen Konzepten (*synsets*) oder zwischen Wortbedeutungen

---

[1] GermaNet wurde unter der Leitung von Helmut Feldweg im Rahmen des SLD-Projektes („Ressourcen und Methoden zur semantisch-lexikalischen Disambiguierung") aufgebaut, das 1996 und 1997 vom Land Baden-Württemberg gefördert wurde. An der Realisierung des Projektes waren im weiteren Valérie Béchet-Tsarnos, Birgit Hamp, Michael Hipp, Claudia Kunze, Karin Naumann, Susanne Schüle, Rosmary Stegmann, Karen Steinicke, Christine Thielen und Andreas Wagner beteiligt.

(einzelnen Synonymen aus den *synsets*). Solche *synsets* werden gleichermaßen für Nomen, Verben und Adjektive implementiert.

Die zentrale Relation ist die Hyponymie-Beziehung, welche die Konzepte aller Wortarten (auch der Adjektive) hierarchisch gliedert.

Zur Zeit enthält die Datenbank, deren Abdeckung sich in kontinuierlicher Erweiterung befindet, ca. 25000 *synsets* und etwa 30000 Wortbedeutungen. Einträge der Datenbank werden mit Frequenzlisten, die aus Korpora extrahiert sind, abgeglichen, um fehlende Konzepte systematisch zu ergänzen.

Im Netz sind nur morphologische Vollformen kodiert und lediglich sehr geläufige Mehrwortlexeme wie *erste Hilfe* oder *gesprochene Sprache*. Eigennamen werden, sofern sie berücksichtigt sind (z.B. im Wortfeld der Geographie die Namen der Städte, Länder und Flüsse), speziell markiert. Einige wichtige Abkürzungen, etwa für die politischen Volksparteien, sind ebenfalls in GermaNet repräsentiert.

Der Datenbestand ist in fünfzehn semantische Felder unterteilt, die weitgehend von WordNet übernommen wurden und die zur Bearbeitung in den sog. ‚lexicographer files' hilfreich sind.

2.2 Relationen

GermaNet unterschiedet zwischen *lexikalischen* und *konzeptuellen Relationen*:

- Lexikalische Relationen wie Synonymie und Antonymie bestehen zwischen verschiedenen lexikalischen Realisierungen von Konzepten und sind bidirektionale Relationen, die für alle drei Wortklassen gelten.
- Konzeptuelle Relationen wie Hyponymie, Hyperonymie, Meronymie, Implikation und Kausation bestehen zwischen gegebenen Konzepten in all ihren Lexikalisierungen.

Ferner gibt es noch eine *semantische Derivationsrelation*, die kategorienübergreifend relevant ist für denominale Adjektive (**finanziell** zu **Finanzen**), deverbale Nominalisierungen (**Entdeckung** zu **entdecken**) und deadjektivische Nominalisierungen (**Müdigkeit** zu **müde**).

Das grundlegende Strukturierungsprinzip stellt die *Hyponymierelation*, wie sie z.B. zwischen **Rotkehlchen** und **Vogel** besteht, dar. Deszendentenketten für Nomen weisen oft eine beträchtliche Hierarchietiefe auf, aber auch im verbalen und adjektivischen Bereich ist die Taxonomie wesentliche Gliederungsrelation.

Die *Teil-Ganzes-Beziehung* (Meronymie) wird nur für Nomen spezifiziert. So ist ein **Arm** nur unzureichend als eine Art **Körper** klassifiziert, sondern zählt als Teil eines Körpers. Teil-Ganzes-Beziehungen liegen auch auf abstrakter Ebene vor, etwa in bezug auf Mitgliedschaft in einer Gruppe (**Vorsitzender** eines **Vereins**) oder als Material in einer Komposition (**Fensterscheibe** aus **Glas**).

Die *Implikationsbeziehung* ist anhand einiger weniger Beispiele kodiert. Hier sind Verbkonzepte in einem logischen Zusammenhang (‚backword presupposition') erfaßt, wie dieser z.B. zwischen **gelingen** und **versuchen** besteht.

Wichtiger und in größerem Ausmaß kodiert ist die klassenübergreifende *Kausationsrelation*, die lexikalische Resultative betrifft und z.B. **töten** und **sterben** oder **öffnen** und **offen** verknüpft.

Die folgenden Abbildungen zeigen einen Verbeintrag und einen nominalen Eintrag mit allen korrelierten Konzepten. Diejenigen Lesarten, die zu den Ausgangskonzepten keine

hyponymische bzw. hyperonymische Relation aufweisen, sind grau markiert. Wir haben die Lesartennummern der *synset*-Varianten aufgeführt, um zu verdeutlichen, daß in GermaNet Wortbedeutungen (‚senses') repräsentiert und semantisch miteinander verknüpft werden.

Das **synset** {**öffnen#3, aufmachen#2**} hat als Hyperonym {**wandeln#4, verändern#2, ändern#2**} sowie die vier Hyponyme {**aufschieben#1**}, {**aufstoßen#2**}, {**aufbrechen#1**} und {**aufsperren#1**}. Es gibt eine kausale Relation zum inchoativen **öffnen** (vgl. {**öffnen#1, aufgehen#1**}). Interessanterweise haben die Varianten im synset unterschiedliche Antonyme: **öffnen#3** hat das Antonym **schließen#7**, **aufmachen#2** das Antonym **zumachen#2**. Zur Verdeutlichung sind die bilateralen Antonym-Pfeile direkt auf die entsprechende Variante gerichtet und nicht auf den gesamten Konzeptknoten.

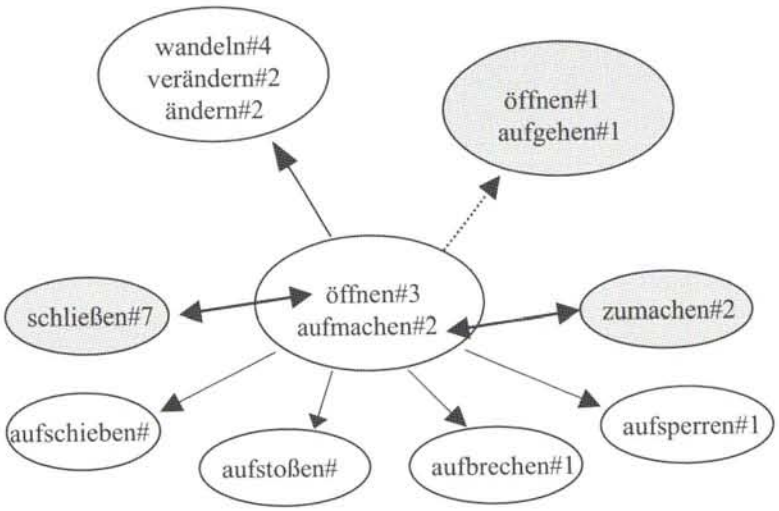

Abb. 1: Semantische Relationen des kausativen Verbs **öffnen**. Die einfachen Pfeile indizieren Überordnung (Pfeilspitze weist nach oben) und Unterordnung (mit der Pfeilspitze nach unten). Antonymie ist durch den Doppelpfeil gekennzeichnet, die kausative Relation mittels des gepunkteten Pfeils.

Abbildung 2 (folgende Seite) zeigt das Beispiel **Atmungsorgan** mit zwei Hyponymen (**Lunge** und **Kieme**), einem Hyperonym (**Organ**) und zwei Holonymen (**Oberkörper** und **Atemsystem**). Ein Meronym (**Luftröhre**) ist ebenfalls kodiert.

2.3 Kreuzklassifikation

In GermaNet werden Konzepte, die unterschiedlichen Hierarchien zugehören, kreuzklassifiziert. Das *Kaninchen* ist als *Haustier*, *Nutztier* und *Hasentier* klassifiziert, der *Hase* lediglich als *Hasentier*, der *Wellensittich* als *Haustier* und *Vogel*, der *Hund* lediglich als *Haustier* und die *Drossel* nur als *Vogel* (Abbildung 3).

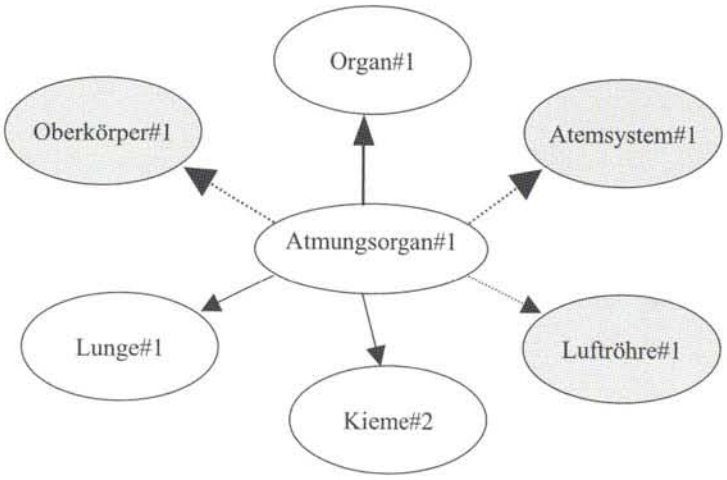

Abb. 2: Semantische Relationen des Nomens **Atmungsorgan**. Die Teil-Ganzes-Beziehung wird durch die gepunkteten Pfeile angezeigt, Meronyme mit nach unten, Holonyme mit nach oben weisender Pfeilspitze.

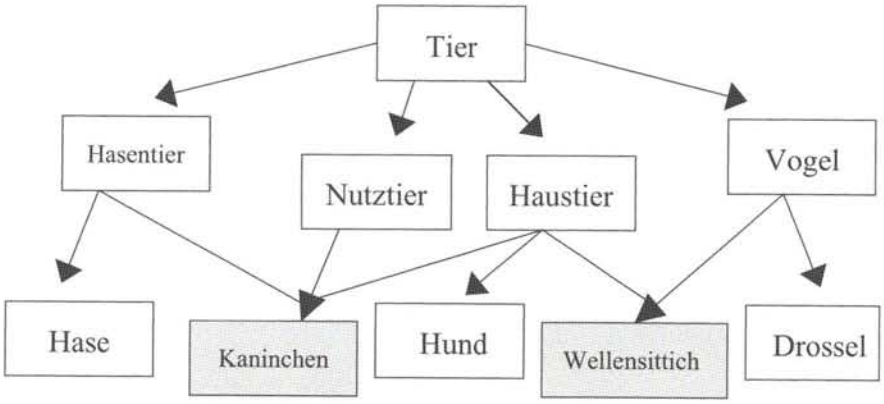

Abb. 3: Kreuzklassifikation im Tierreich: **Kaninchen** und **Wellensitich**

Nicht nur die Zugreifbarkeit der Lexeme unter verschiedenartigen Bedeutungsaspekten ist ein Vorteil des systematischen Kreuzklassifizierens.

Anhand durchgängiger Kreuzklassifikation können wir Muster regulärer Polysemie ausmachen, die für Restrukturierungen des Lexikons nützlich sind, vgl. aus der Klasse der Früchte diejenige Teilmenge, die wie *Banane* zugleich als *Nahrungsmittel* und als *Pflanze* klassifiziert ist. Andere Fälle regulärer Polysemie betreffen *Birke* als *Pflanze* und als *Holzart* oder *Tennis* als *Veranstaltung* und als *Sportart*. Die empirische Analyse deutet auf sehr viele produktive Muster (vgl. Buitelaars Analyse der CoreLex-Pattern 1998).

2.4 Artifizielle Konzepte

Auf der konzeptuellen Ebene tragen eigens eingeführte künstliche Knoten zu einer ausgewogeneren Taxonomie bei. Künstliche Konzepte können auf lexikalische Lücken in der Sprache bezogen sein (z.B. das fehlende Antonym zu *durstig*), aber auch rein konzeptuelle Konstrukte wie etwa **?Charakterbeschaffener** betreffen. Oftmals helfen künstliche Knoten, unmotivierte Kohyponymie zu vermeiden.

Das Beispiel in Abbildung 4 enthält mit **?Schullehrer** und **?hierarchischer Lehrer** zwei künstliche Konzepte, welche das Teilnetz im Wortfeld **Lehrer** symmetrischer strukturieren. Nach Cruse (1986:22) sollten Kohyponyme eines Mutterknotens möglichst inkompatibel zueinander sein. Diese Inkompatibilität operiert auf einer Ebene von Ähnlichkeit, die durch den gemeinsamen Oberbegriff gegeben ist, vgl. **Baby**, **Kleinkind**, **Vorschulkind**, **Schulkind** als Unterbegriffe zu **Kind**, die einander wechselseitig ausschließen.

Da ein Fachlehrer aber an einem Schultyp in einer hierarchischen Position unterrichtet, wären die sechs Endknoten des Beispielnetzes als direkte Deszendenten des **Lehrer**-Knotens nicht inkompatibel genug, so daß die nicht-lexikalisierten Konzepte, die in GermaNet durch ein initiales Fragezeichen gekennzeichnet werden, sinnvoll eingeführt sind.

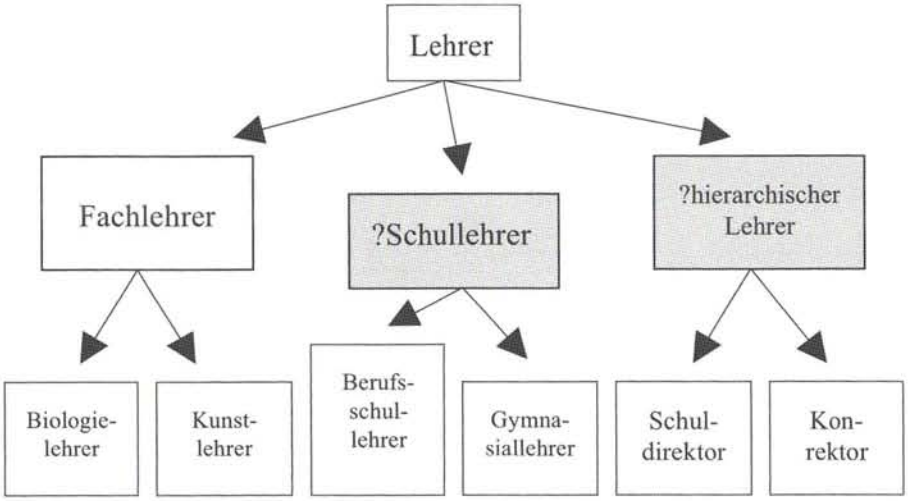

Abb. 4: Künstliche Konzepte im Wortfeld **Lehrer**

2.5 Subkategorisierungsrahmen

Alle Verbeinträge in GermaNet sind mit Subkategorisierungsrahmen und diesbezüglichen Beispielen versehen. Die kodierten Rahmen geben Aufschluß über das syntaktische Komplementierungsverhalten der GermaNet-Prädikate, leisten also einen Beitrag zur Syntax-Semantik-Schnittstelle.

Die Notation orientiert sich an den *Celex*-Frames, ist aber in bezug auf die Kodierung von Subjektphrasen und Reflexivphrasen leicht modifiziert worden. Unterschiedliche Verbrahmen zu einem Konzept helfen bei der Lesartendisambiguierung, vgl. das Beispiel *setzen*:

setzen1	NN.AN	*Er setzt die Fahnen.*
setzen2	NN.AR	*Sie setzt sich.*
setzen3	NN.AN.BL	*Er setzt den Schüler ans Fenster.*
setzen4	NE.AN	*Es setzt Prügel.*
setzen5	NN.AN.PP	*Er setzt seine Hoffnungen auf sie.*
setzen6	NN.An	*Die Häsin hat (Junge) gesetzt. (Jägersprache)*
setzen7	NN.BL	*Das Pferd setzt über die Hürde.*
setzen8	NN.AN.Dn	*Das Protokoll setzt (ihnen) Schranken.*

2.6 Unterschiede zwischen GermaNet und WordNet

Trotz der strukturellen Ähnlichkeit des GermaNet zum Princeton WordNet lassen sich folgende Unterschiede skizzieren:

- GermaNet orientiert sich im Gegensatz zu WordNet an linguistischen und nicht an psychologischen Strukturierungsprinzipien der Daten.
- Um eine ausgewogene Konzepthierarchie zu gestalten und unmotivierte Kohyponymie zu reduzieren, wird in GermaNet systematischer Gebrauch von artifiziellen Konzepten gemacht, die entsprechend markiert sind.
- GermaNet kodiert Partikelverben (vgl. Hamp 1997), die in WordNet nicht berücksichtigt werden.
- Die Kausationsrelation, die in WordNet lediglich als Relation zwischen Verbinstanzen vorgesehen ist, kann in GermaNet zwischen allen Wortarten kodiert werden.
- In GermaNet sind Adjektive taxonomisch strukturiert und unterliegen nicht dem Satelliten-Ansatz des WordNet, einem assoziativen Verbund von Adjektiven, der auch zu wenig intuitiven Konzepten wie *unschwanger* führen kann.

Zunehmend zeigt sich, daß die Großressource WordNet, welche ca. die dreifache Menge an Einträgen enthält, zu feinkörnige Lesartenunterscheidungen vornimmt, um effizient genug in computerlinguistischen Anwendungen zu sein. Für den Eintrag *go* gibt es 32 Lesarten. Der „richtige Polysemiegrad" ist gefragt, um erfolgreich Bedeutungsdisambiguierung leisten zu können (vgl. Buitelaar 1998 zur Modellierung regelgeleiteter Polysemie).

Eine Ressource mittlerer Größe wie GermaNet, die zudem noch von Restrukturierungsansätzen zur Lesartenreduktion mittels des sogenannten ‚Sense Clustering' (vgl. Peters et al. 1998) profitieren kann, kann durchaus leistungsfähiger in der lexikalisch-semantischen Disambiguierung sein als das behäbige WordNet.

3 GermaNet im multilingualen Kontext

Das Basisvokabular des GermaNet ist Bestandteil des multilingualen semantischen Netzes EuroWordNet, das im Rahmen eines Projektes der Europäischen Gemeinschaft für acht europäische Sprachen aufgebaut worden ist.[2] EuroWordNet ist eine wertvolle Ressource für

[2] Genau gesagt handelt es sich um ein zweiteiliges Projekt, EuroWordNet-1 und EuroWordNet-2. Tübingen ist Partner des EuroWordNet-2-Projektes LE4 8328, vgl. Vossen 1998: „Extending EuroWordNet with four languages".

die Sprachtechnologie in bezug auf multilinguale Anwendungen der Informationserschließung.

Durch die kontrastive Analyse der erstellten Daten im Projekt haben wir auch für die monolinguale Weiterarbeit an GermaNet profitieren können, z.B. in Hinblick auf die Datenabdeckung und durch die Verwendung der statistischen Methoden zur Evaluierung der Daten.

3.1 Interlingualer Index und Basiskonzepte

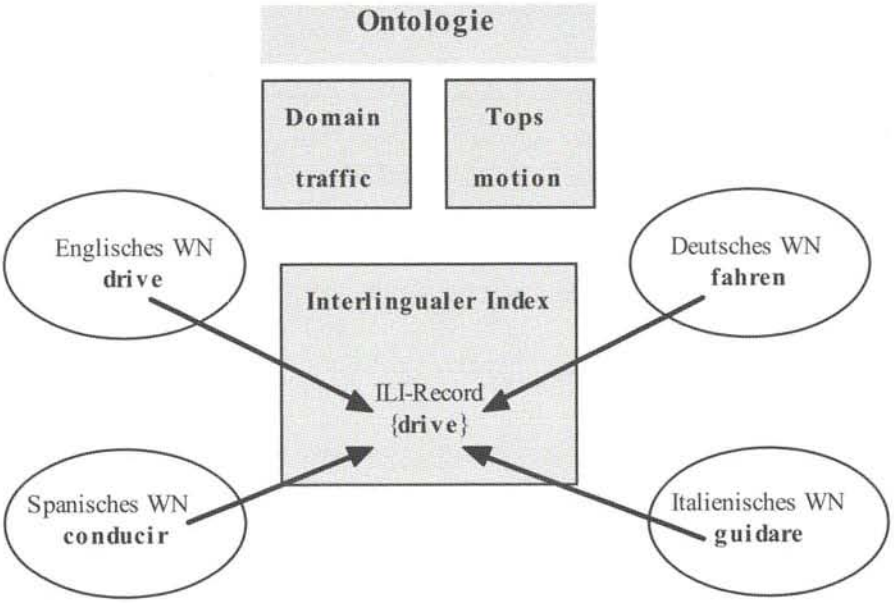

Abb. 5: Ausschnitt aus der EuroWordNet-Architektur. Die sprachunabhängigen Komponenten, zu denen neben dem ILI auch die merkmalsbasierten Ontologien gehören, sind grau markiert. Einzelsprachliche Konzepte werden über Äquivalenzrelationen an den ILI gelinkt.

Die EuroWordNet-Datenbank ist ein multilinguales Netz, das Basiskonzepte acht europäischer Sprachen (Englisch, Holländisch, Spanisch, Italienisch, Deutsch, Französisch, Estnisch und Tschechisch) in ihren semantischen Relationen modelliert. In der Datenbankarchitektur sind die einzelsprachlichen Komponenten über den sprachunabhängigen Interlingualen Index (ILI) korreliert. Trotz seiner Übersprachlichkeit ist der ILI, welcher eine unstrukturierte Liste sogenannter ILI-Records[3] enthält, durch die Dominanz und Vorreiterrolle des Princeton WordNet stark an den englischen Konzepten bzw. den WordNet-Einträgen orientiert. Die sprachspezifischen Konzepte der einzelnen Sprachen werden über eine Äquivalenzrelation an passende ILI-Records gelinkt. Einzelne Sprachpaare zu erfragten Konzepten werden also mittelbar (über den ILI) erzeugt, vgl. Abbildung 5.

---

[3] Ein ILI-Record ist durch einen eindeutigen Code, den *unique identifier*, gekennzeichnet.

Um die Abdeckung der sprachspezifischen Wortnetze kompatibel zu gestalten, müssen alle Sprachnetze anhand der sogenannten ‚Base Concepts' strukturiert sein. Die 1300 ‚Base Concepts' (ca. 1000 Nomen und 300 Verben), die durch einzelsprachliche Selektionen und statistische Evaluierungen dieser Selektionen ermittelt worden sind, müssen folgenden Kriterien genügen:

- ‚Base Concepts' sollten einen hohen Abstraktionsgrad aufweisen, der sich an der Menge der dominierten Unterbegriffe und an der Hierarchietiefe der dominierten Kette manifestiert. Base Concepts sollten spezifischer sein als semantische Merkmale der Top Ontology wie **Funktion, Eigenschaft, Dynamisch**, etc., aber auch abstrakter als Roschs ‚Basic Level Concepts'[4] (z.B. **Tisch** und **Hammer**). Den richtigen Abstraktionsgrad weisen deren semantische Oberbegriffe **Möbel** und **Werkzeug** auf.
- Außerdem sollen Konzepte, die in einer Sprache (und möglichst auch sprachübergreifend) sehr häufig vorkommen, als ‚Base Concepts' berücksichtigt werden, auch wenn sie nicht den gewünschten Abstraktionsgrad aufweisen, wie etwa **lieben** und **mögen**.

Dies Inventar gemeinsamer ‚Base Concepts' ist in einem ersten Schritt über Äquivalenzrelationen an den ILI zu binden, um dann sowohl die leicht erfaßbaren Topknoten als auch die Hyponyme erster Ordnung (meist ‚Basic Level Concepts') zu linken. Die einzelsprachlichen Netze können so unabhängig voneinander, jedoch mit einem Großmaß an Kompatibilität, integriert werden.[5] Durch die Vererbung der semantischen Merkmale der Top Ontology ist es weiterhin auch möglich, die Abdeckung der Netze in den einzelnen semantischen Feldern statistisch zu evaluieren.

3.2 Relationstypen

Nicht immer können äquivalente ILI-Records als Übersetzungen der einzelsprachlichen Konzepte ausgemacht werden. Neben den unterschiedlichen Lexikalisierungspattern, die auf sprachliche und kulturelle Unterschiede zurückgehen, sind dafür auch unterschiedliche Gewichtungen der Konzepte sowie Kodierungslücken verantwortlich. So gibt es im WordNet (das ja weitgehend den ILI prägt) kein Konzept, das dem deutschen Lexem **Lebensgefährte** (als unverheirateter Partner einer eheähnlichen Lebensbeziehung) entspricht. Im deutsch-englischen *COLLINS* hingegen konnten wir das Literal *companion through life* finden. Ein weiteres Beispiel betrifft den Wettbewerbstyp **championships** ‚Meisterschaft', der eine Lexikalisierung im Englischen hat, aber in EuroWordNet nicht als ILI-Record vorhanden ist. Neben der Synonymiebeziehung und der Quasi-Synonymiebeziehung stehen auch nicht-synonymische Äquivalenzlinks der Hyperonymie und Meronymie, ferner Rollenbeziehungen und Kausationsbeziehungen zur Verfügung.

Mitunter ist ein Konzept gut abzubilden, indem mehrere nicht-synonymische Verknüpfungen verwendet werden, vgl. Abbildung 6:

---

[4] Vgl. Rosch (1978).

[5] Mittels dieser Prozedur ist das erste Daten-Ensemble mit ca. 7500 Einträgen entstanden.

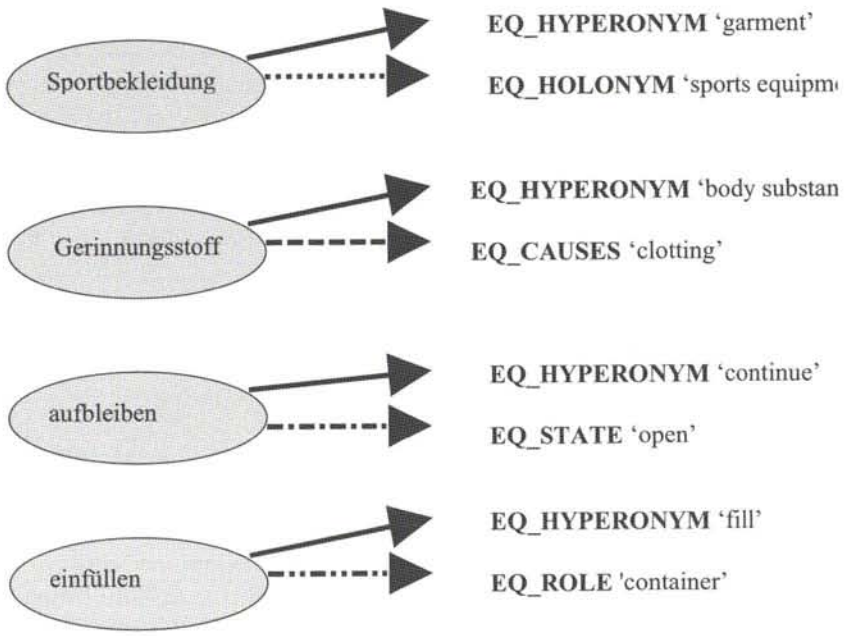

Abb. 6: Beispiele für nicht-synonymische Äquivalenzlinks

Der ebenfalls in der EuroWordNet-Spezifikation vorgesehene Near-Synonym-Link verleitet mitunter zu Ungenauigkeiten. Auf ihn wird oft rekurriert, wenn die Quasi-Äquivalente im Bedeutungsumfang nicht ganz deckungsgleich sind oder wenn zu einem Konzept mehr als ein Synonymlink zu unterschiedlichen ILI-Konzepten etabliert wird, die im WordNet nicht dem gleichen *synset* angehören.

3.3 Synergien für GermaNet

Die Integration des GermaNet in die EuroWordNet hat einige wesentliche Impulse für Optimierungsansätze unserer monolingualen Ressource beigesteuert:

- Im Projektkontext konnten wir ein lexikalisches Werkzeug unserer Amsterdamer Projektpartner für das Deutsche adaptieren, das sehr benutzerfreundlich zum Browsen und Editieren lexikalischer Datenbanken einsetzbar ist.[6]
- Mithilfe der ‚Base Concepts' konnte eine strukturierte, interlingual hinterfragte Überprüfung der Grundwortschatzabdeckung vorgenommen werden. So hatte es im GermaNet u. a. eine Abdeckungslücke im Bereich der Ereignisnominalisierungen gegeben, die nun ausgeglichen werden kann.

---

[6] Hierbei handelt es sich um das ALS (Amsterdam Lexical System), das von Boersma und Vossen zwischen 1992 und 1997 entwickelt worden ist. Die Adaption für das Deutsche von A. Wagner haben wir TALS (Tübinger ALS) genannt. Im Verlauf des Projektes haben wir TALS für die ILI-Anbindung verwendet und nicht zur Editierung des GermaNet.

- Aus der EuroWordNet Spezifikation wollen wir aus dem Inventar innersprachlicher Relationen die Rollentypen für GermaNet übernehmen. Diese sprachinternen Relationen (**ROLE_AGENT, role_patient, role_location**) und deren konverse Relationstypen (**involved_agent, involved_Patient, involved_location**) wollen wir zur Kodierung semantischer Rollen in GermaNet und zur Formulierung von geeigneten Selektionsrestriktionen für Verbeinträge implementieren.

4 Anwendungsperspektiven des GermaNet

In diesem Kapitel werden die Einsatzmöglichkeiten lexikalisch-semantischer Netze wie WordNet und GermaNet in der maschinellen Sprachverarbeitung anhand der Beispiele Lesartendisambiguierung, Informationserschließung und Textkategorisierung exemplarisch aufgezeigt. Insbesondere soll die Ermittlung von Selektionspräferenzen im Mittelpunkt stehen.

4.1 Lesartendisambiguierung

Lesartendisambiguierung (‚word sense disambiguation') ist von entscheidender Bedeutung für maschinelles Sprachverstehen. Gerade häufig verwendete Wörter sind mehrdeutig. Um einen Satz semantisch interpretieren zu können, müssen die mehrdeutigen Wörter in diesem Satz disambiguiert werden. Hierbei ist ein lexikalisch-semantisches Netz wie WordNet oder GermaNet in zweifacher Hinsicht nützlich: Zum einen liefert ein solches Netz (mit seinen Knoten) ein Inventar semantischer Konzepte, mit denen Wortbedeutungen repräsentiert werden können, die *synsets*. Ist ein Wort mehrdeutig, so gehört es zu mehreren *synsets*, z.B. **Ton** mit den *synsets* {**Ton**, **Laut**} und {**Ton**, **Tonerde**}, was den beiden Lesarten dieses Wortes entspricht. Zum anderen modelliert ein lexikalisch-semantisches Netz (mit seinen Kanten) Beziehungen zwischen Konzepten. Diese Beziehungen liefern wichtige Informationen für die Disambiguierung. So nutzt der nun skizzierte Ansatz die durch die Hyponymie-Relationen definierte Hierarchie, um die semantische Ähnlichkeit von Konzepten zu quantifizieren.

Stetina et al. (1998) entwickeln ein Verfahren zur semantischen Disambiguierung von Inhaltswörtern in geparsten Texten. Zur Bestimmung der Lesart eines Wortes werden die Lesarten derjenigen Wörter herangezogen, die mit diesem Wort in einer syntaktischen Relation (z.B. Subjekt–Verb) stehen. Verschiedene Kombinationen von Wortbedeutungen werden mit unterschiedlicher Wahrscheinlichkeit durch bestimmte syntaktische Relationen miteinander verbunden. Den Inhaltswörtern in einem Satz werden nun diejenigen (WordNet-)Lesarten zugewiesen, die gemäß den in diesem Satz vorhandenen syntaktischen Relationen die wahrscheinlichsten sind. Die zu Grunde liegenden Wahrscheinlichkeiten (z.B. die Wahrscheinlichkeit, daß ein Verb mit der Lesart y ein Subjekt mit der Lesart x hat) werden durch eine statistische Analyse des semantisch annotierten Korpus SemCor (Miller u.a. 1993) eingeschätzt. SemCor besteht aus etwa 200000 Wörtern, die jeweils mit ihrer WordNet-Lesart annotiert sind. Hierbei ergibt sich das Problem, daß viele Wörter, die später disambiguiert werden müssen, in diesem Korpus nicht vorkommen, so daß für sie keine Wahrscheinlichkeit geschätzt werden kann. Dies Problem wird dadurch gelöst, daß

für die Einschätzung der Lesarten solcher Wörter die Wahrscheinlichkeiten semantisch ähnlicher Lesarten herangezogen werden, die in SemCor vorkommen. Die semantische Ähnlichkeit zweier Lesarten wird über die Hyponymie-Hierarchie von WordNet ermittelt: Je näher beieinander die entsprechenden Konzepte in der Hierarchie angeordnet sind, desto größer ist die semantische Ähnlichkeit.

4.2 Informationserschließung

Bei der automatischen Informationserschließung (‚Information Retrieval') geht es darum, aus einem umfangreichen Inventar von Dokumenten diejenigen Texte zu finden, die bestimmte, durch eine Anfrage spezifizierte Informationen enthalten. Für diese Aufgabe kann ein semantisches Netz nützliche Hinweise liefern. Wenn sowohl die Anfrage als auch die zu durchsuchenden Dokumente semantisch disambiguiert sind, kann gezielt nach Begriffen in der intendierten Lesart gesucht werden. Wenn z.B. nach Informationen zum Stichwort **Bank** (im Sinne von **Geldinstitut**) gesucht wird, so werden keine Texte über Sitzmöbel geliefert. Dadurch wird die „Treffergenauigkeit" der ermittelten Dokumente (‚precision') erhöht. Ein zweiter Vorteil eines semantischen Netzes ist, daß mit seiner Hilfe die Anfrage um Konzepte erweitert werden kann, die mit den Suchbegriffen in einer semantischen Beziehung stehen. So können bei einer Anfrage nach **Bank** auch Texte gefunden werden, in denen der Begriff selbst nicht vorkommt, jedoch **Geldinstitut** oder **Sparkasse**. Dadurch wird die Anzahl der korrekt ermittelten Dokumente (‚recall') erhöht. Mit einem multilingualen semantischen Netz wie EuroWordNet ist so auch die Durchsuchung von Texten in unterschiedlichen Sprachen möglich, indem die Anfrage um Konzepte aus verschiedenen Sprachen erweitert wird, die zu den Suchbegriffen äquivalent sind.

Gonzalo et al. (1998) haben durch entsprechende Experimente herausgefunden, daß die Performanz von Information Retrieval signifikant erhöht wird, wenn die Anfrage und die zu durchsuchenden Texte mit WordNet-*synsets* indiziert sind (d.h. jedem Wort das entsprechende *synset* zugewiesen wird). Diese Indizierung liefert erstens Lesartendisambiguierung und zweitens die Erweiterung der Anfrage um Synonyme der Suchbegriffe.

4.3 Textkategorisierung

Textkategorisierung befaßt sich mit der Klassifikation von Texten im Hinblick auf eine (vorgegebene) Menge von Kategorien (z.B. Domänen oder Textsorten). Systeme zur automatischen Textkategorisierung werden zunächst mit Hilfe einer Kollektion von Texten trainiert, die manuell mit dem vorgesehenen Inventar von Kategorien klassifiziert wurden. Ein neu zu kategorisierender Text wird mit den Texten in dieser Kollektion bzgl. der Vorkommenshäufigkeit bestimmter Begriffe verglichen. Die ausgewählte Kategorie ergibt sich aus diesem Vergleich.

Buenaga Rodríguez et al. (1997) ziehen WordNet als zusätzliche Informationsquelle heran: In das Kategorisierungsverfahren gehen auch die Vorkommenshäufigkeiten der Kategoriebezeichnungen selbst sowie ihrer Synonyme im zu klassifizierenden Text ein. Die Synonyme werden aus WordNet extrahiert. Auch hier ist die Einbeziehung anderer Konzepte denkbar, die mit den Kategoriebezeichnungen durch semantische Relationen verbunden sind.

4.4 Eine lexikographische Anwendung: Selektionsbeschränkungen

In diesem Abschnitt wird exemplarisch eine lexikographische Anwendung des GermaNet ausführlicher beschrieben: die Akquisition von Selektionsbeschränkungen. Diese Aufgabe stellt gleichzeitig eine Perspektive für die qualitative Weiterentwicklung von GermaNet dar.

Selektionsbeschränkungen sind semantische Beschränkungen, die ein Prädikat (z.B. ein Verb oder Adjektiv) seinen Argumenten (z.B. einer Verbergänzung oder einem durch ein Adjektiv modifizierten Nomen) auferlegt. So fordert beispielsweise das Verb *essen* einen menschlichen oder tierischen Agens und einen Patiens, der ein Nahrungsmittel bezeichnet.

4.4.1 Motivation

Die Akquisition von Selektionsbeschränkungen ist aus mehreren Gründen sinnvoll. Zum einen können sie einen wichtigen Beitrag zur syntaktischen und lexikalischen Disambiguierung leisten. Im Satz

(1) Das Brot schneidet die Mutter.

folgt aus den Selektionsbeschränkungen von *schneiden*, daß *das Brot* die Akkusativ- und *die Mutter* die Nominativergänzung ist, nicht umgekehrt, wie es nach rein morphologischen und syntaktischen Kriterien möglich wäre. Im Beispiel

(2) Der Mann tritt gegen den Ball.

wird *Ball* aufgrund der Selektionsbeschränkungen von *treten* als Spielgerät (und nicht als Tanzveranstaltung) disambiguiert. Selektionsbeschränkungen sind also als eine Informationsquelle zur Disambiguierung für maschinelle Sprachverarbeitungssysteme interessant.

Daneben kann es jedoch auch zweckmäßig sein, Selektionsbeschränkungen in Lexika für menschliche Benutzer aufzunehmen. Vor allem für Fremdsprachenlerner können sie wichtige Hinweise für den Wortgebrauch liefern, die aus der Wortbedeutung nicht unbedingt und aus Verwendungsbeispielen höchstens indirekt hervorgehen. Z.B. verwendet man *tranchieren* nur im Zusammenhang mit Fleisch, nicht mit Fisch, Gemüse oder Holz, was aus der Wortbedeutung nicht zwingend folgt. In einigen deutschen Wörterbüchern sind Selektionsbeschränkungen explizit angegeben (z.B. in VERBEN IN FELDERN und im WÖRTERBUCH ZUR VALENZ UND DISTRIBUTION DEUTSCHER VERBEN), in anderen sind sie implizit in Bedeutungsdefinitionen und Beispielen enthalten.

4.4.2 Akquisition von Selektionsbeschränkungen

Es ist offensichtlich, daß die manuelle Akquisition von Selektionsbeschränkungen zeit- und arbeitsaufwendig ist, wenn sie für ein Lexikon breiteren Umfangs durchgeführt werden soll. Dies ergibt sich nicht zuletzt aus der Tatsache, daß prinzipiell jede semantische Eigenschaft eine Rolle bei Selektionsbeschränkungen spielen kann und es folglich empirisch nicht adäquat ist, Selektionsbeschränkungen mit einem relativ kleinen Inventar semantischer Merkmale wie **belebt** oder **abstrakt** zu modellieren. Manche Prädikate stellen sehr spezielle Selektionsanforderungen an ihre Argumente, die von einem solchen Inventar nicht erfaßt

werden können. So kann man nichts anderes *diagonalisieren* als eine *Matrix*, und nur eine *Geschwulst* kann als *gutartig* charakterisiert werden.

In den letzten Jahren sind Verfahren entwickelt worden, um Selektionsbeschränkungen durch statistische Analyse großer Textkorpora zu ermitteln (Resnik 1993, Ribas 1994, Abe/Li 1996). Diese Verfahren ermitteln WordNet-Konzepte, die von einem Prädikat präferiert werden (z.B. *food* als Objekt von *eat*). Hierbei weisen sie den Konzepten jeweils einen Präferenzwert zu, der die Stärke der Präferenz charakterisiert und auf der Grundlage relativer Häufigkeiten von Prädikat-Argument-Kookkurenzen im untersuchten Korpus berechnet wird. Tatsächlich haben Selektionsbeschränkungen eher den Charakter von Präferenzen als von scharfen Restriktionen. So sind kontextbedingte oder metaphorisch zu interpretierende Abweichungen von Selektionspräferenzen wie in *Angst essen Seele auf* durchaus gängig. Außerdem können auch innerhalb des durch Selektionsbeschränkungen sanktionierten „semantischen Raumes" unterschiedliche Präferenzgrade vorliegen. So lassen die Selektionsbeschränkungen von *lesen* für den Satz

(3) Der Student liest den Artikel.

die Interpretation von *Artikel* sowohl als ein Determinans als auch als Text zu. Jedoch wird die Text-Lesart stärker präferiert, sofern kein spezifischer Kontext die andere Interpretation nahelegt. Die statistischen Verfahren haben also neben der automatischen Akquisition auf breiter empirischer Basis den Vorteil, daß sie durch die Quantifizierung des Präferenzverhaltens eines Prädikats das Phänomen Selektionsbeschränkungen adäquater modellieren.

Mit Hilfe von GermaNet sollen mit diesen Verfahren Selektionspräferenzen für das Deutsche ermittelt werden. Für lexikographische Zwecke ist es hierbei wichtig, daß sich die ermittelten Konzepte auf einer angemessenen Generalisierungsebene befinden. Angenommen, wir stoßen u.a. auf folgende Korpusbelege für das Prädikat *essen*:

(4a) Meine Tochter ißt gern Käsekuchen,
(4b) Max hat schon drei Äpfel gegessen.
(4c) Muslime essen kein Schweinefleisch.

Die Komplemente sollen einerseits möglichst kompakt, andererseits empirisch adäquat repräsentiert werden. Das Konzept, das die Objekte in den genannten Beispielen angemessen zusammenfaßt, ist *Nahrungsmittel*. *Gegenstand* wäre zu generell; Konzepte wie *Backwaren*, *Obst*, *Fleisch*, etc. würden dem Kompaktheitsdesiderat zuwiderlaufen.

Das Problem der angemessenen Generalisierung wird von den oben genannten Verfahren nicht befriedigend gelöst. Abe/Li (1996) nehmen zwar für sich in Anspruch, einen informations-theoretisch motivierten Ansatz zu implementieren, der die angemessene Generalisierungsebene liefert. Jedoch haben eigene Experimente gezeigt, daß der ermittelte Generalisierungsgrad von der Größe des untersuchten Korpus sowie der Häufigkeit des untersuchten Prädikats abhängt: Bei häufigen Verben wird tendentiell untergeneralisiert, bei seltenen Verben tendentiell übergeneralisiert. Dieses Verhalten ist zumindest für lexikographische Zwecke nicht akzeptabel. Hier sind also Modifikationen der Verfahren notwendig.

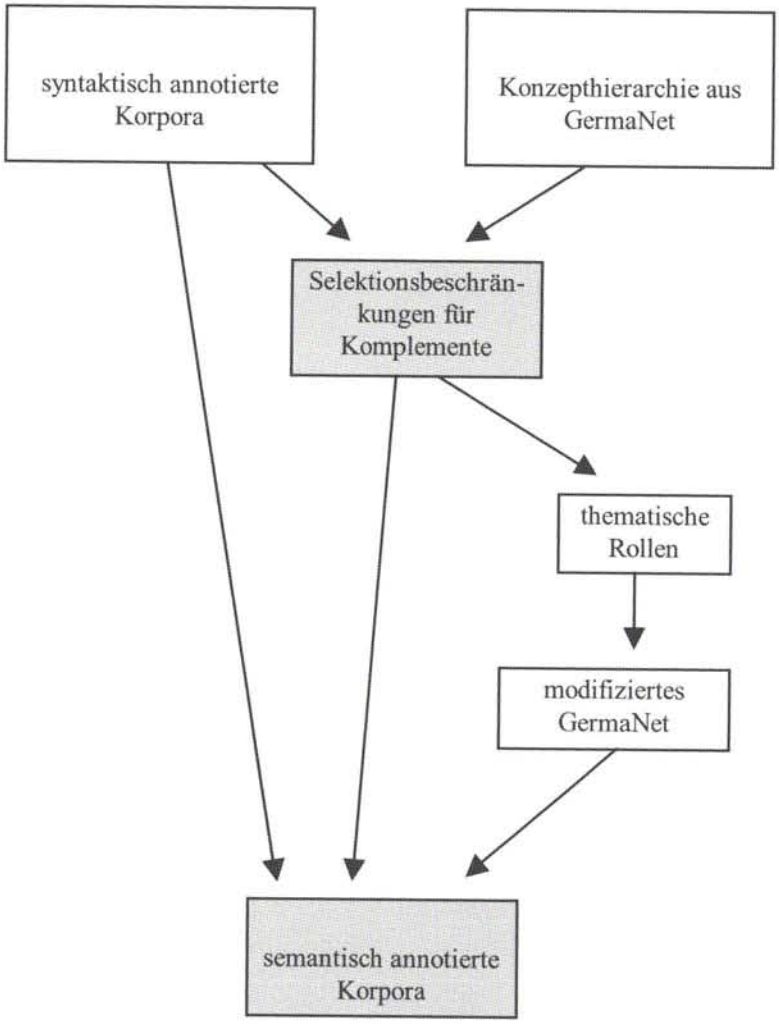

Abb. 7: Selektionspräferenzen und semantische Annotation

4.4.3 Kodierung von Selektionspräferenzen in GermaNet

Die oben genannten Verfahren ermitteln selektionale Präferenzen für syntaktische Verbkomplemente. Jedoch sind Selektionspräferenzen primär mit thematischen Rollen wie Agens, Patiens oder Instrument assoziiert, die auf unterschiedliche Weise syntaktisch realisiert werden können. So kann der Patiens von *kochen* sowohl als Nominativ- als auch als Akkusativergänzung realisiert werden:

(5a) Der Küchenchef kocht die Suppe.
(5b) Die Suppe kocht.

Solche Korpusbelege können dazu führen, daß *Nahrungsmittel* sowohl für das Subjekt als auch das direkte Objekt von *kochen* als präferiertes Konzept ermittelt wird. Beiden Sätzen liegt jedoch zugrunde, daß *kochen Nahrungsmittel* für seine Patiens-Rolle präferiert.

Um das Selektionsverhalten eines Verbs adäquat zu modellieren, müssen also die für seine syntaktischen Komplemente ermittelten Präferenzen auf die zu Grunde liegenden thematischen Rollen abgebildet werden. Um diesbezüglich geeignete Verfahren zu entwickeln, bietet sich z.B. der Ansatz von McCarthy/Korhonen 1998 an.

Die gewonnenen Präferenzen für thematische Rollen sollen in GermaNet kodiert werden. Hierbei soll auf das Inventar entsprechender Relationen zurückgegriffen werden, das im Rahmen der EuroWordNet-Spezifikation festgelegt wurde. Beispielsweise soll die Relation *kochen* **involved_patient** *Nahrungsmittel* in GermaNet aufgenommen werden.

4.4.4 Semantische Annotierung von Korpora

Das um thematische Relationen erweiterte GermaNet soll u.a. für die Erstellung semantisch annotierter Korpora (genauer: Korpora, deren Wörter semantisch disambiguiert sind) genutzt werden, die wiederum für bisher genannte Anwendungen von großem Interesse sind. Nicht nur die Hyponymiebeziehungen sind hierbei nützlich (vgl. 4.1), sondern auch die thematischen Rollenrelationen. So kann Ball in Beispiel (2) durch eine **involved_patient** Relation disambiguiert werden, die zwischen *treten* und der physikalischen Lesart von *Ball* besteht.

Aber auch die statistisch ermittelten Selektionspräferenzen selbst (s. Abschnitt 4.4.2) bilden eine wichtige Informationsquelle für die semantische Disambiguierung, da z.B. die Präferenzstärke, die nicht in GermaNet kodiert werden soll, essentiell sein kann (vgl. Beispiel (3)).

Die von uns vorgesehene Anwendung, das GermaNet zusammen mit syntaktisch annotierten Korpora für die Ermittlung von Selektionspräferenzen und die semantische Korpusannotierung einzusetzen, ist in Abbildung 7 schematisch dargestellt:

5 Schlußwort

In diesem Aufsatz haben wir den Aufbau und die grundlegenden Eigenschaften des GermaNet, eines lexikalisch-semantischen Wortnetzes für das Deutsche, beschrieben und seine Anwendungsperspektiven für die Computerlinguistik dargelegt. Wir haben gezeigt, wie wir von der Kooperation im Rahmen eines multilingualen Projektes in bezug auf die qualitative als auch quantitative Abdeckung profitieren konnten. Die mittlere Größe des Netzes und die Qualität der Daten bieten eine empirische Basis sowohl für theoriebezogene Fragestellungen als auch für praktische Anwendungen.

Der Ausbau unserer Ressource umfaßt nicht nur die korpusbasierte Erweiterung des repräsentierten Wortschatzes, sondern auch die Adaption neuer innersprachlicher Relationstypen, wie sie durch thematische Rollen gegeben sind. So kann GermaNet noch effizienter die Ermittlung von Selektionspräferenzen sowie die semantische Annotierung von Korpora unterstützen.

6 Literatur

Abe, Naoki und Li, Hang (1996): Learning Word Association Norms Using Tree Cut Pair Model. In: Proc. of 13th Int. Conf. on Machine Learning.

Buenaga Rodríguez, Manuel de und Gómez-Hidalgo, José-María und Díaz-Agudo, Belén (1997): Using WordNet to Complement Training Information in Text Categorization. In: Proc. of 2nd Int. Conf. on Recent Advances in NLP.

Buitelaar, Paul (1998): CORELEX: Systematic Polysemy and Underspecification. PhD thesis. Brandeis University.

COLLINS German-English, English-German Dictionary. Hg. Peter Terrell, Veronika Schnorr, Wendy V. A. Morris, Roland Breitsprecher. Glasgow: Harper-Collins ^{2}1991.

Cruse, Alan (1986): Lexical Semantics. Cambridge: Cambridge University Press.

DEUTSCHER WORTSCHATZ. EIN WEGWEISER ZUM TREFFENDEN AUSDRUCK. Hgg. H. Wehrle und H. Eggers. Stuttgart: Ernst-Klett-Verlag 1961.

Dowty, Donald (1988): On the Semantic Content of the Notion Thematic Role. In: G. Chierchia, B. Partee und R. Turner (eds.): Property Theory, Type Theory and Natural Language Semantics. Dordrecht: Kluwer.

DUDEN 8: SINN- UND SACHVERWANDTE WÖRTER. Hgg. Günther Drosdowski, Wolfgang Müller, Werner Scholze-Stubenrecht, Matthias Wermke. Mannheim: Dudenverlag ^{2}1986.

EuroWordNet: Building a multilingual database with wordnets for several European languages. University of Amsterdam. 31. Mai 1999, http://www.let.uva.nl/ewn/.

Fellbaum, Christiane (1998): WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

GermaNet. Universität Tübingen. 31. Mai 1999, http://www.sfs.nphil.uni-tuebingen.de/lsd/.

Gonzalo, Julio und Verdejo, Felisa und Chugur, Irina und Cigarrán, Juan (1998): Indexing with WordNet synsets can improve text retrieval.

Hamp, Birgit (1997): German Particle Verbs in GermaNet. Unveröffentlichtes Arbeitspapier.

– und Feldweg, Helmut (1997): GermaNet – A lexical-semantic net for German. In: Proc. of ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

Harley A. (1994): Cambridge Language Survey: semantic tagger. Technical report. Acquilex II Working Paper 39. Cambridge: Cambridge University Press.

McCarthy, Diana und Korhonen, Anna (1998): Detecting Verbal Participation in Diathesis Alternations. In: Proc. of 36th Annual Meeting of the ACL, Montreal, Vol. 2, 1493–1495.

McCawley, James D. (1968): The role of semantics in grammar. – In: E. Bach, R. Harms (eds.): Universals in Linguistic Theory. 125–169. New York: Holt, Rinehart & Winston.

Miller, George und Beckwith, Richard und Fellbaum, Christiane und Gross, Derek und Miller, Katherine (1990): Five Papers on WordNet. CSL Report, Vol. 43. Cognitive Science Laboratory, Princeton University.

– und Leacock, Claudia und Tengi, Randee (1993): A semantic concordance. In: Proc. of ARPA Human Language Technology Workshop 303–308.

Peters, Wim und Peters, Ivonne und Vossen, Piek (1998): The Reduction of Semantic Ambiguity in Linguistic Resources. In: Proc. of 1st Int. Conf. on Language Resources and Evaluation, Granada.

Resnik, Philip Stuart (1993): Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis. University of Pennsylvania.

Ribas, Francesc (1994): An experiment on learning appropriate selectional restrictions from a parsed corpus. In: Proc. of COLING, Kyoto.

Rosch, Eleanor (1978): Principles of Categorization. – In: E. Rosch, B. Lloyd (eds.): Cognition and Categorization. Hillsdale: Lawrence Erlbaum Associates. 27–48.

Stetina, Jiri und Kurohashi, Sadao und Nagao, Makoto (1998): General Word Sense Disambiguation Method Based on Full Sentential Context. In: Proc. of COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal.

Vossen, Piek (1997): EuroWordNet-2. Extending EuroWordNet with other languages. Annex I to Telematics Application Programme. LE 4–8328.

VERBEN IN FELDERN. VALENZWÖRTERBUCH ZUR SYNTAX UND SEMANTIK DEUTSCHER VERBEN. Hg. Helmut Schumacher. Berlin: Walter de Gruyter 1986.

WÖRTERBUCH ZUR VALENZ UND DISTRIBUTION DEUTSCHER VERBEN. Hgg. Gerhard Helbig, Wolfgang Schenkel. Leipzig: VEB Bibliographisches Institut 1969.

Yarowsky, D. (1992): Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proc. of 15th Int. Conf. on Computational Linguistics. Vol II, 454–460.

Claudia Kunze, Tübingen
Andreas Wagner, Tübingen