Anwendungen des deutschen Wortnetzes in Theorie und Praxis

Beiträge des GermaNet-Workshops
Tübingen, Oktober 2003

LDV/ Impressum

LDV-Forum
ISSN 0175-1336

Zeitschrift für Computerlinguistik und Sprachtechnologie
GLDV-Journal for Computational Linguistics and
Language Technology – Offizielles Organ der GLDV

Herausgeber und Redaktion

Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV)
Prof. Dr. Christian Wolff
Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaft
D-93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Wissenschaftlicher Beirat

Vorstand, Beirat und Arbeitskreisleiter der GLDV
http://www.gldv.org/organe.htm, http://www.gldv.org/AKs/index.htm

Band 19 - 2004
Heft 1/2 - Sonderheft

Beiträge des GermaNet-Workshops, Tübingen, Oktober 2003

Herausgeber des Sonderhefts

Claudia Kunze, Lothar Lemnitzer, Andreas Wagner
Universität Tübingen, Seminar für Sprachwissenschaft
Wilhelmstr. 19, D-72074 Tübingen

Erscheinungsweise

2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind über die Website der GLDV einsehbar (*http://www.gldv.org*).

Einreichung von Beiträgen

Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall elektronisch und zusätzlich auf Papier übermittelt werden. Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind an den Herausgeber zu übermitteln.

Bezugsbedingungen

Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von 25,- € (inkl. Versand), Einzelexemplare zum Preis von 15,- € (zzgl. Versandkosten) bei der Redaktion bestellt werden.

Satz

Christoph Pfeiffer, Regensburg, mit *Adobe InDesign CS 3.0.1*

Druck

Druck TEAM KG, Regensburg

Christian Wolff

Editorial

Liebe GLDV-Mitglieder, liebe Leserinnen und Leser des LDV-Forum,

der 19. Band des LDV-Forum erscheint erneut als Sonderheft, gefüllt mit Beiträgen eines Workshops zu aktuellen Entwicklungen im Umfeld des deutschen Wortnetzes (GermaNet).

Der Workshop fand - mit Unterstützung durch die GLDV - im Oktober 2003 am Seminar für Sprachwissenschaft der Universität Tübingen statt. Die Anregung, die Beiträge dieses Workshops im LDV-Forum zu veröffentlichen, verdanke ich Bernhard Schröder, Bonn.

Der vorliegende Band ist als Doppelheft die 19te Ausgabe des Forum; noch in diesem Jahr wird ihm ein Themenheft mit dem Schwerpunkt Text Mining folgen, das ursprünglich noch vor dem aktuellen Band erscheinen sollte. In Synchronisation mit dem Ausbau der GLDV-Website (*http://www.gldv.org*) wird dabei noch stärker als bisher der Schwerpunkt auf den eigentlichen Fachbeiträgen liegen, da Rubriken wie aktuelle Nachrichten oder Tagungsankündigungen im schnelleren elektronischen Medium besser aufgehoben scheinen.

Nachdem im vergangenen Jahr bereits die Beiträge zur GLDV-Frühjahrstagung an der Hochschule Anhalt in Köthen in leicht modifiziertem Gewand erschienen sind, wurden beim vorliegenden Heft nicht nur der Mantel, sondern auch die Beiträge selbst in neuem Layout und neuer Typographie überarbeitet. Der Vorstand der GLDV hat sich dafür entschieden, nach (fast vollständigem) Abschluss des von Gerhard Knorz gestalteten „Buchstaben-Designs“ auf eine neue Gestaltung umzustellen, da der Abschluss der Serie verbunden mit dem Übergang der Herausgeberschaft einen natürlichen Einschnitt bot.

Nach mehrjähriger Verzögerungsphase, die nicht nur mit dem Wechsel der Herausgeberschaft zusammenhing, sondern zudem noch mit mehrmaligen Ortswechseln des neuen Herausgebers, soll in diesem Jahr wieder der reguläre Publikationstakt des LDV-Forum erreicht werden. Allen Leserinnen und Lesern des LDV-Forum sei jedenfalls für Ihre Geduld ob des verzögerten Publikationstaktes sehr herzlich gedankt.

Die Gestaltungsvorlage wurde von Frank Heckel, Jens Kabisch, Marco Spitzer und Stefan Müller, Studenten der Medieninformatik an der TU Chemnitz im Rahmen eines Praktikums zur Mediengestaltung entworfen. Die Dokumentaufbereitung und den Satz hat freundlicherweise Herr cand. phil. Christoph Pfeiffer übernommen, dem dafür herzlich gedankt sei. Die technischen Schwierigkeiten der Satzvorbereitung, insbesondere die Konvertierung von LaTeX-Vorlagen in MS Word und anschließend in Adobe InDesign hat dabei einige Verzögerungen verursacht.

Regensburg, im April 2004

Christian Wolff

Inhaltsverzeichnis

LDV FORUM - Band 19 - 2004
Beiträge des GermaNet-Workshops, Tübingen, Oktober 2003

LDV/ Inhaltsverzeichnis

Claudia Kunze, Lothar Lemnitzer, Andreas Wagner

Einleitung

Der vorliegende Band enthält alle Beiträge des 1. GermaNet-Anwender-Workshops, der vom 09.–10.10. 2003 in Tübingen stattgefunden hat. Wir blicken auf eine sehr produktive und anregende Veranstaltung mit zahlreichen Diskussionen zurück, Dank des großen Engagements aller Beteiligten. In diesem Vorwort wollen wir kurz die Entwicklung skizzieren, die zur Realisierung unserer Tagung geführt hat.

Seit dem unerwarteten Erfolg, den das Princeton WordNet (vgl. Miller et al. 1990, Fellbaum 1998) als „Mutter aller Netze" in Computerlinguistik und Sprachtechnologie verbuchen konnte, sind zahlreiche einzelsprachliche und auch multilinguale Wortnetze in nationalen und internationalen Projekten entwickelt worden, z.B. *ItalNet, Dutch WordNet, Portuguese WordNet, EuroWordNet, BalkaNet*, um nur einige zu nennen.

Wortnetze im Stile des Princeton WordNets bilden die häufigsten und wichtigsten Wörter einer Sprache und ihre bedeutungstragenden Beziehungen zu anderen Wörtern der Sprache ab. Ein zentraler Anwendungsbereich ist die Lesartendisambiguierung, welche eine unabdingbare Voraussetzung für Anwendungen im Bereich der Informationserschließung und der Maschinellen Übersetzung, für die semantische Annotierung von Sprachdaten und die Entwicklung verschiedener Werkzeuge zum Sprach- und Informationserwerb darstellt.

In Tübingen hatte Helmut Feldweg die Idee, ein Wortnetz für das Deutsche zu erstellen, und konnte 1995 erste Projektmittel vom Land Baden–Württemberg einwerben. Mit großem Engagement begann das erste LexikographInnen-Team um Helmut - darunter Birgit Hamp, Michael Hipp, Susanne Schüle, Rosmary Stegmann, Christine Thielen, Valérie Béchet-Tsarnou - unter der Anleitung Christiane Fellbaums den Aufbau des GermaNet (vgl. Hamp & Feldweg 1997). Durch den Weggang mehrerer Mitarbeiter, aber vor allem durch Helmuts plötzlichen, viel zu frühen Tod, drohte das GermaNet–Projekt zu scheitern, weil niemand mehr zur Weiterentwicklung und Pflege der Daten verfügbar war.

Im Zuge der noch von Helmut akquirierten Projektbeteiligung am EuroWordNet-2 Vorhaben kamen mit Andreas Wagner und Claudia Kunze zwei neue Mitarbeiter, die ab 1998 die Integration des deutschen Wortnetzes in die multilinguale Datenbank EuroWordNet vornahmen und GermaNet selbst in dieser Zeit signifikant erweitern und verbessern konnten - nicht zuletzt durch die Standardisierungsbemühungen des internationalen Konsortiums unter der Leitung von Piek Vossen (vgl. Vossen 1999, Wagner & Kunze 1999) und den Abgleich mit anderen Wortnetzen.

Nach dieser Phase produktiver Kooperation mit europäischen Partnern wurde die Weiterentwicklung des GermaNet durch Karin Naumann und Claudia Kunze in einem zweijährigen Projekt erneut vom Land Baden–Württemberg gefördert (vgl. Kunze 2001). Auf Initiative von Lothar Lemnitzer entstand eine XML-Version der GermaNet ‚Lexicographers' Files', auf dessen Basis ein Visualisierungstool erstellt wurde (vgl. Kunze & Lemnitzer 2002a, Kunze & Lemnitzer 2002b).

GermaNet ist mittlerweile eine stattliche Ressource mit ca. 42 000 Synsets und mehr als 61 000 Lesarten, die von zahlreichen akade-

mischen und industriellen Anwendern genutzt wird. Uns interessierte zunehmend, welche Erfahrungen die GermaNet-Nutzer mit GermaNet in konkreten Anwendungen machen; dagegen ist das Feedback nach der Lizenzierung der Daten seitens der Anwender oftmals nicht mehr allzu groß.

Auf der Language Ressources and Evaluation Conference (LREC) 2002 wurde im Rahmen eines von uns gemeinsam mit der BalkaNet Gruppe durchgeführten Workshops über „Wordnet structures and standardization" (vgl. Lemnitzer et al. 2002) von mehreren Teilnehmern angeregt, in Tübingen einen GermaNet User Workshop zu veranstalten; eine Idee, die wir gern aufgegriffen haben. Dank einer Zusage des Vorstandes der „Gesellschaft für linguistische Datenverarbeitung" konnten wir unser Vorhaben als GLDV-Workshop realisieren und ankündigen.

Die 14 Beiträge haben wir inhaltlich auf drei Sektionen verteilt, wobei diese jeweilig recht heterogene Themen und Ansätze umfassen.

Die erste Sektion fokussiert Anwendungen des GermaNet und des EuroWordNet in verschiedenen, vorwiegend computerlinguistischen Szenarios.

„Using EWN within the Speech Operated System EMBASSI" von Iman Thabet, Bernd Ludwig, Frank-Peter Schweinberger, Kerstin Bücher und Günter Görz thematisiert inkrementelle semantische Konstruktion auf Grundlage der Description Logics für ein Dialogsystem gesprochener Sprache. In diesem Ansatz fungiert EuroWordNet als lexikalische Hintergrundressource zur Modellierung der linguistischen Daten, welche in verschiedenen Systemkomponenten, internen Domänen und Anwendungen wiederverwendet werden sollen. Die AutorInnen geben einen umfassenden Überblick über die semantische Verarbeitung in ihrem Dialogsystem und die Rolle, die EuroWordNet Synsets darin spielen. Sie formulieren Desiderate an ein ideales Wortnetz als geeignete Wissensbasis.

Manuela Kunze und Dietmar Rösner berichten in ihrem Paper „Issues in Exploiting GermaNet as a Resource in Real Applications" über ihre Erfahrungen mit GermaNet als Hintergrundressource für die domänenspezifische Dokumentenanalyse, die sie anhand einer Testsuite von Autopsieprotokollen exemplifizieren. Sie stellen Überlegungen zur Integration des GermaNet in die XML-basierte Dokumentenverarbeitung (XDOC) an und schließen mit einer Wunschliste an die GermaNet-Entwickler, die sich vor allem um einen verbesserten Zugriff auf orthographische und morphologische Varianten der Suchwörter zentriert.

In seinem Beitrag „Die semantische Auswertung von Produktanforderungen mit Hilfe des GermaNet" skizziert Matthias Jörg, wie GermaNet ein industrielles System („ReMaS") unterstützen kann, das für die Analyse auch natürlichsprachlich vorliegender Produktanforderungen als Ergebnis ein semantisches Anforderungsnetzwerk für Produktentwickler, z.B. Automobilhersteller, erzeugen soll. GermaNet wird bei der Auswertung der Anforderungsmodellierung, die im Paper aus Anwenderperspektive umfassend dargestellt wird, zum Einsatz kommen.

Der Nutzen des GermaNet für andere lexikalisch-semantische Ressourcen wird im Beitrag „Die Verwendung zur Pflege und Erweiterung des Computerlexikons HaGenLex" von Rainer Osswald thematisiert. Mittels Lesartenzuordnung zwischen Einträgen aus GermaNet und HaGenLex wird die Abdeckung beider Ressourcen verglichen und die Aufdeckung von lexikon-internen Inkonsistenzen ermöglicht. Diese Methode gestattet die Übernahme der lexikalisch–semantischen Relationen aus GermaNet in die ebenfalls semantisch basierte HaGenLex-Struktur. Der Aufsatz deckt Schwachstellen in den GermaNet-Hierarchien bezüglich der aspektuellen Charakterisierung von Verben und der Kreuzklassifikation von Partikelverben auf und schlägt in diesen Bereichen eine Restrukturierung vor.

Einleitung

Um kognitiv inspirierte Wissensmodellierung geht es beim Mapa (=Mapping Archtecture for People's Association)- Studienprojekt, das Kathrin Beck in ihrem Beitrag „Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa" in Bezug auf die im Titel genannte Beispielanwendung vorstellt. Mapa ist eine Plattform, welche die benutzerfreundliche Modellierung von Wissensnetzen ermöglicht, im Falle des hier vorgestellten Vokabeltrainers die GermaNet Daten als Netzstruktur, auf deren Grundlage deutsche Vokabeln „im Kontext" verstanden und gelernt werden können.

Besonders freuen wir uns über den künstlerisch inspirierten Beitrag von Daniela Alina Plewe, Steffen Meschkat und Manfred Stede, die das Projekt „GeneralNews – ein Metabrowser" präsentieren. In Realzeit ersetzt der beschriebene Metabrowser Wörter auf Webseiten durch ihre Synonyme, Hyperonyme oder Hyponyme und erzeugt dadurch ähnliche, abstraktere oder spezifischere Texte, die somit neue Beschreibungen der Welt generieren bzw. die Perspektive auf „neue Welten" eröffnen. Als Hintergrundressource lexikalischen Wissens fungiert WordNet (bzw. GermaNet für eine deutsche Version des Metabrowsers).

Auch die zweite Sektion thematisiert Anwendungen des GermaNet oder WordNet, wobei hier jeweilig die quantitativen Methoden im Vordergrund stehen.

Sabine Schulte im Walde zeigt mit „GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering", wie statistische Verfahren zu einer verfeinerten automatischen Verbklassifizierung führen können, wenn die Argumentpositionen von diathetischen Verben mit Information über ihre selektionalen Präferenzen in Form von semantischen Topknoten aus GermaNet angereichert werden. Sie vergleicht drei verschiedene Ausgangsebenen für die Verbklassifikation (eine rein syntaktisch motivierte Vorklassifikation, eine syntaktische Vorklassifikation, die geforderte und freie Präpositionalphrasen integriert, und schließlich die zusätzliche Ausstattung der Argumentpositionen mit semantischen Knoten, welche die selektionale Präferenz ausdrücken), und deren Eignung für die Erkennung semantischer Verbcluster.

Andreas Wagner diskutiert in seinem Beitrag „Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies" verschiedene Verfahren zur Häufigkeitsschätzung von Konzepten in Wortnetzen mit Hilfe von Korpusdaten. Insbesondere werden Probleme angesprochen, die sich in diesem Zusammenhang durch die multiple Vererbungsstruktur von Wortnetzen ergeben, und es wird ein neuer Lösungsansatz für diese Probleme vorgeschlagen. Da multiple Vererbung (Kreuzklassifikation) ein wesentliches Strukturierungsprinzip in GermaNet darstellt, ist der vorgestellte Ansatz für das deutsche Wortnetz besonders relevant.

In ihrem Beitrag „Domain Specific Sense Disambiguation with Unsupervised Methods" stellen Diana Steffen, Bogdan Sacaleanu und Paul Buitelaar Experimente zur semantischen Disambiguierung von Wörtern in medizinischen Texten vor. Die Lesarten werden aus GermaNet entnommen und in unüberwachten Verfahren disambiguiert. Die Autoren kombinieren verschiedene Disambiguierungsverfahren (vertikal und horizontal) sowie die Auswertung verschiedener Parameter, z.B. Größe und Struktur der Korpora.

Ein Verfahren zum Lernen von Relationen steht im Mittelpunkt des Papers „Lernen paradigmatischer Relationen auf iterierten Kollokationen", das Christian Biemann, Stefan Bordag und Uwe Quasthoff präsentieren. Relationen zwischen Wörtern wie Synonymie, Hyponymie und Kohyponymie werden auf der Basis einer Trainingsmenge, welche die passenden semantischen Merkmale liefert, gelernt, um in mehreren Schritten lexikalisch-semantische Wortnetze (semi-)automatisch zu erweitern. Dabei

wird im Gegensatz zu anderen Verfahren bei der Trainingsmenge nicht nur von einem relevanten Merkmal, sondern von mehreren geeigneten Features ausgegangen.

Die dritte Sektion ist mit strukturbezogenen Aspekten zu Wortnetzen befasst, und zwar in Hinblick auf Repräsentationsvarianten wie TermNet und TopicMaps, Wortnetzverknüpfungen wie zwischen GermaNet und UniNet und die Integration mit anderen lexikalisch-semantischen Ressourcen wie FrameNet.

Maren Runte, Michael Beißwenger und Angelika Storrer thematisieren mit der „Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet" die aus dem Projekt HyTex erwachsene Konzeption eines domänenspezifischen semantischen Wissensnetzes für Texttechnologie und Hypermedia. Die Struktur des vorgestellten TermNet unterscheidet im Gegensatz zu WordNet zwischen lexikalischen und ontologischen Relationen, wodurch das automatische Linking von Wörtern in Fachtexten zu den geeigneten Domänenkonzepten verbessert werden kann.

Der Beitrag „Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps" von Eva Anna Lenz, Benjamin Birkenhage und Jan Frederik Maas zeigt, wie Wortnetze als XML Topic Maps repräsentiert und visualisiert und zum Hypertext-Linking genutzt werden. Da XML Topic Maps als Repräsentationsformat in Bezug zu anderen Repräsentationsvarianten für Wortnetze betrachtet wird, liefert diese Arbeit Überlegungen zum wichtigen Thema der Standardisierung semantischer Netze.

Die Erweiterung von Wortnetzen wie GermaNet steht im Mittelpunkt des Papers „GermaNet und UniNet: Anknüpfen an semantische Netze" von Simon Clematide. Der Autor argumentiert, dass für einen effizienten Einsatz in sprachtechnologischen Anwendungen das allgemeinsprachlich ausgerichtete GermaNet um Domänensprachen erweitert werden muss. In diesem Zusammenhang wird das UniNet der Universität Zürich, das Begriffe aus dem Bereich der Hochschule kodiert, und seine Integration mit GermaNet problematisiert.

Mit Petra Steiners Beitrag „FrameNet und WordNet, Perspektiven für die Verknüpfung zweier lexikalisch–semantischer Netze" wird ein weiterer Ansatz zur Integration von semantischen Ressourcen diskutiert. Die Autorin stellt das FrameNet-Projekt, Strukturen des FrameNet und die Gewinnung lexikalischer Information in diesem Ansatz vor, und präsentiert Überlegungen zur Verknüpfung der komplementären Informationen, die in FrameNet und GermaNet enthalten sind, in einem integrierten Datenbankmodell.

Wir danken der GLDV für ihre großzügige Förderung, durch welche wir diesen Workshop und seine Proceedings ohne Tagungsgebühren anbieten konnten. Dank gebührt Nicole Maruschka, Daniela Stelle und Zoulia Rakhmatoullina, welche bei der Vorbereitung des Workshops unermüdlich geholfen haben; Cornelia Stoll für ihren organisatorischen Beistand und Erhard Hinrichs für seine ideelle Unterstützung. Wir danken allen Teilnehmern, die zum Workshop nach Tübingen gekommen sind und mit ihren Paper- und Diskussionsbeiträgen diese Tagung erst möglich gemacht haben.

Literatur

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Kunze, C. (2001). „Lexikalisch-semantische Wortnetze." In: Carstensen, K.-U. et al. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: Eine Einführung. Heidelberg: Spektrum Akademischer Verlag, 386-393.

Kunze, C.; Lemnitzer, L. (2002a). "GermaNet - Representation, Visualization, Application." In: Proceedings LREC 2002. 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain, May/June 2002, 1485-1491.

Kunze, C.; Lemnitzer, L. (2002b). "Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet." In: Christodoulakis, Kunze & Lemnitzer (2002), 24-29.

Christodoulakis, D.N.; Kunze, C.; Lemnitzer, L.; (Hrsg.) (2002). Proceedings of the Workshop on Wordnet Structures and Standardizations, and how these Affect Wordnet Applications and Evaluation. 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 28th May 2002.

Miller, G. A. et al. (1990). Five Papers on WordNet. Technical Report 43, Princeton University, Cognitive Science Laboratory, *ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf* [accessed April 2004, first published in: Journal of Lexicography 3(4) (1990), 235-312].

Vossen, P. (Hrsg.) (1999). EuroWordNet: A Multilingual Database with Lexical-semantic Networks. Dordrecht: Kluwer Academic Publishers.

Wagner, A.; Kunze, C. (1999). "Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database." In: Sprache und Datenverarbeitung, 23(2) (1999), 5-20.

Iman Thabet, Bernd Ludwig, Frank-Peter Schweinberger, Kerstin Bücher,Günther Görz

Using EuroWordNet within the Speech Operated System EMBASSI

Abstract

In natural language processing, incremental semantic composition is one of the most prominent issues. In the past, numerous approaches have been developed for assigning meaning to noun and verb phrases and their complements and modifiers. However, their inferential power is often too weak to be applied to practical applications, or the expressiveness of the representation language is so complex, that it leads to intractable inference procedures. As an answer to these problems, we have developed an approach that relies on Description Logics (DL) for handling semantic construction. First, we will discuss this approach and show how a semantic knowledge base can be setup dependant on EuroWordNet[1] (EWN) as a linguistic ontology. Subsequently we will outline our experience with and demands on EWN.

1 What is EMBASSI and its Objective?

Embassi („Elektronische Multimediale Bedien- und Service-Assistenz“) has been a German joint project sponsored by the German Fed. Ministery of Research[2].

Our contribution to this project consists mainly of three components:

- The dialogue manager,
- formal ontologies for several multilingual application domains, and
- the language generation component to communicate system utterances to the user.

The long-term goal of our research is to design and implement a generic dialogue system for rational (spoken) dialogues that helps a user to achieve certain goals in terms of operations of a technical application system – e.g. an information system, a system for controlling devices, or any other kind of problem solving systems. One of its design criteria is the ability to recognize users' intentions in order to establish corresponding subgoals and control their processing. Furthermore, it should enable mixed-initiative, flexible and cooperative conversations, provide a high level of robustness as well as scalability at the linguistic and application dimensions, and easy portability to new domains. In addition, it should be possible to integrate multilingual linguistic interaction with multimodal forms of input and output such as graphical user interfaces, and – by means of appropriate devices – the recognition of deictic actions.

2 DL Models of Applications

Applications are characterized by a DL terminology which models the concepts used for making propositions about application situations. Basically, Embassi's knowledge base is composed of two parts: the EWN ontology, which encodes the linguistic meaning of words determined on an empirical basis, and the Standard Upper Ontology (SUMO) (NP01), which is used as a generic base model for concepts of the application domain (see Ludwig 2002).

3 Semantic Construction

This section discusses the issue of semantic construction during analyzing natural language input. We are using an incremental approach to the composition of semantic representations. The backbone of our approach is λ-DRT (see Fillmore 1969). The parser builds Discourse

Representation Structures (DRSes, see Kamp & Reyle 1993) incrementally and maps them onto ABoxes[3] (see Bücher et al. 2002).

The main question here is how the mapping of domain independent - in terms of EWN - to application specific language usage - in terms of a domain model - is done. In the discourse domain, referents usually refer to instances in the application domain. Such pairs of a discourse referent and a corresponding instance are represented by means of a special role called **has-lex**. For instance, in the definition

$$\texttt{AvEvent} \sqsubseteq \forall \texttt{has-lex.Program1}$$

it is claimed that an `AvEvent`[4] is related to a discourse referent of `Program1`. Consequently, all words that are assigned `Program1` as a meaning in EWN, designate an instance of `AvEvent` in the application domain. The DRS:

$$\left[\begin{array}{c} x\,y \\ \hline \texttt{AvEvent}(x) \\ \texttt{Program1}(y) \\ \texttt{has-lex}(x,y) \end{array}\right]$$

can be mapped onto an ABox asserting:

$$\texttt{Program1}(y) \wedge \texttt{has-lex}(x,y) \wedge \texttt{AvEvent}(x)$$

The set R of possible application specific readings of an instance of `Program1` is the set of all concepts (in the application domain) subsumed by the concept

$$\forall \texttt{has-lex.Program1}$$

Except in trivial cases, a direct mapping from the user utterance to a system command cannot be accomplished. In general, we have to take complex speech acts into account, where the interpretation of the utterance's propositional content is determined by its (local) linguistic-pragmatic context in the first place. This, in turn, is to a large extent influenced by (global) discourse-pragmatic features which provide constraints based on the dialogue history and the actual place of the utterance in the dialogue as being, for instance, the expected answer to a question. In addition, the application provides further constraints to limit the possible meanings of words and phrases to their particular use within the given thematic framework. Therefore, we have to distinguish between several interleaved levels of analysis of user utterances:

- Linguistic analysis on the utterance local level, which in turn consists of several levels of syntactic and semantic construction;
- Semantic evaluation, i.e. evaluation of semantic operators (e.g.disjunction and conditional expressions), reference resolution, and additional transformations of the logical form, augmented by specific computations;
- Application specific constraints on the evaluated semantic representation (see Ludwig 2002);
- Discourse-pragmatic analysis, i.e. determining the underlying speech act and accordingly the user's intention - a proper function of the dialogue manager.

4 Two Parsing Phases

If we want human-computer-dialogues to be natural, we must enable humans to talk to the computer as they do to humans. However, spontaneous speech is often incomplete or incorrect, full of interruptions and self-corrections leading to an ungrammatical input to the parser. Moreover, speech recognizers themselves may produce ungrammatical output even with correct input. Apart from this, parsing German input is difficult because of its fairly free word order and discontinuous constituents. Therefore the grammar cannot rely only on a linear sequence as its main concept.

We tried to overcome these problems by designing a two-phase parsing process (as presented in Bücher 2002). The first phase works with a grammar that employs phrase structure rules to build small phrases called chunks (similar to Abney 1991). Assigning semantic representation to the chunks also takes place in this phase. In the second phase, the interpretation of the whole utterance is derived by relating these chunks and their interpretation to each other.

4.1 Phase 1: Determining the Semantics of Chunks with the help of EWN

In order to ease the adaptability of the dialogue system to different domains and to reflect general and domain independent usage of language as distinct from that of a specific application, the semantics of the chunks is expressed in terms of concept expressions taken from EWN.

In this context, we would like to point out that Embassi is a multilingual system, so we depend not only on GermaNet[5] but also on EWN in our research. Also because the size of EWN is bigger than that of GermaNet we used to search for definitions in EWN if they were missing in GermaNet.

A chunk may consist of either only one element which is normally the head of the chunk C_h, or of a head element and one or more constituents that can be possible fillers of a free position in the head's structure.

After the categorization of the constituents the parser tries to build the chunks by combining the constituents pairwise:

$$C \rightarrow C_1\, C_2$$

The filler is usually a specifier (the determiner in case of a noun phrase *NP* or a modifier (e.g. a prepositional phrase *PP* modifying a verb).

If the chunk consists of only one constituent $C \rightarrow C_1$, which is the head of the chunk and therefore a terminal lexical category, we get the semantics of C from the lexicon, where the semantic information is stored as a λ-DRS (Kuschert 1996) (also referred to as the *extension* of the constituent). If C_1 is an expanded category[6] it contains the head of the chunk, and the semantics of C is derived from C_1. So, if there is only one symbol on the right side of the grammar rule, then the *extension* of the left side is determined as follows:

$$\mathrm{ext}(C) := \mathrm{ext}(\mathrm{head}(C))$$

In the case of a chunk consisting of a head C_h and another constituent C_f ($h \neq f \in \{1,2\}$), C_f is related to the discourse referent of C_h by a role R either taken from the inventory of EWN (see Vossen 1999) or defined by us[7]. Syntactically, the combination of two categories to a chunk is determined by a grammar rule which relates the two constituents via the role *R*. We then get the *extension* of the chunk by λ-composition of the DRSes of both constituents. In this case, the semantic head of the chunk is the one of its DRS:

$$\mathrm{head}(C) := \mathrm{head}(\mathrm{ext}(C))$$

When combining two elements, the parser checks the compatibility of the morphological features (e.g. agreement in case of the combination of a determiner with a *NP* and merges their DRSes resulting in a new DRS for the chunk. In this way, each chunk is assigned an interpretation already at this early stage. This has the advantage that if no further parsing is possible we thereby have means to interpret the whole utterance chunk by chunk.

To illustrate this, consider the utterance "*Kommt Tatort im ZDF?*"[8] taken from our Embassi application domain: To combine the preposition *im* and the *NP* chunk *ZDF* which was built using the (*NP* → *EN*)-rule we apply the following *PP*-rule[9]:

PP: P NP:
head = P:
role = has-value:
P morphfeat position = prepos,
P morphfeat kasrek = NP morphfeat case,
PP vpsynfeat clausetype =
NP vpsynfeat clausetype,
PP = P:

The *PP*-rule contains syntactic as well as semantic information about the chunk-combination. The DRS for the *PP*-chunk is constructed by the use of λ-composition of the DRSes of *ZDF* and *im* obtained from the lexicon:

$$\left[\frac{i}{\texttt{im-SP}(i)}\right] + \left[\frac{l}{\begin{array}{l}\texttt{TVStation1}(l)\\ \texttt{has-name}(l, ZDF)\\ \texttt{Name}(ZDF)\end{array}}\right] + \left[\frac{\emptyset}{\texttt{has-value}(i,l)}\right] = \left[\frac{i\,l}{\begin{array}{l}\texttt{TVStation1}(l)\\ \texttt{has-name}(l, ZDF)\\ \texttt{Name}(ZDF)\\ \texttt{im-SP}(i)\\ \texttt{has-value}(i,l)\end{array}}\right]$$

After having applied all phrase structure rules we get three chunks: The *NP Tatort*, the *PP im ZDF*, and the verb phrase *VP kommt*. Each chunk gets after this first phase a semantic interpretation on its own. The interpretation of the whole utterance is derived by relating these chunks and their interpretation to each other in phase two.

4.2 Phase 2: Applying Case Frames to Chunks

Phase two is different from phase one in that it combines chunks that needn't be adjacent to each other. Consequently, the order of the constituents is not relevant but may be an indicator for preferred readings when disambiguation is required. In this phase we use a kind of dependency grammar that determines for each chunk of phase one a list of possible syntactic functions it may have:

C_1 has C_2 →<synfunc>

<constraint equitation>

e.g.:

VP has *PP* →adverbial

NP has *PP* →attribut

VP has *NP* →subject

NP agr case = nom,

NP agr num = *VP* agr num.

The options are constrained by the morphological features of the chunk, e.g. a *NP*-chunk functions as a subject only if it has nominative case.

For the semantic head of each chunk there is a case frame[10] in which information about the valencies[11] are stored. The valencies of each chunk are filled by combining it with other chunks, e.g. building a *VP* from a verb and a *NP* that functions as its direct object, or expanding a *VP* by an adverb.

The suitability of the combination of two chunks is determined by the semantic constraints of the application domain. For example, consider the case frame for the verb *kommen* in the sense of "running a program":

infinitive: kommen

synt. function	thematic role	EWN concept
subject	agent:	Program1
adverbial	location:	TVStation1

From the case frame we derive hypotheses about possible fillers of a complement position of a chunk using the syntactic functions. Whether a hypothesis is satisfiable or not is determined by the concepts of the chunks. If they fit, the DRS can be computed: For a semantic head C_h, its complement C_k and a theta role R = thema (C_h, synfunc) that C_k can fill, we get the extension of the modified chunk $\tilde{C}_h$ as follows (h := head(C_h), k := head(C_k)):

$$\begin{aligned} \text{ext}(\tilde{C}_h) &= \\ \text{ext}(C_h) + \text{ext}(C_k) &+ \\ \left[\frac{\text{h k}}{\text{thema}(C_1, \texttt{synfunc})(d_1, d_2)}\right] & \end{aligned}$$

In our example, the *VP kommt* can be combined with the adverbial *PP im ZDF* since in the case frame of *kommen* there is a valency for an adverbial with the concept `location`. So we get

$$\left[\begin{array}{l} i\,l\,k \\ \hline \texttt{Run}(k)\ \texttt{TVStation1}(l) \\ \texttt{Name}(ZDF) \\ \texttt{im-SP}(i) \\ \texttt{has-location}(k,l) \\ \texttt{has-name}(l, ZDF) \\ \texttt{has-value}(i,l) \end{array}\right]$$

After λ-composition of the DRS above with the DRS for *Tatort* we get a full DRS for our example utterance.

5 Building a Case Frame Database

In order to encode the semantics of a natural language expression in our DL-domain, we always had to search in EWN for this expression, and if it was found, we had to manually follow up the taxonomic chain until we arrived at a superconcept that was already defined in our domain, and then begin from that point to encode the subtree we expanded in the last step. This task is time consuming and can be a source of errors, like encoding some concepts with their trees more than once, or forgetting subconcepts within a chain, not to mention the typing mistakes, missing parentheses, etc. which makes the domain model inconsistent and the processing difficult or rather impossible.

Moreover we use our approach to semantics construction in different applications. Consequently we gathered a huge amount of semantic definitions (i.e. taxonomic chains) and case frames (i.e. thematic roles) defined by these applications. Some of these data are specific to a given application, whereas others are used by several applications. This made the need for a tool that enables efficient storage and easy and fast access, as well as preparing the data required by the parser be of prime importance.

For this purpose, we have developed a lexicon tool that helps editing semantic data, checks their coherence according to the algorithm presented in sect. 4, and visualizes them as well (see fig.1).

The tool depends on the following resources as a basis for its data:

- EWN Ontology
- SUMO Ontology
- Semantic lexica

In this respect, it is worth highlighting the differences between our frame data base and FrameNet (see Baker et al. 1998). FrameNet is an online lexical resource[12] for English based on the principles of frame semantics and supported by corpus evidence. It can serve as a dictionary, for it includes definitions and grammatical functions of the entries. And hence entries are linked to the semantic frames in which they participate, FrameNet can serve as a thesaurus as well.

However, the information provided by FrameNet is not formal enough to be directly applicable to our system; in other words it is not possible to use FrameNet for parsing utterances directed to the system or constructing the semantic representation for them. So from the practical point of view, what we need is a formal specification for the information represented in FrameNet and which, on the one hand, can directly be encoded in DL notation, and on the other hand, can be used as an efficient inference mechanism. Another difference is that FrameNet is basically constructed for the English language and hence

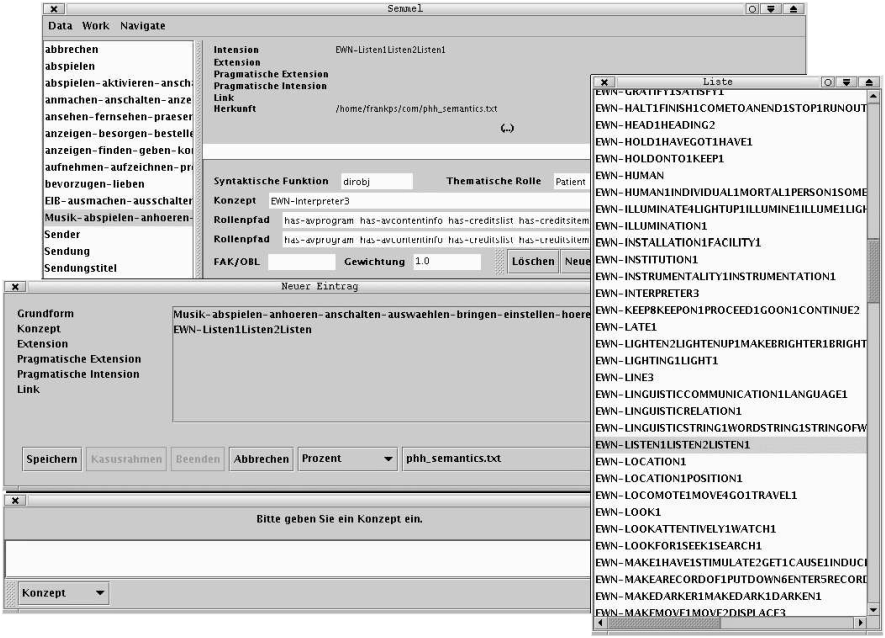

Figure 1: Lexicon-tool

can be used only in English based systems. Since our application is multilingual, our DL is based on the ILI-representation of EWN, which makes our tool language independent.

5.1 The Functionality of the Lexicon Tool

In our semantic lexica each entry has the following structure:

BASEFORM
LEXICAL INTENSION
LEXICAL EXTENSION
PRAGMATIC INTENSION
PRAGMATIC EXTENSION
CASE FRAMES
LINK
COMMENTS

The LEXICAL INTENSION refers to the lexical concept (as presented in EWN) that describes the lexical meaning of the entry, whereas the LEXICAL EXTENSION presents a DRS for the entry, and which has the following schema:

$$\left[\frac{x}{\textbf{lexical-intension}(x)}\right]$$

PRAGMATIC INTENSION and PRAGMATIC EXTENSION[13] provide optional information that can be used by the dialog system.

CASE FRAMES are the valencies assigned to the entry and which need to be filled by other instances that can satisfy the syntactic (e.g. subject), semantic (e.g. agent), and pragmatic ‚also called the thematic role' (e.g. user) constraints.

LINK refers to the name of the corresponding case frame.

The lexicon-tool can be considered as an interface between our application system and the semantic resources mentioned above, because, on the one hand, it stores the expressions used by the different applications and presents them as entries, to which the corresponding case frames are assigned and which are needed by the parser. On the other hand, it stores for each entry the underlying semantic concept as it is represented in EWN together with its taxonomic chain.

It can also be seen as a reproduction of the domain model due to three factors:

1. It maps the pragmatic intension of an entry onto the lexical one. This mapping is essential for determining the fillers (in the syntactic sense) or roles (in the pragmatic sense) specified by the case frame.
2. It maps the roles specified by the application domain onto the concepts obtained from the semantic lexicon. These roles must not violate the conceptual structure.
3. It verifies the consistency between the lexical and pragmatic intensions of the roles.

The interface provides an easy access to the stored information with the help of navigation tools like pop-up menus, text fields, lists, etc. It also enables the user to add new entries to the data base and define their word classes, syntactic functions, thematic roles, and semantic concepts (after obtaining it from EWN). While doing this the lexicon-tool offers lists with options that help the user determinig the most appropriate category by which the selected gap (text field) can be filled, and in the case of ill-formed or inappropriate input it returns detailed error messages with improvement suggestions.

One of the most valuable features of our lexicon-tool is the possibility of controlling and checking the coherence of entries both in terms of complete conceptual hierarchies with regard to our linguistic domain, and appropriate thematic roles with regard to the application domain. So if the user wants to check consistency or dependency relations between some concepts he can do that by typing the required sequence of concepts into the corresponding text field and getting the response after the check. Similarly, on adding new entries to the data base, if the given concept doesn't exist or collides with other concepts it won't be added, and subsequently the tool produces a corresponding error message and proposes possible solutions.

A further feature is the visualization of dependency relations in terms of links and cross references existing in the knowledge base. The possibility of checking whether a concept defined in a semantic lexicon specific to a given application also exists in the domain model or not remains to be done in future development.

6 The Influence of EWN on the Performance of EMBASSI

Our approach distinguishes between discourse and application domains, which in turn leads to a separation between linguistic and application specific knowledge. For this purpose, our knowledge base contains the complete SUMO ontology encoded in DL, the EWN upper ontology, and the concept definitions specific to Embassi applications. However, many SUMO and EWN concepts could be removed from the knowledge base as they were not used by the application specific part. So in an automatic precompilation step, we deleted 862 concepts, which were only defined but not a part of other definitions. This step improved the performance of Embassi a great deal. Nonetheless, the ratio of deleted concepts would be worse in more complex domains. As we didn't include the whole set of EWN concepts, in a more complex application the EWN portion even of the reduced knowledge base would be larger.

7 Experience with and Demands on EWN

In this section, some difficulties that we encountered while using EWN as the linguistic ontology in our knowledge base will be addressed. In the light of these difficulties, we will outline our strategy in dealing with them and consequently our demands on EWN. As EWN is the upper ontology in our system, most of the examples mentioned below are mainly taken from EWN, but the presented problems are valid for both EWN and GermaNet.

- **Missing parts of speech:** EWN is mainly limited to nouns, verbs and adjectives. However, meanings are not just expressed by these elements. Definitions for adverbs, temporal and spatial expressions, function words (e.g. auxiliary verbs, modal verbs, prepositions, etc.), not to mention multi-word elements (e.g. phrasal and prepositional verbs), idioms, collocations, and widely used abbreviations (e.g. ‚CO' for company) are generally not accounted for in EWN. Therefore we had to expand the linguistic domain model to include concepts for temporal and spatial expressions – to mention only the most prominent ones. It is evident that these elements are very important within the domain of Embassi in particular and similar systems in general, because, on the one hand, they function as fillers of roles in the application specific domain, which, in turn, helps determinig the sort of action to be triggerd off as a response to an utterance. On the other hand, in a language like German, prepositions, for instance, determine the case of the following noun. This fact can be used to enhance the mechanisms employed for disambiguation and sense differentiation. In conclusion, these elements are not optional but essential in any natural language system and can play a central role both on the semantic and application level, they shouldn't therefore be ignored in any tool for semantic representation.
- **Missing senses:** Another problem was the case in which the word being searched for already exists in EWN but not in all its senses. A definition of the word ‚part', for instance, in the sense of "member of a group" doesn't exist. Also, the word "subscribe" is only defined in the domain of financial transactions, so when we were searching for the same word in the sense of "being a member of or join" (a mailing list for instance) the corresponding definition couldn't be found. In such cases, we had to get the required sense by using synonymous words, despite the fact that the required word is already defined in EWN but not in all or at least not in the most dominant senses of it.
- **Conceptual gaps:** The definitions of some verbs (e.g. contain, glow, test, treat, sweat, apply, charge, ...) and most adjectives are so short, that they don't lead to the superset of all concepts that already exists in EWN. Concequently, gaps in the conceptual hierarchy may arise. In order to fill in the gaps in the hierarchy, we added general concepts like DO, CHANGE, CAUSE, STATE, QUALITY, MODAL-PROPERTY, MENTAL-PROPERTY and others to our knowledge base. On the one hand, these concepts function as subconcepts of already defined superconcepts in EWN, on the other hand, we can derive the required or rather the missing concepts from them.
- **Long taxonomic chains:** In contrast to verbs and adjectives, some nouns are assigned very long taxonomic chains (see, for example, the definitions of "mall", "tour", "cloth", "stuff"), which makes their encoding in DL and hence the consistency control rather difficult, not to mention the storage place and processing time they may take. We by-passed this problem by taking the definition of the underlying syno-

nym (marked by "=" in EWN), which usually has a shorter taxonomic chain. A side effect of this strategy is that some of the semantic properties of the word get lost, which leads to inacurracy in the semantic representation. Also the synonym definitions always imply a kind of generalization, which may be a source of ambiguity.

- **Antonymy:** Antonyms that can be regularly built by using some negation prefixes like (un-, in-, anti-, dis-,..), in general, are poorly represented in EWN. For example, the word "subscribe" exists but not "unsubscribe", the same holds for "scented" and other words. One would argue that antonyms shouldn't be accounted for in a lexicon, and their semantic representation can be obtained by negating the corresponding affirmative form. However, the negation of a form doesn't always reflect the meaning of the corresponding antonymous form (cp. unsubscribe vs. not subscribe). Apart from the processing perspective, within the foreign language teaching domain, a learner should be able to look up an antonymous form, or at least get information about how to build an antonymous form. In conclusion, it would be helpful, if EWN would pick up the most frequent antonyms either as separate entries or by assigning to every word the corresponding antonymous form or prefix.
- **Derivations:** Like antonyms, many standard[14] derivations are not existent in EWN. To illustrate this, take, for example, the word "moisturizer"; it is not defined, although the verb "moisturize" already exists. So the possibility to account for derivations either statically or dynamically in EWN is essential for building a uniform and balanced taxonomic hierarchy.
- **Insufficient syntactic coverage:** By "syntactic coverage" we mean syntactic features like valencies of a verb; gender, number of nouns, and so forth. Such features are not represented in EWN. In a system for natural language processing (e.g. machine translation system, parser, language generation system,..) these features are very essential not only on the syntactic but also on the semantic level.
- **Compounds:** Like derivations, there are only few entries for compound words in EWN, and there is no way to generate them dynamically. Examples: *Bruttopreis*, *Nettopreis*, *Schutzfolie*, a.o. In our application, we dealt with this problem either by combining the concepts of the individual constituents making up the compound expression, provided the constituents are already defined in EWN, or by searching for synonymous expressions, each consisting of a single word in order to take its definition as a substitute for the compound being actually searched for. The disadvantage of this method is that it makes the semantic construction more difficult and the semantic representation very complex and in some cases even inaccurate as well. This problem becomes more obvious in languages like English, where the constituents of a compound expression are separated by spaces. So it is sometimes difficult to recognize compounds as such. Therefore generating all possible conceptual combinations dynamically would be of a great advantage.
- **Orthographic variants:** As there are no uniform orthographic rules, it would be a big plus for EWN if it would account for possible orthographic variants of an expression like in (*email* / *e-mail*, *anti-perspirant* / *antiperspirant*, *Web-Seite* / *Webseite*), which will accelerate the search and retrieval.

8 Conclusion

The main goal of our research is the design and implementation of a generic dialogue system for spoken language that enables users to achieve specific tasks. This requires an efficient mecha-

nism for incremental semantic construction, in which lexical data can be reused within different DL domains and by several applications. In our system, we have been using EWN as a lexical resource for modelling linguistic data. Our experience with EWN within Embassi showed that the encoding of lexical data in DL and processing them in real-time was so far possible.

However, the practical experience always yields new demands on lexical resources (see sec. 7) and open questions for discussion and further improvement as well. For example:

- Considering FrameNet, which data can be extracted and practically applied to NLP-systems?
- How can they be encoded so that they can be generally used by different systems?
- How can the linguistic data be standardized so that they can be adaptable to several languages?
- Should frequent expressions (like greetings, polite expressions, etc.) be lexicalized to enhance the performance of the system?
- Which linguistic data must be accounted for in a lexicon?

etc.

Most of the above mentioned issues are not really new but as there are no general concepts for handling them they are still relevant both linguistically and practically.

Acknowledgements

The research presented in this paper has been carried out and tested in the framework of the EMBASSI project (Grant No.: 01IL9904F8) providing multi-modal assistance for controlling audio and video equipment.

Notes

[1] EuroWordNet Project (2001). Building a Multilingual Database with Wordnets for Several European Languages. University of Amsterdam, Department of Computational Linguistics, *http://www.illc.uva.nl/EuroWordNet/* [accessed April 2004].

[2] It aims to provide easy access for everybody to complex technical systems (A/V home theatre, car devices, and public terminals), encouraging multimodal as well as multilingual user input.

[3] A general characteristic of DL-Systems is that the knowledge base is made up of two components: the intensional one, called TBox, and the extensional one, called ABox. TBox is a general schema characterizing the classes of individuals to be represented, their general properties and mutual relationships, while ABox is a partial instantiation of this schema, containing assertions relating either individuals to classes, or individuals to each other. So given a concept language L, an ABox-statement in L has one of the forms (Donini et al. 1996): C(a) Concept Membership Assertion R(a, b) Role Membership Assertion where C is an L-concept, R is an L-role, and a, b are individuals.

[4] AvEvent refers to the concept for TV programs in the EMBASSI application.

[5] GermaNet Project (2003). GermaNet Homepage. University of Tübingen, Linguistics Department, Computational Linguistics Division, *http: //www.sfs.uni-tuebingen.de/lsd/* [accessed April 2004].

[6] An example would be a determiner phrase DP that is built from a NP which in turn is built from the lexical category N.

[7] Sometimes we needed a thematic role that was not existent in EWN, consequently we had to define some thematic roles that are required by the application domain in order to facilitate the semantic construction.

[8] Means: *Is Tatort coming on ZDF?* , where *Tatort* is the name of a TV program, and *ZDF* the name of a TV channel.

[9] The fact that this utterance is a Yes/No-question is irrelevant to phase 1, but word order information is stored and made available when the pragmatics of the utterance is computed.

[10] The term case frame here is used in the same way described by Fillmore 1969 and refers to thematic roles of an expression.

[11] The term valency here is used in a broad sense: it doesn't only imply the obligatory elements needed in order to make a phrase syntactically complete; but it also refers to the possible semantic and pragmatic modifications an element may take and their syntactic representations, e.g. attributes for nouns or adverbials for verbs.

[12] FrameNet Project (2004). "FrameNet II Homepage." International Computer Science Institute, Berkeley, CA, *http: //www.icsi.berkeley.edu/framenet/* [accessed April 2004].

[13] The pragmatic extension also presents a DRS for the corresponding pragmatic intension.

[14] "standard" here implies derivations that are built by using productive rules like (e.g. verb + -er →noun adjective + -ly →adverb. etc.).

References

Abney, S. (1991). "Parsing By Chunks." In: Berwick, R. et al. (eds.). Principle-based Parsing. Dordrecht: Kluwer Academic Publishers.

Baker, C.F. et al. (1998). "The Berkeley FrameNet Project." In: Proceedings of the COLING-ACL, Montreal.

Bücher, K. et al. (2002a). "Anything to Clarify? Report Your Parsing Ambiguities!" In: Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-2002), Lyon, July 2002, 465-469.

Bücher, K. et al. (2002b). "Corega Tabs: Incremental Semantic Composition." In: Proceedings of the Workshop on Applications of Description Logics (ADL 2002), Aachen, Germany, September 2002 [= CEUR Workshop Proceedings Vol. 63], *http://CEUR-WS.org/Vol-63/.*

Donini, F. M. et al. (1996). "Reasoning in Description Logics." In: Brewka, G. (ed.) (1996). Principles of Knowledge Representation. Stanford / CA: Center for the Study of Language and Information [= CSLI Publications - Studies in Logic Language and Information], 193-238.

Fillmore, Ch. (1969). Universals in Linguistic Theory. New York: Holt, Rinehart, & Winston.

Fischer, I. et al. (1996). "Incremental Semantics Construction and Anaphora Resolution Using Lambda-DRT." In: Botley, S.; Glass, J. (eds.) (1996). Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-96), Lancaster, July 1996, 235-244.

Kamp, H.; Reyle, U. (1993). From Discourse to Logic. Dordrecht: Kluwer Academic Publishers.

Kuschert, S. (1996). Higher Order Dynamics: Relating Operational and Denotational Semantics for λ-DRT. University of Saarbrücken, CLAUS-Report 84.

Ludwig, B. (2002). "Corega Tabs: Mapping Semantics onto Pragmatics." In: Proceedings of the Workshop on Applications of Description Logics (ADL 2002), Aachen, Germany, September 2002 [= CEUR Workshop Proceedings Vol. 63], *http://CEUR-WS.org/Vol-63/* [accessed April 2004].

Niles, I.; Pease, A. (2001). "Toward a Standard Upper Ontology." In: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), 2001.

Vossen, P. (ed.) (1999). EuroWordNet General Document, Version 3, 1999. (EuroWordNet LE2-4003, LE4-8328, Part A, Final Document). *http://www.illc.uva.nl/EuroWordNet/docs/GeneralDocPS.zip, http://www.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip* [accessed April 2004].

Manuela Kunze, Dietmar Rösner

Issues in Exploiting GermaNet as a Resource in Real Applications

Abstract

This paper reports about experiments with GermaNet as a resource within domain specific document analysis. The main question to be answered is: How is the coverage of GermaNet in a specific domain? We report about results of a field test of GermaNet for analyses of autopsy protocols and present a sketch about the integration of GermaNet inside XDOC.[1] Our remarks will contribute to a GermaNet user's wish list.

1 Introduction

GermaNet – a lexical-semantic net – was developed in the context of the LSD-project: „Ressourcen und Methoden zur semantisch-lexikalischen Disambiguierung" (Hinrichs et al. 1998). This paper describes an experiment about the integration of GermaNet into the Document Suite XDOC. The Document Suite XDOC was designed and implemented as a workbench for flexible processing of electronically available documents in German (Rösner & Kunze 2002).

We currently are experimenting with XDOC in a number of application scenarios. These include:

- Knowledge acquisition from technical documentation about casting technology as support for domain experts for the creation of a domain specific knowledge base.
- Extraction of company profiles from WWW pages for an effective search for products and possible suppliers.
- Information extraction from English Medline abstracts.
- Analysis of autopsy protocols for e.g. statistical investigation of typical injuries in traffic accidents.

The end users of our applications are domain experts (e.g. medical doctors, engineers, ...). They are interested in getting their problems solved but they are typically neither interested nor trained in computational linguistics. Therefore the barrier to overcome before they can use a computational linguistics or text technology system should be as low as possible.

Many of our tools have an extensive need for linguistic resources. Therefore we are interested in ways to exploit existing resources with a minimum of extra work. The resources of GermaNet promise to be helpful for different tasks in our workbench. GermaNet – a German version of the Princeton WordNet (Miller 1990; Fellbaum 1998) – is based on the same design principles, i.e. database structures like WordNet. The intention of GermaNet is defined as the coverage of basic vocabulary of the German language – based on lemmatized frequency lists from text corpora (see Hamp & Feldweg 1997; or Kunze 2001).

The following scenarios for the integration of the GermaNet resource in our work are possible (see also section 5):

- GermaNet as resource for semantic analyses,
- GermaNet for a shallow recognition of implicit document structures,
- GermaNet for compound analysis.

This paper is organized as follows: first we give a short description of the document class 'auto-

psy protocols', because the examples of this paper are based on this corpus. After this we describe our results related to the coverage of GermaNet for our corpus and how the ambiguities inside the results can be resolved. Then we shortly sketch the semantic module of XDOC into which the resources of GermaNet should be integrated. This is followed by the presentation and discussion of results from our experiments. Our remarks will finally contribute to a GermaNet user's wish list.

Characteristics of the Document Class: Autopsy Protocols

Autopsy protocols are especially amenable to processing with techniques from computational linguistics and knowledge representation:

- Forensic autopsy protocols are in most cases written with the clear constraint that they will be used for legal purposes and will have to be interpretable by lawyers and other non-medical experts.
- Autopsy protocols are highly structured and follow a strict ordering.
- The sub-language of the *Findings* section is of a telegramatic style with a preference for 'verbless' structures. The sub-language of other subdocuments is slightly more complex, but still limited due to the communicative requirements (e.g. precision, uniqueness of expression, understandability for non-experts).

2 GermaNet and Autopsy Protocols

In the following we do report about ongoing experiments with a corpus of currently approx. 600 autopsy protocols from Magdeburg. The corpus will soon be extended with protocols from other institutes for forensic medicine from all parts of Germany and shall in the long run be representative for autopsy protocols from all German speaking countries.

The central question for our experiments: Given a corpus with texts from a uniform domain how is GermaNet's coverage as such, i.e. without any investment in extending the available GermaNet resources? We did not attempt lexical analysis of the tokens derived from our test corpus, except comparison with exhaustive lists of tokens from closed word classes of German function words, connectors, prepositions, etc. This is reflecting the situation when a corpus from a new domain is processed for the first time and many domain terms are new and not covered in lexical resources.

2.1 Coverage of GermaNet

First experiments with GermaNet demonstrate the coverage of GermaNet for autopsy protocols.

doc. type	word types	match	percentage coverage
Findings	17520	2591	14,78
Background	8124	2274	27,99
Discussion	8562	1862	21,74

Table 1: Coverage for Different Document Parts.

Table shows the coverage rates for the central document parts of an autopsy protocol. For this evaluation we use tokens, restricted by following parameters:

- The candidates are not function words, like conjunctions, prepositions, etc. Only words that are potential candidates for nouns, adjectives and verbs are tested.
- In autopsy protocols some tokens are 'implicit markup', e.g. enumerations of titles or paragraphs like 'II.' in 'II. Innere Besichtigung'. These tokens were excluded from the test.
- The length of a potential candidate was restricted to greater than three characters.

With these restrictions we reduce the number of different tokens to be evaluated in section *Findings* from 18492 to 17520, in section *Background*

from 8901 to 8124 and in section *Discussion* from 9198 to 8562 tokens.

The least coverage of GermaNet exists in the section *Findings*. This is not astonishing, because there we have many domain specific terms (e.g. 'Thalamus', 'submandibularis', 'Hirnkontusionen' or 'Injektion'). In addition, the medical doctors use their own (subjective) vocabulary for the description of injuries or other findings, like 'weichkäseartig', 'metallstecknadelkopfgroße' or 'teerstuhlartiger'. The best coverage could be achieved in the section *Background*. Here we have many words from common language. This document part describes the case history (e.g. details of a traffic accident). We rarely find domain specific terms in this section. The section *Discussion*, which combines the results of the *Findings* section and the facts from the *Background* section, ranks in the middle with a 21,74 percentage.

A segmentation of the coverage rates into different word classes is shown in table. In these data all hits are counted, without distinction whether a GermaNet entry exists for one or more word classes, therefore the sum of a row is greater than the number of matches in table.

In the coverage summary the word class adverb is ignored, because at the time of writing there are only two synsets for adverbs available in the version of GermaNet and we got zero matches in our corpus for these adverbs.

document type	nouns	verbs	adjectives
Findings	1573	351	806
Background	1622	328	465
Discussion	1162	322	483

Table 2: Coverage for Different Word Classes.

Related to the word class we have uniform results across subdocuments, the largest coverage figure is for nouns, followed by adjectives and verbs. The high ratio of adjectives in the section *Findings* is due to the high frequent usage of adjectives in this section.

2.2 Characteristic of Uncovered Terms

The tokens that had no entry in GermaNet can be divided into two classes. Beside the uncovered lexical terms (like 'Rotor' or 'Klinge') we have a lot of specific terms, which could not be covered by GermaNet. The analysis of these uncovered specific terms, which negatively affect the results above, gives the following classification.

measured values and ranges:
'2cm', '4-9', '120ml',
named entities:
'Beck', 'Otto-von-Guericke-Universität', 'Opel', 'Salvator-Krankenhaus', 'B269', 'Zehringen-Sibbendorf',
truncations:
'-aussenseite', '-wischspuren',
compounds:
'Plastikdreipunktsicherheitsschluessel', 'Oberschenkelspiralmehrfragmentfraktur', 'weisslich-gelblich-roetlich-fleckige',
inflected words:
'Armes', 'besitzt', 'entnommen',
misspellings:
'Herzmnuskulatur', 'Herrrren-T-Shirt', 'Todeseinritt'.

The first category are non-lexical tokens. Depending on the domain and text type their form and frequency is varying. They cannot be expected to be covered by GermaNet and are best treated with special recognizers (e.g. regular expressions).

All items of the first three categories can be preselected by different preprocessing steps, like regular expressions or methods for named entity recognition. The categories *misspellings* and *inflected words* can only successful (in terms of GermaNet) be preprocessed by a complex morphological component, including recognition of inflected words and orthographic similar words. For the processing of compounds in GermaNet it is possible to use the resources of GermaNet itself (see section 5).

3 Resolving Ambiguities

In this section we discuss approaches for resolving ambiguities. The discussion is related to the kind of ambiguity. In our use of GermaNet we found three types of ambiguities. Type one is an ambiguity on the POS level – whether the token to be analysed is for example a noun or a verb. The second type occurs when more than one sense exists for a word class. The last type is a combination of the first two types.

3.1 Part-of-Speech Ambiguity

Table 3 shows the ratios of entries with Part-of-Speech ambiguity.[2]

The first row are results of counting all matches with more than one word class per literal, the percentage rate related to all matches is given in parentheses.

The rows 2 to 5 present the number of matches in which a specific combination of word classes, e.g. noun and verb, occurs. The first value in parentheses displays the percentage rate related to all matches and the second value is the percentage rate related to all matches with POS ambiguity.

In all three document parts the highest case of POS ambiguity occurs between nouns and verbs. For example, the token 'Herzens' in the phrase 'Gewicht des Herzens ...' will be interpreted in GermaNet both as noun and as verb.

Due to the verbless style for this section it is not astonishing that only in the section *Findings* a similar high ratio is given for the case 'nouns and adjectives'.

It can be assumed, that a simple check of capitalisation of a token can probably decrease the rates of POS ambiguity. Taking sentence initial positions into account simple upper-/lowercase distinction could decrease the rate of 'noun-verb' or 'noun-adjective' matches.

Another approach is based on POS information about the tokens to be analysed (using e.g. MORPHIX Finkler & Neumann 1988). With this additional POS information we can directly decide which information we want to retrieve in GermaNet. In addition, we can also use a simple heuristic approach based on the information about the document section. In the section *Findings* readings of adjectives can be preferred over readings as verbs.

	Findings	**Background**	**Discussion**
different word-classes	139 (5,36)	135 (5,93)	104 (5,58)
N and V	72 (2,77; 51,79)	89 (3,9; 65,9)	71 (3,8; 68,2)
N and ADJ	64 (2,47; 46,04)	35 (1,15; 25,92)	31 (1,66; 29,8)
V and ADJ	3 (0,11; 2,15)	5 (0,21; 3,7)	1 (0,05; 0,96)
N, V and ADJ	0 (0; 0)	6 (0,26; 4,44)	1 (0,05; 0,96)

Table 3: POS Ambiguity.

3.2 Sense Ambiguity

The average number of senses for a token of our corpus covered by GermaNet is approx. 1,76. The highest number we get is for verbs with ca 3 senses (average numbers of senses for verbs: 3,18; nouns: 1,49 and adjectives: 1,62). It is apparent that in many cases GermaNet returns more than one sense for an entry. Table shows the number of tokens with more than one sense related to the different document parts.

	ratio	**percentage**
Findings	1034	39,95
Background	914	40,26
Discussion	823	44,27

Table 4: Sense Ambiguities.

A method for resolving the senses is the use of contextual information. The specific structure of our documents (division in three main parts) and content related separation into these parts allowed to exploit this information for the determination of the most likely sense. As a start we

use here the information of the semantic fields of GermaNet. Experiments show (by majorities) clear differences between the parts (see table).

Although the subdocuments may differ slightly in this respect there is a strong preference for medical readings (senses) for potentially ambiguous words in the corpus of autopsy protocols. This is especially true for the subdocument with information about the examination findings. The subdocument with the background is the place where the expectation for medical senses seems to be weakest.

Please note that words may have even conflicting 'medical readings'. 'Blase' may be an organ (bladder) or an injury (e.g., caused by fire).

In the *Findings* section the most frequent GermaNet categories are 'nomen.Körper', 'verb.Veränderung', 'verb.Lokation'. For resolving ambiguities we use this information (majorities) for preselecting senses depending on the current document section. For example, the noun 'Becken' will be classified by GermaNet in the semantic fields 'nomen.Artefakt' (in a sense of 'music instrument') and 'nomen.Körper' (in a sense of 'bone'). In the analysis of the *Findings* section we prefer the sense of 'nomen.Körper'. In the section *Background* the sense of 'nomen.Artefakt' has a higher likelihood than the sense 'nomen.Körper'.

Section	most frequent semantic fields
Findings	nomen.Körper, verb.Lokation, verb.Veränderung, adj.Körper, adj.Perzeption,
Background	nomen.Geschehen, adj.Zeit, adj. Lokation
Discussion	nomen.Geschehen, nomen.Körper, verb.Lokation, verb.Veränderung, adj.Relation

Table 5: Typical Semantic Fields of the Document Parts.

3.3 Combined Ambiguity

These cases are very rare in the corpus: *Findings*: 11 (0,42 %), *Background*: 19 (0,83 %) and *Discussion*: 15 (0,8 %). They could probably be resolved through the approaches that are outlined in section and section .

4 GermaNet inside the Semantic Module of XDOC

The integration of the GermaNet resources takes place for the purposes of semantic analysis. In this section we outline the strategies for semantic analysis within XDOC. The *Semantic Module* in XDOC exploits three analysis techniques for the annotation of documents with semantic information. The results of the analysis are recorded in separate Topic Maps or annotated within documents with a specific XML format. At first we give a short description of the semantic analyses inside XDOC.

Semantic Tagger. The *Semantic Tagger* classifies content words into their semantic categories (different applications may have different organizations of those categories in the form of taxonomies or ontologies). For this function we expect as input data a text tagged with POS tags and we then apply a semantic lexicon. This lexicon contains the semantic interpretation of a token and a case frame combined with the syntactic valence requirements. Similar to POS tagging, the tokens in the input are annotated with their meanings and with a classification into semantic categories (i.e. specific concepts or relations). It is possible that the classification of a token in isolation is not unique. In analogy to the POS tagger, a semantic tagger that processes isolated tokens is not able to disambiguate between multiple semantic categorisations. This task is postponed for contextual processing within case frame analysis (*Semantic Parser*).

Semantic Parser. The *Semantic Parser* is one method in XDOC for the assignment of semantic relations between isolated (but related) tokens. By case frame analysis of a token we obtain details about the type of recognized concepts (resolving multiple interpretations) and possible

relations to other concepts. Fig. 1 contains the results of the analysis of the noun phrase *'Unfallablauf mit Herausschleudern der Koerper aus dem PKW'*. We get here the assignments of the relation *part* between *'Unfallablauf'* and *'Herausschleudern der Koerper aus dem PKW'* and the relations *location* (between *'Herausschleudern'* and *'PKW'*) and *patient* (between *'Herausschleudern'* and *'Koerper'*).

Semantic Interpretation of Syntactic Structure (SIsS). An other step for the recognition of relations between tokens is the *Semantic Interpretation of syntactic Structure* of a phrase or sentence respectively. We exploit the syntactic structure of the language (e.g., structures of noun phrases) and the semantic interpretation of tokens inside the structure to extract relations between several tokens. Fig. 2 is a visualization of the results of the analysis of the noun phrase *'dunkelrote Unterblutung der Schleimhaut der Niere'*. The analysis of this complex noun phrase results in three relations between the separated nouns. The relation *prop* is used to label properties of a concept. Our future work here: The generic relation *gen-attribute* (short for attribute based on a genitive surface case) has to be resolved into the appropriate more specific relations, like *part-of* or *patient*.

The core of all these semantic analyses techniques is a semantic lexicon. This lexicon records the meanings and case frames (only for nouns and verbs) of a word.

Up to now the entries of this lexicon have been manually built up and are partially domain dependent. Now we want to integrate the GermaNet resources into our framework.

4.1 Integration of GermaNet

Currently the integration of GermaNet is realised in the semantic tagger. For the semantic lexicon we use the conceptual relation *hypernym* of GermaNet. The tagger uses the first level of the *hypernym* relation for the annotation of tokens with information about the GermaNet senses:

```
(tag-semantic-xml "<N>Leber</N>
<S-KONJ>und</S-KONJ><N>Niere</N>")

"<CONCEPT TYPE="Innerei;
Verdauungsorgan">Leber
</CONCEPT>
<XXX><S-KONJ>und</S-KONJ></XXX>
<CONCEPT TYPE="Innerei; Harnorgan">Niere
</CONCEPT>"
```

The XML-attribute *TYPE* contains the *hypernym* information from GermaNet. The different senses are separated by a semicolon.

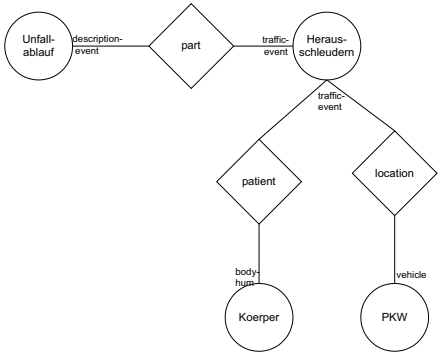

Figure 1: Results of Case Frame Analysis.

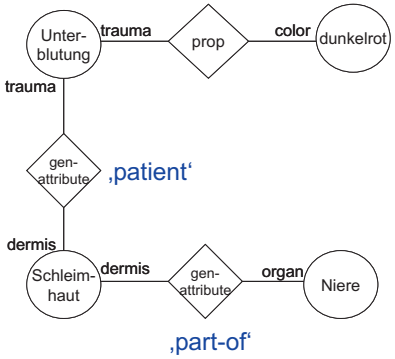

Figure 2: Results of SIsS.

For better results we reconfigured our semantic tagger. In contrast to the early version the semantic tagger now also expects tokens with POS information (word classes), but enriched with additional information about the stem of the tokens. In this way we ask for senses related to a word class with the facility to use a non-inflected word form for the request.

```
(tag-semantic-xml "<N STEM="Gewicht">
Gewicht</N><DETD>des</DETD>
<N STEM="Herz">Herzens</N>")

"<CONCEPT TYPE="?physikalisches
Attribut; Wichtigkeit, Messgeraet,
Messgeraet*o, Messinstrument*o,
Messinstrument; Artefakt, Werk">
Gewicht</CONCEPT>

<XXX><DETD>des</DETD></XXX>
<CONCEPT TYPE="Innerei; Organ; Farbe,
Spielfarbe; Flaeche, Ebene">Herzens
</CONCEPT>"
```

Another integration of the GermaNet resources is possible for the *Semantic Parser*. Here we could use the information of verb frames. Up to now the mapping of GermaNet verb frames to the XDOC case frames could be problematic. For case frames we use in addition to syntactical valency (e.g., noun phrase in accusative) also the description of potential semantic roles for the filler of the frame. This information is not available from GermaNet's verb frames. For this integration of GermaNet it is necessary to complete the additional semantic information manually or by a corpus based approach (learning from corpora). For instance, for the analysis of the sentence 'Sie wurde am Kopf operiert.' we get for the verb 'operieren' the GermaNet sense:

```
Sense 1 operieren
  => medizinisch behandeln
   => wandeln,ändern,mutieren,verändern
```

GermaNet contains for this sense following verb frames:

```
Sense 1
operieren
    *> NN.AN
    *> NN.AN.BL
```

The second verb frame matches our example sentence.

But the usage of these GermaNet's verb frames in the analysis of the sentence 'Sie wurde im KKH xxx am Arm operiert.' is problematic because the *BL* complement could be assigned to the locative preposition phrase 'im KKH xxx' or to the locative preposition phrase 'am Arm'. One of the two prepositional complements gets no direct assignment to a complement defined by GermaNet's verb frames. Other similar problematic examples from our corpus are:

- *Nach polizeilichen Angaben aus der Akte und den klinischen Unterlagen wurde G xxx/xx am Morgen des dd.mm.jj im Krankenhaus X wegen einer knotigen Kropfbildung operiert (Strumaresektion).*
- *Am dd.mm.jj wurde G xxx/xx im KKH xxx am Herzen operiert.*

A detailed description, e.g. additional information about the semantic role of the complement's content, could be helpful for the analysis. Our *Semantic Parser* works with such information. For the usage of the verb frames for the analysis with our *Semantic Parser* we need additional features for the *Adverbial Complement (BL)* of the verb frame: [3]

- semantic role of the filler: body part, for example, organs or extremities,
- possible preposition: am,
- case of PP: dative or not specified.[4]

Other features to be considered in using verb frames are:

- the different complement forms for active or passive usage of a verb and
- the number of a noun phrase: For example, for the verb 'kollidieren' is the possible verb frame 'NN.Pp'. The preposition phrase is defined as an optional complement. A necessary additional feature for the noun phrase is the information about its number (singular or plural). For example, the subject noun phrase in sentences like 'Die Fahrzeuge kollidierten.' must name more than one participant of the accident.

To complete GermaNet's verb frames it is possible on the one hand to add this additional information manually or on the other hand by the analysis of occurrences of similar phrases in the corpus. By the corpus based approach the user gets a list of possible complements for a verb, so that the verb frame of GermaNet can be enriched with the corpus/context related features. GermaNet's verb frames are used as pattern for the search inside the corpus. The basis for this approach is a corpus with syntactic structures annotated by the *Syntactic Parser* of XDOC (Rösner 2000).

One problem inside the SIsS analysis is the correct interpretation of the genitive-relation. One solution is the usage of the conceptual relations *meronym* and *holonym* of GermaNet. For example, the results of the SIsS analysis of the phrase 'unauffaelliger Vorhof des Herzens' is shown in Fig. 3.

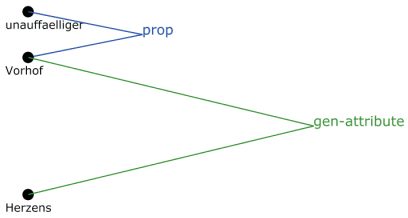

Figure 3: Result for the Phrase 'unauffaelliger Vorhof des Herzens'.

By the SIsS analysis two relations were recognised: first the *prop* relation between the tokens 'unauffaellig' and 'Vorhof', the second recognised relation is the *gen-attribute* relation. This relation can in general be interpreted in several ways (e.g. as *part-of* or *patient-of*). GermaNet results for the token 'Herz' (in the sense of *organ*) give the meronyms: *Vorhof, Herzklappe, Herzkammer, Herzrohr, linke Herzhälfte, rechte Herzhälfte, Herzmuskel, Herzkranzgefäß* and for the token 'Vorhof' the holonym 'Herz'. With *meron* and *holon* information from GermaNet we can decide that the generic relation ('gen-attribute') between the tokens 'Herz' and 'Vorhof' is a *part-of* relation.

4.2 Practical Aspects of the Integration of GermaNet

The technical access to GermaNet was realised in different ways: offline and online usage of GermaNet. In offline usage GermaNet is transformed into an application specific resource. This transformation may be carried out as a compilation step beforehand. Online usage employs GermaNet resources via their API.

For the offline usage of GermaNet we only transform necessary information into the application specific resources. Depending on the task to be performed we do need different information from GermaNet. In one case we need the synsets and hypernyms (*Semantic Tagger)*, in other cases we only work with information about the semantic fields of a token (for example, *scenario: shallow recognition of document sections*). The relations inside GermaNet can also be described as path direction – up, down, horizontal (see Hirst & St-Onge 1998). Within the semantic module of XDOC the following 'path directions' could be useful:

Semantic Tagging: search for a context-based allowed sense (hypernyms, synsets),

Semantic Parser: assignment to semantic roles, search for a filler of a semantic role (hypernyms), syntactic information via verb frames,

SIsS: resolving of semantic interpretation of genitive-structures (e.g. Schleimhaut des Magens) by 'meron' or 'holon' information.

Concerning the coverage of GermaNet we obtained the following results:

- Some tokens of our corpus are not covered by GermaNet, especially in the range of open word classes, like adjectives (e.g. 'quer') or in the range of domain specific words (e.g. 'Fraktur').
- Uncomplete senses of an entry. For instance, for the word 'Abfall' there exists only one sense related to the semantic field 'noun.substance', but in our corpus we often find the word 'Abfall' in phrases like 'Abfall des Blutdruckes'. Please note: For the verb 'abfallen', from which the noun 'Abfall' is derivated by a verb-noun-conversion, we also do not get the right sense related to our domain:

```
1 sense of abfallen
Sense 1 abfallen
=> lösen
 => ?Dauerkontakt
```

From a more technical perspective the following points are relevant:

- A very simple or no morphological component in GermaNet (WordNet is better), e.g. 'Autos' will be found, but 'Organe' will not be found in GermaNet. This is explainable only through the use of an English morphology component (from WordNet). GermaNet uses English flection criteria for the analysis of the input data. By reconfiguration of our *Semantic Tagger* we can avoid this effect (see section).
- Use of umlauts in GermaNet: The documents in our corpus are without umlauts, but GermaNet supports only access via writings with umlauts. Matching of candidates without umlauts to possible candidates in GermaNet with umlauts could be helpful and would lead to a better coverage.

In consideration of the last two points we worked with two additional intermediate steps in our experiment environment. At first we integrated the morphological component MORPHIX (revised results for the different sections are *Background:* 41,39 %, *Findings:* 29,64 %, *Discussion:* 40,57 %) and the second step was the treatment of umlauts which again improved our GermaNet coverage results (*Background:* 43,38 %, *Findings:* 31,02 %, *Discussion:* 42,38 %).

4.3 User's Wish List

Some items for the GermaNet user's wish list:

- It seems that in the case of orthographic variants GermaNet 'knows' sometimes more than it makes available. An example: GermaNet has the information that '4-eckig' is an orthographic variant of 'viereckig', but does only return information when the user (or application program) asks with the (canonic) writing 'viereckig'.
- Flexible match of umlauts and extended writings: Given the fact that in computer written text umlauts are still often represented in the expanded form of 'ae', 'oe', …it would be helpful to increase the flexibility of GermaNet's lexicon access and provide means that search terms in the expanded writing will match existing entries with umlauts (i.e. 'Gebaeude' should match 'Gebäude').
- Avoid artefacts due to English spelling rules from WordNet: WordNet and GermaNet

offer convenience functions to the user for search in the resources in the sense that some but not all inflections, derivations, and alternative spellings can be handled. For example: 'Herzens' matches the verb 'herzen'(!) but not the noun 'Herz'.[5]

- Finally: GermaNet is not error free. In our work we occasionally get messages like 'Error Cycle detected' or 'Synset xxxx not found', which make the user insecure about the results returned by GermaNet.

5 Discussion: Back to the Scenarios

In previous sections we have described some integration aspects of GermaNet for different scenarios. Now we give a concrete outline of the scenarios.

GermaNet as Resource for Semantic Analyses. In section 4.1 we described the integration of GermaNet as resource in the *Semantic Module* of XDOC. There we use the lexical-semantic net for the annotation of tokens with their semantic roles (*Semantic Tagger*). For this task we exploit the different defined relations inside GermaNet (e.g. hypernym or synonym). For the tasks of the *Semantic Parser* and the *SIsS analysis* we additionally use information of verb frames and other conceptual relations, like the 'meron' and the 'holon' relation. The *Semantic Parser* directly uses this information for the analysis, while the *SIsS analysis* uses GermaNet's information in a postprocessing step for the selection of one (possible) interpretation of the different readings resulting from the *SIsS analysis* (e.g. the relation *gen-attribute*).

GermaNet for a Shallow Recognition of Implicit Document Structures. In section 1 we have given a short specification of autopsy protocols. The characteristics of the different document parts can be used for a recognition of these parts. The following parameters describe the different document parts (also related to the available information by GermaNet):

Findings:
high ratio of nouns and adjectives; short specific syntactic (sentence) structures; semantic fields like 'nomen.Körper', 'adj.Körper', 'verb.Veränderung',

Background:
standard distribution of all word classes; regular syntactic structures; semantic fields like 'nomen.Geschehen', 'adj.Zeit', 'adj.Lokation',

Discussion:
standard distribution of all word classes; regular syntactic structures; semantic fields like 'nomen.Geschehen', 'nomen.Körper', 'verb.Lokation', 'verb.Veränderung'.

The distribution of the semantic fields over different document parts can be used for the recognition of these document parts. For example a document part with a high frequent occurrence of tokens, which can be assigned to the semantic fields like 'nomen.Geschehen', 'adj.Zeit', 'adj.Lokation', and no occurrences of tokens with assignments of 'nomen.Körper' etc. can be identified as the *Background* section of an autopsy protocol. For a unique identification we also use information about the word classes by the *POS Tagger* and the information about the kind of syntactic structures by the *Syntactic Parser* to confirm the other characteristic criteria of a document part.

GermaNet for Compound Analysis. In the autopsy protocol corpus – as well as in other medical or technical texts – noun compounds are quite frequent. The question here is: Is it possible to

- safely determine segmentations of noun compounds and to
- construct meaning hypotheses for noun compounds by combining the meaning of the compound's parts if they are covered by GermaNet?

Please note: Segmentation of German noun compounds (i.e. determination of boundaries between parts of a noun or noun compound) may produce artefacts even when the hypothesized compound segments are lexical entries in their own right.

Examples (suggested segmentations indicated with [...]):

```
Transport ... * [Tran][sport]
Lebertransport ... * [Lebertran][sport]
     [Leber][transport]
```

We therefore favour an approach to compound segmentation that additionally takes the corpus and the occurrence frequencies of complex words with common pre- and suffixes into account and thus reduces the dependence on the lexicon and its coverage.

The corpus-based analysis of compounds with GermaNet can be described as follows: The first step is to find all compounds with similar suffixes inside the corpus, like 'Nierentransplantation', 'Lebertransplantation' etc. Then define 'Top Level' relations between possible candidates for compounds, for our example: <organs><medical-operation>, to avoid a wrong interpretation of compounds. Here we can use the semantic field information of GermaNet for the description of relations between possible candidates.

6 Conclusion

We have reported about first experiments in integrating GermaNet resources into XDOC for the processing of autopsy protocols.

Although our results related to the coverage of GermaNet were not as high as in Saito's experiments (Saito et al. 2002), the results for a corpus of autopsy protocols are encouraging. (A parallel experiment with the EUROPARL corpus – available at *http: //www.isi.edu/~koehn* – resulted in a lower coverage. Of 198546 tested tokens only 30344 tokens are covered by GermaNet; this probably is in part due to the high ratio of named entities in the EUROPARL corpus.) The results could be further improved by XDOCs preprocessing steps, like named entity recognition, POS tagger etc., so that an adoption of GermaNet resources into the semantic analyses of XDOC is conceivable.

We use GermaNet's lexical-semantic net for semantic enrichment of documents. GermaNet's resources were primarily integrated into the *Semantic Tagger* of XDOC. In future work we will further extend the integration of GermaNet for the *SIsS* analysis and the *Semantic Parser*.

Notes

1. XDOC stands for XML based DOCument processing.
2. In short: POS ambiguity.
3. When we assumed that BL is a preposition phrase.
4. When no unique assignment to one case is possible.
5. Please note: 'Herzens' can be erroneously derived from the verb 'herzen' under the assumption of an English inflection: 'English' morphological attributes of 'Herzens' are then third person singular.

References

Hinrichs, E. et al. (1998). LSD-Projekt im Forschungsschwerpunkt: Methoden und Ressourcen der lexikalisch-semantischen Disambiguierung. Abschlußbericht, Universität Tübingen, Seminar für Sprachwissenschaft.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Finkler, W.; Neumann, G. (1988). "MORPHIX: A Fast Realization of a Classification-based Approach to Morphology." In: Trost, H. (ed.) (1988). Proceedings der 4. Österreichischen Artificial-Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung. Berlin et al.: Springer [= Informatik-Fachberichte Bd. 176], 11-19.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (eds.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Hirst, G.; St-Onge, D. (1998). "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms." In: Fellbaum, Ch. (ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA / London: MIT Press, 305-333.

Kunze, C. (2001). „Lexikalisch-semantische Wortnetze." In: Carstensen, K.-U. et al. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: Eine Einführung. Heidelberg: Spektrum Akademischer Verlag, 386-393.

Miller, G. A. et al. (1990). Five Papers on WordNet. Technical Report 43, Princeton University, Cognitive Science Laboratory [first published in: Journal of Lexicography 3(4) (1990), 235-312, *ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf,* accessed April 2004].

Rösner, D.; Kunze, M. (2002). "An XML based Document Suite." In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipeh, August / September 2002, 1278-1282.

Rösner, D. (2000). "Combining Robust Parsing and Lexical Acquisition in the XDOC System." In: Proceedings KONVENS 2000 Sprachkommunikation, Berlin / Offenbach:VDE Verlag [= ITG-Fachbericht 161], 75-80.

Saito, J.-T. et al. (2002). "Evaluation of GermaNet: Problem Using GermaNet for Automatic Word Sense Disambiguation." In: Proceedings of the Workshop on Wordnet Structures and Standardizations, and how these Affect Wordnet Applications and Evaluation. 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 28th May 2002, 14-29.

Matthias Jörg

Die semantische Auswertung von Produktanforderungen mit Hilfe von GermaNet

Abstract

Die Entwicklung eines technischen Produktes wird zunehmend aufwändiger und anspruchsvoller. Dies hat vorwiegend zwei Gründe. Die Kunden verlangen nach innovativen und hochwertigen Produkten und können durch den globalen Wettbewerb aus einer Vielzahl von konkurrierenden Produkten auswählen. Der zweite Grund ist die gestiegene Komplexität der Produkte selbst, insbesondere im Automobilbereich. Ohne den Einsatz moderner und effektiver Entwicklungsmethoden lässt sich die Entwicklung eines neuen Fahrzeugs kaum mehr durchführen.

In diesem Artikel wird ein Konzept vorgestellt, mit dem Produktanforderungen systematisch erfasst, verwaltet und geprüft werden können. Der Schwerpunkt liegt auf der Erfassung von qualitativen Anforderungen, die vom Benutzer in textlicher Form beschrieben werden. Betrachtet man die verschiedenen Anforderungen an ein Produkt, so können Abhängigkeiten zwischen diesen erkannt werden. Ziel einer stringenten Anforderungsverarbeitung ist die Darstellung aller Abhängigkeiten in einem sog. *semantischen Anforderungsnetz*. Um ein solch komplexes Netz aufzubauen ist es notwendig, alle Produktanforderungen vereinzelt und in eindeutiger Form vorliegen zu haben. Mit dem beschriebenen Ansatz sollen Anforderungen hinsichtlich ihrer Semantik analysiert und bewertet werden, um sie nachfolgend in das Anforderungsmodell integrieren zu können. Die semantische Auswertung soll mit Hilfe des *GermaNet http://www.sfs.uni-tuebingen.de/lsd/* (2003) erfolgen.

1 Einleitung

Eine der größten Herausforderungen der Produktentwicklung besteht darin, die Wünsche des Kunden zu treffen und somit das Produkt erfolgreich auf dem Markt zu etablieren, um hohe Gewinne zu erzielen. Im Produktentwicklungsprozess stehen oft die technischen Lösungen im Vordergrund und nicht die Anforderungen des Kunden, bzw. des Marktes. Der Verbraucher möchte ein innovatives, preisgünstiges, aber qualitativ hochwertiges Produkt, wobei das Wort *Qualität* nicht allgemeingültig definiert werden kann, da es für jeden Einzelnen eine andere Bedeutung haben kann. Hier spielen die geografische Herkunft, Gewohnheiten, Erfahrungen und persönliche Einstellungen eine entscheidende Rolle, d.h. der pragmatische Kontext der Person muss beachtet werden.

Mit Hilfe verschiedener Marktforschungs- und Qualitätsmethoden können die individuellen Wünsche des Kunden erfasst werden und bilden die Grundlage für die Produktentwicklung. Die Aufgabe der Entwicklungsabteilungen besteht darin, diese *Kunden-* oder *Initial-Anforderungen* in technische *Detail-Anforderungen* zu übersetzen und auf deren Basis eine bestmögliche, wirtschaftliche Lösung zu finden, die im Produkt umgesetzt wird. Der gesamte Prozess ist in *Abbildung 1* dargestellt.

Die *Detail-Anforderungen* werden im Entwicklungsprozess ständig verfeinert und durch Tests mit der aktuellen Lösung abgeglichen. Dieser Iterationsprozess durchläuft mehrere Konkretisierungsstufen und Zuständigkeitsbereiche, bis letztendlich die Serienproduktion beginnen kann.

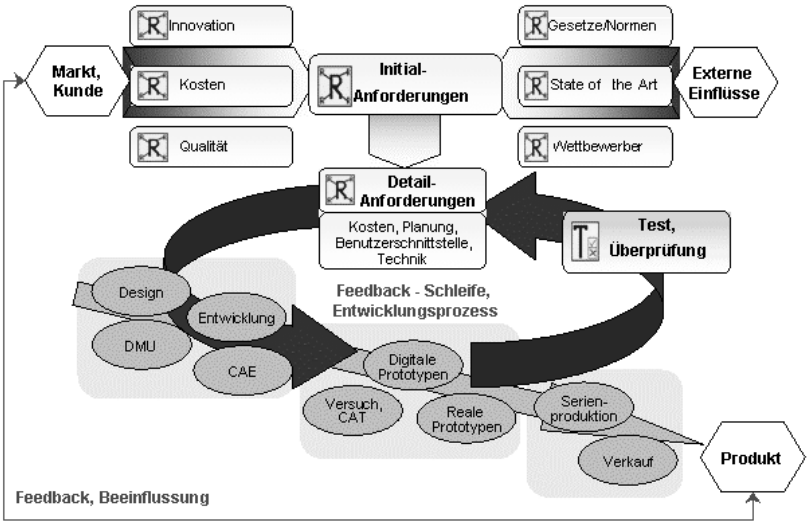

Abbildung 1: Anforderungen im Produktentstehungsprozess.

Die Produktanforderungen werden in der Regel mit *MS Office-Anwendungen* (*Word*, *Excel*, *PowerPoint*) erfasst und verwaltet. Der Vorteil besteht darin, Dokumente einfach und standardisiert austauschen zu können und mit bekannten Tools zu bearbeiten. Um jedoch Daten strukturiert zu erfassen und in den gesamten Entwicklungsprozess zu integrieren, sind diese Anwendungen nicht geeignet. Da die Anforderungen nicht als einzelne Objekte verfügbar sind (vereinzelt), können datentechnisch keine Abhängigkeiten zwischen diesen definiert werden.

Auf dem Markt sind Anwendungen erhältlich, die speziell zur Handhabung von Anforderungen entwickelt wurden (z.B. *Doors*, *Slate*) und auch Schnittstellen zu den bekannten Tools (*MS Office*, *Matlab*, usw.) bieten. Die Bedienung gestaltet sich jedoch deutlich aufwändiger als die der bekannten Office-Anwendungen, weshalb der Anwender geschult werden sollte.

Die Analyse realer Entwicklungsprozesse hat gezeigt, dass *MS Office* ein Standardwerkzeug darstellt, in dem die Anforderungen textlich erfasst werden. Diese Anforderungen werden von unterschiedlichen Personen erstellt und bearbeitet, welche nicht immer den gleichen Fachwortschatz verwenden. Hieraus können zwei Forderungen abgeleitet werden.

- Anforderungen werden als Fließtext erstellt und müssen dann vereinzelt werden, um in ein semantisches Anforderungsnetz überführt zu werden.
- Die verwendeten Worte müssen für alle Projektbeteiligten (Stakeholder) eindeutig und verständlich sein.

Im Rahmen eines internen DaimlerChrysler – Projektes wurde ein Tool (*ReMaS* – „Requirements Management System“) entwickelt, mit dem komplexe Zusammenhänge dargestellt und

ausgewertet werden können. Im folgenden Kapitel wird dieses System kurz vorgestellt.
Für den Benutzer besteht jedoch auch hier der Nachteil, ein neues System einsetzen zu müssen. Anhand von *ReMaS* soll gezeigt werden, dass es prinzipiell möglich ist, textliche, qualitative Produktanforderungen automatisiert zu analysieren und den Benutzer bei deren Definition zu unterstützen, um zu einer eindeutigen, vereinzelten Anforderung zu gelangen, die für alle Prozessbeteiligten verständlich ist. Zur semantischen Analyse soll *GermaNet* verwendet werden. *GermaNet* ist ein lexikalisches Wort netz, basierend auf dem *Princeton WordNet*.
http://www.cogsci.princeton.edu/~wn/ (2003)

2 Anforderungsmodellierung

2.1 Das Software-Tool ReMaS

Merkmale einer Anforderung

Eine Anforderung beschreibt generell eine gewünschte Eigenschaft oder ein Merkmal des zu entwickelnden Produktes, d.h. sie kann auch als *Soll-Eigenschaft* bezeichnet werden, wogegen das Produkt die *Ist-Eigenschaft* repräsentiert.

Die Anforderung kann vom Typ *qualitativ* oder *quantitativ* sein. Der Typ *qualitativ* charakterisiert eine Eigenschaft in sprachlicher Form und kann nicht unmittelbar am Produkt überprüft werden. Eine *quantitative* Anforderung beschreibt eine Produkteigenschaft durch eine eindeutige Beschreibung (Zahl oder Wort) und kann unmittelbar auf ihre Erfüllung getestet werden. Zusätzlich besitzt jede Anforderung verschiedene Attribute, die im Gesamtprozess von Bedeutung sind (z.B. Gewichtung, Status, Erstelldatum, verantwortliche Personen, usw.).

Sichten und Darstellung

Zur übersichtlichen Darstellung werden die Anforderungen dem Benutzer in verschiedenen Sichten visualisiert. Zur Zeit werden dem Anwender drei verschiedene Sichten zur Verfügung gestellt, über die er auf die Anforderungen zugreifen kann.

1. Produktstruktur, Bauteilhierarchie (*Part*): Jedem Bauteil und jeder Baugruppe können Anforderungen zugeordnet werden.
2. Anforderungsstruktur (*Requirement*): Hierarchische Darstellung der Produktanforderungen ausgehend von den *Initial-Anforderungen* bis zu den *Detail-Anforderungen* reichend.
3. Anforderungsklassifizierung (*Classification*): Jede Anforderung ist einer bestimmten Klassifizierung zugeordnet, um den genauen Kontext festzulegen. Die Klassifizierungshierarchie ist fest vorgegeben und gliedert sich in die fünf Hauptbereiche *Kosten* / *Planung*, *Markt*, *Vorschrift*, *Benutzerschnittstelle* und *Technik*. Diese Bereiche untergliedern sich in weitere Teilbereiche.

Links und Tests

Eine Anforderung kann zusätzlich mit einem Objekt der Klasse *Link* und *Test* verknüpft werden.

Ein *Link* kann externe Referenzen (Dateien, Anwendungen, URL, usw.) enthalten, die nicht in *ReMaS* erfasst werden. Dies dient hauptsächlich zur Verknüpfung mit Zusatz- und Hintergrundinformationen und nicht zur Anforderungsspezifikation in externen Dokumenten.

Diese Funktionalität ist aus praktischen Gründen notwendig, da hierdurch existierende Systeme und Datenformate angebunden werden können.

Um eine Anforderung zu prüfen (Soll-Ist-Abgleich), muss dieser ein *Test* zugeordnet werden. Die Tests und Prüfungen sind in der Klasse *Test* hierarchisch gegliedert und beschreiben die Evaluierung und deren Ablauf. Ein *Test* kann eine Berechnung, eine Simulation, ein Fahrversuch oder auch die Bewertung durch eine Kontrollgruppe sein. Somit ist die Vergleichbarkeit mit

anderen Projekten und die objektive Ermittlung des aktuellen Entwicklungsstandes möglich.

Arbeitsweise von ReMaS

Die Funktionalität von *ReMaS* soll anhand eines Beispiels aus der Automobilindustrie erläutert werden.

Der Start eines Projektes beginnt durch die Auswahl einer geeigneten Vorlage (*Template*) für das gewünschte Produkt. Der Benutzer wählt das Template *PKW-Limousine*, und *ReMaS* legt automatisch die Produktstruktur für eine Limousine an. Den einzelnen Baugruppen oder Bauteilen sind Anforderungen an diese zugeordnet, die für jedes Projekt der Kategorie *PKW-Limousine* zu beachten sind. Generelle Abhängigkeiten zwischen den Anforderungen werden ebenso automatisch erzeugt.

Die Aufgabe der Anwender besteht im nächsten Schritt darin, die Anforderungen eindeutig zu definieren, neue projektspezifische Anforderungen hinzuzufügen und zusätzliche Abhängigkeiten zu definieren.

So stellt z.B. die *Tür* eine Baugruppe der *Limousine* dar und besitzt die Anforderungen *Masse* und *Sicherheit*. Die *Tür* wiederum besteht u.a. aus einem *Fenster* und der *Innenverkleidung*. Das *Fenster* besitzt die Anforderung *Glasdicke* und die *Innenverkleidung* die Anforderung *Design*. Der für die *Tür* verantwortliche Bearbeiter definiert nun die Anforderungen *Masse < 50 kg*, *Glasdicke = 3 mm* und *Design soll elegant, hochwertig und luftig* sein. Zusätzlich fügt er der *Innenverkleidung* die neue Anforderung *Leichte Reinigung (nass + trocken)* hinzu.

Aus dem Template wird die Abhängigkeit zwischen den Anforderungen *Masse* und *Glasdicke* automatisch übernommen, da die *Glasdicke* die *Masse* der ganzen *Tür* beeinflusst. Zusätzlich erzeugt der Bearbeiter eine Abhängigkeit zwischen *Design* und *Reinigung*, da nicht alle möglichen Materialien für die *Innenverkleidung* die Anforderung *Leichte Reinigung* erfüllen.

Zusätzlich werden aus dem Template Relationen zu bestimmten Tests und Links übernommen. Der Anforderung *Sicherheit* ist der Test *Crashtest nach Euro-NCAP* zugeordnet. Zum besseren

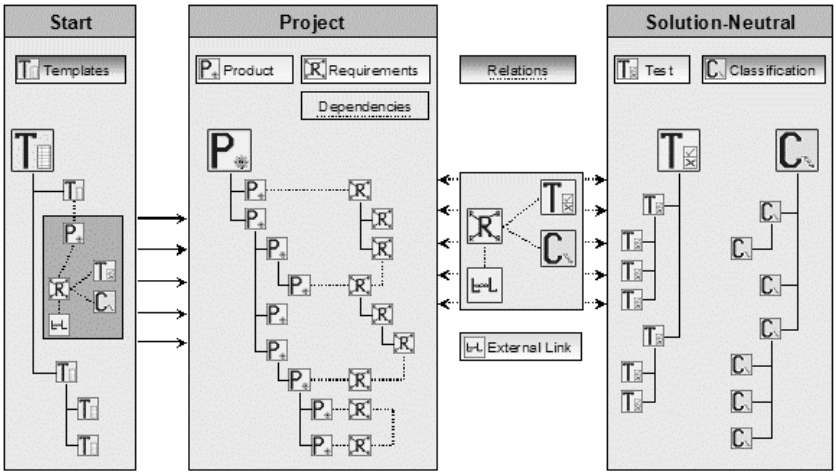

Abbildung 2: Anforderungsmodellierung in ReMaS [Jörg, Klaar, Lossack (2003)]

Verständnis besitzt die Anforderung *Sicherheit* den Link *Euro-NCAP*, über den auf die komplette Crashvorschrift zugegriffen werden kann.

2.2 Anforderungstypen

Wie im vorhergehenden Kapitel erwähnt, gibt es generell zwei verschiedene Typen von Anforderungen.

Quantitative Anforderungen können immer durch einen exakten Wert oder Wertebereich beschrieben und direkt überprüft werden.

Qualitative Anforderungen sind dadurch gekennzeichnet, dass sie nicht explizit auf eine Lösungseigenschaft zeigen und somit auch nicht direkt messbar sind. Eine eindeutige Beschreibung der gewünschten Lösungseigenschaft ist entweder nicht möglich (z.B. *ästhetische Formgebung*) oder zu aufwändig (z.B. *korrosionsbeständig*).

Aus der Analyse verschiedener Produktspezifikationen konnten sechs signifikante Typen von Anforderungen identifiziert werden:

Eindeutig quantitativ
Kann durch die binäre Aussage *wahr* oder *falsch* eindeutig geprüft werden.

1. **Exakt (Zahl)**: durch einen Wert oder Wertebereich beschrieben. Bsp.: *Gewicht*, *Maße*, *Beschleunigung*, *Einbauraum*
2. **Exakt (Wort)**: durch einen sprachlichen Ausdruck beschrieben und eindeutig definiert. Bsp.: *Farbe*, *Antriebsart*, *Bauart*, *Einsatzort*, *Material*

Mischung quantitativ / qualitativ
Kann durch die Kenntnis der Fachdomäne und der referenzierten Objekte objektiv geprüft werden.

3. **Referenzierend**: durch referenzierte Objekte und Terminologien definiert. Bsp.: *wärmedämmendes Glas rundum*, *Gesetze*, *Normen*
4. **Relativ**: durch Vergleich mit allgemeinem oder speziellem Wissen definiert. Bsp.: *schmalere Säulenquerschnitte als Vorgänger*, *weniger Instrumente*

Eindeutig qualitativ
Kann durch einen bestimmten Personenkreis subjektiv geprüft werden.

5. **Subjektiv**: Interpretation personen- und kontextabhängig. Bsp.: *optimales Innenraumklima*, *elegante Formgebung*
6. **Nicht konkretisiert**: Konkretisierung zu aufwändig oder aktuell nicht möglich. Bsp.: *korrosionsbeständig*, *verschleißfrei*

Um die Handhabung speziell von qualitativen Anforderungen zu unterstützen, müssen diese immer im jeweiligen Anwenderkontext gesehen und bewertet werden. Die Absicht des Urhebers ist nicht immer ersichtlich, d.h. die Intension des sprachlichen Ausdrucks ist nicht für alle Personen eindeutig.

2.3 Semantische Abhängigkeiten von Anforderungen

Betrachtet man ein komplexes Produkt und identifiziert die Abhängigkeiten der Anforderungen untereinander ergibt sich ein weitreichendes *semantisches Anforderungsnetz*. Für den Produktentwickler ist ein solches Netz von großem Interesse, da er die Einflüsse, und vor allem solche zu fachfremden Domänen, sehr gut erkennen und damit den Einfluss einer Änderung einer einzelnen Anforderung abschätzen kann. Erfahrenen Ingenieuren sind die groben Abhängigkeiten in der Regel bewusst, so dass sie intuitiv die richtigen Entscheidungen treffen. Mit der wachsenden Komplexität der Produkte steigt auch die Größe des semantischen Anforderungsnetzes und im Gegensatz dazu sinkt die relative Kenntnis der einzelnen Person. Es kann davon ausgegangen werden, dass eine Vielzahl von Abhängigkeiten

unerkannt und nicht dokumentiert ist, so dass die Tragweite einer Änderung der Anforderungen bzw. der angestrebten Lösung nicht vollständig abgeschätzt werden kann.

Werden die möglichen Relationen in einem *semantischen Anforderungsnetz* betrachtet, so können diese in verschiedene Arten eingeteilt werden. In der Arbeit von Gebauer [Gebauer 2001] werden sechs verschiedene Arten von Abhängigkeiten vorgestellt:

1. *Zerlegt sich in*
   Zerlegung und Detaillierung einer Anforderung bis zur Elementaranforderung.
2. *Setzt sich zusammen aus*
   Anforderungen einer Ebene bilden gemeinsam die höhere Ebene.
3. *Erzeugt aus*
   Aus einer Anforderung ergeben sich andere Anforderungen.
4. *Unterstützend*
   Positive Abhängigkeit zweier Anforderungen (Gesamtoptimum).
5. *Konkurrierend*
   Negative Abhängigkeit zweier Anforderungen (nachteilig zum Gesamtoptimum).
6. *Gegensätzlich*
   Anforderungen schließen sich gegenseitig aus.

Für *quantitative Anforderungen* können in *ReMaS* zusätzlich physikalische Zusammenhänge in Form von mathematischen Gleichungen (sog. *Constraints*) definiert werden. Durch die Integration eines leistungsfähigen Moduls (*Constraint-Solver*), das die definierten Gleichungen analysiert, können Inkonsistenzen erkannt und gelöst werden. Ein inkonsistenter Zustand liegt vor, wenn es für die aktuellen Anforderungen (Parameter) nicht mindestens eine Lösung für das Gleichungssystem (*Constraint-Netz*) gibt. Dem Benutzer wird eine Strategie angeboten, mit der er Konflikte erkennen und lösen kann. Hierzu bestimmt er den Lösungsraum, die zu optimierenden Parameter und den Wertebereich. Das System liefert ihm den möglichen Lösungsbereich für das Problem zurück.

Für qualitative Anforderungen ist der *Constraint-Solver* aktuell nicht geeignet, da er eindeutige Eingaben benötigt.

3 Semantische Anforderungs-Analyse

3.1 Ziel und Motivation

In der täglichen Praxis werden *qualitative Produktanforderungen* in textlicher Form in Dokumenten erfasst und bearbeitet. Dies kann als allgemeiner Standard gesehen werden und muss akzeptiert werden. Diese Dokumente werden in einem großen Personenkreis verteilt, so dass sich aufgrund von unterschiedlichem Fachwissen mitunter Verständnisprobleme ergeben.

Mit dem hier beschriebenen Ansatz soll es möglich sein, vorhandene textlich beschriebene Anforderungen zu analysieren und in *ReMaS* zu übernehmen. Bei der Analyse sollen primär drei Bereiche abgedeckt werden.

1. Homonyme und Synonyme sollen erkannt werden, um die Eindeutigkeit zu gewährleisten.
2. Durch bestimmte Begriffe soll die Anforderung einer Domäne und Klassifizierung zugeordnet werden.
3. Unscharfe Begriffe wie Modalverben sollen konkretisiert und auf die quantitative Ebene gehoben werden (z.B. *sollen, können, müssen*).

Prinzipiell können Schnittstellen zu allen existierenden Anwendungen geschaffen werden, wobei hier das Zusammenspiel *ReMaS – GermaNet* untersucht werden soll.

3.2 Relationen im *GermaNet*

Das *GermaNet* ist ein Wortnetz, in dem semantische Relationen zwischen einzelnen Wörtern erfasst sind (Kunze 2003). Diese Relationen sol-

len dazu verwendet werden, textlich beschriebene Anforderungen zu analysieren und auf ihre Konsistenz und Eindeutigkeit zu überprüfen. Im Folgenden werden die in *GermaNet* erfassten Relationen beschrieben und anhand einiger technischer Beispiele erläutert.

Für die Anforderungsverarbeitung können die Relationen drei Zielbereichen zugeordnet werden. Die Begriffe müssen eindeutig sein, sie müssen in einer Hierarchie gegliedert sein, und der Kontext muss interpretiert werden.

Eindeutigkeit der Definitionen
Für einen Begriff kann es mehrere Benennungen geben, oder eine Benennung hat verschiedene kontextabhängige Bedeutungen (DIN 2330; DIN 2342). Es muss sichergestellt werden, dass die Stakeholder sich auf den gleichen Begriff beziehen und die Intention des Urhebers erkennen. Zur Sicherstellung der Eindeutigkeit muss das System prüfen, ob ein Begriff für den Anwender verständlich ist. Hierzu können die folgenden Relationen verwendet werden:

- **Synonymie**: Bedeutungsgleichheit zweier oder mehrerer sprachlicher Zeichen. Bsp.: *LKW, Truck*
- **Homonymie, Polysemie**: Formgleichheit zweier oder mehrerer sprachlicher Zeichen. Bsp.: *Gang* (Homonym), *Sitz* (Polysem)

Hierarchisierung, Einordnung
Über die Hierarchisierungsrelationen kann eine Einordnung des verwendeten Begriffes erfolgen und somit das Umfeld der sprachlichen Bedeutung eingeschränkt werden.

- **Hyperonymie** (⇑), **Hyponomie** (⇓), **Kohyponomy**: Vererbungsbeziehung der Über- / Unterordnung ("gehört zu"). Bsp.: *Fahrzeug* ist der Überbegriff zu *LKW, PKW, Motorrad*, ...
- **Meronymie** (⇓), **Holonymie** (⇑): Beziehung der Über- / Unterordnung im Sinne einer Aggregation ("besteht aus"). Bsp.: *PKW* besteht aus *Motor, Karosserie, Türen*, ...

Interpretation, Bedeutung
Um inhaltliche Abhängigkeiten und Widersprüche zu erkennen, soll die Anforderung vom System interpretiert werden können. Unsinnige, widersprüchliche Wortkombinationen und gegenseitige Abhängigkeiten sollen erkannt werden, um die Konsistenz der Anforderung sicherzustellen.

- **Antonymie**: Gegensätze lexikalischer Einheiten. Komplementäre, graduierbare und relationale Opposition. Bsp.: *dynamisch–statisch* (komplementär), *groß–klein* (graduierbar), *größer–kleiner* (relational)
- **Kausation**: inhaltlicher Zusammenhang zwischen lexikalischen Resultativen. Bsp.: *öffnen–offen*, *lackieren–lackiert*
- **Implikation, Entailment**: logische Folgerung zwischen zwei Begriffen. Bsp.: *gelingen–versuchen, fahren–starten*
- **Ähnlichkeit: assoziative** Verknüpfung im bestimmten Kontext. Bsp.: *Bremse–Pedalkennlinie, Motor–Fahrdynamiktest*
- **Pertonymie**: semantische Derivationsbeziehung, Wortursprung. Bsp.: *rostfrei–Rost*

Benutzerinteraktion und Ablauf
Um eine textlich beschriebene Anforderung zu analysieren, gibt es generell zwei Möglichkeiten. Das System kann den Benutzer direkt bei der Eingabe unterstützen und interaktiv mit einbinden (Nachfragen, Vorschläge unterbreiten) oder die Eingaben nachträglich prüfen (parsen). Bei beiden Methoden sollen Inkonsistenzen aufgedeckt und die Eindeutigkeit der Anforderung erhöht werden. Dies kann, wenn möglich, automatisch geschehen, wobei es immer eine menschliche Kontrollinstanz geben sollte.

Da eine interaktive Analyse des eingegebenen Textes nur sehr aufwändig zu realisieren ist, soll in der prototypischen Implementierung nur die Analyse eines vorhandenen Textes realisiert werden.

In Abbildung 3 ist die mögliche Arbeitsweise des Systems am Beispiel zweier Anwender dargestellt. Prinzipiell kann der Anwender die Anforderung in beliebiger sprachlicher Form beschreiben oder semantisch und pragmatisch korrekt definieren. Der Begriff *korrekt* meint in diesem Kontext: *eindeutig, klassifiziert* und *bewertbar*. *Eindeutig* meint die klare Zuordnung einer Benennung zum beabsichtigten Begriff. Durch die *Klassifizierung* wird der Kontext und somit die Terminologie der Anforderung festgelegt. *Bewertbar* bedeutet eine Transformation eines qualitativen Begriffes auf eine quantitative Ebene durch exakte Definitionen. Ist die Anforderung *korrekt*, so muss keine Analyse erfolgen. Im Folgenden wird der Ablauf beschrieben, wenn eine Anforderung in beliebiger sprachlicher Form erstellt wird.

Anwender 1 hat die Möglichkeit einen Text zu erstellen, der entweder eine Anforderung oder mehrere Anforderungen enthält. Ist die Anforderung vereinzelt, so kann eine direkte semantische Analyse erfolgen. Ist sie nicht vereinzelt, so muss der Text in vereinzelte Anforderungen aufgespaltet werden. Anhand eines Beispiels soll der Ablauf erläutert werden.

Der Anwender erstellt den Text:

Die Tür soll aus Stahl sein, max. 50 kg wiegen, muss die Crashvorschriften erfüllen und schön aussehen.

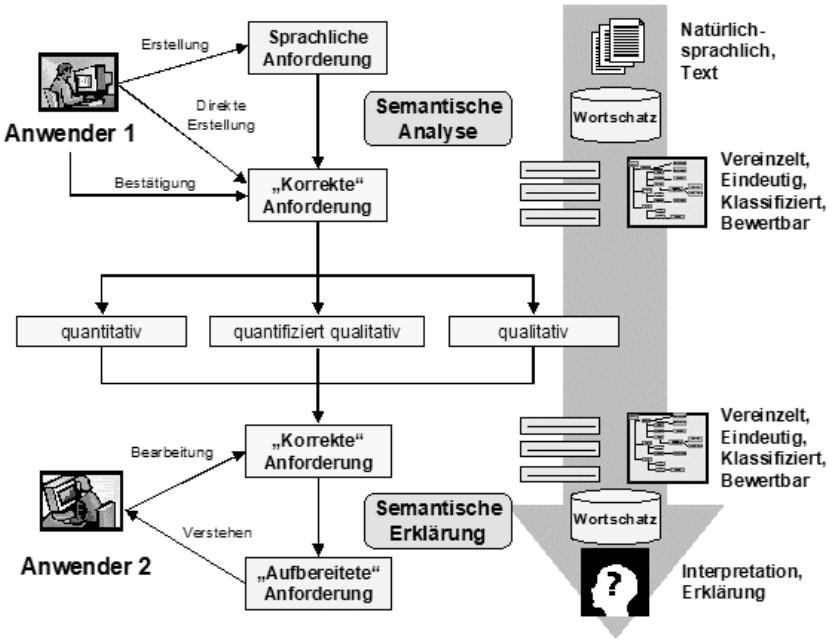

Abbildung 3: Semantische Analyse - Anwenderszenario

Dieser Text enthält vier einzelne Anforderungen:

- *Die Tür soll aus Stahl sein.*
- *Die Tür soll max. 50 kg wiegen.*
- *Die Tür muss die Crashvorschriften erfüllen.*
- *Die Tür soll schön aussehen.*

Korrekt definiert würden die Anforderungen im System wie folgt hinterlegt, bzw. abgeändert.

- *Die Tür (Bauteil=08_Tür_V) soll (Gewichtung=3) aus Stahl sein (Klassifizierung=5221_Werkstoff), (Test=2112_Materialprüfung).*
- *Die Tür (08_Tür_V) soll (3) max. 50 kg wiegen (5254_Gewicht), (2112_Wiegen).*
- *Die Tür (08_Tür_V) muss (5) die Crashvorschriften erfüllen (3121_Gesetz), (2122_Crashtest).*
- *Die Tür (08_Tür_V) soll (4) schön aussehen (4122_Design), (2112_Bewertung_Kontrollgruppe).*

Für eine automatische Aufspaltung des Textes in einzelne, sinnvolle Anforderungen muss die Satzstruktur analysiert und interpretiert werden, was einen hohen Aufwand voraussetzt. Im Beispielszenario soll davon ausgegangen werden, dass die Anforderungen schon vereinzelt vorliegen.

Anwender 1 erstellt eine sprachliche Anforderung (*Die Tür soll max. 50 kg wiegen*) und lässt diese vom System analysieren. Das System prüft den Text auf Homonyme und Synonyme, sucht Schlüsselwörter, die eine Klassifizierung ermöglichen und erkennt unscharfe Begriffe. Der Anwender bekommt als Ergebnis eine Liste der Homonyme und Synonyme angezeigt, eine Klassifizierungskategorie wird vorgeschlagen und unscharfe Begriffe müssen spezifiziert werden.

Im Beispiel erkennt das System den Begriff *Tür*, wobei es als Untergruppen *Vorder-* *Hinter-* und *Heck-Türen* gibt. *Anwender 1* wählt die Vordertür *08_Tür_V*. Der Begriff *soll* wird als unscharf erkannt und der Anwender bekommt die *Gewichtung=3* vorgeschlagen, welche er bestätigen muss. Über den Zahlenwert *50* mit der Einheit *kg* schlägt das System eine Zuordnung der Anforderung zur Klassifizierung *5254_Gewicht* vor. Für jede Klassifizierung gibt es in *ReMaS* eine Auswahl an Tests, mit denen eine Anforderung dieser Art geprüft werden kann. *Anwender 1* entscheidet sich für den Test *2112_Wiegen*. Somit ist die Anforderung *korrekt* definiert.

Da jeder Anwender einen unterschiedlichen Wissensstand besitzt, sollten einem Anwender einer fachfremden Domäne die semantisch eindeutigen Anforderungen erläutert werden können. Hierzu kann *Anwender 2* einzelne Wörter markieren und bekommt Hintergrundinformationen zu diesen. Dies kann z.B. eine Liste von Synonymen sein, die Anzeige der Begriffshierarchie oder mögliche Klassifizierungen für diesen Begriff.

Zum Aufbau eines semantisches Anforderungsnetzes ist es unabdingbar, möglichst viele eindeutige Anforderungen in vereinzelter Form vorliegen zu haben.

3.4 Implementierung und Verknüpfung mit *GermaNet*

Der Datenaustausch mit *GermaNet* soll über XML-Dateien erfolgen. Die genaue Realisierung muss noch untersucht werden. Eine Online-Kopplung ist für den Prototypen nicht notwendig, da die Funktionsfähigkeit des Ansatzes an einigen exemplarischen Beispielen gezeigt werden soll.

Die vorhandenen Datenbankeinträge müssen um fachspezifische Bibliotheken, bzw. lexikografische Files erweitert werden. Hierzu werden typische Begriffe und Benennungen aus der Domäne *Automobilbranche* zusammengetragen und erfasst.

Die folgenden Relationen aus *GermaNet* sollen verwendet werden, um die Anforderung in einen *korrekten* Zustand zu überführen.

Synonymie, Homonymie, Polysemie
Die in der Anforderung enthaltenen Worte werden überprüft, ob es mehrere Begriffe oder Benennungen zu diesem Wort gibt. Der Anwender bekommt die möglichen Begriffe oder Benennungen angezeigt und wählt einen verbindlich aus. Über diesen Mechanismus sollen zum einen die Begriffe eindeutig gemacht, zum anderen unscharfe Begriffe durch standardisierte ersetzt werden. In *ReMaS* werden nur die eindeutigen und standardisierten Begriffe übernommen und Begriffe, die den Erfüllungsgrad einer Anforderung beschreiben (z.B. *soll, kann, muss*), werden in eine Gewichtung umgesetzt.

Wünschenswert wäre die Möglichkeit, eine Benennung bzw. einen Begriff als Standard zu definieren und dem Anwender vorzuschlagen.

Meronymie, Holonymie
Über diese Relationen soll die Produktstruktur aufgebaut werden, damit das System die Begriffe erkennt, die ein Bauteil repräsentieren. Taucht der Begriff in einer Anforderung auf, so kann diese in *ReMaS* dem genannten Bauteil zugeordnet werden.

Die Bauteilbezeichnungen sollten als zusätzlicher Wortschatz definiert werden, da sie abhängig von der Branche und Firma sind. Zusätzlich müssen *Synonyme* der Bauteile abgeprüft werden, da nicht jeder Anwender die exakt gleiche Benennung verwendet.

Hyperonymie, Hyponomie, Ähnlichkeit
Über die *Hyperonymie* und *Hyponomie* soll die Hierarchie der Anforderungsklassifizierung aufgebaut werden. Die *Ähnlichkeits-Relation* verknüpft die Begriffe der Klassifizierung mit den möglichen zugehörigen Worten (z.B. besitzen

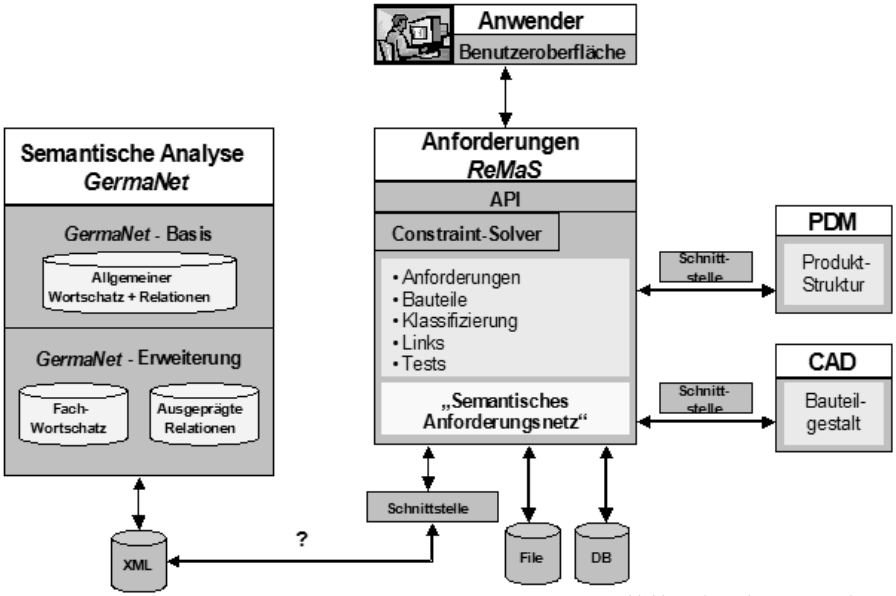

Abbildung 4: Implementierungskonzept

die Worte *kg* und *Masse* eine Relation zu 5254_ Gewicht oder *Design* zu 4122_Design).

Die Anforderungsklassifizierung sollte als zusätzlicher Wortschatz definiert werden, da sie durch eine Nummerierung ergänzt ist und erweitert werden kann. Die *Ähnlichkeits-Relationen* zwischen den Wörtern und der Klassifizierung zu erstellen wird eine umfangreiche Aufgabe sein, da im Prinzip der gesamte relevante Wortschatz eingeordnet werden muss.

4 Zusammenfassung

Die im Rahmen dieser Arbeit durchgeführten Untersuchungen haben gezeigt, dass die sprachliche und semantische Analyse von technischen Produktanforderungen ein weites Forschungsfeld darstellt. Auf dem Markt gibt es keine professionellen Tools oder Systeme, die eine Untersuchung und Bewertung von textlich beschriebenen Anforderungen unterstützen.

Auf dem Gebiet der Linguistik beschäftigen sich weltweit verschiedene Forschungseinrichtungen mit der menschlichen Sprache und Schrift (Müller 2002). Es gibt einige Ansätze und Lösungen, die Sprache zu erfassen und automatisiert auszuwerten. Die Verknüpfung mit dem Gebiet der Anforderungsmodellierung soll durch *GermaNet* realisiert werden.

Die Analyse mehrerer Produktspezifikationen hat gezeigt, dass diese zum Großteil textlich und wenig strukturiert erstellt sind, so dass ein System, das diese Beschreibungen analysieren kann, von großem Nutzen ist. Der Benutzer soll im Sprachgebrauch nicht eingeschränkt, sondern unterstützt werden. Mögliche Fehlinterpretationen und semantische Abhängigkeiten sollen aufgezeigt werden.

Erst durch die Vereinzelung und Verfeinerung der textlichen Anforderungen kann der Übergang zum *semantischen Anforderungsnetz* geschaffen werden. Es muss eine möglichst vollständige und eindeutige Anforderungsmenge vorliegen, um den ganzen Nutzen auszuschöpfen.

In dieser Arbeit werden erste Ansätze präsentiert und entwickelt, die eine Verbindung zwischen den Domänen der Anforderungsmodellierung und der Linguistik schaffen, so dass sich ein weiteres interessantes und umfangreiches Forschungsgebiet eröffnet.

Literatur

Deutsches Institut für Normung (1993). DIN 2330. Begriffe und Benennungen. Berlin: Beuth-Verlag.

Deutsches Institut für Normung (1992). DIN 2342. Begriffe der Terminologielehre. Berlin: Beuth-Verlag.

Gebauer, M. (2001). Kooperative Produktentwicklung auf Basis verteilter Anforderungen. Dissertation, Universität Karlsruhe, Institut für Rechneranwendung in Planung und Konstruktion. Aachen: Shaker-Verlag.

GermaNet-Projekt (2003). GermaNet Homepage. Universität Tübingen, Seminar für Sprachwissenschaft. *http://www.sfs.uni-tuebingen.de/lsd/* [Zugriff April 2004].

Jörg, M.; Klaar, O.; Lossack, R.-S. (2003). "Requirement Driven Engineering in Collaborative Environments." In: Proceedings eChallenges 2003, Bologna, Oktober 2003.

Kunze, C. (2001). „Lexikalisch-semantische Wortnetze." In: Carstensen, K.-U. et al. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: Eine Einführung. Heidelberg: Spektrum Akademischer Verlag, 386-393.

Müller, H. M. (2002). Arbeitsbuch Linguistik, Paderborn: Schöningh Verlag.

Wordnet-Projekt (2003). Wordnet Homepage. Princeton University, Cognitive Science Laboratory. *http://www.cogsci.princeton.edu/~wn/* [accessed April 2004].

Rainer Osswald

Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons HaGenLex

Abstract

Dieser Beitrag soll am Beispiel des semantikbasierten Computerlexikons HaGenLex aufzeigen, wie GermaNet für die Pflege und Erweiterung anderer lexikalisch-semantischer Ressourcen eingesetzt werden kann. Ausgangsbasis ist dabei eine Lesartenzuordnung zwischen GermaNet- und HaGenLex-Einträgen, welche die Übertragung der sinnrelationalen Zusammenhänge von GermaNet auf HaGenLex erlaubt. Auf der Grundlage dieser Kopplung lassen sich beispielsweise Inkonsistenzen in der semantischen Klassifikation von HaGenLex-Einträgen aufdecken. Neben weiteren Anwendungen werden einige sich dabei ergebende Probleme sowie der mögliche Nutzen für die Aufdeckung von Fehlern in GermaNet angesprochen.

1 Das Lexikon HaGenLex

HaGenLex (**Ha**gen **Ge**rma**n** **Lex**icon) ist ein semantikbasiertes Computerlexikon für das Deutsche, das seit 1996 an der FernUniversität Hagen am Lehrgebiet Praktische Informatik VII entwickelt wird. Momentan umfasst es circa 20.000 Lesart-Einträge (etwa 9.200 Substantive, 6.500 Verben und 3.000 Adjektive). Die Einträge wurden primär auf der Grundlage von Frequenzlisten erstellt, mit Unterstützung diverser Wörterbücher des Deutschen. Die Erstellung durch den Lexikographen wird maßgeblich durch eine Werkbank unterstützt, die zum einen die Eingabe leitet, und zum anderen die interne Repräsentation der Einträge als Merkmal-Wert-Strukturen vor dem Nutzer verbirgt bzw. in leicht verständlicher Umschreibung darbietet.

Um Missverständnissen vorzubeugen sei darauf hingewiesen, dass sich HaGenLex von GermaNet in der Gebrauchsweise des Konzeptbegriffs unterscheidet: Während GermaNet, in der Tradition von WordNet, Konzepte durch Synsets repräsentiert sieht, wird in HaGenLex davon ausgegangen, dass jedes lexikalisierte Konzept genau einem Lexem entspricht. Ferner macht HaGenLex, im Gegensatz zu GermaNet, bislang nahezu keinen Gebrauch von künstlichen Konzepten.

Im Folgenden soll der Aufbau von HaGenLex kurz skizziert werden; eine ausführlichere Beschreibung findet sich in (Hartrumpf et al. 2003).

1.1 Der MultiNet-Formalismus

Die Mittel zur Darstellung semantischer Information in HaGenLex sind dem sogenannten *MultiNet-Paradigma* entnommen. Bei diesem handelt es sich um einen Formalismus zur Darstellung der Semantik natürlicher Sprache mittels mehrschichtiger erweiterter semantischer Netze.[1] Grob gesprochen besteht ein solches semantisches Netz aus Knoten, die Konzepte repräsentieren, und Kanten, welche die semantischen Beziehungen zwischen den Konzepten zum Ausdruck bringen.[2] Zur Charakterisierung der Beziehung zwischen Konzepten stellt der MultiNet-Formalismus ein vordefiniertes und ausführlich dokumentiertes Repertoire von weit über hundert Relationen und Funktionen bereit.

Darüber hinaus ist jeder Konzeptknoten von MultiNet hinsichtlich mehrerer Merkmale spezifiziert, die unter anderem zum Ausdruck bringen, ob das Konzept generisch zu interpretieren ist, ob seine Referenz bestimmt oder unbestimmt ist, ob es faktischen oder hypothetischen Charakter hat, und in welcher Weise es einer Quantifikation unterliegt.[3] HaGenLex wurde in

erster Linie zu dem Zweck entwickelt, die automatische Transformation natürlichsprachlicher Ausdrücke in MultiNet-Repräsentationen zu unterstützen.[4] Die hierzu erforderliche syntaktische und semantische Information ist weitgehend lexikalisiert, wobei die Semantik im Lexikon ebenfalls durch MultiNet-Darstellungsmittel geprägt ist. Dazu zählen insbesondere die ontologische Sorte des zugehörigen Konzepts sowie die semantischen Relationen, in denen das Konzept zu anderen Konzepten steht.

1.2 Semantische Klassifikation

Im Rahmen von MultiNet steht eine Hierarchie von 45 ontologischen Sorten zur Klassifikation von Konzepten und damit von Lexemen zur Verfügung. Auf oberster Ebene wird zwischen Objekten, Sachverhalten, Sachverhaltsdeskriptoren, Qualitäten, Graduatoren, Quantitäten und formalen Entitäten unterschieden.

Insbesondere um die Überprüfung von Selektionsrestriktionen zu unterstützen, sind HaGenLex-Lexeme außerdem hinsichtlich 16 binärer semantischer Merkmale klassifiziert. Da zwischen diesen Merkmalen, wie im Fall von HUMAN und ANIMATE, semantische Abhängigkeiten bestehen, sind zulässige Kombinationen von ontologischer Sorte und semantischen Merkmalen zu sogenannten *semantischen Sorten* zusammengefasst. Beispielsweise verbergen sich hinter der semantischen Sorte *con-info* (für 'konkretes Informationsobjekt') die ontologische Sorte *d* (für 'Diskretum') sowie (unter anderem) die semantischen Merkmale [ANIMATE –] (nicht belebt), [ARTIF +] (Artefakt), [INFO +] (Informationsträger) und [MOVABLE +] (beweglich). Subsumierte Lexeme wären in diesem Fall *Abbildung* und *Zeitung*.

1.3 Valenz und Kasusrahmen

HaGenLex spezifiziert die Valenzen von Lexemen sowohl in syntaktischer als auch in semantischer Hinsicht. So ist zu jedem Verb angegeben, in welcher semantischen Beziehung die Partizipanten der vom Verb bezeichneten Situation zu letzterer stehen. Als Ausdrucksmittel finden hierfür wiederum die im Rahmen von MultiNet vorgegebenen semantischen Relationen Verwendung, die insbesondere ein Inventar an thematischen Rollen beinhalten.

In erster Näherung hat der Kasusrahmen für das Verb *informieren* in HaGenLex die folgende Form:

AGT	**OBJ**	**MCONT**
[POTAG +]	[POTAG +]	
np / nom	np / acc	'über'-*pp* / *acc*
	optional	optional

Die erste Zeile listet die thematischen Rollen der Argumente auf, die zweite enthält Selektionsrestriktionen (wobei POTAG für 'potential agent' steht), die dritte gibt (unvollständig) die syntaktischen Valenzen wieder und die letzte Zeile zeigt an, ob es sich um obligatorische oder fakultative Valenzen handelt. Der vollständig spezifizierte Lexikoneintrag ist in Abbildung 2 des Anhangs dargestellt.

1.4 Lexikonstruktur und Datenformat

Die lexikalische Information in HaGenLex ist in Form von *typisierten Merkmal-Wert-Strukturen* repräsentiert. Zugrunde liegt eine baumförmige *Typhierarchie* sowie zu jedem Typ eine *Merkmalsdeklaration*. Die für HaGenLex-Einträge verwendete Merkmal-Wert-Architektur ist im Anhang kurz skizziert.

Zur Strukturierung des Lexikons werden ferner sogenannte *Klassen* eingesetzt, die bestimmte, linguistisch relevante Merkmal-Wert-Kombinationen bündeln. Die Hierarchie der HaGenLex-Klassen ist vererbungsbasiert und erlaubt zudem die Verwendung von Default-Angaben.

2 Vergleich der Darstellungsmittel

Die semantischen Beschreibungsmittel von GermaNet sind darauf ausgerichtet, lexikalisch-se-

mantische Beziehungen zwischen lexikalischen Einheiten bzw. Konzepten darzustellen. Da das MultiNet-Paradigma den Anspruch eines universellen Semantikformalismus erhebt, ist zu erwarten, dass seine Beschreibungsmittel die von GermaNet umfassen.[5]

2.1 Relationen

Die für die Strukturierung von GermaNet und WordNet zentralen, der traditionellen lexikalischen Semantik entliehenen Sinnrelationen werden durch die folgenden MultiNet-Relationen abgedeckt:

SUB	Subordination, Hyponymie
SYNO	Synonymie
ANTO	Antonymie
COMPL	Komplementarität
CONTR	Kontrarität
CNVS	Konversbeziehung

Die drei letztgenannten Relationen sind Unterfälle der Antonymie. Konträr sind zwei Eigenschaften, wenn sie sich gegenseitig ausschließen, wie *heiß* und *kalt*. Komplementäre Eigenschaften müssen sich nicht nur ausschließen, sondern das gesamte Spektrum abdecken; Beispiel: *anwesend* vs. *abwesend*. Die Konversbeziehung zielt dagegen auf Argumentumordnung bei mehrstelligen Prädikaten, wie etwa beim Wechsel zwischen *geben* und *erhalten*. Aus Sicht der lexikalischen Semantik wird man solche Fälle eher unter einen allgemeineren Ansatz zur Behandlung von Diathesen fassen wollen.

Als Ausdrucksmittel vorgesehen sind in GermaNet außerdem eine Kausationsrelation, eine Implikationsbeziehung, eine Teil-Ganzes-Beziehung sowie eine kategorienübergreifende semantische Derivationsrelation (*Pertonymie*).[6] Die folgende Tabelle listet entsprechende, als Substitut taugliche MultiNet-Relationen auf:

CAUS	Kausalbeziehung
IMPL	Implikationsbeziehung
PARS	Teil-Ganzes-Beziehung, Meronymie
CHEA	Wechsel von Ereignis zu Abstraktum

Die letztgenannte Relation CHEA bedarf einer kurzen Erläuterung: Während GermaNet bislang nur eine unspezifische Ableitungsrelation bereitstellt, werden in MultiNet derartige Beziehungen semantisch differenziert ausgedrückt. So bringt CHEA die Beziehung zwischen einem verbalen Ereigniskonzept und dem durch die nominalisierte Form (*nomen actionis*) vermittelten abstrakten Gegenstandskonzept zum Ausdruck, wie etwa zwischen *herstellen* und *Herstellung*. Für weitere derartige „Sortenwechselrelationen" siehe (Helbig 2001).

In diesem Zusammenhang sei angemerkt, dass die semantische Beziehung zwischen einem Verb und seiner Subjekts- bzw. Objektsnominalisierung (*nomen agentis* bzw. *nomen patientis*) in HaGenLex keiner gesonderten semantischen Derivationsrelation bedarf, sondern natürlicherweise durch die entsprechende thematische Rolle des Verbs gegeben ist. Da z.B. im Kasusrahmen von *prüfen* das Subjekt die Rolle AGT (Handelnder) und das direkte Objekt die Rolle OBJ (neutrales Objekt) innehat, ist im Lexikon vermerkt, dass zwischen der Subjektsnominalisierung *Prüfer* und *prüfen* die Relation AGT besteht, und zwischen *Prüfling* und *prüfen* die Relation OBJ.

Abschließend sei noch hervorgehoben, dass der eingangs vermerkte Unterschied zwischen HaGenLex und GermaNet hinsichtlich der Gebrauchsweise des Konzeptbegriffs unter einem formalen Gesichtspunkt vernachlässigbar ist: HaGenLex macht von der Synonymie als *Äquivalenzrelation* Gebrauch, wohingegen GermaNet direkt die daraus resultierenden *Äquivalenzklassen* repräsentiert. Dass sich Hyponymie und Hyperonymie auf ganze Synsets erstrecken, lässt

sich dann durch geeignete MultiNet-Axiome erzwingen:

$\mathrm{SUB}(c_1,c_2) \wedge \mathrm{SYNO}(c_2,c_3) \rightarrow \mathrm{SUB}(c_1,c_3)$,
$\mathrm{SYNO}(c_1,c_2) \wedge \mathrm{SUB}(c_1,c_3) \rightarrow \mathrm{SUB}(c_2,c_3)$.

Entsprechendes gilt für die Reflexivität, Symmetrie und Transitivität der SYNO-Relation.

2.2 Lesartenzuordung

Aufgrund der unterschiedlichen Abdeckung gibt es sowohl HaGenLex-Einträge, die keine GermaNet-Entsprechung haben, als auch den umgekehrten Fall. Momentan beinhalten knapp die Hälfte der HaGenLex-Einträge einen Verweis auf GermaNet-Lesarten (kodiert im lexikalischen Merkmal G-ID; vgl. Anhang).[7]

Bei der Zuordnung der Lesarten eines lexikalischen Wortes lassen sich folgende Fälle unterscheiden:

1. Die Zuordnung ist eineindeutig, d.h. jeder GermaNet-Lesart entspricht genau eine HaGenLex-Lesart, und umgekehrt.
2. Eine GermaNet-Lesart hat keine Entsprechung in HaGenLex.
3. Eine HaGenLex-Lesart hat keine Entsprechung in GermaNet.
4. Eine HaGenLex-Lesart fasst mehrere GermaNet-Lesarten zusammen. Beispielsweise wird *Aal* in GermaNet sowohl als Nahrungsmittel, d.h. als essbare Substanz, als auch als Lebewesen geführt, wohingegen beide Fälle in HaGenLex unter eine Lesart subsumiert sind. (Die diesem Fall zugrunde liegende reguläre Polysemie bzw. Metonymie ist in HaGenLex noch nicht hinreichend erfasst.)
5. Eine GermaNet-Lesart fasst mehrere HaGenLex-Lesarten zusammen. Ein Beispiel hierfür ist der transitive und intransitive Gebrauch von *baden*.

Sei G die Menge der lexikalischen Elemente von GermaNet und H die Menge der mit GermaNet verknüpften HaGenLex-Einträge. Ferner stehe Lgh dafür, dass einer GermaNet-Lesart g die HaGenLex-Lesart h entspricht. (Im Eintrag zu h ist demnach unter dem Merkmal G-ID die Menge $\{ g \in G \mid Lgh \}$ kodiert.) Die auf G gegebene Synonymierelation lässt sich vermittels L auf H projizieren: Sei $\sim$ die kleinste Äquivalenzrelation auf H derart, dass $h \sim h'$ wenn Lgh und $Lg'h'$ für zwei zueinander synonyme Elemente g und g' von G. Mit anderen Worten, $\sim$ ist die transitive Hülle derjenigen symmetrischen und reflexiven Relation auf H, in der zwei Elemente h und h' genau dann stehen, wenn sie GermaNet-Entsprechungen haben, die synonym zueinander sind. (Die Bildung der transitiven Hülle ist aufgrund von Fall 4 erforderlich.)

Die Projektion $\pi(S)$ eines GermaNet-Synsets S auf HaGenLex ist nun folgendermaßen definiert: Steht ein Element von S in der Beziehung L zu einem HaGenLex-Eintrag h, dann sei $\pi(S)$ die zu h gehörige Äquivalenzklasse bzgl. $\sim$; andernfalls sei $\pi(S)$ leer. (Man beachte, dass die Äquivalenzklasse dabei unabhängig von der Wahl von h ist.)

3 Nutzen von GermaNet für HaGenLex

Die Lesarten-Zuordnung erlaubt es, GermaNet in Anwendungen von HaGenLex bei der Informationsrecherche als unterstützende Ressource einzusetzen. Insbesondere können dabei auch GermaNet-Hyperonyme eines in der Anfrage vorkommenden HaGenLex-Lexems herangezogen werden, die keine Entsprechung in HaGenLex haben. Zudem ist über die GermaNet-Verbindung der interlinguale Index von EuroWordNet zugreifbar, wodurch multilinguale Anwendungen, wie die Recherche fremdsprachiger Dokumente mittels deutschsprachiger Anfragen, unterstützt werden.

Im Folgenden wollen wir jedoch nicht die Einsatzmöglichkeiten von GermaNet als ergän-

zende Ressource bei HaGenLex-Anwendungen thematisieren, sondern die Verwendung von GermaNet bei der Pflege und Erweiterung von HaGenLex selbst an zwei Beispielen illustrieren.

3.1 Konsistenzüberprüfung

Eine naheliegende Anwendung der GermaNet-Kopplung besteht darin, die Synset-Projektionen in HaGenLex auf semantische Konsistenz zu überprüfen. Da synonyme Lexeme identische oder zumindest kompatible semantische Sorten aufweisen sollten, ergeben sich dadurch Hinweise auf mögliche Kategorisierungsfehler in HaGenLex. Die auf diese Weise automatisch gewonnenen Hinweise sind dann vom Lexikographen im Einzelnen zu überprüfen. Für eine Inkompatibilität in einer Synset-Projektion kommen als mögliche Fehlerquellen in Betracht:

1. Die semantische Kategorisierung von HaGenLex-Lexemen,
2. die Lesart-Zuordnung zwischen HaGenLex und GermaNet,
3. die Synset-Bildung in GermaNet.

Die ersten beiden Fälle ziehen eine Korrektur von HaGenLex-Einträgen nach sich, wobei der erste Fall eine Qualitätsverbesserung von HaGenLex im engeren Sinne bewirkt.

Der dritte Fall kann auf eine Schwachstelle in GermaNet hindeuten. Das geschilderte Verfahren führt beispielsweise bei der Projektion des GermaNet-Synsets {*Nervosität.1*, *Hektik.1*, ...} zu einer Sorten-Inkompatibilität, da der HaGenLex-Eintrag zu *Nervosität* im Gegensatz zu dem von *Hektik* als (*mentaler*) *Zustand* klassifiziert ist. In der Tat ist die Nervosität einer Person keineswegs mit hektischem Verhalten gleichzusetzen, sodass das Postulat der Synonymie von *Nervosität* und *Hektik* unangemessen scheint. (Man beachte, dass GermaNet keine weiteren Lesarten zu *Nervosität* oder *Hektik* anbietet, und dass {*Gefühl.1, Emotion.1, Empfindung.1, Gemütsbewegung.1*} als Hyperonym fungiert.)

Hinweise auf mögliche Defekte können sich auch aus Fällen ergeben, in denen zwar Kompatibilität gewährleistet ist, aber Unterschiede in der Sortenspezifizität vorliegen. Dies ist etwa bei der Projektion des Synsets {*Nachkomme.1, Kind.1, Nachfahre.1*, ...} der Fall: Die entsprechende Lesart von *Kind* ist in HaGenLex als menschlich klassifiziert, wohingegen *Nachkomme* und *Nachfahre* allgemeiner als Lebewesen eingeordnet sind. Hier ist festzustellen, dass die von GermaNet als synonym zu *Nachkomme* angenommene Lesart von *Kind* zumindest fragwürdig ist. (Als Hyperonym ist übrigens {*Verwandter.1, Angehöriger.1*, ...} vorgesehen.) Zwar ist jedes Kind ein Nachkomme und jeder Nachkomme ist ein Kind von jemandem, um aber als synonym zu gelten, müssten *Nachkomme* und *Kind* in Kontexten wie *Hans ist ein ____ von Adam und Eva* ohne große Bedeutungsverschiebung austauschbar sein, was offenbar nicht der Fall ist. Mit anderen Worten, die Synonymie inhärent relationaler Begriffe muss die zum Ausdruck gebrachte Relation respektieren.

3.2 Restriktion freier Ergänzungen

Neben der Angabe obligatorischer und fakultativer Valenzen schränken Lexeme in HaGenLex auch die möglichen freien Ergänzungen ein, was speziell die Disambiguierung von Präpositionalphrasen unterstützen soll. Zu diesem Zweck sind im Lexikoneintrag diejenigen MultiNet-Relationen aufgelistet, die mit dem zugehörigen Konzept kompatibel sind und damit einer freien Ergänzung semantisch zugrunde liegen können. Beispielsweise sind bei punktuellen Verben wie *platzen* oder *aufwachen* Dauerangaben (ohne iterative Umdeutung) ausgeschlossen, was sich im Lexikoneintrag dadurch niederschlägt, dass die Relation DUR unter dem lexikalischen Merkmal COMPAT-R (kurz für 'compatible relations') nicht aufgeführt ist. Die erforderliche aktionsartliche

resp. aspektuelle Klassifizierung von Verben wird allerdings von der GermaNet-Hierarchie nicht hinreichend unterstützt. So ist *platzen.1*, ebenso wie *reißen.1*, *abkühlen.1* und *trocknen.1*, Hyponym von *?Mat_Zustands_Veränderung.1* ('Veränderung der materiellen Beschaffenheit'); und *aufwachen.1* ist, ebenso wie *altern.1* und *?Mat_Zustands_Veränderung.1*, Hyponym von {*wandeln.3, ändern.1*, ...}. Offenbar geht die Punktualität von Geschehnissen (bzw. deren Ingressivität, Egressivität oder Semelfaktivität) bisher nicht als Strukturierungskriterium in die Verbklassifikation von GermaNet ein.[8]

Im Fall *direktionaler* Ergänzungen, die auf semantischer Ebene durch die MultiNet-Relation DIRCL ('direction/local goal') zum Ausdruck gebracht werden, erscheint dagegen die Heranziehung der GermaNet-Hierarchie auf den ersten Blick vielversprechend. Da die Klasse der Lokationsverben in GermaNet relativ gut strukturiert ist, besteht die Hoffnung, die Klasse der Verben mit möglichen direktionalen Ergänzungen durch eine geeignete Kombination von Einschluss- und Ausschlussklassen herauszufiltern.

Es ergeben sich jedoch verschiedene Schwierigkeiten. Als erstes wäre ein Problem zu nennen, das in der Natur der Sache selbst liegt: Bei Verben, die eine gerichtete Bewegung zum Ausdruck bringen, wird die Richtungsangabe häufig nicht als freie Fügung, sondern als Valenz angesehen. Beispiele aus HaGenLex sind *laufen* und *fahren*. Positiv gewendet scheint sich damit ein Hilfsmittel zur Aufdeckung direktionaler Valenzen anzudeuten. Hier muss aber sofort einschränkend darauf hingewiesen werden, dass Verbpartikeln wie *herum* (oder auch *vorbei* und *entlang*) direktionale Angaben blockieren:[9] **Peter fuhr in die Stadt herum*. Andererseits ist aber *herumfahren.1* Hyponym von *fahren.4*. Hier wäre eine sorgfältige Kreuzklassifikation bezüglich der durch Partikeln ausgedrückten Art und Weise der Bewegung hilfreich.[10]

Abschließend sei exemplarisch auf ein nicht untypisches Einzelproblem hingewiesen. Der Versuch, diejenigen Verben einzugrenzen, die direktionale Ergänzungen als freie Fügungen erlauben, führt unter anderem zum Ausschlusskonzept *?bewegen_auf_Stelle.1*. Hierunter findet man *?iterative_Bew.2* und darunter *tanzen.2*, das als einzig mögliche Lesart für *Das Paar tanzte ins Nachbarzimmer* in Frage kommt. In derartigen Fällen ist es nicht immer unmittelbar einsichtig, ob nur eine Fehlklassifikation in GermaNet vorliegt, oder ob es sich um ein tiefergehendes Problem in der semantischen Hierarchie handelt.

4 Zusammenfassung

Am Beispiel des semantikbasierten Lexikons HaGenLex wurde illustriert, wie eine Lesarten-Kopplung mit GermaNet die Pflege und Erweiterung anderer lexikalischer Ressourcen unterstützen kann, und wie sich dabei gleichzeitig Schwachstellen von GermaNet aufdecken lassen.

Insbesondere hat sich gezeigt, dass eine Restrukturierung resp. Ergänzung der GermaNet-Hierarchie hinsichtlich Aktionsart und Aspekt aus Sicht von HaGenLex sehr wünschenswert wäre. Das Gleiche lässt sich von einer Kreuzklassifikation von Partikelverben hinsichtlich des semantischen Beitrags der Partikeln sagen.

Anhang: Interne Repräsentation

Die interne Repräsentation von HaGenLex-Einträgen basiert auf einem Merkmal-Wert-Formalismus im Stile von (Carpenter 1992). Das Kerngerüst bildet eine baumförmige Typhierarchie, wobei zu jedem Typ angegeben ist, welche Merkmale mit welchen Werten für Strukturen dieses Typs zulässig sind. In Abbildung 1 sind einige der wichtigsten in HaGenLex verwendeten Merkmalsdeklarationen (in leicht vereinfachter Form) aufgeführt.

Alle lexikalischen Einträge sind vom Typ *word*. Da *word* in der Typhierarchie unter *sign* angeordnet ist, „erbt" jede Struktur vom Typ *word* per Konvention die für *sign* deklarierten

Merkmale. Die oberste Merkmalsebene eines Eintrags ergibt sich somit aus der Zusammenfassung der für *word* und *sign* deklarierten Merkmale; vgl. Abbildung 2. Neben Angaben zu Morphologie (MORPH) und Syntax (SYN) findet sich hier das Merkmal SEMSEL, das eine Struktur vom Typ *semsel* zum Wert hat, die ihrerseits die Semantik (SEM) und die Valenz (SELECT) des Eintrags näher spezifiziert. Die Valenzinformation wiederum besteht aus einer Liste von Strukturen des Typs *select-element*. Jede dieser Strukturen bringt die semantische Rolle des jeweiligen Arguments zum Ausdruck (REL), dessen syntaktische Notwendigkeit (OBLIG) sowie seine weitere Charakterisierung als Struktur vom Typ *sign*. Strukturen vom Typ *sem* schließlich kodieren die Semantik eines Eintrags durch seine semantische Sorte (ENTITY), zusätzliche MultiNet-Spezifikationen (NET), MultiNet-Merkmale des Konzeptknotens (LAY) sowie durch den Hinweis, ob eine bestimmte Ausprägung regulärer Polysemie vorliegt (MOLEC).

sign

MORPH	*morph*
SYN	*syn*
SEMSEL	*semsel*

word

G-ID	*set*(*integer*)
ORIGIN	*string*

semsel

SEM	*sem*
C-ID	*string*
DOMAIN	*domain*
SELECT	*list*(*select-element*)
COMPAT-R	*set*(*rel*)

sem

ENTITY	*entity*
NET	*net*
LAY	*lay*
MOLEC	*boolean*

select-element

REL	*set*(*rel*)
OBLIG	*boolean*
SEL	*sign*

Abbildung 1: Ausschnitt der in HaGenLex verwendeten Merkmalsdeklarationen.

Gegenwärtig sind Merkmal-Wert-Strukturen im Rahmen von HaGenLex als Scheme-Strukturen implementiert, auf die sowohl die Lexikonwerkzeuge als auch der Parser über gemeinsame Schnittstellen zugreifen.

Anmerkungen

1 Für eine detaillierte Darstellung sei der Leser auf HELBIG 2001 verwiesen.

2 Man beachte, dass hier nicht nur generische Konzepte gemeint sind, sondern dass etwa auch *der diesjährige GermaNet-Workshop in Tübingen* als ein Konzept aufgefasst wird, das einem Knoten in der zugehörigen semantischen Repräsentation entspricht.

3 Eine Beschreibung des dabei verwendeten Parsers gibt HARTRUMPF 2003, Kap. 3. Eine Anwendung zur natürlichsprachlichen Informationsrecherche wird in LEVELING & HELBIG 2002 vorgestellt.

4 Alle Verweise beziehen sich auf GermaNet 4.0.

5 Vgl. etwa KUNZE & WAGNER 2001.

6 Nach anfänglichen Versuchen einer vollautomatischen Lesartenzuordnung hat sich letztlich die Software-unterstützte Zuordnung durch den Lexikographen als effektivste und verlässlichste Methode erwiesen.

7 Als Nebenprodukt einer solchen, an übergeordneten semantischen Gesichtspunkten orientierten Sichtung der Struktur von GermaNet ergeben sich wiederum Hinweise auf Schwachstellen derselben. Problematisch erscheint es beispielsweise, *aufwachen.1* und *einschlafen.3*, neben *schlummern.1* und *dösen.1*, als Hyponyme von *schlafen.2* anzusehen. Ferner sind *platzen.1* und *zerschellen.1* nicht nur Hyponyme von *Mat_Zustands_Veränderung.1*, sondern auch von {*zerstören.1, destruieren.1*}, was wiederum *Mat_Zustands_Veränderung.2* untergeordnet ist. Andererseits soll aber *Mat_Zustands_Veränderung.2*, im Gegensatz zu *Mat_Zustands_Veränderung.1*, gerade *kausative* Veränderungen der materiellen Beschaffenheit erfassen.

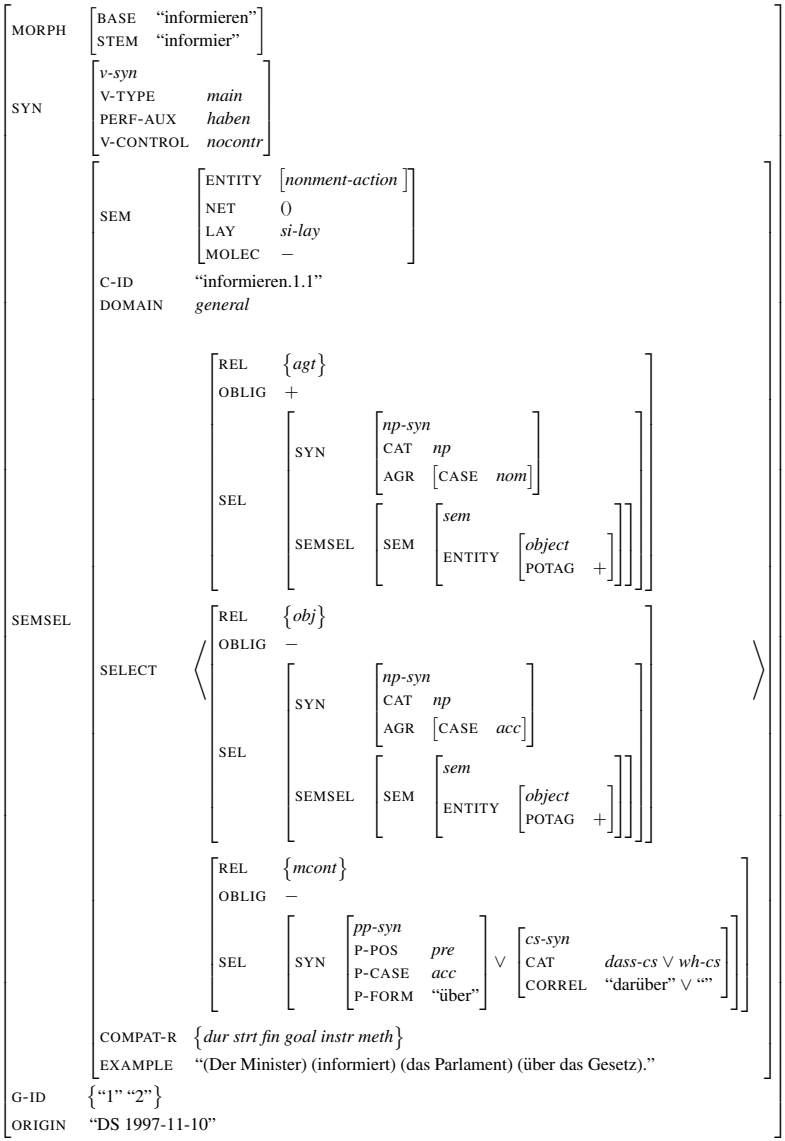

Abbildung 2: Merkmal-Wert-Struktur des HaGenLex-Eintrags für *informieren*.

[8] Ähnlich verhält es sich übrigens mit bestimmten Kognitionsverben wie *blicken*.

[9] Die unter *http: //www.sfs.nphil.uni-tuebingen.de/lsd/* zu findende Online-Dokumentation zu GermaNet führt eine derartige Klassifikation im Abschnitt "Future Work" auf – ebenso wie die Klassifikation nach Aktionsarten [Zugriff April 2004].

Literatur

Carpenter, B. (1992). The Logic of Typed Feature Structures. Cambridge Tracts in Theoretical Computer Science 32. Cambridge University Press.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hartrumpf, S.; Helbig, H.; Osswald, R. (2003). "The Semantically Based Computer Lexicon HaGenLex – Structure and Technological Environment." In: Traitement Automatique des Langues 44(2) (2003) [erscheint].

Hartrumpf, S. (2003). Hybrid Disambiguation in Natural Language Analysis. Osnabrück: Der Andere Verlag.

Helbig, H. (2001). Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet. Berlin et al.: Springer.

Kunze, C.; Wagner, A. (2001). „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche." In: Lemberg, I.; Schröder, B.; Storrer, A. (Hrsg.) (2001). Chancen und Perspektiven computergestützter Lexikographie. Tübingen: Niemeyer [= Lexicographica Series Maior Vol. 107], 229-246.

Leveling, J.; Helbig, H. (2002). "A Robust Natural Language Interface for Access to Bibliographic Databases." In: Callaos, N.; Margenstern, M.; Sanchez, B. (eds.) (2002). Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), volume XI. Orlando, FL: International Institute of Informatics and Systemics (IIIS), 133-138.

Kathrin Beck

Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa - Mapping Architecture for People's Associations

Abstract
Dieser Beitrag zeigt unser studentisches interuniversitäres Projekt[1] eines prototypischen Programms zur Konstruktion von „Wissensnetzen". Dabei soll der Benutzer Wörter und andere Daten aller Art verlinken können, um sein Wissen auf eine kognitiv adäquate Art und Weise zu repräsentieren.

Als Beispielapplikation zu diesem Projekt haben wir einen Vokabeltrainer konzipiert, der die Daten von GermaNet[2] (Kunze 2001) als Netzstruktur darstellt, mit dessen Hilfe Lerner deutsche Wörter verstehen und lernen können.

1 Einführung

1.1 Projektrahmen

Das vom Projekt MiLCA unterstützte Studienprojekt Mapa[3] (Mapping Architecture for People's Associations), fand an drei Universitäten statt.

Sieben Studenten des Studiengangs „Cognitive Science" an der Universität Osnabrück, eine Studentin der Linguistik und der Sprachlehrforschung an der Universität Bochum und zwei Studentinnen des Studienganges „Allgemeine Sprachwissenschaft und Nebenfächer" an der Universität Tübingen arbeiteten für dieses Projekt ein Jahr lang zusammen.

Im ersten Semester wurde in wöchentlichen Chats, im Whiteboard, in Telefonkonferenzen und per Email diskutiert und das Konzept des Gesamtprojekts und die Aufgabenverteilung spezifiziert. Die Aufgabe der Osnabrücker Studenten war dabei die Ausarbeitung des Frames von Mapa, die Tübinger und Bochumer Studentinnen erarbeiteten den Vokabeltrainer dazu. Im zweiten Semester folgte der Programmier-Teil des Projekts. Die Koordination und Kommunikation erfolgte weiterhin über Telefon-Konferenzen und per Email, außerdem gab es einen gemeinsamen Workshop und zwei Programmierwochen, in denen sich alle Projektteilnehmer ausführlicher absprechen konnten.

Das Master-Projekt der Osnabrücker Studenten begann im Oktober 2002 und wird am 15. Oktober 2003 enden.

1.2 Konzept von Mapa

Das Mapa-Konzept wurde von den Studierenden selbst entwickelt[4]. Ziel dieses Projekts war ein Programm, mit dem Nutzer ihre eigenen Wissensnetze erstellen und darin Daten aller Art einbetten können, z.B. Bilder, Videos, Dokumente oder Emails.

Darüber hinaus soll es durch Kollaboration möglich sein, über das Internet Mapa-Netze von anderen Benutzern zu empfangen und das eigene Netz um deren Informationen zu erweitern. Gibt es in dem empfangenen Netz einen Knoten, über der Benutzer schon verfügt, wird dieser nicht neu erstellt. Seine Relationen werden dem schon existierenden Knoten hinzugefügt und so ins Netz aufgenommen.

1.2.1 Funktionalität

Mapa stellt ein Werkzeug zum „kognieren" (überlegen, phantasieren, assoziieren, denken …) dar, das möglichst viele verschiedene Ansprüche von Benutzern erfüllen soll. Deshalb können auch unterschiedliche Informationstypen (Text, Bilder, Videos, Emails…) integriert werden.
Das Tool soll Benutzern ermöglichen, ihr Wissen so zu strukturieren, dass sie ihre Informationen am intuitivsten und für sie sinnvollsten

anordnen können. Ziel ist es, Wissen so anzuordnen, wie es möglicherweise auch im Gehirn angeordnet ist; das Programm versucht, die Arbeitsweise des Gehirns kognitiv adäquat wiederzugeben (Vester 2002).

Der Benutzer soll möglichst viel Freiheit bei der Gestaltung haben. Statt Restriktionen, die die Möglichkeiten einengen, sollen Konventionen für Verständlichkeit sorgen. Wir gehen davon aus, dass der Benutzer nur sinnvolle Einträge macht, die er im Nachhinein rekonstruieren kann.

Das Netzwerk, das man mit Mapa konstruieren kann, hat nicht den Anspruch, Wissen zu repräsentieren sondern Stichworte für Wissen. Man muss also nicht das gesamte Wissen ins Netz schreiben. Es genügt, ein Wort oder einen Satz zu schreiben, durch den das Wissen zum Thema aktiviert wird.

Durch weltweite Vernetzung der Anwender und deren Kollaboration kann man das Wissen anderer Benutzer verwenden und sein eigenes Netz weiterentwickeln.

Jedes Jahr wird 1 Milliarde Gigabyte neuer Daten generiert. Durch die Zunahme verfügbaren Wissens wird es immer wichtiger, dies zusammenzufassen und zu strukturieren. Darüber hinaus müssen Möglichkeiten gefunden werden, das Wissen möglichst einfach zu erwerben und dann zu behalten (Tergan 2003).

Als wesentliche Bausteine kognitiver Kompetenz gelten:

- Wissensmanagement
- möglichst schnelle Übersicht über viele Daten
- möglichst schnelle Analyse der relevanten Daten
- Verständnis komplexer Sachverhalte
- möglichst langes Behalten der Informationen

1.2.2 Visualisierung

Einer der Vorteile von Mapa ist, dass durch die graphische Zusammenfassung jedes Konzept nur einmal eingeführt wird. Anders als bei Texten, wo jedes Konzept in jeder Aussage genannt werden muss, reicht es durch die Relationspfeile bei einer Graphik, jedes Konzept nur einmal aufzuschreiben.

Im Gegensatz zu Texten werden Zusammenhänge graphisch deutlicher. Bei mehreren Aussagen über mehrere Konzepte können diese verstreut im Gesamttext sein. Durch die Relationspfeile in einer Graphik dagegen sieht man stets auf einen Blick alle Relationen eines Konzepts.

Es ist auch wissenschaftlich nachgewiesen, dass graphische Darstellungen reichhaltigere Gedächtnisspuren hinterlassen als textuelle (Paivio 1971). Man kann Informationen also besser behalten, wenn man sie in einem Netz darstellt, als wenn man sie nur als Text aufschreibt.

Modell und Ansicht des Modells sind voneinander unabhängig und getrennt. Das bedeutet, man kann je nach momentaner Präferenz sein abstraktes Netz auf verschiedene Arten visuell dargestellt betrachten, als hyperbolischen Graphen oder als Touchgraphen[5], siehe Abb. 1,2:
Wichtig für optimales Lernen und eine gute Wissensorganisation ist es, Inhalte auf eine Weise darzustellen, die der mentalen Repräsentation dieser Inhalte entgegenkommt. Das muss notwendigerweise graphisch sein, da das Gehirn Bilder und Zusammenhänge abspeichert, keine „Texte“.

Die Vorteile einer graphischen Darstellung (Map) sind:

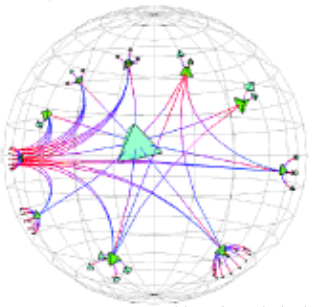

Abb. 1: hyperbolischer Graph

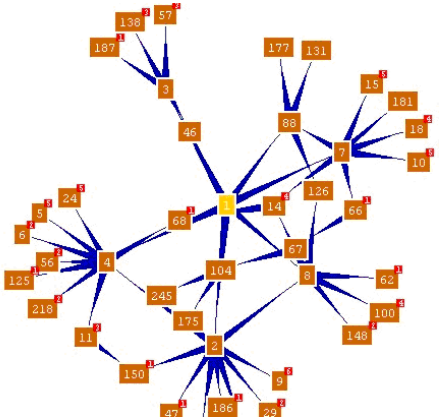

Abbildung 2: Touchgraph

- Man kann automatisch den Kontext einer Informationseinheit sehen, da er in ihrem Umfeld dargestellt ist.
- Man hat Freiheiten in der Reihenfolge der Exploration der Informationen einer Map.
- Mapping erfordert Verringerung der Komplexität, der Benutzer sieht automatisch die wichtigsten Punkte des Themas komprimiert.
- Der Benutzer bekommt sofort einen schnellen Überblick über den ihm unbekannten Stoff.

Das von uns entwickelte Lern- und Organisationstool besitzt folgende visuelle Features:

- Visualisierung einzelner Knoten in der Visualisierung des Netzes (Bilder, Videos etc, die ein Konzept verdeutlichen)
- Typisierung durch Farben/ Formen
- Zooming (man kann sich einen ausgesuchten Teil des Netzes vergrößert darstellen lassen)
- Chunking (man kann sich nur ausgewählte Knoten anzeigen lassen).

Große Vorteile dieses Konzepts gegenüber Mindmaps sind seine Dreidimensionalität, seine unhierarchische Oberflächenstruktur und die beliebig vielen Relationen, die ein Knoten haben kann. So können auch große Mengen an Knoten (und ihren konzeptuellen Strukturen) anschaulich dargestellt werden.

1.2.3 Benutzbarkeit

Wir haben versucht, allgemeine Kriterien für die Benutzbarkeit von Software zu berücksichtigen (ISO 9241[6]):

- einfacher Aufbau und Transparenz, um die Anwendung des Programms zu erleichtern
- Skalierbarkeit in möglichst vielen Dimensionen
- leichte Erweiterbarkeit des Programms für persönliche Veränderungen
- Konventionen-Bildung und minimale Restriktivität für möglichst vielseitige Benutzbarkeit
- Unterstützung verteilter Datenhaltung (Kollaboration) für Kommunikation zwischen Benutzern, ohne einen Administrator und einen Server zu benötigen (Peer to Peer)

2 Der Vokabeltrainer

2.1 Konzept

Unser Vokabeltrainer basiert auf dem Konzept von Mapa und ist eine Beispielanwendung mit lexikalischen Daten. Er verwendet also die Ansatzmöglichkeiten und die Netzwerkstruktur von Mapa.

Die zugrunde liegende Idee ist, dass linguistisches Wissen, das ja semantisches Wissen beinhaltet, als Netzwerk von verknüpften Konzepten repräsentiert ist.

In das Netzwerk von Mapa werden Daten von GermaNet, das diese Informationen implementiert hat, eingebaut. Knoten sind Lemmata, Verbindungen sind semantische und morphologische Relationen wie Synonymie, Hyperonymie, Derivationen etc. Ein lexikalischer Eintrag wird als Knoten mit dem ihn umgebenden Feld dargestellt, siehe Abbildung 3.

Wir verwenden die XML-Version von GermaNet, in dem deutsche Wörter als Synsets[7] (Fellbaum 1998) kodiert und die Relationen zwischen den Wörtern und Konzepten gespei-

chert sind. Mapa extrahiert diese Daten und speist sie in sein Netz-Schema ein. Sie sind die Basis des Vokabeltrainers.
Aus GermaNet importierte Relationen sind:

- Semantische Relationen: Hypo-/ Hyperonymie, Synonymie, Antonymie, Meronymie, Holonymie, Cause und Assoziation
- Lexikalische Relationen, z.B. Derivationen

Die Idee des Vokabeltrainers ist, dass in natürlicher Sprache Wörter, also Konzepte in semantischem Kontext auftreten. Aus der kognitiven Psychologie ist bekannt, dass im Gehirn Bedeutungen netzwerkartig abgespeichert werden. Und aus der kognitiven Informationsverarbeitung wissen wir, dass semantische Verarbeitung die Erinnerungsleistung verbessern kann (Vester 2002). Folglich müsste sich auch der Fremdsprachenerwerb verbessern, wenn Vokabeln im Kontext gelernt werden. Wir versuchen also, dem Vokabeltrainer ein möglichst kognitiv adäquates Modell der Sprache zugrunde zu legen.

Die Zielgruppe dieses Vokabeltrainers sind Lerner mit einem gewissen Vorwissen an Wortschatz und Grammatik, die ihr Vokabelwissen vertiefen möchten. Ein Vorteil dieses Vokabeltrainers ist, dass er einsprachig deutsch ist. Weder zum Verständnis der Wörter noch zur Abfrage sind Übersetzungen in eine andere Sprache notwendig, sodass die Anwender den Vokabeltrainer ohne Vorkenntnisse anderer Fremdsprachen benutzen können.

Da bei dieser Art von Vokabeltrainer die Bedeutung eines Wortes vorerst nur aus dem Kontext seiner Nachbarwörter und der Relationen zwischen ihnen erschlossen werden kann, muss der Lerner notwendigerweise schon Wörter kennen, mit deren Hilfe er die Bedeutung von Nachbarwörtern verstehen kann. Außerdem hat der Vokabeltrainer keine Komponenten, mit denen man Grammatik und Satzbau trainieren könnte, da wir uns auf den Netzwerkaspekt des Wortschatzes konzentriert haben.

2.2 Surf-Modus

Um sich im Netz zu bewegen, wählt der Lerner ein für ihn interessantes Einstiegswort aus, das ihm mit seinen Nachbarwörtern präsentiert wird. Von dort kann er eine Sequenz von Knoten, also Wörtern im Graphen, besichtigen und sich von Nachbarwort zu Nachbarwort klicken. Die Pfade, die ihm zur Verfügung stehen, sind die semantischen Relationen, selbstverständlich kann er aber auch zu einem ganz anderen Wort springen.
Diese Methode der Exploration bietet zwei Vorteile: Zum einen wird eine Kohärenz in der Abfolge der besuchten Wörter sichergestellt, da der Lerner sich ja die Wörter zu einem bestimmten Thema anschaut, das ihn gerade interessiert und kein Wort unabhängig von seinem Nachbarwort ist. Und zum anderen prägt sich der Lerner die Wörter schon ein, während er das Netz exploriert, weil er jedes Wort semantisch verarbeiten muss, um seine Bedeutung zu verstehen.

2.2.1 Auswahl von zu lernenden Vokabeln

Während der Lerner durch das Netz „surft“, kann er durch einfachen Mausklick Wörter, die

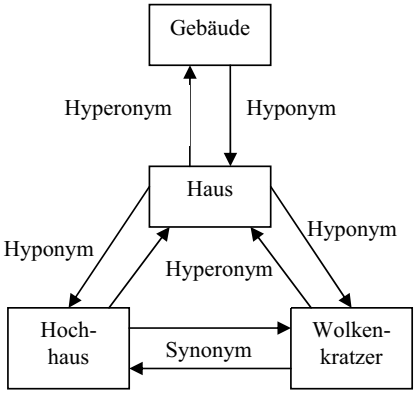

Abbildung 3: Knoten „Haus“ mit Nachbarknoten

er lernen möchte, in seine Vokabelliste aufnehmen. Dabei hat er die Wahl zwischen zwei Arten der Vokabelauswahl.

Er kann jedes Wort einzeln auswählen, um eine kohärente Lernliste zu bekommen und nur explizit gewünschte Wörter in seine Lernliste aufzunehmen.

Er kann aber auch automatisch alle Nachbarwörter eines bestimmten Wortes mit zum Lernen auswählen, die 1, 2,… Relationen von ihm entfernt sind. Die vom Anwender auszuwählende „Suchtiefe" n bestimmt, dass alle Wörter, die maximal n Relationen vom zentralen Wort entfernt sind, ausgewählt werden.

Im besten Fall erreicht er damit nur semantisch verwandte Wörter und kann schnell eine große Menge an Wörtern, ähnlich Wortfeldern, auswählen, z.B. zu einem bestimmten Stichwort wie „Bahnhof", „essen" etc. Es besteht allerdings schon bei relativ geringer Suchtiefe die Gefahr, auch semantisch unabhängige Wörter mit auszuwählen, z.B. Student → Vorlesung → Tafel → Kreide → weiß (Suchtiefe 4).

2.2.2 Individuelle Organisation

Lerner können eigene Lektionen anlegen, in die sie Vokabeln abspeichern und die sie unabhängig von anderen Lektionen lernen können.
Außerdem wird der Pfad, den ein Lerner beim Surfen benutzt hat, abgespeichert und kann von ihm wieder angeschaut werden.

2.2.3 Erweiterung des Netzes

Wie vollständig das semantische Netz ist, hängt von den Kenntnissen und Interessen des Lerners ab. Er kann neue Wörter hinzufügen, z.B. Fachwörter, umgangssprachliche Wörter etc.

Manchmal können auch persönliche Lernhilfen, Eselsbrücken nützlich sein.

Des Weiteren gibt es die Möglichkeit, Meta-Informationen zu einem Wort, z.B. Beispielsätze, mit ihm zu verknüpfen.

2.3 Vokabel-Abfrage

Dieser Vokabeltrainer-Prototyp bietet drei Abfragearten an, die vollständig auf GermaNet und seiner Netz-Idee basieren: Zentrales Wort einfügen, Relationen einfügen und Nachbarwörter einfügen.

2.3.1 Zentrales Wort einfügen

Bei dieser Übungsart bekommt der Lerner alle Nachbarwörter eines gesuchten Wortes und ihre Relationen gezeigt. Wird beispielsweise „Haus" gesucht, sieht er „Hyperonym: Gebäude, Hyponym: „Hochhaus", Hyponym: Wolkenkratzer", siehe Abbildung 4.

Diese Übung sollte die einfachste der drei Übungen sein. Der Lerner erhält hier viel Kontextinformation, er muss nicht alle Nachbarwörter verstehen. Manchmal kann er das gesuchte Wort sogar aus einem Nachbarwort ableiten, z.B. „Haus" aus „Hochhaus".

Ein Problem bei der Konzeption war hier, eine geeignete Abfragemethode zu finden für Wörter, die mit mehreren anderen Konzepten durch dieselbe Relation verbunden sind. Bei Abbildung 4 beispielsweise enthält jedes der beiden unteren Kästchen eine mögliche Antwort auf die Frage nach dem Hyponym von „Haus".
Wir haben das Problem gelöst, indem das Programm jedes Wort akzeptiert, das durch diese Relation mit dem gegebenen Wort verbunden ist, aber so lange nach einer anderen Lösung gefragt wird, bis der Lerner die momentan gesuchte Vokabel eingibt.

2.3.2 Relationen einfügen

Bei dieser Übungsart bekommt der Benutzer ein zentrales Wort mit all seinen Nachbarwörtern und muss die Relationsarten zwischen ihnen bestimmen. Diese Übungsart ist leichter als die nachfolgende Übung, da man alle Nachbarwörter zur Verfügung hat und aus bekannten Relationsarten andere erschließen kann, siehe Abbildung 5.

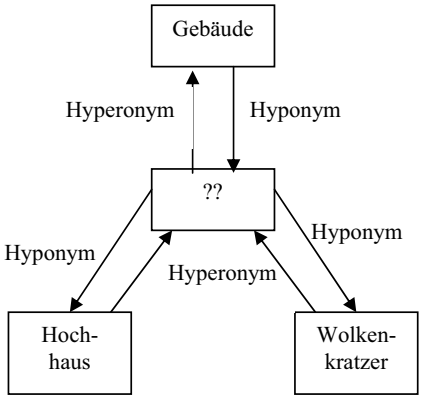

Abb. 4: Übungsmodus mit fehlendem zentralen Knoten

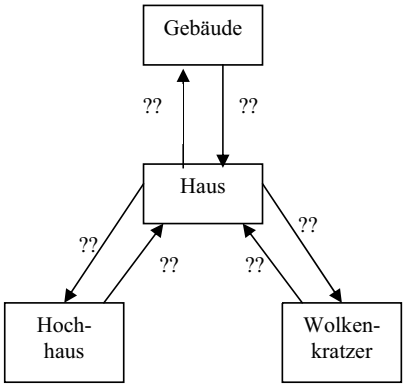

Abb. 5: Übungsmodus mit fehlenden Relationen

2.3.3 Nachbarwörter einfügen

Hier sieht der Lerner das zentrale Wort mit seinen Relationen, siehe Abbildung 6. Er muss nun zu jeder Relation das passende Nachbarwort eingeben. Gibt es mehrere Nachbarwörter, die über die gleiche Relationsart mit dem zentralen Wort verbunden sind (z.B. Synonym von „teuer“ = „wertvoll“, Synonym von „teuer“ = „kostspielig“), so spielt die Reihenfolge der Eingabe natürlich keine Rolle.

Genau wie bei der Übungsart mit fehlenden Relationen gibt es hier das Problem der Definition, wann der Lerner eine Vokabel gewusst hat und wann nicht. Mangels psychologischer Kenntnisse zu diesem Thema haben wir entschieden, vorerst zu definieren, dass der Lerner ein Wort gewusst hat, wenn er

a) bei einem einzigen gesuchten Nachbarwort dieses und
b) bei mehreren gesuchten Nachbarwörtern mindestens zwei von ihnen richtig benennen kann.

2.3.4 Ablauf der Vokabel-Abfrage

Für jedes gesuchte Wort bekommt der Lerner Informationen zu seinem Verständnis dargestellt. Er nennt daraufhin seinen Lösungsvorschlag. Ist seine Antwort falsch, bekommt er noch einen zweiten Versuch, um sich selbst zu korrigieren und Alternativen auszuprobieren. Findet er beim zweiten Versuch die korrekte Lösung, wird das trotzdem als falsch abgespeichert, da er das Wort auf Anhieb wissen soll.

Hat er nur Groß- und Kleinschreibung nicht beachtet, wird er darauf hingewiesen.

Wurde der Lerner alle momentan zu lernenden Vokabeln einmal abgefragt, so werden sie neu sortiert. Das heißt, Wörter, die jetzt als gelernt gelten, werden aussortiert, die verbleibenden werden, je nachdem, ob der Lerner sie gewusst hat oder nicht, in eine neue Reihenfolge nach Priorität der Abfrage sortiert, und der Lerner kann sich weiter abfragen lassen, so lange, bis er das Programm abbricht oder alle Wörter gelernt hat.

Am Ende der Sitzung bekommt er dann ein Feedback über seine Leistung:

- Anzahl der abgefragten Wörter
- Anteil der Wörter, die er bei dieser Sitzung richtig beantwortet hat

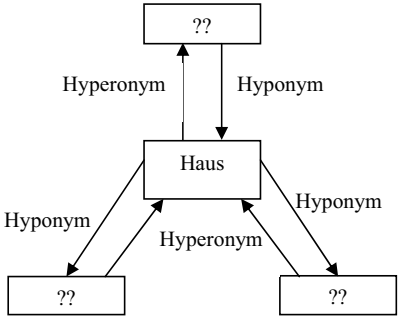

Abb.6 Übungsmodus mit fehlenden Nachbarknoten

- Anteil der Wörter, die er so oft in Folge richtig beantwortet hat, dass sie als gelernt gelten und nicht mehr abgefragt werden
- Gesamtanzahl der Wörter in der momentan abgefragten Lektion
- Anteil der schon vollständig gelernten Wörter dieser Lektion

So hat der Lerner einerseits ein Feedback über seine Leistung der letzten Abfrage und sieht, ob sein Lernen ausreicht, oder ob er mehr lernen sollte.
Außerdem bekommt er auf Wunsch ein Feedback in Bezug auf die gelernte Lektion. Er kann abschätzen, wie viel Arbeit wohl noch nötig sein wird, bis er die Lektion vollständig gelernt haben wird.

Wir glauben, dass die Selbsteinschätzung, die dem Lerner durch dieses Feedback ermöglicht wird, seine Motivation, Vokabeln zu lernen, stark fördert im Vergleich zu Vokabeltrainern ohne Feedback

2.3.5 Problem der nicht verbundenen Knoten

Bei der Vokabelabfrage nur auf der Grundlage von GermaNet können mit dem derzeitigen Programm Wörter, die der Lerner selbst eingefügt hat, und die keine Verbindung zu Nachbarknoten haben, nicht abgefragt werden. Man könnte hier beispielsweise anbieten, mit Hilfe eines Übersetzungsprogramms die englische oder französische Übersetzung des Wortes zu suchen und mit dem deutschen Begriff zu verknüpfen.

2.4 Benutzer-Profil

Für jeden Benutzer des Vokabeltrainers wird ein persönliches Profil angelegt.

Es speichert Informationen über die Abfragesitzungen, das Lernverhalten des Benutzers und über alle besuchten Wörter, also über alle Wörter, über die der Lerner schon einmal „gesurft“ ist.

Da der Vokabeltrainer auf dem Karteikastenprinzip beruht, orientiert sich die Benutzermodellierung an diesem Prinzip.

Ein solcher Vokabel-Karteikasten besteht aus 6 Fächern (Leitner 1977); dabei kommen noch nie gewusste Wörter ins 1. „Fach“, einmal gewusste ins 2. „Fach“ usw. bis zum letzten „Fach“. Diese Wörter gelten als gelernt und werden nicht mehr abgefragt. Wird ein Wort einmal nicht gewusst, kommt es unabhängig von dem „Fach“, in dem es vorher war, wieder ins 1. „Fach“ zurück.

Ein Wort, das man erst selten richtig beantwortet hat, vergisst man schneller als ein Wort, das man schon als richtige Lösung eingegeben hat, und es muss früher wieder abgefragt werden. So wird eine Abfragereihenfolge erstellt nach der Dringlichkeit der erneuten Abfrage. Diese Methode bietet den Vorteil, dass Vokabeln so spät wie möglich abgefragt werden können, kurz bevor der Lerner sie evtl. wieder vergessen würde. Wir erwarten, dass sich dadurch die Anzahl der momentan abzufragenden Vokabeln und die Abfragezeit verringern, vorausgesetzt, man verwendet den Vokabeltrainer in genügend kurzen Abständen. So kann sich der Lerner auf die in dem Moment relevanten Vokabeln konzentrieren.

Hat der Lerner schon alle Wörter einer Lektion gelernt, wird ihm das gemeldet. Er kann sie aber noch einmal lernen, wenn er beispielsweise sein Wissen wieder auffrischen will. Dann kom-

men alle Wörter zurück ins 1. „Fach“ des „Karteikastens“.

2.4.1 Vokabel-Profil

Im Benutzerprofil werden folgende Daten zu jeder Vokabel gespeichert:

- „Fach“ des „Karteikastens“, in dem es sich befindet
- Datum der letzten Abfrage
- Ob es bei der letzten Abfrage gewusst wurde oder nicht

Aus dem Datum der letzten Abfrage eines Wortes und seinem „Karteikasten-Fach“ wird berechnet, wann das Wort zum nächsten Mal abgefragt werden muss.

2.4.2 Surf-Profil

Das Benutzer-Profil speichert außerdem:

- alle besuchten Pfade (Wörter in der besuchten Reihenfolge mit Datum
- Anzahl der besuchten Wortarten
- Anzahl der besuchten Relationsarten

Dadurch kann einerseits der Lerner alle abgespeicherten Pfade noch einmal anschauen, um seine Erinnerungen an den Wortschatz zu vertiefen oder z.B. ein Wort wieder zu finden. Diese Daten liefern aber auch Aufschluss über das Verhalten und die Kenntnisse des Benutzers. Man kann daraus eine Statistik der bevorzugten Wort- und Relationsarten ableiten, und evtl. auch Probleme erkennen. Wenn ein Lerner beispielsweise gewisse Relationsarten kaum benutzt, könnte das bedeuten, dass er ihre Bedeutung nicht verstanden hat.

2.4.3 Vokabellern-Statistk

Die dritte Art gespeicherter Daten ist für das Feedback des Lerners gedacht und enthält Informationen über sein Abfrage- und Lernverhalten:

- Anzahl der gelernten Wörter
- Anzahl der noch nicht gelernten, aber schon mehrmals gewussten Wörter
- Anzahl der gar nicht gewussten Wörter
- Anzahl der bei der letzten Abfrage gewussten Wörter
- Anzahl der Wörter, die in den nächsten Tagen wiederholt werden müssen

Diese Daten sind als Feedback für den Lerner gedacht, wie oben erklärt.

Die letzte Information ist besonders nützlich, weil sie den Lerner rechtzeitig daran erinnert, Vokabeln zu lernen, bevor er die schon halb gelernten Wörter wieder vergessen hat und er zu viele Wörter wiederholen muss.

3 Ausblick

Wir haben viele Erweiterungsvorschläge für dieses Projekt, jedoch weder die Zeit noch die Software, finanziellen Mittel oder andere Ressourcen, sie hatten im Rahmen unseres Studentenprojekts in die Realität umzusetzen.

3.1 Fehleranalyse und Feedback

Ein wichtiger Vorschlag wäre, statt nur zwischen richtiger und falscher Lösung zu unterscheiden, eine detaillierte Fehleranalyse mit Unterscheidung zwischen Rechtschreibfehlern und semantisch falschen Wörtern. Sie sollte einerseits Verfahren des approximativen String Matching (Baeza-Yates 1998) berücksichtigen können – um die Eingabe mit den Vokabeln des Netzes zu vergleichen-, andererseits auch das Fehlerverhalten beim Fremdsprachenlernen modellieren können (Chanier 1992).

Zusätzlich könnte man ein angemessenes informatives und motivierendes Feedback geben, z.B. „Du hast dich vertippt.“ oder „Du hast das Antonym der gesuchten Lösung eingegeben.“

3.2 Vorgefertigte Lektionen

Man könnte auch relevante Teile des Netzes als Lektion abspeichern und dem Lerner anbieten, z.B. Wörter zu einem bestimmten Thema, oder alle Nachbarwörter einer bestimmten Tiefe zu einem zentralen Wort.

3.3 Kollokationen

Genauso wichtig für den korrekten Sprachgebrauch sind Kollokationen (Lewis 2000), denn ohne ihre Kenntnis macht der Deutsch-Lerner im besten Fall Fehler, im schlimmsten Fall wird er aber missverstanden. Möchte beispielsweise ein Engländer „eine Rede machen", so will er sie wahrscheinlich halten. Der Deutsche versteht spontan vielleicht aber, dass er eine Rede vorbereiten, also „machen" will.

Die Problematik bei Kollokationen ist aber einerseits, dass es keine genaue Grenze zwischen ihnen und offenen syntagmatischen Relationen gibt, und andererseits, dass es sehr viele Kollokationen gibt. Da sie in GermaNet nicht mit aufgenommen sind, müssten sie alle einzeln von Hand eingetragen werden, was einen immensen Aufwand darstellen würde.

3.4 Mündliche Abfrage

Man könnte zusätzlich zur schriftlichen eine mündliche Abfrage mit den gleichen Übungsarten anbieten. Diese Abfrageart wäre besonders geeignet, wenn der Lerner in kurzer Zeit seine Vokabeln auffrischen will. Er müsste seine Antwort nur laut aussprechen und dann selbst mit der richtigen Lösung vergleichen. Hier wäre es sinnvoll, ein Hörbeispiel oder zumindest die phonetische Lautschrift anzuzeigen, sodass der Lerner auch die korrekte Aussprache lernen kann. Optimal wäre natürlich eine automatische Spracherkennung, die die Antwort des Lerners analysiert. Solange diese Technologie aber nicht wirklich funktioniert, sollte die Selbstkorrektur des Lerners nicht in sein Benutzerprofil eingehen, da ihre Korrektheit nicht vom Programm überprüft werden kann.

3.5 Detaillierte Abfrage

Mit dem momentanen Programm kann man die in GermaNet kodierten, bedeutungstragenden Wortarten wie Nomina, Verben und Adjektive lernen. Das bedeutet, man lernt vielleicht in kurzer Zeit, sich mehr recht als schlecht verständlich zu machen, aber man lernt nicht, grammatikalisch korrekt zu formulieren. Für Lerner, die daran interessiert sind, muss ein zusätzlicher Modus angeboten werden, in dem sie grammatische Details der Wörter lernen. Das wären beispielsweise bei Nomina Artikel und Deklination, bei Adjektiven Flexion und bei Verben die Konjugation.

3.6 Einbindung von anderen Ressourcen

Außer der Netz-Struktur von GermaNet gäbe es noch viele andere Arten, die Bedeutung eines Wortes zu zeigen, z.B. über Bilder und Videos. So könnte man ein Nomen durch (ein) Bild(er) und ein Adjektiv durch Bilder der Gegenteile (z.B. eine große Frau vs. eine kleine Frau) darstellen. Verben könnte man mit einem Video zum Vorgang erklären.

Wort-Definitionen, wie man sie in einsprachigen Wörterbüchern findet, stellen eine andere, schwerer verständliche Art dar, ein Wort zu erklären. Der Lerner könnte dabei aber andere Wörter, mit denen es häufig auftritt oder Kontexte, in denen es verwendet wird, lernen und zur Sprachproduktion verwenden.

Zusätzlich wäre ein Übersetzungsprogramm hilfreich, um auch abstrakte Wörter zu verstehen.

3.7 Automatische Auswahl des Übungsprogramms

Mit all den oben vorgeschlagenen Erweiterungen könnte man die Vokabelabfrage so erweitern, dass das Programm je nach Lerngrad einer

Vokabel im Karteikastensystem einen anderen Übungsmodus auswählt.

Das hätte zum einen den Vorteil, dass durch verschiedene Hinweise auf das gesuchte Wort (lexikalisch durch verwandte Wörter oder semantisch durch Bilder, Videos) beim Lerner verschiedene Assoziationen geweckt werden und sich das Wort besser einprägen kann. Zum anderen gäbe es aber sicher einen großen pädagogischen Vorteil. Durch eine abwechslungsreich variierte Abfrage würden die Aufmerksamkeit, die Motivation und das Interesse am Vokabellernen vermutlich signifikant erhöht.

Anmerkungen

[1] Das diesem Projekt zugrunde liegende Projekt MiLCA (Medienintensive Lehrinhalte in der Computerlinguistik-Ausbildung) wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 NM 167 A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin.

[2] Die in Tübingen erstellte deutsche Version von WordNet: *http://www.sfs.uni-tuebingen.de/lsd/* [Zugriff April 2004].

[3] Mapa (spanisch: Landkarte): strukturierte Visualisierung von Wissensinhalten. Universität Osnabrück, Institut für Kognitionswissenschaft, Studentisches Projekt Mapa, *http://www.cogsci.uni-osnabrueck.de/~mapa, mapa@cogsci.uni-osnabrueck.de* [Zugriff April 2004].

[4] Die Tübinger Dozentin Dr. Karin Krüger-Thielmann und die Osnabrücker Dozenten Dr. Petra Ludewig, Dr. Claus Rollinger und Dr. Veit Reuer betreuten das Projekt (Reuer 2004).

[5] TouchGraph LLC (2004). TouchGraph Website. http://*www.touchgraph.com* [accessed April 2004].

[6] International Organization for Standardization (ISO) (2004). ISO Online Website, *http://www.iso.ch* [accessed April 2004].

[7] Ein Synset ist ein aus synonymen Wörtern bestehender Knoten, der die Bedeutung dieser Wörter beschreibt.

Literatur

Baeza-Yates, R., Navarro, G. (1998). Fast Approximate String Matching in a Dictionary. University of Chile, Santiago de Chile, Dept. of Computer Science, *http://citeseer.nj.nec.com/1593.html* [Zugriff April 2004].

Chanier, T.; Pengelly, M.; Twidale, M.; Self, J. (1992). "Conceptual Modeling in Error Analysis in Computer-Assisted Language learning Systems." In: Swartz, M. L.; Yazdani, M. (eds.) (1992). Intelligent Tutoring Systems for Foreign Language Learning. Berlin et al.: Springer.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Kunze, C. (2001). „Lexikalisch-semantische Wortnetze." In: Carstensen, K.-U. et al. (eds.) (2001). Computerlinguistik und Sprachtechnologie: eine Einführung. Heidelberg; Berlin: Spektrum, Akademischer Verlag, 386-393.

Leitner, S. (1977). So lernt man richtig. Freiburg: Herder KG.

Lewis, M. (2000). Teaching Collocation: Further Developments in the Lexical Approach. Hove, England: LTP.

Paivio, A. (1971). Imagery and Verbal Processes. New York: Holt, Rinehart & Fisher.

Reuer, V. et al. (2003). „Studienprojekte in den Bereichen Computerlinguistik und Cognitive Science." In: Sprache und Datenverarbeitung Heft 27(1/2) (2003), 185-202.

Tergan, S.-O. (2003). „Lernen und Wissensmanagement mit Hypermedien." In: Unterrichtswissenschaft 31 (4), 334-358.

Vester, F. (2002). Die Kunst, vernetzt zu denken. Ideen und Werkzeuge für einen neuen Umgang mit Komplexität. Der neue Bericht an den Club of Rome. München: DTV.

Daniela Alina Plewe, Manfred Stede, Steffen Meschkat

GeneralNews – An Interactive Metabrowser

Abstract

GeneralNews (*http://www.generalnews.de* [1]) is a meta-browser that substitutes in real-time the words on websites. The substitutions may be synonymous to, more abstract, or more specific than the original expressions. By these variations, new descriptions of the world and descriptions of new worlds are created.

1 GeneralNews: The Idea

The system takes an existing web page - in principle, an arbitrary one, but the idea works best with news - and replaces words in real-time, as desired by the user. Specifically, the user can change the level of abstraction with a graphical slider (see Figure 1) and thereby investigate the emerging 'space of possibilites' around a text.

- Synonyms preserve the original meaning - more or less! - and indicate the variety of linguistic expressions.
- Abstractions (hypernyms) lead to generalizations and possibly open a bigger picture. When taken to the extreme, though, they trivialize the content.
- Specifications (hyponyms) create similar but alternative worlds to the original text.

The system relies on an electronic lexicon (WordNet from Princeton University [2]), which contains the various kinds of related words, such as synonyms, hyper- and hyponyms. The systems displays these entries when the user activates the detail mode. The system then analyzes the average level of abstraction for each original text and indicates the value on the the scale.

A German version of the system, using GermaNet [3], is currently in preparation. Furthermore, we are gradually improving the coverage of the system by adding linguistic analysis, moving from PoS-tagging and a simple morphology module to a more elaborate one that can deal more properly with verb inflections.

2 Background: Reflection of Language

GeneralNews reflects language. When we speak or write, the manifoldness of the world is reduced to the linearity of language. The mechanism of GeneralNews enriches this linearity again with the semantic variety of possible other descriptions. These variations are displayed as animated text on the screen.

3 Text-mediated Reality

When applied to news-sites, GeneralNews alters the text-mediated reality, thereby roughly keeping the same meaning. With extreme abstractions, however, the content is trivialized, since only abstract objects relate to each other. On a slightly lower level of abstraction, the general structures of the content can result. In the other, specific extreme, we get various alternative stories about the original incident. Concerning the selection of information and its strategic presentation, GeneralNews is exposed to the mechanisms of mass media communication.

As a special application called ArtAbstracts, working exclusively on the website of the ZKM (Center for Media and Arts, Karlsruhe), the system changes the interpretations of art-pieces (meaning-preserving substitutions and abstractions) and refers to alternative 'artworks' (specifications).

4 Abstraction – a Powerful Invention

Thinking in alternatives is important for progress since we develop new ideas by changing existing descriptions. Here, abstraction as a thought pattern plays a major role, even though it is often used in the pejorative sense: 'too abstract'. Yet, a plead for abstraction would include the following topics:

Language and the structure of knowledge According to many cognitive theories, we group our concepts in taxonomies. With these hierarchies, properties from the generic terms are inherited to the specific terms. This also allows us to derive conclusions about the unknown. With the aid of abstractions we organize our knowledge and create a big picture.

Science According to our notion of science, we need the understanding of general laws and structure, in order to create good, i.e. probable, prophecies. Most science is the attempt to formulate theses laws - and thereby abstraction captures the variety of single cases and exceptions. Thus abstraction helps to organize the relevant information within the abundance of data.

Creative and artistic processes Our actions may be described as moving up and down between abstract, diffuse and intuitive goals and concrete actions following these intentions.

The ability to abstract can only partially be automated with the contemporary means of computational linguistics and artificial intelligence. GeneralNews is the computational attempt to generalize the incoming information purely on the basis of lexical knowledge. Empirical, causal or even strategic insights can not be implemented by that method. Yet, as a meta browser GeneralNews aims at triggering our fantasies - via abstraction.

5 Implementation

Technically, GeneralNews is implemented as a relocating HTTP proxy for HTML documents. HTML documents are requested through the GeneralNews proxy by prepending their origin URL by the URL of the GeneralNews proxy.

Thus, assuming the proxy being located at *http://www.generalnews.de/proxy/* for instance, the document at *http://www.iht.com/* is requested through the proxy under the URL *http://www.generalnews.de/proxy/www.iht.com/*. The proxy requests the document from its origin URL, processes its content, and sends the processed document in its response. The processing of the document consists of these functions:

Compositional references All relative references to external ressources that compose the document by means of transclusion – such as images, style sheets, script files – are resolved relative to the origin location of the document such that they are not requested through the proxy. Absolute references to the resources are not affected. This processing results in the resources being loaded directly from their origin location and not through the proxy, besides assuring that relative references that consist of absolute path names can be correctly resolved in the relocated document. Excepted from this processing are relative references in framesets, which are treated like navigational references, as described below.

Navigational references All navigational references to other documents i.e. hyperlinks, image-map areas, and form actions, as well as references in frame elements, are changed to point through the proxy. This causes hyperlinks in the relocated document to link to a relocated document as well, that is, following links from GeneralNews processed pages leads again to GeneralNews processed pages.

Lexical filtering Words in the text on the filtered page are looked up in WordNet. Data about the synsets as well as hypernyms and hyponyms are inserted as Javascript data structures into the filtered document. These data structures are associated with the occurrences of the words in the text via HTML element ids, which are added to the HTML text.

User interface User interface code in HTML and Javascript is added to the filtered document as HTML text inserted into the *body* element and references to CSS style sheet and a JavaScript script document as a *link* and a *script* element respectively that are inserted into the *head* element. Initialization code for the user interface and the text manipulation code is placed in the *onload* eventhandler of the *body* element such that the user interface is initialized after loading the filtered document in the browser. This added Javascript code uses the lexical information provided in the data structures generated in the lexical filtering step to affect the text display of the page in the browser.

The GeneralNews proxy is written in Perl [4] and runs as mod perl [5] application under the Apache http server [6]. The WordNet database [7] is stored in a MySQL [8] relational database and is accessed using the DBI and DBD::mysql [9] Perl modules.

Figure 1: The slider blended into a filtered web page. It can be interactively dragged to any convenient position in the browser window. Shown is the front page of the International Herald Tribune [10]

Figure 2: The slider blended into the filtered web page in different modi of operation, shown as clippings from figure 1. From top: the slider operated to highest abstraction, to most specificity, and the detail view of one of the words. The gap on the scale marks the average abstraction level of the text measured as the average length of the WordNet hypernym graph starting at each word in the text.

Credits

Idea, Production: Daniela Alina Plewe

Software: Steffen Meschkat

Computational Linguistics: Manfred Stede, Peter Tauter, Uwe Küussner

Layout, Interface: Daniela Alina Plewe, Daniel van Alphen

Website: Karl Gampper

Partner: ZKM, Center for Art and Media
University of Potsdam - Institute for Linguistics

Lexical Database: WordNet, GermaNet

Contact: Daniela Alina Plewe,
mail@danielaplewe.de,
+49-30-6141497 or +49-172-3116388.

Online Ressources

The Apache Software Foundation (2004). mod perl Homepage. *http://perl.apache.org/*, accessed April 2004 .

The Apache Software Foundation (2004). Apache HTTP Server Project. *http://httpd.apache.org/* [accessed April 2004].

GermaNet-Projekt (2003). GermaNet Homepage. Universität Tübingen, Seminar für Sprachwissenschaft. *http://www.sfs.uni-tuebingen.de/lsd/* [accessed April 2004].

Fielding, R. et al. (1999). Hypertext Transfer Protocol -- HTTP/1.1. The Internet Society, Internet Engineering Task Force (IETF), Network Working Group, Request for Comments Nr. 2616, June 1999. *http://www.ietf.org/rfc/rfc2616.txt* [accessed April 2004].

International Herald Tribune (2004). International Herald Tribune. IHT Online Homepage. *http://www.iht.com/* [accessed April 2004].

Lippan, R. (2003). DBD-mysql. *http://search.cpan.org/dist/DBD-mysql/* [accessed April 2004].

Miller, G. A. et al. (2004) WordNet - a Lexical Database for the English Language. *http://www.cogsci.princeton.edu/ wn/* [accessed April 2004].

MySQL AB (2004). MySQL Database Server Homepage. *http://www.mysql.com/* [accessed April 2004].

The Perl Foundation (2004). The Perl Directory. *http://www.perl.org/* [accessed April 2004].

Plewe, D. A. (2004). GeneralNews Website. *http://www.generalnews.de/*, [accessed April 2004].

Rennie, J. (2003). WordNet::QueryData Module. *http://ai.mit.edu/people/jrennie/WordNet/*, *http://search.cpan.org/dist/WordNet-QueryData/* [accessed April 2004].

World Wide Web Consortium (W3C) (2003). Naming and Addressing: URIs, URLs, ... *http://www.w3.org/Addressing/* [accessed April 2004].

World Wide Web Consortium (W3C) (2004). HyperText Markup Language (HTML) Home Page. *http://www.w3.org/MarkUp/* [accessed April 2004].

Sabine Schulte im Walde

GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering

Abstract

WordNet and its German version GermaNet have widely been used as source for fine-grained selectional preference information, focusing on but not restricted to verb-object relationships (Resnik 1997; Ribas 1995; Li & Abe 1998; Abney & Light 1999; Wagner 2000; McCarthy 2001; Clark & Weir 2002). In contrast, this paper presents an approach where argument slots of variable verb-frame combinations are refined by coarse selectional preferences as obtained from the top-level GermaNet nodes. The selectional preference information is applied to an alternation-like verb description and successfully utilised for an automatic clustering of German verbs (Schulte im Walde 2003b).

1 Introduction

This work is concerned with the definition and benefit of selectional preferences as used in an alternation-like verb description for the automatic induction of German semantic verb classes. Semantic verb classes are an artificial construct of natural language which generalises over verbs according to their semantic properties; the class labels refer to the common semantic properties of the verbs in a class at a general conceptual level, and the idiosyncratic lexical semantic properties of the verbs are either added to the class description or left underspecified. Examples for the conceptual classes are Position verbs such as *liegen* 'to lie', *sitzen* 'to sit', *stehen* 'to stand', and *Manner of Motion with a Vehicle* verbs such as *fahren* 'to drive', *fliegen* 'to fly', *rudern* 'to row'. On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs. On the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class; under this aspect, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is provided for rare events.

But how can one obtain a semantic classification of verbs, avoiding a tedious manual definition of the verbs and the classes? A semantic classification demands a definition of semantic properties, but it is difficult to automatically induce semantic features from available resources, both with respect to lexical semantics and conceptual structure. Therefore, the construction of semantic classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its behaviour: To a certain extent, the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its arguments, cf. Levin 1993. We can utilise this meaning-behaviour relationship in that we induce a verb classification on basis of verb features describing verb behaviour (which are easier to obtain automatically than semantic features) and expect the resulting behaviour-classification to agree with a semantic classification to a certain extent.

A widely used approach to define verb behaviour is captured by the diathesis alternation of verbs (see for example Levin 1993; Dorr & Jones 1996; Lapata 1999; Schulte im Walde 2000; Merlo & Stevenson 2001; McCarthy 2001; Joanis 2002). Alternations are alternative constructions at the syntax-semantic interface which express the same or a similar conceptual

idea of a verb. In Example (1), the most common alternations for the *Manner of Motion with a Vehicle* verb *fahren* 'to drive' are illustrated. The participants in the conceptual structure are a driver, a vehicle, a driven person or thing, and a direction. In (a), the vehicle is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction of the movement. In (b), the driver is expressed as subject in a transitive verb construction, again with a prepositional phrase indicating the direction. In (c), the driver is expressed as subject in a transitive verb construction, with an accusative noun phrase indicating the vehicle. And in (d), the driver is expressed as subject in a ditransitive verb construction, with an accusative noun phrase indicating a driven person, and a prepositional phrase indicating the direction of the movement. Even if a certain participant is not realised within an alternation, its contribution might be implicitly defined by the verb. For example, in (a) the driver is not expressed overtly, but we know that there is a driver, and in (b) and (d) the vehicle is not expressed overtly, but we know that there is a vehicle.

(1)
(a) *Der Wagen fährt in die Innenstadt.*
'The car drives to the city centre.'
(b) *Die Frau fährt nach Hause.*
'The woman drives home.'
(c) *Der Filius fährt einen blauen Ferrari.*
'The son drives a blue Ferrari.'
(d) *Der Junge fährt seinen Vater zum Zug.*
'The boy drives his father to the train.'

Assuming that the verb behaviour can be captured by the diathesis alternation of the verb, which are the relevant syntactic and semantic properties one would have to obtain for a verb description? The verbs are distributionally described on three levels, each of them refining the previous level by additional information. The first level $D1$ encodes a purely syntactic definition of verb subcategorisation, the second level $D2$ encodes a syntactico-semantic definition of subcategorisation with prepositional preferences, and the third level $D3$ encodes a syntactico-semantic definition of subcategorisation with prepositional and selectional preferences. The most elaborated description comes close to a definition of verb alternation behaviour. The benefit of each information level can be determined with respect to the lower levels of information.

This paper concentrates on the definition and benefit of selectional preferences at $D3$, the alternation-like verb description. The selectional preferences are based on the German noun hierarchy in GermaNet (Hamp & Feldweg 1997; Kunze 2000), by specifying a coarse generalisation on the top-level synsets for argument slots of variable verb-frame combinations. Section 2 introduces the alternation-like verb descriptions, and Section 3 describes the automatic induction of semantic verb classes as based on the verb descriptions. Finally, Section 4 discusses the usage of the selectional preference information in semantic verb clustering with respect to the demands of German verbs and verb classes.

2 Alternation-Like Verb Descriptions for Verb Clustering

I have developed a statistical grammar model for German which provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs (Schulte im Walde 2002; Schulte im Walde 2003a). The grammar model serves as source for a German verb description at the syntax-semantic interface. For $D1$, it provides frequency distributions of German verbs over 38 purely syntactic subcategorisation frames, which comprise maximally three arguments. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), subor-

dinated non-finite clauses (i), subordinated finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). For example, subcategorising a direct (accusative case) object and a non-finite clause would be represented by 'nai'.

In addition to a purely syntactic definition of subcategorisation frames, the grammar provides detailed information for *D*2 about the types of PPs within the frames. For each of the prepositional phrase frame types in the grammar, the joint frequency of a verb and the PP frame is distributed over the prepositional phrases, according to their frequencies in the corpus. Prepositional phrases are defined by case and preposition, such as 'mit_{Dat}' and '$für_{Akk}$'. The total number of features on *D*2 is 183.

For *D*3, the verb-frame combinations are refined by selectional preferences, i.e. the argument slots within a subcategorisation frame type are specified according to which 'kind' of argument they require. The grammar provides selectional preference information on a fine-grained level: it specifies the possible argument realisations in form of lexical heads, with reference to a specific verb-frame-slot combination. I.e. the grammar provides frequencies for heads for each verb and each frame type and each argument slot of the frame type. For example, the most frequent nominal argument heads for the verb *verfolgen* 'to follow' and the accusative NP of the transitive frame type 'na' are *Ziel* 'goal', *Strategie* 'strategy', *Politik* 'policy', *Interesse* 'interest', *Konzept* 'concept', *Entwicklung* 'development', *Kurs* 'direction', *Spiel* 'game', *Plan* 'plan', *Spur* 'trace'.

Obviously, we would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions on such a specific level. We are provided with rich information on the nominal level, but we need a generalisation of the selectional preference definition. *WordNet* (Miller et al. 1990; Fellbaum 1998) and its German version *GermaNet* (Hamp & Feldweg 1997; Kunze 2000) have widely been used as source for fine-grained selectional preference information (Resnik 1997; Ribas 1995; Li & Abe 1998; Abney & Light 1999; Wagner 2000; McCarthy 2001; Clark & Weir 2002). In contrast to these approaches, I utilise the German noun hierarchy in GermaNet for a *coarse* generalisation of selectional preferences. The hierarchy is realised by means of synsets, sets of synonymous nouns, which are organised by multiple inheritance hyponym/hypernym relationships. A noun can appear in several synsets, according to its number of senses. Figure 1 illustrates the (slightly simplified) GermaNet hierarchy for the noun *Kaffee* 'coffee', which is encoded with two senses, (1) as a beverage and luxury food, and (2) as expression for an afternoon meal. Both senses are subsumed under the general top-level node *Objekt* 'object'. My approach is as follows. For each noun in a verb-frame-slot combination, the joint frequency is split over the different senses of the noun and propagated upwards the hierarchy. In case of multiple hypernym synsets, the frequency is split again. The sum of frequencies over all top synsets equals the total joint frequency. For example, we assume that the frequency of the noun *Kaffee* 'coffee' with respect to the verb *trinken* 'to drink' and the accusative argument in the transitive frame 'na' is 10. Each of the two synsets containing *Kaffee* is therefore assigned a value of 5, and the node values are propagated upwards, as Figure 1 illustrates. Repeating the frequency assignment and propagation for all nouns appearing in a verb-frame-slot combination, the result defines a frequency distribution of the verb-frame-slot combination over all GermaNet synsets.

To restrict the variety of noun concepts to a general level, I consider only the frequency distributions over the top GermaNet nodes. Since GermaNet had not been completed at the point of time I have used the hierarchy, I have manu-

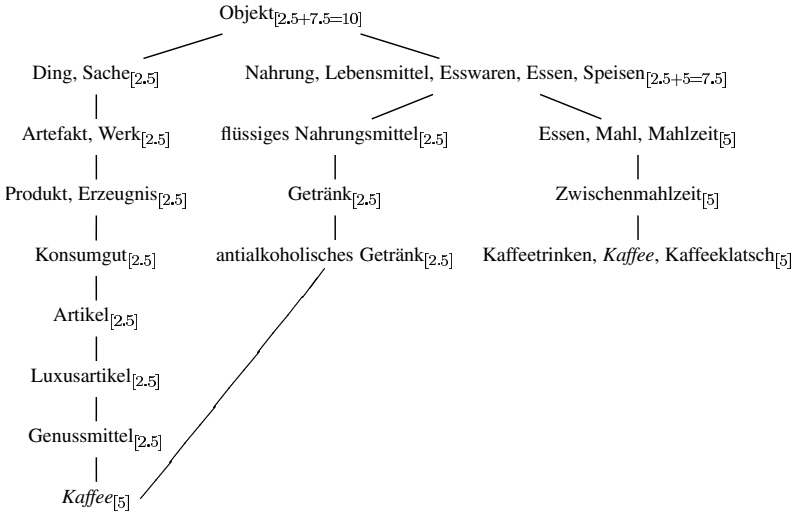

Figure 1: Propagating frequencies through the GermaNet hierarchy

ally added few hypernym definitions, such that most branches are subsumed under the following 15 conceptual top levels. Most of them were already present; the additional links might be regarded as a refinement.

Since the 15 nodes exclude each other and the frequencies sum to the total joint verb-frame frequency, we can use the frequencies to define probability distributions. Therefore, the 15 nodes provide a coarse definition of selectional preferences for a verb-frame-slot combination. Table 1 presents three example verb-frame-slot combinations (the relevant frame slot is underlined) with their preferences. This coarse selectional preference information is provided for each verb-frame-slot combination in the grammar model (trained on 35 million words of German newspaper corpora).

- Lebewesen 'creature'
- Sache 'thing'
- Besitz 'property'
- Substanz 'substance'
- Nahrung 'food'
- Mittel 'means'
- Situation 'situation'
- Zustand 'state'
- Struktur 'structure'
- Physis 'body'
- Zeit 'time'
- Ort 'space'
- Attribut 'attribute'
- Kognitives Objekt 'cognitive object'
- Kognitiver Prozess 'cognitive process'

Table 2 summarises the verb distributions and presents three verbs from different verb classes and their ten most frequent frame types with respect to the three levels of verb definition, accompanied by the probability values. On D_2 frame types including PPs are specified for the PP type, and on D_3 the frame slot for selectional preference refinement is underlined, and the top-level synset is given in brackets. D_1 for *beginnen* 'to begin' defines 'np' and 'n' as the most probable frame types. Even by splitting the 'np'

Verb	Frame+Slot	Top-Level Synset	Freq	Prob
verfolgen	na	Situation	140.99	0.244
'to follow'		Kognitives Objekt	109.89	0.191
		Zustand	81.35	0.141
		Sache	61.30	0.106
		Attribut	52.69	0.091
		Lebewesen	46.56	0.081
		Ort	45.95	0.080
		Struktur	14.25	0.025
		Kognitiver Prozess	11.77	0.020
		Zeit	4.58	0.008
		Besitz	2.86	0.005
		Substanz	2.08	0.004
		Nahrung	2.00	0.003
		Physis	0.50	0.001
essen	na	Nahrung	127.98	0.399
'to eat'		Sache	66.49	0.207
		Lebewesen	50.06	0.156
		Attribut	17.73	0.055
		Zeit	11.98	0.037
		Substanz	11.88	0.037
		Kognitives Objekt	10.70	0.033
		Struktur	8.55	0.027
		Ort	4.91	0.015
		Zustand	4.26	0.013
		Situation	2.93	0.009
		Besitz	1.33	0.004
		Mittel	0.67	0.002
		Physis	0.67	0.002
		Kognitiver Prozess	0.58	0.002
beginnen	n	Situation	1,102.26	0.425
'to begin'		Zustand	301.82	0.116
		Zeit	256.64	0.099
		Sache	222.13	0.086
		Kognitives Objekt	148.12	0.057
		Kognitiver Prozess	139.55	0.054
		Ort	107.68	0.041
		Attribut	101.47	0.039
		Struktur	87.08	0.034
		Lebewesen	81.34	0.031
		Besitz	36.77	0.014
		Physis	4.18	0.002
		Substanz	3.70	0.001
		Nahrung	3.29	0.001

Table 1: Selectional preference definition with GermaNet top nodes.

Verb	Distribution					
	*D*1		*D*2		*D*3	
beginnen	np	0.43	n	0.28	n(Situation)	0.12
'to begin'	n	0.28	np:um$_{Akk}$	0.16	np:um$_{Akk}$(Situation)	0.09
	ni	0.09	ni	0.09	np:mit$_{Dat}$(Situation)	0.04
	na	0.07	np:mit$_{Dat}$	0.08	ni(Lebewesen)	0.03
	nd	0.04	na	0.07	n(Zustand)	0.03
	nap	0.03	np:an$_{Dat}$	0.06	np:an$_{Dat}$(Situation)	0.03
	nad	0.03	np:in$_{Dat}$	0.06	np:in$_{Dat}$(Situation)	0.03
	nir	0.01	nd	0.04	n(Zeit)	0.03
	ns-2	0.01	nad	0.03	n(Sache)	0.02
	xp	0.01	np:nach$_{Dat}$	0.01	na(Situation)	0.02
essen	na	0.42	na	0.42	na(Lebewesen)	0.33
'to eat'	n	0.26	n	0.26	na(Nahrung)	0.17
	nad	0.10	nad	0.10	na(Sache)	0.09
	np	0.06	nd	0.05	n(Lebewesen)	0.08
	nd	0.05	ns-2	0.02	na(Lebewesen)	0.07
	nap	0.04	np:auf$_{Dat}$	0.02	n(Nahrung)	0.06
	ns-2	0.02	ns-w	0.01	n(Sache)	0.04
	ns-w	0.01	ni	0.01	nd(Lebewesen)	0.04
	ni	0.01	np:mit$_{Dat}$	0.01	nd(Nahrung)	0.02
	nas-2	0.01	np:in$_{Dat}$	0.01	na(Attribut)	0.02
fahren	n	0.34	n	0.34	n(Sache)	0.12
'to drive'	np	0.29	na	0.19	n(Lebewesen)	0.10
	na	0.19	np:in$_{Akk}$	0.05	na(Lebewesen)	0.08
	nap	0.06	nad	0.04	na(Sache)	0.06
	nad	0.04	np:zu$_{Dat}$	0.04	n(Ort)	0.06
	nd	0.04	nd	0.04	na(Sache)	0.05
	ni	0.01	np:nach$_{Dat}$	0.04	np:in$_{Akk}$(Sache)	0.02
	ns-2	0.01	np:mit$_{Dat}$	0.03	np:zu$_{Dat}$(Sache)	0.02
	ndp	0.01	np:in$_{Dat}$	0.03	np:in$_{Akk}$(Lebewesen)	0.02
	ns-w	0.01	np:auf$_{Dat}$	0.02	np:nach$_{Dat}$(Sache)	0.02

Table 2: Examples of most probable frame types.

probability over the different PP types in *D*2, a number of prominent PPs are left, the time indicating *um$_{Akk}$* and *nach$_{Dat}$*, *mit$_{Dat}$* referring to the begun event, *an$_{Dat}$* as date and *in$_{Dat}$* as place indicator. It is obvious that adjunct PPs as well as argument PPs represent a distinctive part of the verb behaviour. *D*3 illustrates that typical selectional preferences for beginner roles are *Situation*, *Zustand*, *Zeit*, *Sache*. *D*3 has the potential to indicate verb alternation behaviour, e.g. 'na(Situation)' refers to the same role for the direct object in a transitive frame as 'n(Situation)' in an intransitive frame. *essen* 'to eat' as an object drop verb shows strong preferences for both intransitive and transitive usage. As desired, the argument roles are strongly determined by *Lebewesen* for both 'n' and 'na' and *Nahrung* for 'na'. *fahren* 'to drive' chooses typical manner of motion frames ('n', 'np', 'na') with the refining PPs being directional (*in$_{Akk}$*, *zu$_{Dat}$*, *nach$_{Dat}$*) or referring to a means of motion (*mit$_{Dat}$*, *in$_{Dat}$*, *auf$_{Dat}$*). The selectional preferences represent a correct alternation behaviour: *Lebewesen* in the object drop case for 'n' and 'na', *Sache* in the inchoative/causative case for 'n' and 'na'.

3 Induction of Semantic Verb Classes

The selectional preference information is applied to an alternation-like verb description in automatic verb clustering. The clustering of the German verbs is performed by the k-Means algorithm, a standard unsupervised clustering technique as proposed by Forgy 1965. Based on the distributional verb descriptions and standard notions of similarity between distributional vectors, k-Means iteratively re-organises initial verb clusters by assigning each verb to its closest cluster and re-calculating cluster centroids until no further changes take place. For details on the clustering setup and experiments, the reader is referred to Schulte im Walde 2003b.

The clustering experiments are performed on 168 partly ambiguous German verbs. Before the experiments, I manually classified the verbs into 43 semantic classes. The purpose of the manual classification is to evaluate the reliability and performance of the clustering experiments. In the following, I present representative parts of a cluster analysis which uses the alternation-like verb description on *D*3. For each cluster, the verbs which belong to the same gold standard class are presented in one line, accompanied by the class label. I compare the respective clusters with their pendants under *D*1 and *D*2, to demonstrate the effect of the feature refinements.

(a) nieseln regnen schneien – *Weather*

(b) dämmern – *Weather*

(c) kriechen rennen – *Manner of Motion*: *Locomotion*
eilen – *Manner of Motion: Rush*
gleiten – *Manner of Motion: Flotation*
starren – *Facial Expression*

(d) klettern wandern – *Manner of Motion*: *Locomotion*
fahren fliegen segeln – *Manner of Motion: Vehicle*
fließen – *Manner of Motion: Flotation*

(e) beginnen enden – *Aspect*
bestehen existieren – *Existence*
liegen sitzen stehen – *Position*
laufen – *Manner of Motion: Locomotion*

(f) festlegen – *Constitution*
bilden – *Production*
erhöhen senken steigern vergrößern verkleinern – *Quantum Change*

(g) töten – *Elimination*
unterrichten – *Teaching*

The weather verbs in cluster (a) strongly agree in their syntactic expression on *D*1 and do not need *D*2 or *D*3 refinements for a successful class constitution. *dämmern* in cluster (b) is ambiguous between a weather verb and expressing a sense of understanding; this ambiguity is idiosyncratically expressed in *D*1 frames already, so *dämmern* is never clustered together with the other weather verbs on *D*1-*D*3. *Manner of Motion, Existence, Position* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together on *D*1, but adding PP information distinguishes the respective verb classes: *Manner of Motion* verbs primarily demand directional PPs, *Aspect* verbs are distinguished by patient mit_{Dat} and time and location prepositions, and *Existence* and *Position* verbs are distinguished by locative prepositions, with *Position* verbs showing more PP variation. The PP information is essential for successfully distinguishing these verb classes, and the coherence is partly destroyed by *D*3: *Manner of Motion* verbs (from the sub-classes *Locomotion, Rotation, Rush, Vehicle, Flotation*) are captured well by clusters (c) and (d), since they inhibit strong common alternations, but cluster (e) merges the *Existence, Position* and *Aspect* verbs, since verb-idiosyncratic selectional preferences destroy the *D*2 class demarcation. Admittedly, the verbs in cluster (e) are close in their semantics, with a common sense of (bringing into vs. being in) existence. *laufen* fits into the cluster with its sense of 'to function'. Cluster (f) contains most verbs

of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The semantics of the cluster is therefore rather pure. The verbs in the cluster typically subcategorise a direct object, alternating with a reflexive usage, 'nr' and 'npr' with mostly *auf*$_{Akk}$ and *um*$_{Akk}$. The selectional preferences help to distinguish this cluster: the verbs agree in demanding a thing or situation as subject, and various objects such as attribute, cognitive object, state, structure or thing as object. Without selectional preferences (on $D1$ and $D2$), the change of quantum verbs are not found together with the same degree of purity. There are verbs as in cluster (g), whose properties are correctly stated as similar on $D1$-$D3$, so a common cluster is justified; but the verbs only have coarse common meaning components, in this case *töten* and *unterrichten* agree in an action of one person or institution towards another.

4 Discussion

Which exactly is the nature of the meaning-behaviour relationship in the constitution of semantic verb classes? And, more specifically, which is the benefit of the selectional preferences in the alternation-like verb description as based on GermaNet top-level nodes?

Addressing the nature of the meaning-behaviour relationship in the clustering, (a) already a purely syntactic verb description allows a verb clustering clearly above the baseline. The result is a successful (semantic) classification of verbs which agree in their syntactic frame definitions, e.g. most of the *Support* verbs *dienen*, *helfen*, *folgen*. The clustering fails for semantically similar verbs which differ in their syntactic behaviour, e.g. *unterstützen* which does belong to the *Support* verbs but demands an accusative instead of a dative object. In addition, it fails for syntactically similar verbs which are clustered together even though they do not exhibit semantic similarity, e.g. many verbs from different semantic classes subcategorise an accusative object, so they are falsely clustered together. (b) Refining the syntactic verb information by prepositional phrases is helpful for the semantic clustering, not only in the clustering of verbs where the PPs are obligatory, but also in the clustering of verbs with optional PP arguments. The improvement underlines the linguistic fact that verbs which are similar in their meaning agree either on a specific prepositional complement (e.g. *glauben/denken an*$_{Akk}$) or on a more general kind of modification, e.g. directional PPs for manner of motion verbs. (c) Defining selectional preferences for arguments once more improves the clustering results, but the improvement is not as persuasive as when refining the purely syntactic verb descriptions by prepositional information. For example, the selectional preferences help demarcate the *Quantum Change* class, because the respective verbs agree in their structural as well as selectional properties. But in the *Consumption* class, *essen* and *trinken* have strong preferences for a food object, whereas *konsumieren* allows a wider range of object types. On the contrary, there are verbs which are very similar in their behaviour, especially with respect to a coarse definition of selectional preferences, but they do not belong to the same fine-grained semantic class, e.g. *töten* and *unterrichten*.

The description of the clustering examples has shown that the dividing line between the common and idiosyncratic features of verbs in a verb class defines the level of verb description which is relevant for the class constitution. The meaning components of verbs to a certain extent determine their behaviour, but this does not mean that all properties of all verbs in a common class are similar and we could extend and refine the feature description endlessly. The meaning of verbs comprises both (i) properties which are general for the respective verb classes, and (ii) idiosyncratic properties which distinguish the verbs from each other. As long as we define the verbs by those properties which represent the common

parts of the verb classes, a clustering can succeed. But step-wise refining the verb description by including lexical idiosyncrasy, the emphasis of the common properties vanishes. Some verbs and verb classes are distinctive on a coarse feature level, some need fine-grained extensions, some are not distinctive with respect to any combination of features. There is no unique perfect choice and encoding of the verb features; the feature choice rather depends on the specific properties of the desired verb classes.

The usage of selectional preference information in semantic verb clustering is a particular challenge for the verb description. On the one hand, one would want a selectional preference description as fine-grained as possible, to e.g. distinguish the verbs *töten* and *unterrichten* which are similar on a coarse selectional preference level (agreeing in an action of one person or institution towards another), but distinguished on a fine-grained level: in a transitive construction, *töten* appears with subjects such as *Soldat* 'soldier', *Angreifer* 'attacker', *Schütze* 'shooter', *Terrorist* 'terrorist', *Jäger* 'hunter' and direct objects such as *Soldat* 'soldier', *Zivilist* 'civilian', *Rebell* 'rebel', *Nebenbuhler* 'rival', *Tier* 'animal', and *unterrichten* appears with subjects such as *Lehrerschaft* 'community of teachers', *College* 'college', *Professor* 'professor' and direct objects such as *Kind* 'child', *Schüler* 'pupil', *Klasse* 'class', *Fach* 'subject', *Grammatik* 'grammar'. Assuming that we use GermaNet as source for the preference definition, in the example case we would need an algorithm comparable to those by Resnik 1997; Ribas 1995; Li & Abe 1998; Abney & Light 1999; Wagner 2000; McCarthy 2001; Clark & Weir 2002 which is able to filter selectional preferences of arbitrary depth in the hierarchy. On the other hand, one would want a selectional preference description on a more general level. Consider the most specific conceptual level of semantic classes, a classification with classes of verb synonyms.[1] But even the verb behaviour of synonyms does not overlap perfectly, since e.g. selectional preferences of synonyms vary. For example, the German verbs *bekommen* and *erhalten* 'to get, to receive' are synonymous, but they cannot be exchanged in all contexts, cf. *einen Schnupfen bekommen* 'to catch a cold' vs. *einen Schnupfen erhalten*. This means that even for synonyms a fine-grained definition of selectional preferences would not provide a perfect overlap of the distributional features and that some generalisation is desirable.

In addition to the linguistic conflict in clustering when defining selectional preferences for verbs, a clustering algorithm has to pay attention to the technical issue of feature encoding. We would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions on a fine-grained level. Again, this means that some generalisation level of selectional preferences is adequate.

Summarising, both the theoretical assumption of encoding features of verb alternation as verb behaviour and the practical realisation by encoding syntactic frame types, prepositional phrases and selectional preferences have proven successful. But the exact feature choice for verb descriptions in verb clustering depends on the specific properties of the desired verb classes. And even if classes are perfectly defined on a common conceptual level, the relevant level of behavioural properties of the verb classes might differ. This insight is especially problematic for the definition of selectional preferences, since numerous variations for their encoding are possible, but each choice would present advantages for some verb classes and disadvantages for others. This work has presented evidence for the usefulness of GermaNet top levels nodes as coarse generalisation of selectional preferences, but the issue of improving the level of GermaNet preference definitions is subject to further work.

Acknowledgements

The work reported here was performed while the author was a member of the DFG-funded PhD program 'Graduiertenkolleg' *Sprachliche Repräsentationen und ihre Interpretation* at the Institute for Natural Language Processing (IMS), University of Stuttgart, Germany.

Note

[1] In this context, synonymy refers to 'partial synonymy' where synonymous verbs cannot necessarily be exchanged in all contexts, as compared to 'total synonymy' where synonymous verbs can be exchanged in all contexts – if anything like 'total synonymy' exists at all (Bussmann 1990).

References

Abney, S.; Light, M. (1999). "Hiding a Semantic Class Hierarchy in a Markow Model." In: Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, College Park, MD, 1-8.

Bussmann, H. (1990²). Lexikon der Sprachwissenschaft. Stuttgart: Alfred Kröner Verlag.

Clark, S.; Weir, D. (2002). "Class-Based Probability Estimation using a Semantic Hierarchy." In: Computational Linguistics, 28(2) (2002), 187-206.

Dorr, B. J.; Jones, D. (1996). "Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues." In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, August 1996, 322-327.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Forgy, E. W. (1965). "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications." In: Biometrics 21 (1965), 768-780.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (eds.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Joanis, E. (2002). Automatic Verb Classification using a General Feature Space. Master's thesis, University of Toronto, Department of Computer Science.

Kunze, C. (2000). "Extension and Use of GermaNet, a Lexical-Semantic Database." In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, May 2000, 999-1002.

Lapata, M. (1999). "Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations." In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99), College Park, MD, 397-404.

Levin, B. (1993). English Verb Classes and Alternations. Chicago, IL: The University of Chicago Press.

Li, H.; Abe, N. (1998). "Generalizing Case Frames Using a Thesaurus and the MDL Principle." In: Computational Linguistics 24(2) (1998), 217-244.

McCarthy, D. (2001). Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Ph.D. thesis, University of Sussex, Brighton, UK.

Merlo, P.; Stevenson, S. (2001). "Automatic Verb Classification Based on Statistical Distributions of Argument Structure." In: Computational Linguistics 27(3) (2001), 373-408.

Miller, G. A. et al. (1990). "Introduction to Wordnet: An On-line Lexical Database." In: International Journal of Lexicography 3(4) (1990), 235-244.

Resnik, P. (1997). "Selectional Preference and Sense Disambiguation." In: Proceedings of the ACL SIGLEX / ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D. C., April 1997.

Ribas, F. (1995). "On Learning More Appropriate Selectional Restrictions." In: Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95), Dublin, March 1995.

Schulte im Walde, S. (2000). "Clustering Verbs Semantically According to their Alternation Behaviour." In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbrücken, Germany, August 2000, 747-753.

Schulte im Walde, S. (2002). "A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG." In: Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, May/June 2002, vol. IV, 1351-1357.

Schulte im Walde, S. (2003a). "A Collocation Database for German Nouns and Verbs." In: Proceedings of the 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003), Budapest, April 2003.

Schulte im Walde, S. (2003b). Experiments on the Automatic Induction of German Semantic Verb Classes. Ph.D. thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung [= AIMS Report 9(2)].

Wagner, A. (2000). "Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis." In: Proceedings of the ECAI-2000 Workshop on Ontology Learning, Berlin, August 2000, 37- 42.

Andreas Wagner

Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies

Abstract

This paper deals with methods for estimating frequencies of concepts in wordnets from corpus data. In particular, it addresses issues which multiple inheritance structures in wordnets raise regarding this task. One of the discussed approaches (tree cut) is problematic in this respect, because it requires a pure tree hierarchy. Applying this approach to a wordnet requires that its DAG structure is transformed into a tree. I propose a mathematically sound method for that purpose and compare this method to a commonly used ad-hoc strategy. This strategy leads to biases in the estimated frequencies which are avoided by the approach proposed here. Experiments with GermaNet demonstrate that these biases have significant impacts.

1 Introduction

Wordnets, i.e. lexical-semantic hierarchies in the style of WordNet (cf. Fellbaum 1998), have commonly been employed in NLP applications which involve quantitative methods. In particular, within the paradigm of statistical corpus linguistics, approaches have been proposed which combine the quantitative evidence provided by word frequencies obtained from a corpus with the symbolic knowledge provided by a wordnet. To establish this combination, the frequencies of words in the corpus are propagated to the respective concepts that subsume these words. In this way, concept frequencies are estimated from word frequencies. For example, the frequency of the word 'person' in the corpus plays a role for the frequency estimates for the concepts <person>, <life_form>, and <entity> in the semantic hierarchy. Concept frequencies, in turn, are used to estimate concept probabilities, which then can be employed for the NLP task in question.

A fundamental issue in this context is how concept frequencies can be adequately estimated from word frequencies. This paper is concerned with this issue. In principle, there are several possible ways to achieve that goal. In section 2, I will sketch three basic methods and discuss suitability conditions for their application by considering approaches to a particular NLP task. It turns out that different acquisition approaches – even if they serve the same task – demand different methods of estimating concept frequencies.

The rest of the paper focuses on a general incompatibility that arises if one of the methods described in section 2 is applied to a wordnet. This method requires that the concept hierarchy has a pure tree structure. However, a wordnet generally has the structure of a DAG, i.e. a concept may have more than one parent (immediate hyperonym). To overcome this conflict, a simple ad-hoc strategy to (virtually) convert the DAG structure into a tree structure has been largely used. In section 3, I will point out that this strategy introduces undesirable biases into the frequency estimations. Treating multiple inheritance in an ad-hoc manner has been justified (if at all) by the fact that multiple inheritance (multiple parents) in WordNet is rare: Less than 1% of the noun concepts in WordNet have more than one parent, most of which are very specific, i.e. located at low levels of the hierarchy (cf. McCarthy 2001). However, for other wordnets, the situation is different. For example, for GermaNet (cf. Hamp & Feldweg 1997; Kunze & Wagner 2000), cross-classification of con-

cepts has been a major design principle, and thus multiple inheritance is common; 11.5% of the GermaNet concepts have more than one parent. Hence, when applying a frequency estimation method which requires a tree-shaped hierarchy to a hierarchy like GermaNet, a principled solution to that conflict is highly desirable. Therefore, I propose a more sophisticated method for propagating word frequency counts to concepts. This method converts a wordnet DAG structure into a tree structure, but avoids the drawbacks mentioned above.

Finally, in section 4, I report some experiments performed with GermaNet. These experiments show that the biases introduced by the abovementioned ad-hoc strategy have significant impacts.

2 Basic Methods

2.1 An Exemplary Task

In order to exemplify the use of different ways to estimate concept frequencies, I will discuss their role in a particular task: learning selectional preferences. Selectional preferences are semantic preferences that a predicate (e.g. a verb) exhibits for its arguments. For example, the verb 'eat' prefers a subject referring to a human being or animal and an object denoting food. Such preferences can be represented by wordnet concepts. Statistical approaches for acquiring selectional preferences using WordNet retrieve for each concept a preference value which quantifies the degree of preference (or dispreference) of that concept (with regard to a certain argument slot of a certain verb). The computation of such preference values is based on concept probabilities, which are derived from concept frequencies.

In this section, I describe the basic approaches for concept frequency estimation which have been proposed in the literature that deals with learning selectional preferences by combining statistical corpus analysis and WordNet. Furthermore, I sketch how these frequency counts are employed for preference acquisition. It turns out that different ways to choose the concepts that should *represent* the selectional preferences of a verb (e.g. <food> for the object of 'eat') require different frequency estimation strategies.

The training data that are used by the approaches discussed here are extracted from a parsed corpus. They comprise pairs of the form (v, n), where v is a verb and n is the head noun of a certain fixed argument type (e.g. the object) of v. From these data, the verb–noun pair frequencies $freq(v,n)$ as well as the marginal frequencies $freq(v)$ and $freq(n)$ (the overall frequencies of v and n in the data) are extracted and employed to estimate noun concept frequencies $freq(ncpt)$ and $freq(v,ncpt)$, respectively, where $ncpt$ is a concept subsuming n. Based on these concept counts, concept probabilities are usually obtained by maximum likelihood estimation:

$$p(ncpt \mid v) = \frac{freq(v, ncpt)}{freq(v)} \quad (1)$$

$$p(ncpt) = \frac{freq(ncpt)}{N} \quad (2)$$

where N is the size of the training data.

These probabilities are used to obtain the preference value of $ncpt$ (w.r.t. v). There are several ways to quantify selectional preference. Here, I shortly mention the most common ones. The simplest possibility is to immediately use $p(ncpt|v)$ (the probability that $ncpt$ occurs as complement of v) as preference score (pursued e.g. in Li & Abe 1998). An alternative possibility (proposed in Li & Abe 1996), is to compute the preference value by the ratio

$$\frac{p(ncpt \mid v)}{p(ncpt)} \quad (3)$$

This quantity measures the probability that $ncpt$ co-occurs with v relative to the general probability of $ncpt$ in the data. This definition offers an obvious way to distinguish between preference and dispreference: If the ratio is greater than 1,

then *v* selects *ncpt* with higher probability than by chance, and thus *ncpt* is preferred by *v*. Conversely, a ratio smaller than 1 indicates dispreference.

A third possibility (proposed e.g. in Resnik 1998 and Ribas 1995a) combines the above mentioned alternatives:

$$p(ncpt \mid v) \log \frac{p(ncpt \mid v)}{p(ncpt)} \qquad (4)$$

Here, the logarithm of the ratio in (3) (which corresponds to the mutual information between *v* and *ncpt*) is weighted by $p(ncpt|v)$. Due to the factor $\log \frac{p(ncpt|v)}{p(ncpt)}$, this measure also distinguishes between preferred concepts (preference value > 0) and dispreferred ones (preference value < 0). In addition, the magnitude of the preference score is scaled by the probability that *v* selects *ncpt*.

2.2 The Word-to-Concept Approach

The method I refer to as word-to-concept approach was proposed by Resnik 1998. This method immediately divides the frequency count of a noun *n* equally among all concepts which subsume *n* (denoted as *concepts*(*n*)).

Figure 1 illustrates how the word-to-concept approach works. There are four WordNet concepts that subsume the word 'person': <person>, <life_form>, <causal_agent>, and <entity>. Thus, each of these four concepts receives ¼ of the frequency of 'person' in the corpus (100/4 = 25 in the example).[1]

Formally, the frequency of a concept *ncpt* is calculated as

$$freq(ncpt) = \sum_{n \in words^{+}(ncpt)} \frac{freq(n)}{concepts(n)} \qquad (5)$$

where words^{+}(*ncpt*) is the set of words which are subsumed by *ncpt*, i.e. which are a member either of the synset of *ncpt* or of the synset of one of its hyponyms. (The joint frequency *freq*(*v*,*ncpt*) of a verb *v* and a noun concept *ncpt* is computed analogously; one just replaces *freq*(*n*) by *freq*(*v*,*n*) in equation (5).)

The word-to-concept approach yields a probability distribution *over all concepts in the hierarchy*, i.e. the probabilities $p(ncpt)$ of all concepts sum to 1. The same holds for the conditional probabilities $p(ncpt|v)$. This property corresponds to Resnik's approach to represent the selectional preferences of a verb by *all* WordNet concepts (and their preference values), rather than to retrieve a subset of 'representative concepts' for that purpose. Moreover, he uses the distributions $p(ncpt|v)$ and $p(ncpt)$ and to quantify the overall *preference strength* of *v*. The selectional preference strength quantifies how strong

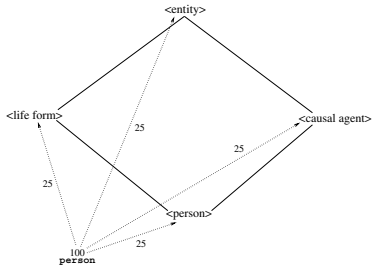

Figure 1: Frequency propagation by the word-to-concept approach

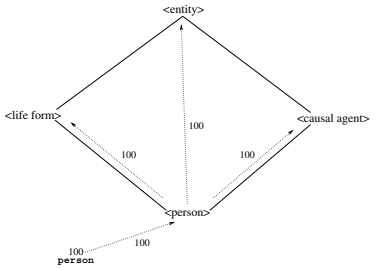

Figure 2: Frequency propagation by the word-to-sense approach

the predicate semantically constrains its arguments. For example, 'eat' has a greater selectional preference strength for its object then 'have', because 'eat' strongly prefers objects denoting food, whereas 'have' can select almost any noun as its object. Resnik's approach of quantifying the overall preference strength is to measure to what extent the probability distribution $p(ncpt|v)$ deviates from the general distribution $p(ncpt)$. The larger the difference between the two distributions, the higher the preference strength. Resnik calculates this difference by the well-known information-theoretic distance measure of *relative entropy*. In fact, Resnik 1998 reports a low preference strength for 'have' (0.43) and a comparably high preference strength for 'eat' (3.51).

2.3 The Word-to-Sense Approach

While Resnik divides the frequency count of a noun *n* among all concepts *concepts*(*n*) which subsume *n*, Ribas (cf. Ribas 1995a) proposes a different approach: He divides *freq*(*n*) among the concepts which represent an immediate sense of *n*, i.e. those concepts whose synsets contain *n* (denoted as *senses*(*n*)). I refer to this strategy as the word-to-sense approach. However, as a noun does not only provide evidence for its senses, but also for the hyperonyms of these senses, the frequency count obtained for a noun sense is completely propagated to each of its ancestors in the hierarchy.

Figure 2 takes up the example in figure 1, this time illustrating the word-to-sense approach. The frequency of 'person' (100) is mapped to the synset <person>, which represents the corresponding word sense.[2] This count is completely propagated to all concepts that subsume person.

In general, the frequency of a concept is estimated as the sum of the counts of those word senses which the concept subsumes. More formally, let $senses_{ncpt}(n)$ be the set of senses of *n* which are subsumed by *ncpt*. Then, the frequency of a concept is estimated by the equation

$$freq(ncpt) = \sum_{n \in words^+(ncpt)} \left|senses_{ncpt}(n)\right| \frac{freq(n)}{|senses(n)|} \quad (6)$$

The word-to-sense approach views the WordNet hierarchy as an inventory of concepts with implication relations among each other. A hyponymy/hyperonymy relation between two concepts indicates that one concept (the hyponym) implies the other (the hyperonym). This means that a concept inherits all the probability mass of its hyponyms. In particular, since the root of the hierarchy is implied by all concepts, its probability is 1. In contrast, the word-to-concept approach views the WordNet hierarchy as a pool of concepts which represent a smaller or larger set of nouns. In this model, hyponymy/hyperonymy relations between concepts indicate a common (sub)set of nouns providing evidence for these concepts. This model is required for quantities which are based on probability distributions over the whole inventory of concepts, like Resnik's overall preference strength. A consequence of this model which might be somewhat counterintuitive is that the probability of the root concept is below 1. This is because probability mass is not completely inherited by, but equally divided among hyperonyms.

As noted above, Ribas quantifies selectional preference according to formula (4). In contrast to Resnik, he does not keep all noun concepts, but extracts a 'representative set' of concepts to model the preferential behaviour of a verb. To induce this set, he uses a greedy approach which can be sketched as follows: Initially, consider all noun concepts as 'candidates' for inclusion into the representative set. Among them, select that concept *ncpt* which has the highest preference value and insert it into the target set. After that, remove *ncpt* and all its hyponyms and hyperonyms from the set of candidates. (This is done to avoid redundancy.) Repeat these steps until the candidate set is empty. In this way, Ribas yields a non-redundant set of highly preferred

concepts. For example, Ribas 1995a reports that this approach acquired (among others) the following concepts for the subject of 'present': <causal_agent> (4.15), organization (0.45), <administrative_district> (0.26), <and life_form> (0.14). Ribas' simple heuristic for retrieving a representative set of concepts does not depend on a particular approach for estimating concept frequencies. All methods discussed in this paper are compatible with it.

2.4 The Tree Cut Approach

The tree cut approach is a more sophisticated way of retrieving a collection of 'representative' concepts from a semantic hierarchy. It was developed by Li and Abe (cf. Abe & Li 1996; Li & Abe 1998) for the task of acquiring selectional preferences. Li and Abe represent the selectional preferences of a verb by a *tree cut model*. Such a model provides a horizontal cut through the noun hierarchy so that the concepts which are located along this cut form a partition of the noun senses covered by the hierarchy. A tree cut model consists of the concepts specified by a cut and the preference values for these concepts. Figure 3 shows a portion of the WordNet hierarchy—with preference values attached to the individual concepts, computed according to formula (3)—and two of the possible cuts across the hierarchy (indicated by a solid and a dashed line, respectively). The difference between the corresponding models is that one model contains the concept <animal>, whereas the other model contains the more specific concepts <bird>, <insectivore>, and <primate>. This is an artificial example intended to illustrate plausible preference values and tree cut models for the subject of 'fly'.

The tree cut approach aims at finding a cut at the appropriate level of generalisation. In this respect, the cut indicated by the solid line in figure 3 is more appropriate than the more general cut indicated by the dashed line, because the latter one does not capture the fact that some kinds of animals (birds, insects) normally fly, while others do not. The cut at the adequate abstraction level is selected by the *Minimum Description Length* (MDL) Principle. I will not go into details concerning this information-theoretic principle; cf. Li & Abe 1998 and Abe & Li 1996 for its motivation and application for the given task. In our context, it is important to note that the MDL approach requires that every possible tree cut model exactly captures the probability mass that represents the whole training data. In other words, the sum of the frequency counts of the concepts on the cut has to correspond to the size of the data.[3]

To ensure this requirement, the frequency of a noun sense has to be completely propagated to its superconcepts so that the frequency of a concept on the cut (and hence its probability) encompasses the frequencies (probabilities) of all senses it subsumes. Therefore, concept frequencies have to be estimated according to the word-to-sense approach. However, there is a further constraint: It is necessary that each noun sense is subsumed by *one and only one* concept on the cut. Therefore, the structure of the hierarchy must exhibit two properties: Firstly, the noun senses must be modelled by leaf nodes in the hierarchy, while the inner nodes model more abstract concepts. This is required to ensure that all noun senses are below the cut and thus captured by it. Secondly, the hierarchy must be a pure tree, i.e. all concepts (except the root) must have exactly one parent. This is necessary to guaran-

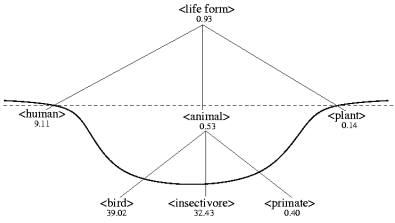

Figure 3: Two possible tree cut models for the subject of 'fly'

tee that no noun sense is represented by multiple concepts on the cut.[4] Obviously, the structure of wordnets deviates from these requirements. Word senses are not only represented by leaves, but by all nodes in the hierarchy. Furthermore, as noted, a wordnet generally exhibits a DAG structure with multiple inheritance.

Thus, to be able to apply the tree cut approach to a wordnet, its structure has to be adapted to meet the two abovementioned properties. To account for the first requirement, a widely used strategy is to create for each inner node an additional node that represents the sense of those words which belong to the synset corresponding to that node. This additional node becomes a hyponym of the original node. In this way, all word senses are captured by leaf nodes. The second requirement is much more complex, since it necessitates a (virtual) transformation of the wordnet DAG structure into a pure tree structure. The core of such a transformation is propagating frequency counts upwards in the hierarchy in a way which 'mimics' a tree structure. The next section addresses this issue.

3 Transforming the Wordnet DAG Structure

One crucial part of the virtual transformation of the wordnet structure can be performed as a side effect of processing the hierarchy. If a wordnet is processed top-down (as is done by the tree cut acquisition algorithm developed by Li and Abe), then its DAG structure is 'resolved' into a tree structure. Nodes that have multiple parents are processed multiple times, once for each parent. For example, as <person> is a hyponym of both <life_form> and <causal_agent>, this concept (and thus its hyponyms) is processed twice, once as a child of <life_form>, and once as a child of <causal_agent>. In this way, a 'virtual copy' of such a node (and its descendants) is created for each of its parents, and the DAG is 'broken into a tree' (cf. figure 4; virtual copies are indicated by a dashed link). Thus, if the task in question involves top-down processing, a tree structure is virtually simulated. Otherwise, the wordnet structure (i.e. the database) has to be altered accordingly.

In any case, one has to ensure that the estimated concept frequencies are consistent with that tree structure. As mentioned in section 2.4, the tree cut approach employs the word-to-sense method to obtain concept frequencies, i.e. the frequency of each word sense is propagated to all its ancestors in the hierarchy, and for each concept, the frequencies accumulated at it add up to its count. In fact, there are several possibilities of how to perform this propagation. Following Ribas' approach explained in section 2.3, the frequency of a concept is the sum of the frequencies of the word senses which are subsumed by that concept (cf. equation (6)). If the hierarchy is a tree structure, then this frequency is equivalent to the sum of the frequencies of the immediate hyponyms (i.e. the children) of the concept:

$$freq(ncpt) = \sum_{ncpt_c \in children(ncpt)} freq(ncpt_c) \quad (7)$$

However, if the hierarchy is a DAG, then equation (7) might yield different values than equation (6). For example, in figure 2, <entity> would receive the count of <life_form> plus the count of <causal_agent>, i.e. the count of <entity> would be 200 instead of 100.

A straightforward way to obtain frequency counts consistent with the tree structure is to employ equation (7) instead of equation (6) for frequency estimation. Li and Abe as well as other researchers adopted this solution. Here, the duplication of subtrees is reflected by the corresponding counts. The drawback of this approach is that multiplying certain subtrees corresponds to multiplying that portion of the data which is covered by the concepts in that subtree.

Figure 4 shows an example. Here, as the concept <person> is processed twice, all instances in the data denoting a person are counted twice. Thus, the relative proportion of these instances is increased. In particular, the frequency of the top node <entity> contains the count of <person> twice.

In order to avoid such biases, I propose a different approach for retrieving concept frequencies. The general idea of this approach is as follows: As in the work of Li and Abe, the count of a concept is directly determined by the counts of its children. This simulates a tree structure. However, a concept does not necessarily inherit the *total* count from each of its children. If a concept has multiple parents, then the count of that concept is divided among its parents. In this way, counts are not duplicated, and thus no bias towards certain parts of the sample is created. The frequency portion that a child concept $ncpt_c$ passes to each of its parents is determined by a probability distribution $p(ncpt_p|ncpt_c)$ where $ncpt_p$ is a parent of $ncpt_c$. Thus, the frequency of a concept is given by

$$freq(ncpt_p) = \sum_{ncpt_c \in children(ncpt_p)} freq(ncpt_c) p(ncpt_p|ncpt_c) \quad (8)$$

The crucial question is how to estimate the distribution $p(ncpt_p|ncpt_c)$in this equation. I decided to guide this estimation by the frequencies of the parents: The count of a concept is apportioned among its parents according to their respective frequency, relative to the frequencies of the other parents. Formally, for a concept $ncpt_c$, the distribution $p(ncpt_p|ncpt_c)$ is estimated by the ratio of the frequency of $ncpt_p$ and the sum of the frequencies of all parents of $ncpt_c$:

$$p(ncpt_p|ncpt_c) = \frac{freq(ncpt_p)}{\sum_{ncpt' \in parents(ncpt_c)} freq(ncpt')} \quad (9)$$

In the trivial case in which $ncpt_c$ has only one parent, $p(ncpt_p|ncpt_c)$ is 1, i.e. the complete concept frequency is propagated to that parent.

The equations (8) and (9) depend on each other. The probability of the parent given a child concept in equation (8) is estimated by equation (9), whereas the parent frequencies in equation (9) are obtained by equation (8). Therefore, to make these equations applicable, it is necessary to assume certain initial values. It is quite straightforward to initialise the parent probabilities by assuming uniform distributions:

$$p(ncpt_p|ncpt_c) = \frac{1}{|parents(ncpt_c)|} \quad (10)$$

In this way, the count of a concept is equally apportioned to its parents in the initial iteration. As the parents of a concept have different (additional) children, this iteration yields different counts for them. Thus, in the following iterations, equation (9) will estimate differing probabilities for the parents of a concept. In general, an iteration step changes the counts and probabilities. The approach proposed here can be viewed as an instance of the EM algorithm: equation (8) corresponds to the E-step and equation (9) to the M-step.

For example, in figure 5, the initialisation step equally apportions the count of <person> to its two parents; each parent inherits the count 100½ = 50. Then, in the reestimation step, the person

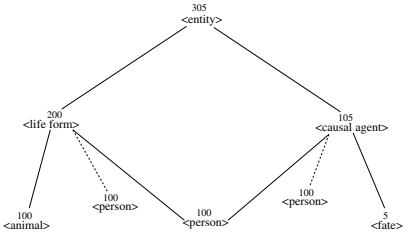

Figure 4: Breaking a DAG into a tree structure

count is divided relative to the frequencies of the parents:

<life_form> gets 100 x 150 / (150+55) = 73.13,
while <causal_agent> receives
100 x 150 / (150+55) = 26.83 from <person>.

(The counts for <animal> and <fate> are completely propagated to their respective parents.) Note that the count for the top node <entity> does not change. It corresponds to the unbiased total frequency of the data.

In addition, the count of a child concept $ncpt_c$ has to be apportioned among the different (virtual or real) copies of it which emerge from breaking the DAG into a tree. In the tree structure, each copy of $ncpt_c$ has exactly one parent $ncpt_p$ and receives that portion of the original frequency $freq(ncpt_c)$ that has been propagated to $ncpt_p$, i.e. $freq(ncpt_c)p(ncpt_p|ncpt_c)$. Likewise, the corresponding copies of the descendants of $ncpt_c$ have to be scaled by $p(ncpt_p|ncpt_c)$. Figure 5 illustrates this adjustment for the copies of <person>.

A possible intuitive access to the general idea that the count of a concept is divided among its parents might be to understand hyperonymy in a more 'subjective' manner than usual: Instead of „is a kind of', a hyperonymy relation could be interpreted as 'is perceived / referred to as'. This means that multiple hyperonyms represent different aspects of a concept which might have different salience. For example, a person might be primarily referred to as a life form in some situations (e.g. in an utterance like 'How many persons died? '), and as a causal agent in other situations (e.g. in 'This person caused the accident.'). The probabilities p(<life_form>|<person>) and p(<causal_agent>|<person>), together with the corresponding split of the count of <person>, model the relative salience of these two aspects w.r.t. <person>. The way proposed here to estimate these probabilities employs the only empirical quantitative information about the parent concepts that is available: their overall frequency. A parent that has a high frequency (compared to the other parents) gets a high probability, while a parent with a (comparably) low frequency is assigned a low probability. The count of a parent concept reflects its 'global' salience (w.r.t. the training data); the comparison with the counts of the other parents reflects peculiarities of their common child.

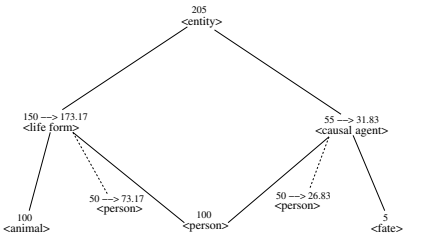

Figure 5: Reestimating frequencies

More formally, the approach described here can be viewed as performing a soft classification of noun senses. The concepts can be regarded as soft classes of senses, and multiple hyperonymy corresponds to graded membership. For example, all instances of <person> are graded members of both classes <life_form> and <causal_agent>. The degree of membership is represented by p(<life_form>|<person>) and p(<causal_agent>|<person>), respectively.

4 Experiments

This section describes experiments I carried out to test the effect of employing the two frequency estimation methods sketched in section 3 for acquiring selectional preferences using the tree cut approach. As mentioned, the method using equation (7) (henceforth called 'Simple') multiplies frequency counts of noun senses which are covered by duplicated concepts, while the approach using equations (8)–(10) (henceforth called 'Reestimation') avoids such a bias. For the experiments, I used GermaNet as semantic hierarchy. As noted in section 1, multiple inheritance

Reestimation			Simple		
concept	pref. value	prob.	concept	pref. value	prob.
<?kognitives_Objekt>	1.25	0.21	<Entität>	0.69	0.32
<Objekt>	0.42	0.13			
<Verhältnis#Relation>	0.94	0.02	<Verhältnis#Relation>	0.61	0.03
<Stelle#Ort#Stätte>	0.36	0.02	<Stelle#Ort#Stätte>	0.42	0.02
<Motiv#Intention>	0.43	0.004	<Motiv#Intention>	0.45	0.003
<Menge>	0.66	0.02	<Menge>	0.50	0.03
<Situation>	0.65	0.16	<Situation>	0.83	0.20
<Besitz>	0.02	0.0006	<Besitz>	0.04	0.001
<Zustand>	0.51	0.009	<Zustand>	1.13	0.02
<Attribut#Eigenschaft>	5.10	0.40	<Attribut#Eigenschaft>	4.77	0.37

Table 1: Tree cut models for 'wissen'

is a common structural property in this resource. This suggests that the bias which the Simple approach imposes on the frequency estimates is significant when applied to GermaNet. The experiments described below aim at verifying this hypothesis.

4.1 Setting

The experiments acquired selectional preferences for the object of several verbs. The training data I used were extracted from parsed relative clauses and verb-final clauses originating from a large German newspaper corpus. This parsed corpus was created at the IMS, University of Stuttgart. From these sentences, I extracted verb–noun (object) pairs (666,831 altogether). To avoid the problem of data sparseness, I acquired selectional preferences for those verbs which occur at least 500 times in the training set (261 verbs). For preference acquisition, I used a modified version of the tree cut approach described in Abe & Li 1996.[5] This modification involves an additional parameter that can be varied to influence the generalisation level of the induced cut (cf. Wagner 2000 or Wagner 2002) for details of this modified approach). With this parameter, I forced the algorithm to select the cut at or close to the highest possible level of abstraction, which comprises the top concepts of GermaNet. This is a conservative proceeding, since differences in tree cuts are much more likely if they tend to be located at low levels in the hierarchy, capturing peculiarities of very specific concepts.

Concerning frequency estimation, I carried out the experiments once using the Simple approach and once using the Reestimation approach (after the initial iteration using equation (10), I performed one reestimation iteration).

4.2 Results

The results show considerable differences between the selectional preferences acquired using the Simple and the Reestimation approach, respectively. First of all, it turned out that Simple yielded significantly higher total frequency counts at the hierarchy root for each verb than Reestimation: The average total count per verb was 1300.35 for Simple vs. 1149.55 for Reestimation. This means that Simple artificially increased the total count of the data by 13%. A more interesting question is to what extent the preferences acquired with the two approaches are different. Comparing the individual concepts which are classified as being preferred (preference value > 1),

the difference is considerable. For the whole set of 261 test verbs, Simple acquired 1085 preferred concepts, Reestimation 1087 preferred concepts altogether. Of these, 924 concepts were equal. This amounts to a difference of 15%. At first glance, this does not seem too much. But taking into account that the cuts comprise concepts at a very high generalisation level, the difference is remarkable. Looking at the complete preference profiles acquired for each verb, the picture becomes much more clear-cut. Only for 99 verbs, i.e. 38% of the test verbs, the two methods yielded the same set of preferred concepts.

As an example, table 1 shows the tree cut models acquired for 'wissen' (to know). Both models classify the concept <Attribut#Eigenschaft > (attribute, property) as preferred. The Reestimation cut also models <?kognitives_Objekt> (cognitive object) as preferred concept, which is in accordance with human intuition. The Simple cut does not contain this concept, since it is located one level higher, at <Entität> (entity), which subsumes <?kognitives_Objekt> and <Objekt> (object). However, the Simple model classifies <Zustand> (state) as preferred, which is much less intuitive. The probability distributions $p(ncpt|wissen)$ of the concepts on the two cuts are rather similar, though some differences (e.g. 0.16 versus 0.20 for <Situation>) might matter when employed for a particular application.

Altogether, the experiments verify that the Simple results differ significantly (though not dramatically) from the Reestimation results.

5 Conclusion

In this paper, I discussed different methods for estimating frequencies of concepts in wordnets from corpus data. Based on an example NLP task (selectional preference acquisition), I illustrated that the selection of an appropriate frequency estimation method largely depends on the statistical methods that employ the induced frequencies. In particular, this paper focused on the problems which multiple inheritance in wordnets impose on concept frequency estimation. Two of the discussed methods, word-to-concept and word-to-sense, are suitable for multiple inheritance hierarchies without modification. These approaches rest on the subsumption relation between words and concepts rather than the immediate hyperonymy relation and thus are compatible with DAG structures. However, the tree cut approach requires a concept hierarchy that exhibits a pure tree structure. To apply this approach to a wordnet requires a transformation of the wordnet's DAG structure. I discussed the most commonly used ad-hoc strategy for this transformation. This strategy leads to biases of the estimated frequency counts, which are evoked just by the multiple inheritance structure. Therefore, I proposed a more sophisticated EM-style strategy which involves the adjustment and reestimation of frequency counts. Experiments showed that the bias imposed by the ad-hoc approach is significant.

For future work, it will be interesting to test the performance of the different frequency estimation approaches w.r.t. particular NLP tasks. For example, selectional preferences acquired by the two approaches tested in section 4 could be employed for lexical or structural disambiguation. A priori, it is not clear whether the mathematically sound approach which I proposed performs better than the simple ad-hoc approach. This has to be examined empirically. In any case, the issue of concept frequency estimation should not be disregarded.

Notes

[1] This is a simplification because it does not take into account that 'person' is ambiguous. The example only takes the 'human' sense of the word into account. If the data are not lexically disambiguated, which is mostly the case, then the frequency of "person" has to be equally divided among all concepts which subsume any sense of the word.

[2] Again, this simplified example does not take ambiguity into account. If a word is ambiguous (in fact, 'person' is) and the data are not disambiguated, then the count of a word is equally divided among its senses.

[3] This requirement follows from the peculiarity that the MDL approach employs tree cut models for efficiently encoding the training data, in order to compare the performance of alternative models w.r.t. data compression. This only works properly if all possible tree cut models capture the whole amount of data.

[4] For example, if the cut contained <life_form> and <causal_agent>, then, assuming the WordNet structure depicted in figures 1 and 2, the senses subsumed by <person> would be represented twice.

[5] As mentioned, this approach employs equation (3) to compute preference values.

References

Abe, N.; Li, H. (1996). "Learning Word Association Norms Using Tree Cut Pair Models." In: Proceedings of 13th International Conference on Machine Learning, Bari, Italy, July 1996.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Kunze, C.; Wagner, A. (2001). „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche." In: Lemberg, I.; Schröder, B.; Storrer, A. (eds.) (2001). Chancen und Perspektiven computergestützter Lexikographie. Tübingen: Niemeyer [= Lexicographica Series Maior Vol. 107], 229-246.

Li, H.; Abe, N. (1998). "Generalizing Case Frames Using a Thesaurus and the MDL Principle." In: Computational Linguistics 24(2) (1998), 217-244.

McCarthy, D. (2001). Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Ph.D. thesis, University of Sussex, Brighton, UK.

Resnik, P. (1998). "WordNet and Class-based Probabilities." In: Fellbaum (1998), 239-263.

Ribas, F. (1995). On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy. PhD thesis, Universitat Politecnica de Catalunya, Barcelona, Spain.

Wagner, A. (2000). "Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis." In: Proceeding of ECAI-2000 Workshop on Ontology Learning. Berlin, August 2000, 37-42.

Wagner, A. (2002). "Learning Thematic Role Relations for Wordnets." In: Proceeding of ESSLI 2002 Workshop on Machine Learning Approaches in Computational Linguistics. Trento, Italy, 99-113.

Diana Steffen, Bogdan Sacaleanu, Paul Buitelaar

Domain Specific Sense Disambiguation with Unsupervised Methods

Abstract

Most approaches in sense disambiguation have been restricted to supervised training over manually annotated, non-technical, English corpora. Application to a new language or technical domain requires extensive manual annotation of appropriate training corpora. As this is both expensive and inefficient, unsupervised methods are to be preferred, specifically in technical domains such as medicine. In the context of a project in the medical domain, we developed and evaluated two unsupervised methods for sense disambiguation.

1 Introduction

Although a lot of work on sense disambiguation has been reported in recent years (for an overview, see: Ide & Véronis 1998; Kilgarriff & Palmer 2000; Preiss & Yarowsky 2001), most of these approaches are restricted to supervised training over manually annotated, non-technical, English corpora like SEMCOR (Fellbaum 1997) and DSO (Ng & Lee 1996). Application of such systems to a new language or technical domain requires extensive manual annotation of appropriate training corpora. As this is both expensive and inefficient, unsupervised methods are to be preferred, specifically in technical domains such as medicine.

In the context of a project on cross-language information retrieval (CLIR) in the medical domain, we developed two unsupervised methods for sense disambiguation. The project is concerned with a systematic comparison of concept-based and corpus-based methods in medical CLIR. Primary goals of the project are: 1. to develop and evaluate methods for the effective use of multilingual semantic resources in the semantic annotation of English and German medical texts; 2. to subsequently evaluate and compare the impact of semantic information on the retrieval of these annotated texts.

The semantic resources used are UMLS[1] (Unified Medical Language System), a multilingual database of medical terms, and EuroWordNet (Vossen 1997), which interconnects a number of wordnets for several European languages. However, given that the size of the German part in EuroWordNet is rather small, all our experiments reported here on development of a sense disambiguation system use the considerably larger GermaNet (Hamp & Feldweg 1997) database instead.

For our experiments we used a corpus of German medical scientific abstracts, obtained from the Springer Link web site[2]. The corpus consists approximately of 1 million tokens. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.).

The two unsupervised methods are described in the next section, followed by a detailed overview of experiments and results in Section 3., and an outlook on future work in Section 4.

2 Methods

2.1 Domain Specific Sense

Within the context of a specific technical domain, one of the senses of an ambiguous word may be more appropriate in the context of this domain than the other senses (Cucchiarelli & Velardi 1998; Magnini & Strapparava 2000; Magnini et al. 2001; Buitelaar 2001; Buitelaar & Sacaleanu 2001). Here, we describe

a method that automatically determines such a domain specific sense on the basis of its statistical relevance across several domain specific corpora.

For this purpose, we first compute the domain relevance of each term and use this information to compute the cumulative relevance of each sense. As senses in GermaNet correspond to sets of similar terms (i.e. *synsets*), we may compute the relevance of each synset in which domain specific terms occur. This allows for a ranking of synsets (senses) according to domain relevance.

The relevance measure used in this process is a slightly adapted version of standard *tf.idf*, as used in vector-space models for information retrieval (Salton & Buckley 1988):

$$rlv(t \mid d) = (1 + \log(tf_{t,d})) * \log(\frac{N}{df_t})$$

where t represents the term, d the domain (corpus), N is the total number of domains (corpora) taken into account. Term frequency has been scaled logarithmically because more occurrences of a word indicate higher importance, but not as much importance as the count solely would suggest. By scaling domain frequency as well, this formula gives full weight to terms that occur in just one domain and a weight of zero to those occurring in all domains.

Given term relevance, we are now able to compute the relevance of each synset. This is simply the sum of the relevance of each term in the synset, which may be defined as follows:

$$rlv(c \mid d) = \sum_{t \in c} rlv(t \mid d)$$

However, suppose we want to compute the relevance for the following senses (i.e. the synsets in which this term occurs) of `Zelle`:

```
[Zelle,Gefängniszelle]
     prison cell
[Zelle] living cell
```

Although `Zelle` will have a high relevance in the medical domain, the occurrence of `Gefängniszelle` in this domain is very unlikely and therefore the relevance value of both concepts will be equal. Although the latter concept is more relevant to the medical domain, we would not be able to automatically determine this by merely adding up the relevance of the terms in each of the synsets. Therefore we reconsidered the concept relevance definition to take into account the number of terms in the synset that actually occur in the domain corpus:

$$rlv(c \mid d) = \frac{T}{|c|} \sum_{t \in c} rlv(t \mid d)$$

where T represents the lexical coverage, and $|c|$ is the length of synset c. This relevance measure reflects the intuition that if many terms in the synset occur in the domain, then the more likely it is that the synset is relevant for that domain.

To increase the number of terms to be found within a domain corpus, we considered adding hyponyms to a given synset as these are often directly related. For example, the two synsets for *Zelle* can be extended with hyponyms as follows:

```
[Zelle,Gefängniszelle,
     Todeszelle]
[Zelle,Körperzelle,
     Pflanzenzelle]
```

Adding hyponyms changes the relevance formula accordingly:

$$rlv(c+ \mid d) = \frac{T}{|c|} \sum_{t \in c+} rlv(t \mid d)$$

where c+ is the extended synset. Note that T (number of terms in the concept that occur in the domain) and $|c|$ (number of terms in the synset) have not changed. That is, hyponyms do not affect lexical coverage, but only add to the summed weight of the synset.

2.2 Instance-Based Learning

2.2.1 Introduction

The second method we used in our experiments implements a k-nearest neighbor instance-based learning algorithm using the WEKA[3] suit of machine-learning tools (Witten & Eibe 2000). In this method, sense disambiguation is seen as a classification task, in which an ambiguous word needs to be classified to the appropriate class given a particular context.

There have been several reports on the use of instance-based learning in sense disambiguation (Ng & Lee 1996; Mihalcea 2002). However, all of these approaches were supervised, based on a manually annotated training corpus. Here we report on the use of instance-based learning in an unsupervised manner by generalizing over Resnik`s work on selection restrictions (Resnik 1997).

The basic idea is as follows. Consider these (ambiguous and non-ambiguous) instances of the verb `drink` in the context of the semantic classes (i.e. senses) `FOOD` and `LIQUID`:

```
He drank coffee <LIQUID>
He drank tea <LIQUID>
He drank chocolate
     <FOOD,LIQUID>
```

From these examples we may infer that the verb drink has a preference for the semantic class `LIQUID`. We can apply this in the disambiguation of the following ambiguous instance (and select `LIQUID`):

```
He drank Java
     <GEOGAPHICAL,LIQUID>
```

2.2.2 Algorithm

An instance-based learning algorithm consists of a training step and an application step:

Training: Collecting classified instances from a training corpus (as our method is unsupervised, this corpus has not been previously annotated).

An instance is a set of attribute-value pairs, one of which identifies the class attribute. Classifying an instance then means finding the missing value for this class attribute.

Constructing an instance involves the following. Let *w* be a word in the training corpus. We can build instances for *w*, where the values of the attributes are always its left and right neighbor words in a context of size n, and the value of the class attribute varies over its senses.

Collecting classified instances from the training corpus may now be defined as follows. Given a training corpus annotated with part-of-speech and morphology, for any ambiguous word *w* and its set of senses *S*:

- Determine all contexts, in which *w* occurs organized according to part-of-speech pattern.
- For every part-of-speech pattern *p* collect all instances corresponding to contexts of *w* of pattern *p* in the training corpus, under the constraint that the value of the class attribute belongs to *S*.

To illustrate the construction of particular instances, consider the following sentence from our corpus:

In dem Fall, sind korrigierende Eingriffe nur eingeschränkt möglich.
(In this case the possibility of corrective surgery is limited.)

The ambiguous word *Eingriff* has the following two senses (identified by their GermaNet synset ids):

460326: Operation, Eingriff
388935: Eingriff, Intervention, Eingreifen

From the sentence we may now derive the following instances for *Eingriff* with context size 5 (2

words on the left, 2 words on the right) of the part-of-speech pattern

-, ADJ, NOUN, -, VERB

where *Eingriff* takes the position of the *NOUN* and '-' stands for other parts-of-speech:

[sein, korrigieren, nur, einschränken,388935]
[sein, korrigieren, nur, einschränken,460326]

Application: Classifying an occurrence of an ambiguous word *w* by finding the *k* most similar training instances:

- Determine its part-of-speech pattern *p*.
- Extract the corresponding set of instances *I(w,p)* as found in the training step.

For instance, the set:

[und,therapeutisch,werden,vorstellen,388935]
[und,therapeutisch,werden,vorstellen, 460326]
[ein,chirurgisch,nicht,profitieren ,388935]
[sein,korrigieren,stets,ermoeglichen , 460326]
[oder,offen,zu,erfassen, 460326]
[sein, korrigieren,nur,einschraenken,388935]
[sein, korrigieren,nur,einschraenken,460326]

- Delete all instances corresponding to the occurrence (i.e. instances for the occurrence that correspond to each sense – the last two instances in the example set), resulting in the set of instances *I´(w,p).*
- Create an instance for the occurrence, with the class attribute missing:

  [sein, korrigieren,nur,einschränken,?]

- Classify the instance using *I´(w,p).*

The algorithm does not return a specific sense, but a probability distribution over all senses of the ambiguous word. We assign the sense with highest probability to the corresponding word occurrence. If such a sense does not exist, no decision is made.

3 Evaluation

3.1 Evaluation Corpus

An important aspect in the development of a word sense disambiguation system is the evaluation of different methods and parameters. Unfortunately, there is a lack of test sets for evaluation, specifically for languages other than English and even more so for specific domains like medicine. Given that our work focuses on German text in the medical domain, we needed to construct an evaluation corpus specifically for this purpose.

Selection of ambiguous GermaNet terms to be included in the evaluation corpus proceeded in several steps. First, we calculated relevance values regarding the medical domain for all GermaNet noun synsets occurring in the medical corpus, using the method described in Section 2.1. Given these relevance values, we compiled a list of terms with high relevance, at least 100 occurrences in the medical corpus and with more than one synset in GermaNet. This produced a list of 40 terms, for each of which we then automatically extracted 100 occurrences at random.

Three annotators (a medical expert and two linguistics students) annotated the occurrences of the 40 ambiguous terms. They were allowed to annotate an occurrence with more than one sense if needed or with *undef*, if GermaNet did not contain any appropriate sense. With a further arbitration step to settle any disagreement cases they then produced together a gold standard. Removing the occurrences annotated with *undef* from the gold standard gave us the final evaluation corpus, which we used in our experiments.

3.2 Experiments

The evaluation corpus was used to experiment with the previously mentioned methods and a combination thereof. For each experiment we computed *recall* (number of correctly disambiguated occurrences divided by the number of occurrences to be disambiguated) and *precision* (number of correctly[4] disambiguated occurrences divided by the number of disambiguated occurrences). A theoretical baseline for the evaluation corpus was computed as follows, where GS means gold standard and GN means GermaNet:

$$prec_{random} = \frac{1}{|GS|} \sum_{o \in GS} \frac{|GS_{sense(s)}|}{|GN_{senses}|}$$

For every occurrence in the gold standard, the probability of assigning it the correct sense is computed by dividing the number of senses in the gold standard by the number of corresponding GermaNet senses. The average precision is the sum of all probabilities divided by the number of all occurrences. For our evaluation corpus the precision (= recall) is 36%, by a coverage of 100% (F-measure F1: 0.36).

3.2.1 Domain Specific Sense

The identification of domain specific senses has been evaluated as an individual component in (Buitelaar & Sacaleanu 2001). Here we evaluated this method as part of a broader sense disambiguation system. For all GermaNet senses in the training corpus we computed a domain relevance score, according to the method described in Section 2.1. We experimented with different sets of domain specific corpora and with different corpora sizes. The corpora used are:

sp	Springer (medical abstracts)
dp	Deutsche Presse Agentur (news)
fb	Fussball (soccer game reports)
wr	Wirtschaftswoche (economic news)
rd	Radiology (examination reports)

In disambiguation, the sense with the highest domain relevance was selected. If no sense had a relevance value, no decision was made. Table 1. shows the evaluation results for different corpora sets and sizes:

Corpora	Size	Rec	Prec	F1
sp-dp-fb-wr	2Mb	4%	77%	0.08
sp-dp	2Mb	6%	99%	0.11
sp-dp	10Mb	4%	26%	0.07
rd-dp-fb-wr	2Mb	17%	44%	0.24
rd-dp	2Mb	9%	50%	0.15
rd-dp	10Mb	3%	34%	0.05

Table 1: Domain Specific Sense.

Unfortunately, F-measure results show that none of the experiments actually improve on the baseline mentioned above. However, as will be discussed in the next section, a combination of this method with the instance-based learning method does result in an improvement if compared to the use of instance-based learning by itself.

In terms of recall and precision we can observe the following. Precision reaches highest values when the domain specific corpora are small (i.e. 2 Mb). Large corpora have a correspondingly large set of common terms, for which the relevance score will be zero[5] – see Section 2.1.

3.2.2 Instance-Based Learning (IBL)

In the training and application steps we experimented with four parameters.

- **Training Corpus:**
  Springer (S) vs. Radiology (R)
  We were interested to see how well our system performs when training and application use the same corpus compared to when the training corpus (Radiology reports) is different from the test corpus (Springer medical abstracts), but still belonging to the same domain.

- **Context Size:** 3 vs. 5 words
  We were interested to measure the effect of larger vs. smaller contaxt sizes. Larger contexts give a higher precision, but will have less instances – with correspondingly fewer occurrences that can be disambiguated .

- **Part-of-Speech Selection:** all PoS (all) vs. only nouns, verbs or adjectives (N/V/A)
  We wanted to find out if words with little content have any influence on the disambiguation result. In order to discard them in some experiments we gave all attributes corresponding to parts-of-speech other than N/V/A the value *null*.

- **Attribute-Values:** lemmas vs. lemmas and synsets
  It is hard to classify instances with attribute-values (i.e. particular lemmas), which do not occur in the training corpus. We introduced synsets (i.e. senses) as values for these attributes. This in effect maps a particular lemma to a set of lemmas, thereby reducing this sparse data problem.

Corpus; Context Size, PoS		Lemma			Lemma + Synsets		
		Rec	Prec	F1	Rec	Prec	F1
S	3-all	30%	49%	0.37	29%	48%	0.36
	3-N/V/A	17%	43%	0.24	17%	43%	0.24
	5-all	18%	54%	0.27	17%	53%	0.26
	5-N/V/A	21%	47%	0.29	21%	48%	0.29
R	3-all	21%	43%	0.28	21%	43%	0.28
	3-N/V/A	13%	44%	0.20	13%	44%	0.20
	5-all	13%	42%	0.20	13%	43%	0.20
	5-N/V/A	16%	48%	0.24	16%	46%	0.24

Table 2: Instance-Based Learning

Training Corpus: As we expected, precision and recall are better when the training corpus is the same with the test corpus.

Context Size: We cannot say much about recall if we only consider the context size. This is only relevant together with the part-of-speech selection. Best recall values are reached with context size 3 and all parts-of-speech, followed by context size 5 with nouns, verbs and adjectives. On the other hand, precision will be highest when using larger contexts (5), as these will contain more words that contribute to the selection of a particular sense.

Part-of-Speech Selection: With contexts of size 3 precision values are better when using all parts-of-speech. This makes sense, because very often in small contexts no noun, verb or adjective occurs and therefore we can not build any useful training instances. With context size 5, different training corpora produce different results. For Springer, the results are better when using all parts-of-speech (54% vs. 47%), while for Radiology the use of only nouns, verbs and adjectives will do a better job (48% vs. 42%).

Attribute-Values: Using synsets as values in addition to lemmas does not bring any improvement, but rather some slight degradation of results.

3.2.3 Combination of Methods

Here we used the domain relevance values which led to the best results in the first set of experiments and the sets of training instances generated for the second set of experiments. For every occurrence of an ambiguous word we applied the two methods disjunctively, that is, if the first method could not make any decision, the second one was applied. The order in which the methods were applied is an extra parameter.
Table 3 shows the results.

It is interesting to note that instance-based learning produces better results (precision as well as recall) in combination with the domain specific sense method. In these experiments we used the domain specific senses that were computed with the corpus set *sp-dp-fb-wr* and a corpus size of *2Mb*.

Corpus; Context Size, PoS		IBL → DOMSpecSense			DOMSpecSense → IBL		
		Rec	Prec	F1	Rec	Prec	F1
S	3-all	35%	52%	0.42	35%	53%	0.42
	3-N/V/A	23%	50%	0.31	24%	51%	0.33
	5-all	25%	60%	0.35	25%	60%	0.35
	5-N/V/A	27%	53%	0.36	27%	53%	0.36
R	3-all	29%	45%	0.35	28%	44%	0.34
	3-N/V/A	24%	44%	0.31	24%	43%	0.31
	5-all	24%	46%	0.31	24%	44%	0.31
	5-N/V/A	27%	49%	0.35	26%	46%	0.33

Table 3: IBL and Domain Specific Sense

In comparison to the domain specific sense method, recall is much better in combination with instance-based learning, which was of course to be expected. However, precision (highest at 60%) does not reach the highest result (77%) that we saw when using the domain specific sense method by itself.

Finally, we note that the order of applying the two methods has some significance. Applying the instance-based learning method first produces slightly better results than applying the domain specific sense method first. This may result from the fact that the domain specific sense method always selects the same sense for every occurrence whereas the instance-based learning method selects a sense depending on a particular context.

3.3 Related Work

Unfortunately, a straightforward comparison of our work with other related work in sense disambiguation is not possible, as German medical language has not been studied widely in this respect. Nevertheless, some work has been done on sense disambiguation in other languages, primarily for English (Rindflesch et al. 1994; Weeber et al. 2001; Liu et al. 2001), but also for instance for French (Bouillon et al. 2000). The work most similar to our work is that of (Liu et al. 2001), who also report on an unsupervised method for sense disambiguation in medical text The object of this study, however, is the ambiguity of medical terms as specified in the medical semantic resource UMLS whereas we report on the disambiguation of more general terms as used in medical text.

4 Future Work

Sense disambiguation is concerned with the selection of the appropriate interpretation of a word in respect of a given semantic lexicon. Obviously this implies that the word at hand is represented in the lexicon, which is often not the case. However, in semantic tagging, the task of mapping words to semantic classes, we would like to tag each word and not only those occurring in the given semantic lexicon.

Therefore, in addition to sense disambiguation of *known* words we need to classify *unknown* words. As senses may be simply viewed as semantic classes, these tasks can also be combined. In this respect we intend to treat classification of unknown words as sense disambiguation between a dynamically selected set of domain specific senses (i.e. semantic classes).

5 Conclusions

In this paper we describe two unsupervised methods to sense disambiguation of terms in medical text. The first method automatically determines a domain specific sense on the basis of its statistical relevance across several domain specific corpora. The second approach implements a k-nearest neighbor instance-based learning algorithm. Experiments with an evaluation corpus built specifi-

cally for this task show that a combination of the two methods produces the best results.

Acknowledgements
This research has in part been supported by EC/NSF grant IST-1999-11438 for the MUCHMORE project.

Notes

1 National Library of Medicine (2004). Unified Medical Language System (UMLS) Homepage. *http://www.nlm.nih.gov/research/umls/umlsmain.html* [accessed April 2004].

2 Springer Verlag (2004). SpringerLink Homepage. *http://link.springer.de/* [accessed April 2004].

3 Weka Machine Learning Project (2004). Weka Homepage. The University of Waikato, New Zealand, *http://www.cs.waikato.ac.nz/~ml/* [accessed April 2004].

4 If an occurrence was assigned several senses by the human annotators and the system delivered one of them, we counted the occurrence as correct.

5 In further experiments we intend to adjust the term relevance measure so as to assign a non-zero weight even to those terms occurring in all domains.

References

Bouillon, P. et al. (2000). "Indexing by Statistical Tagging." In: Proceedings of 5es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2000), Lausanne, Switzerland, March 2000.

Buitelaar, P.; Sacaleanu, B. (2001). "Ranking and Selecting Synsets by Domain Relevance." In: Proceedings of the NAACL Workshop on WordNet and Other Lexcical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, June 2001.

Buitelaar, P. (2001). "The SENSEVAL-II Panel on Domains, Topics and Senses." In: Preiss, J.; Yarowsky, D. (eds.) (2001). Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-II), Toulouse, France, June 2001.

Cucchiarelli, A.; Velardi, P. (1998). "Finding a Domain-Appropriate Sense Inventory for Semantically Tagging a Corpus." In: Natural Language Engineering 4(4) (1998), 325-344.

Fellbaum, Ch. (1997). "Analysis of a hand-tagging task." In: Proceedings of the ACL SIGLEX / ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D. C., April 1997.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Ide, N.; Véronis, J. (eds.) (1998). "Word Sense Disambiguation". In: Computational Linguistics 24(1) (1998) [Introduction to a Special Issue on Word Sense Disambiguation].

Kilgarriff, A.; Palmer, M. (2000). "Introduction." In: Computers and the Humanities 34(1/2) (2000), 1-13 [Introduction to the special issue on SENSEVAL].

Liu, H.; Lussier, Y.; Friedman, C. (2001). "Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method." In: Journal of Biomedical Informatics, 34(4) (2001), 249-261.

Magnini, B.; Strapparava, C. (2000). "Experiments in Word Domain Disambiguation for Parallel Texts." In: Proceedings of the ACL-SIGLEX Workshop on Word Senses and Multi-linguality, Hong Kong, October 2000.

Magnini, B.; Strapparava, C.; Pezzulo G.; Gliozzo, A. (2001). "Using Domain Information for Word Sense Disambiguation." In: Preiss, J.; Yarowsky, D. (eds.) (2001). Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-II), Toulouse, France, June 2001.

Mihalcea, R. (2002). "Instance Based learning with Automatic Feature Selection Applied to Word Sense Disambiguation." In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipeh, August / September 2002.

Miller, G.; Chodorow, M.; Landes, S.; Leacock, C.; Thomas, R. (1994). "Using a Semantic Concordance for Sense Identification." In: ARPA Workshop on Human Language Technology, Plainsboro, NJ, March 1994.

Miller, G. A. (1995). "WordNet: A Lexical Database for English." In: Communications of the ACM 38(11) (1995), 39-41.

Ng, H.T.; Lee, H.B. (1996). "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-based Approach." In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), Santa Cruz, CA, June 1996, 40-47.

Preiss, J.; Yarowsky, D. (eds.) (2001). Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-II), Toulouse, France, June 2001.

Resnik, P. (1997). "Selectional Preference and Sense Disambiguation." In: Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What. and How?, Washington D.C., April 1997.

Rindflesch, T. C.; Aronson A. R. (1994). "Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus." In: Ozbolt, J.G. (ed.) (1994). Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 240-244, *http://nls5.nlm.nih.gov/pubs/scamc94.pdf* [accessed April 2004].

Salton, G.; Buckley, Ch. (1988). "Term-Weighting Approaches In Automatic Text Retrieval." In: Information Processing & Management 24(5) (1998), 515-523.

Vossen, P. (1997). "EuroWordNet: A Multilingual Database for Information Retrieval." In: Proceedings of the 3rd Delos Workshop. Cross-Language Information Retrieval. Zurich, Switzerland, March 1997 [= ERCIM Workshop Proceedings - No. 97-W003].

Witten, I. H. & Eibe F. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco / CA: Morgan Kaufmann [The Morgan Kaufmann Series in Data Management Systems].

Weeber M.; Mork J., Aronson A.R. (2001). "Developing a Test Collection for Biomedical Word Sense Disambiguation." In: Proceedings of The American Medical Informatics Association 2001 Symposium (AMIA 2001), Washington / DC, November 2001, 746-50, *http://lhncbc.nlm.nih.gov/lhc/docs/published/2001/pub2001054.pdf* [accessed April 2004].

Christian Biemann, Stefan Bordag, Uwe Quasthoff

Lernen paradigmatischer Relationen auf iterierten Kollokationen

Abstract

Das Lernen paradigmatischer Relationen wie Synonymie, Homonymie, Antonymie und Hyponymie ist Thema verschiedener statistischer Ansätze. Die bisherigen Ansätze verwenden nur je ein statistisches Feature, um derartige Relationen aus großen Textkorpora zu extrahieren. In diesem Papier soll eine Architektur vorgestellt werden, die es ermöglicht, Relationen zwischen Wörtern durch eine Trainingsmenge zu lernen, um weitere in der Relation stehende Wörter zu erhalten, um schließlich lexikalisch-semantische Wortnetze automatisch oder halbautomatisch zu erweitern. Hierzu wird zunächst eine passende Menge von Features aus einer großen Menge vorhandener Features aufgrund der Trainingsmenge ausgewählt, statistisch getestet und zum Erweitern des Wortnetzes verwendet.

1 Einleitung

Ein in den letzten Jahren wiederholt angesprochenes Problem der linguistischen Datenverarbeitung ist das „acquisition bottleneck“: Die meiste Zeit wird darauf verwendet, lexikalische Ressourcen manuell aufzubauen. Wird auch die Wiederverwendung durch Standards wie MLEXd (siehe Ahmad 1994) oder TEI (Sperberg-McQueen et al. 2002) unterstützt, so benötigt dennoch beinahe jede neue Anwendung andere Angaben in den jeweiligen Lexika oder Wortnetzen.

Wir wollen einen Ansatz vorstellen, wie vorhandene Wortnetze automatisch bzw. semiautomatisch unter Zuhilfenahme großer Korpora erweitert werden können.

Eine Teilaufgabe bei der Erweiterung von Ressourcen wie GermaNet (*http://www.sfs.uni-tuebingen.de/lsd/Intro.html*) ist das Zuordnen von bisher nicht aufgenommenen Wörtern zu vorhandenen Synsets oder Gruppen von Synsets.

Vorhandene Verfahren zum Auffinden von Synonymen oder Quasi-Synonymen in großen Dokumentenkollektionen wie Ruge 1997 oder Rapp 2002 nehmen an, dass ähnliche Wörter in ähnlichen Umgebungen zu finden sind und benutzen als Umgebung einerseits Dependenzen, andererseits das gemeinsame Vorkommen in Sätzen.

Nach Eingabe eines Wortes liefern derartige Verfahren eine gerankte Liste von ähnlichen Wörtern. Die Qualität ist jedoch für unsere Zwecke nicht ausreichend, da zwar Synonyme im Allgemeinen gefunden werden, aber auch viele anderweitig verwandte Wörter.

Unser Ansatz versucht, durch den Vergleich mehrerer derartiger Listen die Synonyme von den übrigen Wörtern zu trennen.

Der Ansatz ähnelt dem in Chiaramita 2002, wo versucht wird, mit Hilfe von morphologischen Merkmalen Wörter zu weit gefassten Wordnet-Synsets zuzuordnen, benutzt jedoch mehr Features.

2 Verfahren

Unsere Aufgabe ist das Erweitern von Wortmengen, z.B. Synsets aus GermaNet. Das Vorgehen gliedert sich in zwei Schritte: Im ersten Schritt werden mittels verschiedener Verfahren weitere Wortmengen erzeugt, deren Elemente sich möglicherweise zur Erweiterung einer vorgegebenen Menge eignen. Da möglicherweise diese Mengen von unterschiedlicher Qualität sein können, werden im zweiten Schritt daraus geeignete Mengen selektiert und aus den Mengen geeignete Elemente zur Erweiterung extrahiert.

2.1 Mengen konstruieren

Die uns interessierenden Synsets bestehen in der Regel aus (Quasi-)Synonymen. D.h., die enthaltenen Wörter stimmen in vielen wichtigen semantischen Eigenschaften überein, unterscheiden sich möglicherweise aber auch in einem Attribut.

Zur Konstruktion der im zweiten Schritt zur Erweiterung verwendeten Mengen benutzen wir nun Verfahren, die korpusbasiert möglichst semantisch homogene Wortmengen liefern. Semantische Homogenität soll hier bedeuten, das sich die Elemente einer solchen Menge möglichst einfach beschreiben lassen, beispielsweise durch Variation nur eines semantischen Attributes.

Die automatische Konstruktion solcher Mengen ist selbstverständlich schwierig, in der Regel entstehen Mengen, die diese Eigenschaft nur näherungsweise erfüllen. Bei vielen Verfahren entsteht zusätzlich ein Ranking, d.h. ein numerischer Wert, welcher die Rangordnung in der Menge beschreibt. Hier kann mit einem Schwellwert Einfluss auf die Größe der erzeugten Mengen genommen werden. Mit einem hohen Schwellwert kann möglicherweise eine höhere Qualität der erzeugten Mengen gesichert werden. Im Abschnitt 2 werden mehrere Verfahren zur Erzeugung solcher Mengen vorgestellt.

2.2 Synsets erweitern

Im zweiten Schritt sollen aus unserem eben erzeugten Reservoir an semantisch möglichst homogenen Mengen diejenigen ausgewählt werden, die sich zur Erweiterung eines gegebenen Synsets eignen. Das Problem besteht darin, dass zwar auch das Synset semantisch homogen ist (oder wenigstens sein sollte), aber die semantische Homogenität kann sich auf ein anderes Attribut beziehen. Zur Erweiterung eignen sich also nur solche Mengen, die bezüglich desselben Attributs semantisch homogen sind wie das vorgegebene Synset. Da weder die semantischen Attribute noch ihre Werte bekannt sind, muss die Übereinstimmung an Hand gleicher Elemente getestet werden: Zwei semantisch homogene Mengen haben diese Eigenschaft bezüglich desselben Attributs, falls sie möglichst viele Elemente gemeinsam enthalten.

Unser Vorgehen gestaltet sich damit folgendermaßen: Um ein gegebenes Synset $S=\{s_1, ..., s_n\}$ zu erweitern, suchen wir in unserem Reservoir an semantisch homogenen Mengen eine Menge $M=\{m_1, ..., m_k\}$ mit möglichst großem Durchschnitt $S \cap M$. Die zusätzlichen Elemente $M \setminus S$ sind die Kandidaten für die Erweiterung von S.

Falls sich die erzeugte Erweiterungsmenge M als nicht ausreichend semantisch homogen erweist, lassen sich alternativ zunächst mehrere Mengen $M_1, ..., M_r$ auswählen und zur tatsächlichen Erweiterung nur die Durchschnittsmenge $M=M_1 \cap ... \cap M_r$ benutzen.

Die Schnittmengenoperation kann an dieser Stelle auch derartig abgeschwächt werden, dass ein Wort auch dann in der Ergebnismenge zu finden ist, wenn es in den meisten Mengen M_i mit hohem Rang vorkommt.

Für das Verfahren sind an zwei Stellen ausreichende Datenmengen nötig: zum einen ein ausreichend großes Korpus, um nicht am *data sparseness problem* für statistische Verfahren zu scheitern, zum anderen eine genügend große zu erweiternde Wortmenge, um die Art der Homogenität automatisch erfassen zu können.

Letzteres Problem lässt sich für GermaNet derart behandeln, dass nicht z.B. Synsets als Wortmenge erweitert werden, sondern Vereinigungen von Synsets, die in der Hierarchie benachbart liegen.

Große Korpora stehen bereit, siehe dazu Abschnitt 2.3.1.

2.3 Wortmengen verschiedener Eigenschaften

2.3.1 Kollokationen

Seit 1993 wird im Rahmen des Projekts *Deutscher Wortschatz* eine umfangreiche Textsamm-

lung (siehe Quasthoff 1998) gepflegt, für welche u.a. statistische Kollokationen auf Satzbasis und Nachbarschaftsbasis berechnet werden, siehe dazu (Heyer et al. 2001).

Zu jedem Wort kann eine nach Signifikanz geordnete Liste von Wörtern extrahiert werden, die mit diesem Wort statistisch auffällig gemeinsam auftreten. Dieses gemeinsame Auftreten wird dabei unterschieden nach unmittelbaren linken und rechten Nachbarn, sowie dem Auftreten im gleichen Satz.

Durch solche Kollokationen werden zwar menschliche Assoziationen reflektiert, also semantische Zusammenhänge ausgedrückt, aber in der Regel ist eine solche Kollokationsmenge semantisch recht inhomogen, weil verschiedenartige Assoziationen möglich sind. Diese Inhomogenität wird bei einer möglichen Polysemie des Ausgangswortes natürlich noch verstärkt.

Aus diesen Gründen sind reine Kollokationsmengen für die Erweiterung von Synsets zunächst ungeeignet, wie die Kollokationsmenge für „Witterung“ aufzeigt (Die Zahlen in Klammern geben Signifikanzen an):

Kollokationen für Witterung:

kühler (115), ungünstiger (85), schlechter (81), milde (77), kühle (72), kalte (69), milden (69), kühlen (66), wegen (62), feuchte (57), schlechten (48), warme (43), naßkalten (40), naßkalter (39), günstiger (38), warmen (38), kalten (37), aufgenommen (36), feuchter (35), kalter (35), trotz (35), Unbilden (34), ungünstigen (29), warmer (28), ungünstige (27), Wegen (26), aufgrund (24), günstige (24), Bei (22), Feldberg (22), anhaltend (22), zuläßt (21), Jahreszeit (20), Temperaturen (20), Ts (20), ausgesetzt (20), naßkalte (20), ungewöhnlich (19), je (18), schlechte (17), trockener (17), Sommer (16), angenehmer (16), günstigen (16), Regen (15), geschützt (15), wechselhafte (15), Heizöl (14), feuchten (14), vergangenen (14), Begünstigt (13), Grad (13), besonders (13), Pilz (12), Trauben (12), begünstigt (12), verlegt (12), widrigen (12), Baaderplatzes (11), Bauarbeiten (11), Bäume (11), Heizgradtage (11), Meteorologen (11), Sobald (11), Trotz (11), Volksheilkundliche (11), extrem (11), kühlere (11), nasse (11), selbstgefressenen (11), waldfreundliche (11), widriger (11), winterlichen (11), Freien (10), Schnee (10), Schneekappe (10), Tharaus (10), Warenhaus-Manager (10), Wetter (10), Wildkräuterführungen (10).

2.3.2 Filtern von Kollokationsmengen

In vielen Fällen drückt sich die eben beschriebene semantische Inhomogenität einer Kollokationsmenge auch durch die Wortart der entsprechenden Wörter aus. Eine einfache Filterung nach Wortarten erzeugt häufig gute Mengen zur Weiterverarbeitung. Solche Filterkriterien sind beispielsweise:

- Für Adjektive: spezielle Substantive als rechte Nachbarn
- Für Substantive: spezielle Adjektive als linke Nachbarn
- Für Substantive: spezielle Verben als linke oder rechte Nachbarn
- Für Substantive: Substantive, die Satzkollokationen, aber keine Nachbarschaftskollokationen sind.

2.3.3 Iterierte Kollokationen

Aus diesen Kollokationen erster Stufe können iterativ Kollokationen höherer Stufe erzeugt werden: Für die Kollokationen zweiter Stufe wird ein Korpus benutzt, welches aus Kollokationsmengen von Kollokationen erster Stufe besteht. Diese Kollokationsmengen übernehmen die Rolle der Sätze des Korpus. Da die Reihenfolge der Wörter in den Kollokationsmengen nicht aussagekräftig ist, wohl aber ihr gemeinsames Vorkommen, sind hier nur Satzkollokationen sinnvoll. In der dritten Stufe werden statt Sätzen die Kollokationsmengen zweiter Stufe ausgewertet und so weiter.

Hohe Signifikanzen zweier Wörter in den Kollokationen zweiter Stufe bedeuten, dass diese Wörter häufig gemeinsam in Kollokationsmengen erster Stufe auftreten. Dies wiederum bedeutet, dass die entsprechenden Wörtern in vielen Kontexten gemeinsam auftreten und damit zu

einem größeren Kontext gehören. Diese Eigenschaft erinnert sofort an die Zusammenfassung von Wörtern entsprechend zu Bedeutungsgruppen, oder andersherum, an die Zerlegung von Kollokationsmengen erster Ordnung entsprechend verschiedener Bedeutungen.

Die inhaltliche Bedeutung von Kollokationen höherer Ordnung ist zunächst nicht klar. Aus den Berechnungen bis Stufe 10 (um evtl. reine Mengen zu erhalten) ist beobachtbar, dass die Kollokate höherer Stufe für viele Wörter semantisch relativ homogene Wortmengen darstellen. Allerdings ist nicht unbedingt vorhersehbar, auf welches semantische Attribut sich die Homogenität bezieht. Dies kann abhängen von der Art der verwendeten Kollokationsmengen (Satzkollokationen oder Nachbarschaftskollokationen) im ersten Schritt, ebenso von der Größe der verwendeten Kollokationsmengen. Hier scheinen größere Kollokationsmengen allgemeinere Zusammenhänge zu extrahieren.

Selbstverständlich können auch diese iterierten Kollokationsmengen mit Filtern wie aus Abschnitt 2.3.2 behandelt werden.

warm	**kühl**	**kalt**
abgekühlt	abgeklärt	abgekühlt
abkühlen	abgekühlt	abkühlen
angestiegen	abkühlen	angestiegen
anzeigt	ablehnend	anzeigt
aufgeheizt	abstrakt	aufgeheizt
eingefroren	aggressiv	aushalten
erhitzt	ähnlich	eingefroren
erwärmt	altmodisch	einstellen
fertig	anders	erhitzt
gebrannt	archaisch	ernst
gefallen	aufgeheizt	erwärmt
gehalten	aushalten	frei
geklettert	bedrohlich	gebrannt
gekühlt	bescheiden	gefallen
gelagert	bitter	gehalten
gemessen	blaß	geklettert
gesenkt	blutleer	gekühlt
gestiegen	distanziert	gelagert
gesunken	eingefroren	gemessen
gut	empfindlich	genug
Heiß	empört	gesenkt
heruntergekühlt	entrüstet	gestiegen
hoch	entsetzt	hart
höher	entspannt	heiß
kalt	erhitzt	heruntergekühlt
kalte	erleichtert	hoch
kalten	erschöpft	höher
...	...	...

Tabelle 1: Nachbarschaftskollokationen zweiter Stufe mit Adjektivfilter für „warm“, „kühl“ und „kalt“

2.4 Beispiele

Im verbleibenden Teil wird anhand zweier Beispiele illustriert, wie Kohyponyme mit Kookkurrenzen verschiedener Stufen automatisch extrahiert werden können.

2.4.1 Kohyponyme über Nachbarschaftskollokationen zweiter Stufe

Als Beispiel betrachten wir die Kohyponyme „warm“, „kalt“ und „kühl“. Zu diesen Wörtern berechnen wir die Nachbarschaftskollokationen zweiter Stufe, also Wörter, die in einem größeren Kontext zusammen mit den Startwörtern auftreten. Ferner filtern wir die drei Wortmengen, so dass nur Adjektive übrigbleiben. Tabelle 1 zeigt einen Ausschnitt aus den drei alphabetisch sortierten Wortmengen.

Die Schnittmenge dieser drei Mengen enthält die Wörter *abgekühlt, aufgeheizt, eingefroren, erhitzt, erwärmt, gebrannt, gelagert, heiß, heruntergekühlt, verbrannt* und *wärmer*, also bis auf ein Wort Abstufungen von Temperatur. Die emotionale Lesart von „kalt“ und „kühl“, die sich in Wörtern wie *abgeklärt* ablesen lässt, wird vollständig durch den Schnitt mit der Menge zu „warm“ eliminiert.

2.4.2 Kohyponyme mit Kollokationen erster Stufe

In diesem Abschnitt wird eine weitere Möglichkeit gezeigt, semantische Relationen zu extrahieren, und zwar insbesondere linguistische Kollokationen, Kohyponyme oder Hyperonyme, wo-

bei hier zwischen diesen Arten unterschieden werden kann. Diese Möglichkeit ergibt sich aus der Parallele zwischen syntagmatischen sowie paradigmatischen Beziehungen de Saussures, siehe (Saussure 1916), und den Mengen von Satzkollokationen einzelner Wortformen. Während gemeinsames Auftreten in einem Satz grob als syntagmatische Relation eingeschätzt werden kann, soll im Folgenden eine paradigmatische Relation beschrieben werden. Zu jedem Wort einer Menge von Satzkollokationen werden wiederum dessen Satzkollokationen ermittelt. Diese Mengen werden verglichen, indem der Anteil gemeinsam auftretender Elemente bestimmt wird. Damit haben wir für zwei Wörter A und B zwei Zahlenwerte ermittelt: Die Kollokationsstärke zwischen A und B sowie die Ähnlichkeit der Kollokationsmengen von A und von B. Trägt man für ein festes Wort A alle seine Kollokate entsprechend dieser beiden Werte in einem Koordinatensystem ab, so beobachtet man folgendes: Kohyponyme weisen einen hohen direkten Signifikanzwert auf und besitzen auch ähnliche Kollokationsmengen und befinden sich damit im rechten oberen Bereich der graphischen Darstellung. Hyperonyme hingegen dürften selten im gleichen Satz auftreten, ansonsten aber ein sehr vergleichbares Kollokationsprofil aufweisen und damit im äußersten unteren Bereich der Abbildung zu finden sein. Linguistische Kollokationen hingegen müssten geradezu umgekehrt, zwar häufig im gleichen Satz auftretend, ansonsten aber sehr unterschiedliche Kollokationprofile besitzen und sich damit im linken oberen Bereich der Abbildung aufhalten.

Eine prototypische Implementierung dieses Verfahrens hat gezeigt, dass es wie erwartet die drei genannten Relationen extrahiert. Es hat aber auch gezeigt, dass es Wörter gibt, für die fälschlicherweise Kohyponyme extrahiert werden, obwohl diese keine besitzen. Es wird demnach noch ein weiteres Feature benötigt, um die richtigen Wörter von den unpassenden unterscheiden zu können. Weiterhin ist die auf diese Weise erhaltene Kohyponymmenge oft durch in der ‚part-of'-Relation stehende Wörter verunreinigt. Schließlich aber befinden sich in der Nähe der erwarteten Hyperonyme oft noch andere, unbeteiligte Wörter, wofür ebenfalls ein anderes Filterkriterium gefunden werden muss.

Es sei jedoch angemerkt, dass es nicht wichtig ist, die Grenzen zwischen den einzelnen räumlich getrennten Bereichen für eine optimale Zuordnung zu kennen, zudem sich diese Grenzen von Wortform zu Wortform leicht unterscheiden können. Da jedoch eine Wortform mehrere Kohyponyme hat und idealerweise lediglich ein Hyperonym, können sich die Ergebnisse der Berechnungen der Kohyponyme gegenseitig verifizieren und so zu einem meist eindeutigen Ergebnis kommen.

Abbildung 1 zeigt ein Beispiel anhand „Elefant".

3 Implementierung und Evaluation

In diesem Abschnitt stellen wir Ergebnisse vor, die wir mit einer prototypischen Implementierung eines Verfahrens wie in 2.2. beschrieben erzielen konnten.

Zum Einsatz kamen Nachbarschaftskollokationen zweiter und dritter Stufe, sowie ein Wortartenfilter.

Intuitiv enthält die Nachbarschaftskollokationsmenge zweiter Stufe eines Wortes diejenigen Wörter, die ähnliche Bigrammkontexte aufweisen, der Wortartenfilter stellt sicher, dass nur Wörter mit derselben Wortart wie die Wörter der zu erweiternden Menge ausgegeben werden.

Zum Testen verwendeten wir einerseits Synsets aus drei verschiedenen lexikalischen Feldern aus GermaNet, andererseits Synonyme aus unserer Wortschatz-Sammlung.

Evaluiert wurde, wie viele Synonyme und wie viele ‚Xonyme' (dies soll hier verwendet werden als Sammelbezeichnung für Antonyme, Hyponyme, Hyperonyme und Kohyponyme) unter

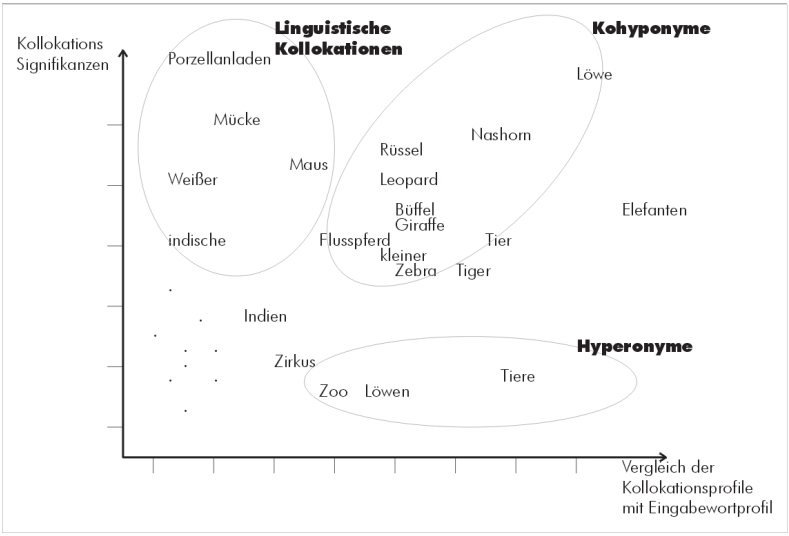

Abbildung 1: Koordinatensystem für Elefant mit Zuweisung von linguistischenKollokationen, Hyperonymen und Kohyponymen

den ersten 5 bzw. 10 ausgegebenen Wörtern zu finden waren. Die Reihenfolge der Ausgabe erfolgte nach Signifikanz der Kollokation wie in (Quasthoff & Wolff 2002) beschrieben.

Getestet wurde mit Synset-Mengen der Mächtigkeit 1 und 2.

Tabelle 3 zeigt Synsets, die Top 10 der automatisch erzeugten Ergebnismenge, sowie deren Klassifizierung in Synonyme (S) und ander Xonyme (X) .

Die Abbildungen 2a – 2b zeigen die Anteile von Synonymen, Xonymen und nicht in diesem Sinne verwandten Wörtern für verschiedene Arten von Startmengen, die Werte sind jeweils gemittelt für 10 zufällig ausgewählte Synsets. Die folgende Tabelle erklärt die in der Abbildung verwendeten Kürzel:

Kürzel	**Bedeutung**
WS_ADJ	Adjektive aus dem Wortschatz
WS_NN	Nomen aus dem Wortschatz
WS_VV	Verben aus dem Wortschatz
GER_ADJK	Adjektive aus GermaNet/Körper
GER_ADJO	Adjektive aus GermaNet/Ort
GER_VVL	Verben aus GermaNet/Lokation

Tabelle 2: Erklärung der Kürzel. Eine „1“ bzw. „2“ am Ende bedeutet, das das zu erweiternde Synset aus einem bzw. zwei Elementen bestand.

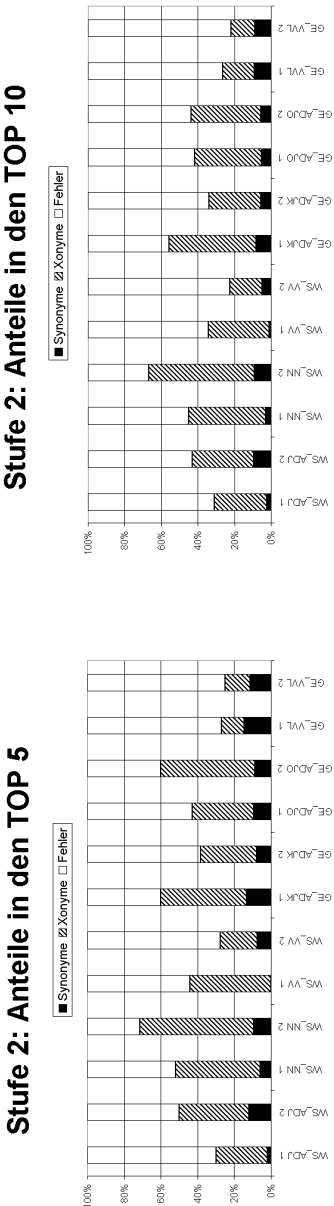

Abbildung 2a: Kollokationen zweiter Stufe.

Abbildung 2b: Kollokationen dritter Stufe.

Synset	Ergebnismenge
erstklassig, prima	besser(X), exzellent(S), glänzend(S), gut(X)
Mörder, Killer	Täter(X), Verbrecher(X), Attentäter(X), Kriegsverbrecher(X), Straftäter(X), Räuber(X), Mann, Einbrecher(X), Terrorist(X), Brandstifter(X)
verrückt	gerne, enttäuscht, bekannt, albern(S), begeistert, bescheuert(S), erschöpft, frustriert, dünn
festigen, stabilisieren	sichern(S), sorgen, auswirken, beitragen(X), gefährden, stützen(S), erwarten, fördern(X), profitieren, bringen, führen
dunkelbraun	braun(X), hellbraun(X), rotbraun(X), orange(X), rot(X), rosa(X), violett(X), religiös, blutrot(X), graubraun(X)

Tabelle 3: Synsets und deren automatische Erweiterungen. Synonyme sind mit (S), Xonyme mit (X) gekennzeichnet.

Der Synonym- und Xonymanteil in den ersten 5 Rängen ist in den meisten Fällen höher als in den ersten 10, was die Wirksamkeit unserer Ranking-Methode bestätigt.

Unsere Erwartungen, dass größere Startmengen bessere Ergebnisse bringen sollten, erfüllten sich nur teilweise: Während bei den Mengen aus dem Wortschatz die Qualität bei den Zweiermengen deutlich höher liegt als bei den Einermengen, zeigt sich bei den Germanet-Synsets ein ausgeglichenes Bild.

Insgesamt scheinen die Verben für nachbarschaftskontextbasierte Features wie die hier evaluierten am schlechtesten geeignet zu sein, was in Anbetracht ihrer komplexeren Argumentstruktur nicht verwunderlich ist. Die unterschiedliche Qualität bei den verschiedenen Wortarten motiviert das Lernen von Featurekombinationen wie in 2.2 beschrieben.

In der vorliegenden Form können die erzielten Daten lediglich als Vorschläge dienen, um GermaNet zu erweitern. Die endgültige Entscheidung über die Einordnung an der vorgeschlagenen Stelle muss manuell vorgenommen werden. Trotzdem entsteht bei dieser Methode durch das Vorschlagen eines Synsets schon eine Zeitersparnis von ca. 90% gegenüber der reinen Handarbeit.

4 Ergebnis und Ausblicke

Wir haben Möglichkeiten vorgestellt, korpusbasierte Erweiterungen von beliebigen Wortmengen durchzuführen.

Bei der Frage auf Anwendbarkeit auf semantische Netze wie GermaNet, muss nun folgendes beachtet werden: Hat das zur Erweiterung vorgelegte semantische Netz eine gewisse Größe erreicht, so geht es bei der Einordnung nur noch um mittel- und niederfrequente Wörter. Die Verfahren zur Erzeugung der Kandidatenmengen müssen also auch mit solchen Wörtern umgehen können.

Um zu verdeutlichen, dass das verwendete Korpus (36 Millionen deutsche Sätze) hierfür ausreichend ist, zeigt Abbildung 3 die durchschnittliche Größe der Kollokationsmengen für absteigend nach Frequenz geordnete Wörter (Vollformen).

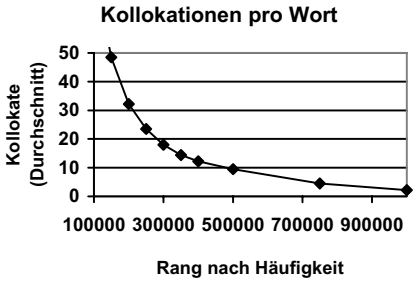

Abbildung 3: Durchschnittliche Anzahl Kollokate nach Häufigkeitsrang.

Um sinnvolle Ergebnisse zu erhalten, benötigen wir mindestens 10 Kollokate pro Wort, was bis

Rang 500.000 der Fall ist. Unter der Berücksichtigung der Größe von GermaNet (etwa 61.000 Lexical Units) und der Struktur (es werden Lemmas gespeichert, während das Korpus mit Vollformen umgeht) erwarten wir, GermaNet in etwa um die Hälfte erweitern zu können.

Um auch niederfrequente Wörter extrahieren zu können, für die zu wenig Kollokate existieren, ist es nötig, auf musterbasierte Verfahren wie (Biemann 2003) zurückzugreifen, die auf Beispielsätzen arbeiten und auch mit Wortmengen trainiert werden können.

Literatur

Ahmad, K. (ed) (1994). MULTILEX: Final Report. Guildford: University of Surrey.

Biemann, Ch. (2003). „Extraktion von semantischen Relationen aus natürlichsprachlichem Text mit Hilfe von maschinellem Lernen." In: Seewald-Heeg, U. (ed.) (2003): Sprachtechnologie für multilinguale Kommunikation. Beiträge der GLDV-Frühjahrstagung 2003. Sankt Augustin: Gardez! Verlag, 12-25 [zugl. erschienen in: LDV-Forum 18(1/2) (2003), 12-25].

Bordag, S. (2003). "Sentence Co-occurrences as Small-World Graphs: A solution to Automatic Lexical Disambiguation." In: Gelbukh, A. (ed.) (2003). Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2003), Mexico City, Februar 2003. Berlin et al.: Springer [= LNCS 2588], 329-332 .

Ciaramita, M. (2002). "Boosting Automatic Lexical Acquisition with Morphological Information." In: Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX 2002), Philadelphia, PA, July 2002, 17-25.

Heyer, G. et al. (2001). "Learning Relations using Collocations." In: Maedche, A. et al. (eds.) (2001). Proceedings Workshop on Ontology Learning. 17th International Conference on Artificial Intelligence (IJCAI '2001), Seattle, WA, August 2001, 19-24.

Quasthoff, U. (1998). „Projekt Der Deutsche Wortschatz." In: Heyer, G.; Wolff, Ch. (Hrsg.) (1998). Linguistik und neue Medien, Proceedings GLDV-Jahrestagung, Leipzig, März 1997. Wiesbaden: Deutscher Universitätsverlag, 93-99.

Quasthoff, U.; Wolff, Ch. (2002). "The Poisson Collocation Measure and its Applications." In: Proceedings 2nd International Workshop on Computational Approaches to Collocations, Wien, July 2002.

Rapp, R. (2002). "The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches." In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipeh, August / September 2002.

Ruge, G. (1997). "Automatic Detection of Thesaurus Relations for Information Retrieval Applications." In: Freksa, Ch.; Jantzen, M.; Valk, R. (eds.) (1997). Foundations of Computer Science: Potential - Theory - Cognition. Berlin et al.: Springer [= LNCS 1337], 499-506.

Saussure de, F. (1916). Cours de Linguistique Générale, Paris: Payot.

Sperberg-McQueen, C.M. Burnard, L. (eds.) (2002). TEI P4: Guidelines for Electronic Text Encoding and Interchange. XML Version. Oxford: Text Encoding Initiative Consortium / University of Oxford, Humanities Computing Unit.

Michael Beißwenger, Angelika Storrer, Maren Runte

Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet

Abstract

Dieser Beitrag skizziert die Konzeption eines im Projekt „Hypertextualisierung auf textgrammtischer Grundlage“ (*HyTex*) modellierten terminologischen Wortnetzes (*TermNet*) zu den Fachtextdomänen Texttechnologie und Hypermedia. Schwerpunkt des Beitrags ist es zum einen, die Modellierung von *TermNet* in Hinblick auf fachsprachen- und domänenspezifische Merkmale vorzustellen, und zum anderen, die Anwendung von *TermNet* für die Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen zu erörtern.

1 Projektrahmen und Motivation des Ansatzes

Unter *Hypertextualisierung* versteht man die Aufbereitung von Dokumenten für die selektiven, interaktiven Nutzungsformen in einem Hypertextsystem, z.B. dem World Wide Web. Das Projekt *HyTex*[1] (*Hypertextualisierung auf textgrammatischer Grundlage*) sucht nach textgrammatisch geleiteten Verfahren für eine hypertextadäquate Aufbereitung selektiv organisierter Fachtexte (wissenschaftliche Artikel, technische Spezifikationen), d.h. eine Aufbereitung, die hypertexttypischen selektiven Rezeptionsformen optimal entgegenkommt. Auf der technischen Seite benötigt man für diese Aufgabe Konversionstools; auf der konzeptionellen Seite benötigt man Strategien und Verfahren für die folgenden beiden Teilaufgaben der Hypertextualisierung:

- *Segmentierung* (Zerlegung der Dokumente in Module).
- *Linking* (Verknüpfung der Module durch Hyperlinks).

Für die in *HyTex* entwickelte textgrammatisch geleitete Herangehensweise an diese Aufgaben gibt es zwei Leitlinien: (a) *Reversibilität* und (b) *Hypertextualisierung nach Kohärenzkriterien*.

Ad a): Reversibilität bedeutet, dass wir Hypertext-Sichten auf lineare Dokumente als zusätzliche Sichten generieren, die regelgeleitet aus textgrammatischem Markup und anderen Wissensquellen – z.B. aus dem in diesem Papier beschriebenen Terminologienetz – abgeleitet werden. Die sequentielle Struktur und der Originalwortlaut eines Dokuments bleiben dabei als eine mögliche Sicht auf das Dokument erhalten.[2] Damit geben wir dem Rezipienten die Möglichkeit, einen Text in der ursprünglichen linearen Form und Abfolge zu rezipieren, wenn er die Zeit dazu hat; die Hypertextsichten sind als zusätzliche Angebote für den eiligen Querleser gedacht.

Ad b): Das Ziel der Hypertextualisierung in unserem Ansatz ist es, Kohärenzbildungsprozesse beim selektiven Querlesen besser zu unterstützen als dies in Printmedien möglich ist und damit das Mehrwertpotenzial von Hypertexten auszureizen. Im Hinblick auf diese Zielsetzung spielen bei der Segmentierung und beim Linking *Kohärenzkriterien* eine zentrale Rolle. *Hypertextualisierung nach Kohärenzkriterien* ist eine Strategie, die Rainer Kuhlen (Kuhlen 1991) einer Strategie gegenüberstellt, die als „Hypertextualisierung nach formalen Texteigenschaften“ bezeichnet wird. Bei der Hypertextualisierung nach formalen Texteigenschaften erfolgt die Segmentation ausschließlich anhand der typographisch angezeigten Unterteilung in Kapitel, Unterkapitel und Abschnitte. Diese werden dann in Nachbildung der hierarchischen Dokumentenstruktur wieder durch Links verknüpft, d.h. die

Teil-Ganzes-Bezüge zwischen Kapiteln, Unterkapiteln und Abschnitten werden als Links nachgebildet und mit einem Inhaltsverzeichnis auf der Einstiegsseite verlinkt. Zusätzlich wird häufig ein Lesepfad gelegt, der in einer Tiefe-vor-Breite-Strategie auf genau demjenigen Weg durch den Hypertext führt, der der Abfolge im gedruckten Pendant entspricht. Der in *HyTex* verfolgte Ansatz hingegen legt den Schwerpunkt bei der Strategiebildung auf textgrammatisch geleitete, verfeinerte Segmentations- und Linkingtechniken, die die Kohärenzbildung des selektiv und quer lesenden Nutzers optimal unterstützen. Eine wichtiger Strategietyp ist dabei das sog. *Linking nach Wissensvoraussetzungen*, welches darauf abzielt, mit automatischen Verfahren Links zu genau denjenigen Textsegmenten zu generieren, deren Inhalte für das Verständnis des von einem selektiv zugreifenden Hypertextrezipienten aktuell rezipierten Moduls benötigt werden[3].

Hierbei orientieren wir uns an einem Szenario, das wir als *Hypertext-Rezeptionsumgebung für Fachtexte* bezeichnen. Das Szenario ist zugeschnitten auf Nutzungssituationen, in welchen sich ein Nutzer unter Zeitdruck und mit einer ganz speziellen Zielsetzung Wissen zu einem Fachgebiet erarbeiten muss, für welches er zwar bereits Wissensvoraussetzungen mitbringt, in dem er aber kein Experte ist. Beispiele, in denen solche Situationen auftreten, sind interdisziplinäre Projektarbeit, Wissenschaftsjournalismus, Fachlexikographie, sowie interdisziplinäres Arbeiten in Studium und Weiterbildung. Ganz unabhängig von WWW und Hypertext lesen Nutzer in solchen Situationen quer und partiell und rezipieren nur selektiv Teiltexte. Hypertextsichten kommen dieser Rezeptionsform nun prinzipiell entgegen, indem sie längere Dokumente bereits in modularisierter Form präsentieren und darin verschiedene Such- und Navigationsoptionen zur Verfügung stellen. Werden sequentiell organisierte Texte aber nur nach formalen Texteigenschaften hypertextualisiert, so besteht die Gefahr, dass dem selektiven Querleser wichtige Voraussetzungen für das korrekte Verständnis eines aktuell rezipierten Textausschnitts fehlen; schließlich sind die Teiltexte der einzelnen Module ja weiterhin auf die Ganzlektüre auf einem vorgebenen Leseweg hin formuliert. Dieses Kohärenzproblem bei der Hypertextrezeption[4] soll der bereits benannte Strategietyp eines „Linkings nach Wissensvoraussetzungen“ durch Generierung von Links und zusätzlichen Sichten (z.B. der Glossarsicht, vgl. Abschnitt 4) kompensieren. Die hierbei entwickelten Strategien verarbeiten Informationen aus drei verschiedenen Ebenen, die in Abb. 1 visualisiert sind und die sich folgendermaßen skizzieren lassen:

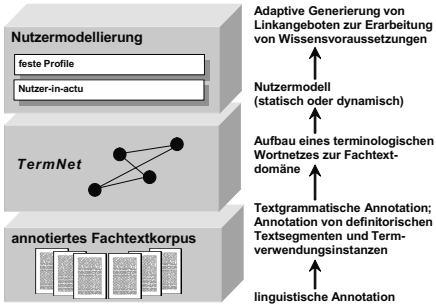

Abbildung 1: Der 3-Ebenen-Ansatz in HyTex

Auf der *Dokumentenebene* werden deutsche Fachtexte zum Thema „Texttechnologie“ und „Hypermedia“ textgrammatisch annotiert. Unser Korpus umfasst insgesamt 96 Dokumente, wobei neben Fachartikeln und normativ („Spezifikation“) oder didaktisch motivierten Dokumenttypen (z.B. „Tutorial“, „Einführung“, „Überblicksdarstellung“) auch diskursiv geprägte Textsorten wie FAQs, Mailinglist- oder Foren-Postings und Chat-Protokolle berücksichtigt sind. Das Korpus wird in Kooperation mit dem Tübinger *DEREKO*-Projekt automatisch linguistisch annotiert (Lemmatisierung, POS-Tagging, Chunk-Parsing). Auf dieser linguistischen Annotation aufsetzend erfolgt dann die textgrammati-

sche Annotation, d.h. die Auszeichnung von Koreferenzstrukturen, von rhetorischen und thematischen Strukturen und – für das im Beitrag fokussierte Thema besonders relevant – von Termverwendungsinstanzen und definitorischen Textsegmenten. Die Annotationen auf dieser Ebene erfassen also sowohl Aspekte der syntaktischen Kohärenz (z.B. Koreferenzbezüge zwischen Textelementen) als auch Aspekte der durch thematische und rhetorische Muster geleiteten globalen semantischen Textkohärenz.

Auf der *Ebene der Nutzermodellierung* werden in der laufenden Projektphase fixe Nutzerprofile verarbeitet; geplant ist im weiteren Verlauf aber auch die Verarbeitung von Nutzungsprotokollen, aus denen sich Anhaltspunkte über den aktuellen Wissensstand und den aktuellen Aufmerksamkeitsbereich des Nutzers ergeben können.

Im Mittelpunkt dieses Beitrags steht die *Ebene der Modellierung von terminologischem Wissen zu grundlegenden Konzepten und Lexemen der im Korpus repräsentierten Fachdomäne*[5]. Auf dieser Ebene nutzen wir eine erweiterte Form des in *WordNet* für semantisch-lexikalische Wortnetze entwickelten Beschreibungsmodells für die Modellierung einerseits zentraler Konzepte und andererseits der (terminologisch eingeführten) Lexeme, mit denen diese Konzepte in verschiedenen Publikationen und von verschiedenen Autoren versprachlicht werden, für das Linking und für die Generierung von mit den Dokumenten verlinkten Glossarsichten. Im folgenden Abschnitt werden wir zunächst auf einige verwandte Ansätze eingehen, um dann die theoretischen Grundlagen der Modellierung unseres terminologischen Netzes (*TermNet*) zu erläutern. Abschließend werden wir anhand eines Anwendungsbeispiels skizzieren, wie *TermNet* im Zusammenspiel mit den anderen der in Abb. 1 skizzierten Ebenen zur Hypertextualisierung eingesetzt wird.

2 Wortnetze und Ontologien für das Linking

Bislang wurden Wortnetze vor allem für statistisch basierte Strategien zur Hypertextualisierung eingesetzt. Der in Green (Green 1998) beschriebene Ansatz nutzt das englische *WordNet* (vgl. Fellbaum 1998) als lexikalische Ressource zur vollautomatischen Hypertext-Konversion. *WordNet* dient dabei vor allem dazu, aus den Ausgangstexten lexikalische Ketten („lexical chains"), d.h. Folgen von semantisch verwandten Textwörtern, zu extrahieren und dabei ggf. mehrdeutige Lesarten zu disambiguieren. Als Grundlage für das vollautomatische Linking dient die Berechnung der relativen Wichtigkeit der extrahierten Ketten für das jeweilige Textmodul (Paragraph) und ein Berechnungsmaß für die Ähnlichkeit zwischen Modulen, das auf der Ähnlichkeit der jeweiligen lexikalischen Ketten beruht. Überschreitet das Ähnlichkeitsmaß einen bestimmten Schwellenwert, so werden die entsprechenden Module verlinkt. Durch ein ähnliches Verfahren werden auch Links zwischen Modulen verschiedener Dokumente und zwischen ganzen Dokumenten erzeugt. Die Links, die auf statistischer Grundlage aufgrund von Ähnlichkeitsberechnungen ohne Berücksichtigung des Nutzungskontextes und der Wissensvoraussetzungen des Nutzers generiert werden, sind assoziative Links, die semantisch nicht weiter typisiert sind.

Im Gegensatz zu diesem Ansatz nutzen wir in *HyTex WordNet* nicht als lexikalische Ressource[6], sondern als Strukturierungskonzept für die Modellierung von Wissen über die Bedeutung und Verwendungsweise von zentralen Fachtermini der im Fachtextkorpus behandelten Domäne (also „Texttechnologie" und „Hypermedia"). Auf der Grundlage einer über dem Fachtextkorpus erzeugten Liste von Termini wird ein Terminologienetz im Stil lexikalisch-semantischer Wortnetze[7] erstellt und mit den Termverwendungsinstanzen in den Dokumenten verlinkt. Dieses Netz dient der Generierung semantisch

typisierter Links, anhand derer sich Nutzer genau diejenigen Wissensvoraussetzungen erarbeiten können, die für die korrekte Semantisierung der Verwendung von Termini in je konkreten Kontexten notwendig sind.

Insofern ist unsere Verwendung des Wortnetzes verwandt mit der Verwendung sog. „Ontologien" in *ontologiebasierten Hypertexten* (vgl. Miles-Board et al. 2001). In diesem Ansatz, der in verschiedenen Projekten zum Einsatz gebracht wurde[8], werden die Objekte sowie die Beziehungen zwischen ihnen in einer als Ontologie bezeichneten Wissensrepräsentation abgebildet, die dann die Strukturierung und die Verlinkung von Hypertextdokumenten motiviert. Die Ansätze verwenden Metadaten, um Dokumenteninhalte zu beschreiben, und generieren daraus automatisch Links zwischen den Dokumenten und der Ontologie. Im Vergleich zum *HyTex*-Ansatz, der zunächst nur auf dem geschlossenen Fachtextkorpus operiert, zeichnen sich diese Ansätze durch einen hohen Grad an Automatisierung und die Anwendbarkeit auf offene Korpora aus. Das zur Modellierung der Ontologien genutzte Inventar von Entitäten und Relationen ist allerdings im Vergleich zu dem Beschreibungsinventar für lexikalisch-semantische Wortnetze beschränkter, enthält aber dafür „ontologietypische" Entitäten (Instanzen) und entsprechende Relationen (z.B. *class-instance*), die für Inferenzen (d.h. die Gewinnung nicht explizit gespeicherten Wissens aus dem Modell durch logische Inferenzmechanismen) genutzt werden können.

Die im Folgenden dargestellten Modifikationen und Erweiterungen des Beschreibungsrahmens von *WordNet*, das in seiner ursprünglichen Form nicht ohne Weiteres für Inferenzen verwendet werden kann und auch nicht als Ontologie konzipiert wurde (vgl. Fischer 1998, Carr et al. 2001), versucht die Stärken des *WordNet*-Ansatzes bei der Modellierung von Sprachwissen mit den Vorteilen von Ontologien beim Durchführen von Inferenzen auf Wissensrepräsentationen zu verbinden und den „klassischen" Wortnetz-Ansatz auf die Erfordernisse des oben skizzierten Anwendungsszenarios hin zuzuschneiden.

3 Modellierung terminologischen Wissens in *TermNet*

Die in *TermNet* verfolgten Modellierungsprinzipien folgen im Grundsatz dem *WordNet*-Ansatz. Zwar ist der *WordNet*-Ansatz primär auf die Gemeinsprache ausgerichtet; doch obwohl sich für Fachsprachdomänen in verschiedenerlei Hinsicht (sowohl lexikalischer, konzeptueller wie auch funktionaler Art) „besondere" (d.h.: von der Gemeinsprache verschiedene) Regularitäten der Prägung, Etablierung und Verwendung lexikalischer Einheiten feststellen lassen, kann die prinzipielle Unterscheidung zwischen *words* (Lexemen) und *synsets* (Konzepten) auch für die Modellierung fachdomänenspezifischer Wortnetze vorteilhaft sein (insbesondere in Hinblick auf den in *HyTex* anvisierten Anwendungsbereich einer automatischen Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen beim selektiven Zugriff auf den Dokumentenpool). Allerdings waren bei der Konzeption von *TermNet* – eben in Hinblick auf den in Rede stehenden Anwendungsbereich – einige Modifikationen des in *WordNet* enthaltenen Relationeninventars notwendig. Diese (sowie die damit in Zusammenhang stehenden fachsprachdomänenspezifischen Besonderheiten) sollen im folgenden skizziert und anhand von Beispielen erläutert werden.

3.1 Behandlung der Synonymie in Fachtextdomänen

Die beiden grundlegenden Typen von Entitäten bei der Modellierung von Wortnetzen im Stile von *WordNet* werden auch in *TermNet* unterschieden: das *Lexem* („word" in der Terminologie von *WordNet*) und das *Synset*, welches eine (ein- oder mehrelementige) Menge von Lexe-

men darstellt, denen jeweils paarweise und restfrei ein Verbundensein über eine Synonymierelation zugesprochen werden kann. In *TermNet* umfasst ein Synset terminologische Ausdrücke, die in der Fachtextdomäne in ungefähr dasselbe Konzept lexikalisieren. So werden z.B. die Lexeme Hyperlink und Verweis demselben Synset zugeordnet, da sie von unterschiedlichen Autoren der Fachdomäne für identische oder zumindest *thematisch* identische Konzepte eingeführt und genutzt werden. Wie auch in *WordNet* wird der Ermittlung der Mitglieder eines Synsets ein weit gefasster Synonymiebegriff zu Grunde gelegt, welcher die Austauschbarkeit der Lexeme in mindestens einem Kontext postuliert. Hierzu ist natürlich anzumerken, dass ein Synonymiekonzept, wie es bei der Modellierung gemeinsprachlicher Wortnetze in durchaus begründbarer Weise angewendet werden kann, nicht ohne weiteres auch auf die Modellierung fachsprachlicher Wortnetze übertragbar ist. Während die Regeln der Verwendung gemeinsprachlicher Ausdrücke prinzipiell gebrauchs- und kontextabhängig sind und bedeutungsähnliche lexikalische Einheiten unter Absehung von je konkreten Verwendungskontexten in semantischer Hinsicht nicht trennscharf von einander geschieden werden können, ist es eine Besonderheit des Sprachgebrauchs in fachsprachlicher Kommunikation, dass jeder Autor prinzipiell als Herr seiner eigenen Benennungs- und Konzeptsysteme in Erscheinung treten kann. Typisch für Fachtexte ist, dass anhand definitorischen Sprachhandelns Konzepte explizit mit einer bestimmten Benennung versehen (und somit vom Autor terminologisiert) werden beziehungsweise Regeln für die Verwendung bestimmter Bezeichnungen fixiert werden (durch explizite Beschreibung der Konzepte, zu deren Benennung sie – zumindest im Textuniversum des betreffenden Autors – fungieren sollen). Unterschiedliche Autoren (und bisweilen sogar derselbe Autor in unterschiedlichen Abschnitten seiner wissenschaftlichen Biographie) können (a) dieselben Benennungen für unterschiedliche Konzepte oder (b) unterschiedliche Benennungen für gleiche oder ähnliche Konzepte oder aber auch (c) unterschiedliche Benennungen für unterschiedliche, aber thematisch identische Konzepte einführen und verwenden. Deshalb ist die konzeptuelle wie lexikalische Geordnetheit fachsprachlichen Konzept- und Bedeutungswissens somit immer textabhängig. Ein fachsprachliches Wort- und Konzeptnetz, welches den Anspruch erhebt, die *tatsächlichen* Verhältnisse abzubilden, müsste daher im Grunde für jeden einzelnen Fachtext (oder zumindest für die Fachsprache jedes einzelnen Autors der Domäne) individuell modelliert werden.

Unter fachsprachenlinguistischem Aspekt wäre die „autorsensitive" Modellierung fachsprachlicher Wortnetze (und insbesondere ihr Vergleich) zweifelsohne von großem Interesse. In Hinblick auf den im *HyTex*-Projekt verfolgten Anwendungsrahmen (der auf die Unterstützung der selektiven Fachtextrezeption und nicht auf Untersuchungen zur lexikalischen und konzeptuellen Struktur von Fachsprachen abhebt; siehe Abschnitt 1) wäre eine solche „autorsensitive" Modellierung allerdings weder praktikabel noch zielführend. Unser Anwendungsszenario geht davon aus, dass ein selektiv auf eine große Menge von (Fachtext-)Dokumenten zugreifender Nutzer mit Semi-Experten-Status sich möglichst schnell darüber informieren möchte, (a) welches Konzept einem in einem Textmodul verwendeten Terminus autorseitig unterliegt, und (b) in welcher Beziehung dieses Konzept zu den übrigen in der Fachtextdomäne relevanten Konzepten steht. Anforderung (a) versuchen wir dadurch nachzukommen, dass wir über ein Verfahren zur pragmatischen Gewichtung von Definitionen in Fachtexten dem Nutzer genau diejenige Definition im Vortext über ein Linkangebot zugänglich machen, von welcher wir annehmen, dass sie diejenige ist, an die sich der Autor des betreffenden Textes auch in seinem eigenen

Sprachhandeln hält[9]. Da Nutzer mit Semi-Experten-Status zum Verständnis der in der Fachtextdomäne behandelten Gegenstände zunächst einmal eines systematischen Überblicks über einzelne Konzepte in ihrer Beziehung zu anderen Konzepten der Domäne bedürfen, primär also an der Geordnetheit des in der Domäne behandelten Wissensauschnitts interessiert sind, versuchen wir Anforderung (b) dadurch nachzukommen, dass wir bei der Modellierung von *TermNet* ein ähnlich weitgefasstes Synonymiekonzept wie in *WordNet* ansetzen. Damit können wir den Nutzer (der in dem anvisierten Szenario primär an *Themen* und erst sekundär an deren ggf. variabler Konzeptualisierung und Benennung interessiert ist) von verschiedenen bedeutungs*ähnlichen* Termini, die von einzelnen Autoren zu Zwecken der Konzeptualisierung und Benennung desselben Themas geprägt wurden, zu ein- und demselben Synset unseres Wortnetzes führen. Von diesem Synset aus kann er sich dann über die Beziehung des damit repräsentierten Themas zu anderen Themen der Domäne orientieren. Dass Bedeutungs*ähnlichkeit* (im Sinne einer graduellen *Unähnlichkeit*) von Lexemen in Fachsprachen letztlich auch zugleich eine Unähnlichkeit der damit bezeichneten Konzepte bedeuten kann, wird dabei nicht geleugnet. Letztlich sollen sowohl *TermNet* als auch die auf seiner Basis generierten Linkangebote als Orientierungshilfen bei der selektiven Fachtextrezeption dienen. Für ein differenzierteres Eindringen in die Fachdomäne (zumindest, wenn sie – wie im *HyTex*-Korpus – durch linear, d.h. nicht speziell in Hinblick auf eine selektive Rezeption hin organisierte Texte repräsentiert ist) empfiehlt sich nach wie vor die Lektüre eines Textes von Anfang bis Ende (also gemäß dem vom Autor geplanten Rezeptionsverlauf). Da in vielen Situationen der Beschäftigung mit Fachtexten (Studium, Fachjournalismus, interdisziplinäres Arbeiten) die Bewältigung großer Dokumentenmengen unter Zeitdruck erfolgt und daher nur selektiv möglich ist, erscheint uns ein Wort- und Konzeptnetz, wie wir es mit *TermNet* aufbauen, dennoch als eine wertvolle Verständnishilfe (welche aber nicht beansprucht, die lineare und ausführliche Textrezeption qualitativ zu ersetzen).

3.2 Lexikalische Relationen

Als *lexikalische Relationen* fassen wir bei der Modellierung von *TermNet* solche Relationen, die zwischen Einheiten bestehen, die insgesamt ein Synset konstituieren. Aufgrund der oben beschriebenen fachsprachenspezifischen Abhängigkeit der strukturellen Gliederung des Konzeptbereichs von autorindividuellen Terminologisierungsoperationen sind die Synsets in *TermNet* so flexibel konzipiert, dass sie zwar im Idealfall Konzepte repräsentieren, im Falle konzeptueller Varianz aber auch für *thematische Ausschnitte* aus dem insgesamt in der Fachtextdomäne behandelten Gegenstandsbereich stehen können. Kriterium für die Zugehörigkeit zweier terminologischer Einheiten zu ein- und demselben Synset ist somit im Zweifelsfalle nicht nur die *konzeptuelle*, sondern auch die *thematische Identität*. Für die praktische Modellierungsarbeit bedeutet dies, dass zwei Termini, die (bei Betrachtung der von den jeweiligen Autoren qua Definition zugeordneten Konzepte) zwar partiell unterschiedliche Konzepte lexikalisieren, trotzdem demselben Synset zugewiesen werden können. So sind beispielsweise die Termini `Verknüpfung` (nach Kuhlen 1991) und Verweis (nach Tochtermann 1995) zwar konzeptuell unterschiedlich, aber thematisch identisch und in *TermNet* daher als Mitglieder desselben Synsets LINK[10] ausgewiesen.

Um Zusammenhänge zwischen den Termini, die demselben Synset angehören, in Hinblick auf die Besonderheiten des Sprachgebrauchs in der Fachtextdomäne differenzierter beschreiben zu können, haben wir in Erweiterung des *WordNet*-Modells (bei welchem die Mitglieder eines Synsets ganz allgemein über die Relation

der Bedeutungsähnlichkeit verbunden sind) einige weitere Relationen eingeführt, die als Spezifizierungen der Relation der Bedeutungsähnlichkeit aufgefasst werden können (siehe Abb. 2). So beschreibt die (symmetrische) Relation `ist_orthographische_Variante_zu` eine Synonymiebeziehung, die sich durch Varianz in der Schreibung eines Terminus ergibt (Beispiele: `referenzieller Link ist_orthographische_Variante_zu referentieller Link` und `Hyper-Link ist_orthographische_Variante_zu Hyperlink`). Die Relationen `ist_Akronym_zu` (mit der Konverse `ist_Vollform_zu`) und `ist_Abkürzung_zu` (mit der Konverse `ist_Expansionsform_zu`) beschreiben zwei terminologische Ausdrücke als synonym hinsichtlich der Tatsache, dass sich die Differenz ihrer Ausdrucksseiten auf wortbildungsmorphologische Prozesse zurückführen lässt (Beispiele: `HTML ist_Akronym_zu Hypertext Markup Language` und `Link ist_Abkürzung_zu Hyperlink`).

Eine weiteres Phänomen, dessen Berücksichtigung speziell für solche Domänen (wie die in unserem Korpus dokumentierten) relevant ist, deren Konzepte im Englischen entwickelt und terminologisiert und dann auf das Deutsche übertragen wurden, ist das der *sprachkontaktbedingten Lexemkonkurrenz*. Hierunter fassen wir solche Fälle der Synonymie, die sich daraus ergeben, dass ein englischer Ausdruck im Deutschen sowohl als Lehnwort als auch in Form einer oder mehrerer Lehnübersetzungen existiert, die zwar im Ausdruck verschieden sind, aber identisch verwendet werden. Da *TermNet* dem Nutzer die Möglichkeit bieten soll, zu einem gesuchten Konzept oder Thema sämtliche Ausdrücke zu ermitteln, die dieses Konzept oder Thema terminologisch lexikalisieren, werden Lehnwort- und Lehnübersetzungsbeziehungen explizit als solche modelliert. Zu diesem Zweck werden bei Bedarf

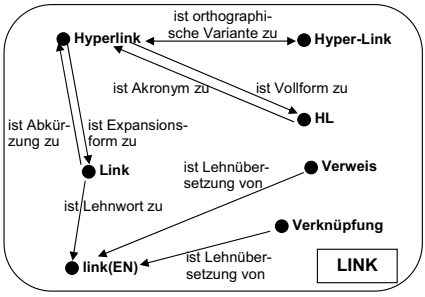

Abbildung 2: Lexikalische Relationen zwischen den Mitgliedern des Synsets LINK.

die entsprechenden englischen Lexeme in die betreffenden Synsets mitaufgenommen, um dann mit ihren deutschen Äquivalenten in folgender Weise verbunden werden zu können: `Link ist_Lehnwort_zu link(EN)` bzw. `Verweis ist_Lehnübersetzung_von link(EN)` bzw. `Verknüpfung ist_Lehnübersetzung_von link (EN)`. Für den Nutzer können die Termini `Link`, `Verweis` und `Verknüpfung` dann paarweise als *Lokalisierungsvarianten* ausgewiesen werden, die sich auf unterschiedliche Arten der lexikalischen Übernahme des mit dem englischen Ausdruck `link` verbundenen Konzepts in die deutsche Fachsprache zurückführen lassen. Durch die explizite Darstellung solcher durch Sprachkontakt verursachter Fälle von Synonymie kann *TermNet* zugleich eine „Brückenfunktion" erfüllen, insofern Lehnwörter und Lehnübersetzungen zu ihren jeweiligen englischen Ursprungswörtern in Beziehung gesetzt werden und sich somit auch die Möglichkeit eines Zugangs zur englischen Fachsprache eröffnet.

3.3 Konzeptuelle Relationen

Anhand der vorgestellten lexikalischen Relationen strukturieren wir Mengen von Termini als Mitglieder einzelner Synsets. Die Synsets wiederum repräsentieren – in unserer Modellierung wie auch in *WordNet* – *Konzepte* (mit der Einschränkung, dass in Fällen terminologiebil-

dungsbedingter konzeptueller Varianz in *TermNet* Termini auch unter dem weitergefassten Kriterium der thematischen Identität zu Synsets gruppiert werden können). Die Menge der vorhandenen Synsets wird über *konzeptuelle Relationen* strukturiert. Zentral ist hierbei die *Hyponymie*-Relation (`ist_hyponym_zu` mit der Konverse `ist_hyperonym_zu`), anhand derer der Konzeptbereich in hierarchischer Gliederung dargestellt werden kann und aus deren Anwendung sich paarweise Kohyponymie-Beziehungen für demselben Mutterknoten untergeordnete Geschwisterkonzepte ableiten lassen.

Zur vertikalen Strukturierung reicht die Hyponymie-Relation oft nicht aus, daher wird auch in *TermNet* die *Meronymie* als weitere, hierarchisierende Relation modelliert. Zur Beschreibung von Meronymiebeziehungen orientieren wir uns an den in *EuroWordNet* unterschiedenen Mernoymietypen (vgl. Vossen 1998), verwenden aber letztlich nur zwei davon, da für die zu modellierende Fachtextdomäne sowie in Hinblick auf das in *HyTex* verfolgte Anwendungsszenario lediglich Meronymiebeziehungen des Typs *Konstituenz* und *Gruppenzugehörigkeit* relevant sind. Die Bezeichnungen für die betreffenden Relationen (und ihre jeweiligen Konversen) wurden aus *EurWordNet* übernommen. So beschreiben wir die Konstituenzbeziehung zwischen `MODUL` und `HYPERTEXT` als `MODUL has_holo_part HYPERTEXT` und die Konverse als `HYPERTEXT has_mero_part MODUL`. Die Gruppenzugehörigkeit von `XHTML`, `XTM` und `NITF` zu `XML-SPRACHENFAMILIE` wird beschrieben als `XHTML, XTM, NITF has_holo_member XML-SPRACHENFAMILIE`, die Konverse als `XML-SPRACHENFAMILIE has_mero_member XHTML, XTM, NITF`.

Neben der Hyponymie-/Hyperonymie- sowie der Meronymie-/Holonymie-Relation umfasst unser Inventar an konzeptuellen Relationen auch noch die Relation der *Antonymie*. In Hinblick auf die anvisierte Anwendung (Linking nach Kohärenzkriterien) hat für die *TermNet*-Modellierung eine Opposition zwischen Konzepten ein stärkeres Gewicht als eine Opposition zwischen einzelnen Lexemen. Während es im Standard-*WordNet*-Konzept der Antonymie insbesondere um die Erfassung stilistischer Feinheiten geht (z.B.: `ascend` steht in lexikalischer Opposition zu `descend` und nicht zu `go down`), ist es für unser Anwendungsszenario wichtiger, dem Rezipienten zu vermitteln, dass zwei Konzepte – z.B. `WOHLGEFORMT` und `NICHT-WOHLGEFORMT` aus der Domäne „Texttechnologie" – sich wechselseitig ausschließen, als ihm zu vermitteln, ob der Terminus `nicht-wohlgeformt` eher mit dem Terminus `wohlgeformt` oder dem synonymen Lehnwort `well-formed` kontrastiert.

In Hinblick auf die Modellierung fachsprachlicher Domänen reicht die aus der Hyponymie-/Hyperonymie-Beziehung abgeleitete Kohyponymie-Beziehung in einigen Fällen nicht aus, um spezielle Ausschlussbeziehungen zwischen Paaren oder Gruppen von Kohyponymen, die aus typologisch motivierten terminologischen Konzeptualisierungsprozessen resultieren, zu beschreiben. So stehen beispielsweise bestimmte Hyponyme zu `LINK` nicht gleichrangig nebeneinander, sondern schließen sich bezüglich der Möglichkeit ihres Zutreffens auf einen konkreten Vertreter des Typs `LINK` gruppenweise wechselseitig aus: `INTRAHYPERTEXTUELLER LINK`, `INTERHYPERTEXTUALLER LINK` und `EXTRAHYPERTEXTUELLER LINK` schließen sich wechselseitig aus, während dies z.B. für `INTERHYPERTEXTUELLER LINK` und `UNIDIREKTIONALER LINK` nicht der Fall ist. Die dem Konzept `LINK` untergeordneten Hyponyme bilden Gruppen, die in sich jeweils durch ein bestimmtes (durch typologische Differenzierung bei der terminologischen Konzeptualisierung motiviertes) klassifikatorisches Merkmal bestimmt sind. So kann man

die Klasse der Links (als Klasse von Objekten in der Welt) nach dem Merkmal der Direktionalität z.B. in UNIDIREKTIONALE LINKS und BIDIREKTIONALE LINKS unterteilen, während man nach dem Verhältnis von Ausgangsanker und Zielanker zwischen INTRAHYPERTEXTUELLEN LINKS, INTERHYPERTEXTUELLEN LINKS und EXTRAHYPERTEXTUELLEN LINKS unterscheiden kann. Wenn man nur die Hyponymie-Relation kodiert, so wird ein für das Folgern wichtiger Aspekt verdeckt, nämlich derjenige, dass ein individueller Link zwar zugleich eine Instanz der Konzepte INTERTEXTUELLER LINK und UNIDIREKTIONALER LINK sein kann, aber nicht zugleich Instanz der Konzepte UNIDIREKTIONALER LINK und BIDIREKTIONALER LINK. Dieses Phänomen tritt natürlich nicht nur in Fachdomänen auf, sondern auch in der Allgemeinsprache: Ein individuelles Pferd kann zwar gleichzeitig zu den Klassen HENGST und RAPPE gehören; es kann aber nicht gleichzeitig RAPPE und SCHIMMEL sein, zumindest nicht in der uns vertrauten nicht-fiktionalen Tierwelt. Die Kohyponyme auf derselben Hierachiestufe können also durch klassifikatorische Merkmale weiter struktuiert sein, was dazu führt, dass den Kohyponymen mit demselben Klassifikationsmerkmal extensionional disjunkte Teilmengen von Instanzen entsprechen.

Das Standard-Modell von *WordNet* bietet bislang keine Lösung, um diesen Sachverhalt zu erfassen. Aus diesem Grund haben wir das Modell erweitert, indem wir Attribute eingeführt haben, über die wir inferieren können, welche Gruppen (Mengen) von Kohyponymen in einer Relation wechselseitiger *Disjunktivität* stehen (siehe Abb. 3). Durch die Zuweisung von Attributen, anhand derer sich einzelne Mengen von alternativen Konzepten über einen einmalig zu vergebenden Attributwert (z.B. „Direktionalität“) definieren lassen, kann Disjunktivität zwischen Kohyponymenmengen modelliert werden, *ohne* dass dabei auf die Einführung von künstlichen Konzepten (Pseudokonzepten in der hierarchischen Modellierung) zurückgegriffen werden muss, die anschließend auf der Präsentationsebene wieder herausgefiltert werden müssen.

4 Nutzung von TermNet für das automatische Linking: Ein Anwendungsbeispiel

Das richtige Verständnis von Fachtexten hängt zu einem nicht unerheblichen Anteil davon ab, dass die in der Domäne eingespielten terminologischen Einheiten mit den entsprechenden Konzepten verbunden und Unterschiede und Ähnlichkeiten zwischen verschiedenen Benennungen für dasselbe Konzept erkannt werden. Im Rahmen des im ersten Abschnitt skizzierten Strategietyps „Linking nach Wissensvoraussetzungen“ versuchen wir deshalb, Wissen über die Verwendungsregeln von Termini und über Bezüge zwischen den Konzepten unserer Fachtextdomäne über Linkangebote rekonstruierbar zu machen.[11] Die Generierung dieser Linkangebote erfolgt auf der Grundlage einer teilautomatischen Annotation sowohl von definitorischen Textsegmenten (also Textsegmenten, in denen Verwendungsregeln für Termini spezifiziert werden), als auch von Verwendungen der entsprechenden Termini in den Dokumenten (den Termverwendungsinstanzen). Um den Nutzern die für sie relevanten Informationen geben zu können, werden Termverwendungsinstanzen verlinkt mit automatisch generierten Glossarsichten, die Aufschluss über die spezifische Verwendung eines Terminus in der Fachtextdomäne geben.

Das in Abb. 4 gezeigte Beispiel für die nach diesem Prinzip generierten Sichten soll illustrieren, wie die Linkingstrategien in einem konkreten Nutzungskontext zusammenspielen. Ein Nutzer, der im Zuge der selektiven Textlektüre im oben rechts angezeigten Modul einsteigt, trifft auf eine Verwendungsinstanz des Terminus Link, die in der Hypertextsicht als Linkanzeiger

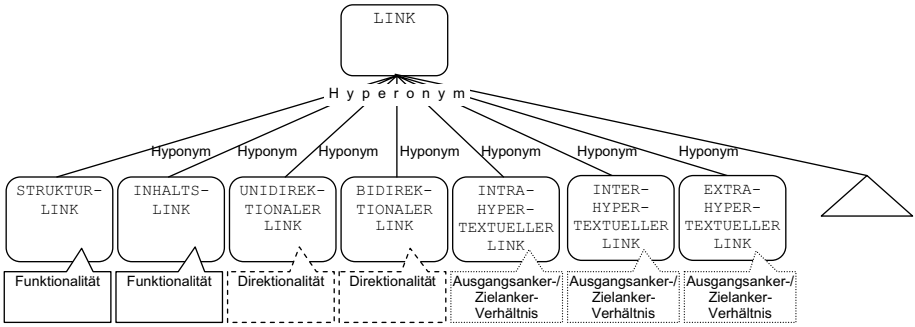

Abbildung 3: Veranschaulichung zur attributbasierten Modellierung der Disjunktivität in *TermNet*

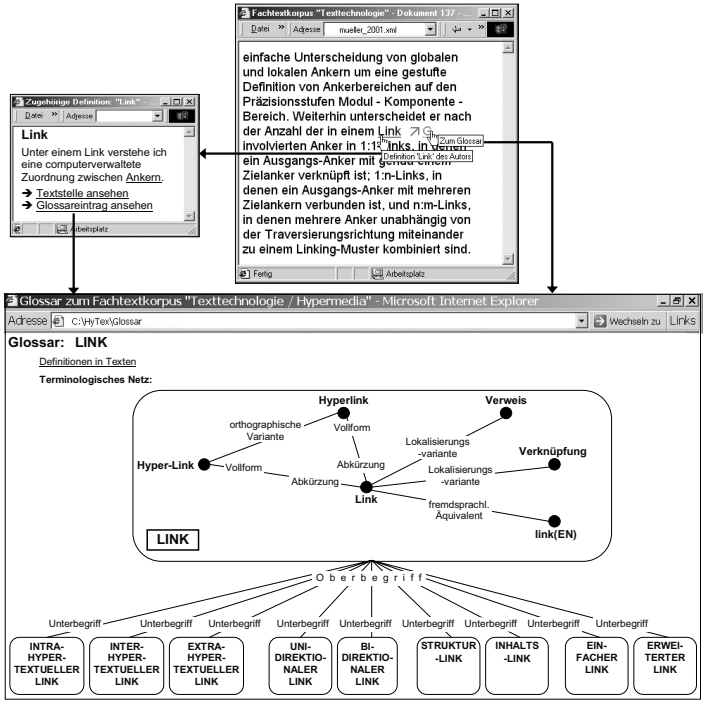

Abbildung 4: Veranschaulichung zur Vernetzung einer Termverwendungsinstanz (hier: „Link“) in einem Dokumentenmodul mit der zugehörigen Definition des Autors sowie mit einer automatisch generierten Glossarsicht, welche Ausschnitte aus dem terminologischen Netz (hier: das Synset LINK mit den zugehörigen lexikalischen Einheiten und seiner konzeptuellen Umgebung) in grafisch visualisierter Form präsentiert.

repräsentiert ist. Will sich der Nutzer über die Bedeutung des Terminus informieren, so erhält er durch Aktivierung des Linkanzeigers zunächst das oben links angezeigte Definitionsfenster, das die im Vortext vom Autor des Dokuments eingeführte Definition präsentiert. Diese enthält wiederum den Terminus `Anker`, der ebenfalls mit einer entsprechenden Definition und einem Glossareintrag verknüpft ist. Durch Aktivieren des Linkanzeigers „Textstelle ansehen" kann sich der Nutzer bei Bedarf das textuelle Umfeld der Definition anzeigen lassen; ein Mausklick auf den Linkanzeiger „Glossareintrag ansehen" erzeugt den im Fenster unten dargestellten Glossareintrag. Dieser Glossareintrag ist für Nutzer gedacht, die in der Lektüre des ursprünglichen Textes pausieren und sich zunächst vertiefend mit dem hinter dem Terminus `Link` stehenden Konzept vertraut machen möchten. Der Glossareintrag bietet als weitere Konzeptualisierungshilfen Verknüpfungen zu Definitionen des Terminus `Link` in verschiedenen Fachtextdokumenten an (diese Verknüpfungen operieren über dem Lexem `Link`). Des weiteren präsentiert der Glossareintrag einen grafisch visualisierten Ausschnitt aus dem terminologischen Netz. Dieser zeigt die lexikalische und konzeptuelle Umgebung des Terminus, wobei die Kanten durch Relationennamen gelabelt und die beiden Knotentypen (Konzept und Lexem) in der Darstellung unterschieden sind. Die Knoten des Netzes sind ebenfalls als Linkanzeiger gestaltet, bei deren Aktivierung ein entsprechender neuer Glossareintrag generiert wird. Auf diese Weise kann sich der Nutzer mit den jeweiligen Konzepten in derjenigen Detailtiefe auseinandersetzen, die seinem aktuellen Informationsbedarf und seinem Vorwissensstand am besten entspricht.

Unser Strategieansatz bietet dem Nutzer bei der selektiven Textlektüre somit zwei Arten von terminologiebezogenen Informationsangeboten:

- *Informationen über den Terminus selbst*. Dazu gehört sowohl die Angabe der Definition des Terminus durch den Autor des Primärtextes selbst als auch die Möglichkeit, andere mit demselben terminologischen Ausdruck verbundene Konzeptualisierungen in den Dokumenten des Fachtextkorpus einzusehen.
- *Informationen über die Beziehungen des Terminus zu anderen Termini*. Im Glossar werden zu jedem Terminus dessen lexikalische und konzeptuelle Relationen zu anderen Termini dargestellt. Hierbei ist es auch möglich, zu den jeweils „verwandten" Termini zu gelangen und von dort aus Informationen über diese zu erhalten.

TermNet spielt dabei sowohl bei der Verknüpfung der Termini mit den Termverwendungsinstanzen und den dafür relevanten Definitionen eine Rolle, als auch bei der Generierung der Glossarsichten und der Einbettung eines Konzepts in sein jeweiliges konzeptuelles und lexikalisches Umfeld.

Anmerkungen

[1] HyTex wird seit April 2002 an der Universität Dortmund durchgeführt (*http://www.hytex.info*) und ist ein Teilprojekt der Forschergruppe „Texttechnologische Informationsmodellierung" (*http://www.text-technology.de*), die sich mit den theoretischen Grundlagen und Methoden der Modellierung von Sprachdaten mit Markup-Sprachen (insbesondere XML und Tochterstandards) beschäftigt.

[2] Streng genommen handelt es sich nicht um Reversibilität im Sinne eines Umkehrprozesses, sondern die Hypertextualisierung erfolgt „on the fly" auf der Basis der textgrammatischen Annotationen der weiterhin in der ursprünglichen Form verfügbaren Ausgangstexte. Wir verwenden den Ausdruck „reversibel", um uns von Ansätzen zur Hypertextkonversion abzugrenzen, in denen der ursprüngliche Text irreversibel in Form und Struktur umgestaltet wird.

[3] Vgl. Beisswenger et al. 2002 und Lenz et al. 2002.

[4] Vgl. Storrer 2002.

[5] Da Fachdomänen, wenn man sie auf der Grundlage von Textkorpora (und nicht etwa wissenschaftssoziologisch oder in Hinblick auf die mündliche Fachkommunikation) betrachtet, stets nur über die in den jeweiligen Korpora (die grundsätzlich immer nur einen Ausschnitt der für die Fachdomäne relevanten bzw. konstitutiven Texte umfassen) enthaltenen Textdokumente repräsentiert werden, sprechen wir im Folgenden in bezug auf den Bezugsbereich unserer Modellierung von einer Fachtextdomäne. Dies ist zu lesen als „Die Fachdomäne, so wie sie sich in den zugrunde gelegten Texten zeigt". Den Ausdruck Fachdomäne verwenden wir nur dann, wenn wir uns ganz allgemein auf das Thema unserer Modellierung, nicht aber auf deren – mit dem Zuschnitt unseres Korpus gegebenen – Bezugsbereich beziehen.

[6] Für unser deutsches Korpus käme ein Nachbau des Ansatzes von Green ohnehin nur mit Hilfe des deutschen GermaNet (vgl. z.B. Kunze & Wagner 2001) in Frage.

[7] Der Ausdruck lexikalisch-semantische Wortnetze bezeichnet nach Kunze 2001 generell einen Modellierungsansatz für Sprach- und Konzeptwissen im Stile des WordNet-Projekts, der zunächst an der Universität Princeton für das Englische entwickelt und später für viele andere Sprachen ausgebaut wurde.

[8] Z.B. COHSE (Carr et al. 2001), OntoPortal und ESKIMO (Miles-Board et al. 2001), On2broker (Decker et al. 1998).

[9] Siehe hierzu Beisswenger et al. 2002.

[10] Um die unterschiedlichen Modellierungseinheiten in TermNet, wenn sie im Rahmen von Beispielen oder Erwähnungen benannt werden, typographisch von einander zu unterscheiden, geben wir hier und im Folgenden Lexeme (Termini) in Normalschrift, die Namen von Konzepten hingegen in Versalien an.

[11] Die verschiedenen Teilaspekte der Strategien sind in Beisswenger et al. 2002; Lenz et al. 2002 und Lenz et al. 2004 beschrieben.

Literatur

Beisswenger, M.; Lenz, E. A.; Storrer, A. (2002). „Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen." In: Busemann, S. (Hrsg.) (2002). Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002), Saarbrücken, September / Oktober 2002. Saarbrücken: Deutsches Institut für Künstliche Intellgenz (DFKI) [DFKI Document D-02-01], 187-191.

Carr, L.; Bechhofer, S.; Goble, C.; Hall, W. (2001). "Conceptual Linking: Ontology-based Open Hypermedia." In: Proceedings of the 10th International World Wide Web Conference, Hong Kong, May 2001, *http://www.ecs.soton.ac.uk/~lac/WWW10/ConceptualLinking.html* [accessed April 2004].

Decker, S. et al. (1998). "Ontobroker in a Nutshell." In: Constantine, N.; Stephanidis, C. (eds.) (1998). Research and Advanced Technologies for Digital Libraries. Proceedings of the 2nd European Conference on Research and Advanced Technology For Digital Libraries (ECDL'98). Berlin et al.: Springer [= LNCS 1513], 540-651, *http://citeseer.ist.psu.edu/89427.html* [accessed April 2004].

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Green, S. J. (1998). "Automated Link Generation: Can We Do Better than Term Repetition?" In: Proceedings of the 7th International World Wide Web Conference. Computer Networks and ISDN Systems, 30 (1-7) (1998), 87-84, *http://citeseer.nj.nec.com/green98automated.html* [accessed April 2004].

Kuhlen, R. (1991). Hypertext: ein nicht-lineares Medium zwischen Buch und Wissensbank. Berlin et al.: Springer [Edition SEL-Stiftung].

Kunze, C. (2001). „Lexikalisch-Semantische Wortnetze." In: Carstensen, K.-U. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: eine Einführung. Heidelberg, Berlin: Spektrum Akademischer Verlag, 386-393.

Kunze, C.; Wagner, A. (2001). „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche." In: Lemberg, I.; Schröder, B.; Storrer, A. (eds.) (2001). Chancen und Perspektiven computergestützter Lexikographie. Tübingen: Niemeyer [= Lexicographica Series Maior Vol. 107], 229-246.

Lenz, E. A.; Beisswenger, M.; Storrer, A. (2002). „Hypertextualisierung mit Topic Maps – ein Ansatz zur Untersützung des Textverständnisses bei der selektiven Rezeption von Fachtexten." In: Tolksdorf, R.; Eckstein, R. (Hrsg.) (2002). Proceedings Workshop XML-Technologien für das SemanticWeb (XSW 2002), Berlin, Juni 2003. Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-14], 151-159.

Lenz, E. A.; Birkenhake, B.; Maas, J. F. (2004). Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps. In diesem Band, 127-136.

Miles-Board, T.; Kampa, S.; Carr, L.; Hall, W. (2001). "Hypertext in the Semantic Web." In: Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HT '01), 237-238.

Storrer, A. (2002). "Coherence in Text and Hypertext." In: Document Design 3(2) (2002), 156-168.

Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

Eva Anna Lenz, Benjamin Birkenhake, Jan Frederik Maas

Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps

Abstract

Wir beschreiben exemplarisch anhand eines im Projekt „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex)[1] entwickelten Wortnetzes, wie ein nach dem WordNet-Modell strukturiertes Wortnetz mit texttechnologischen Methoden erstellt, gewartet, als XML Topic Map (XTM) repräsentiert, visualisiert und zum Hypertext-Linking genutzt werden kann. Im Mittelpunkt steht dabei die Präsentation im XTM-Format, die zu anderen technischen Repräsentationsformalismen für Wortnetze in Beziehung gesetzt wird.

1 Einleitung

Mittlerweile gibt es eine große Anzahl von WordNet-Derivaten (im folgenden „Wortnetze“ genannt) für unterschiedliche Sprachen und Anwendungen, denen allen die Struktur aus Lexemen, Konzepten (Synsets), lexikalischen Relationen und konzeptuellen Relationen gemeinsam ist. Sie werden technisch unterschiedlich repräsentiert, z. B. durch Datenbanken, verschiedenartige XML-Repräsentationen, oder RDFS.

Wir stellen eine weitere Repräsentation vor, die sich als Austauschformat für Wortnetze sehr gut eignet: **XML Topic Maps**.

XML Topic Maps können als standardisiertes Format für semantische Netze aufgefasst werden, das einige interessante Eigenschaften aufweist, darunter die Möglichkeit der Anbindung von Dokumenten oder Dokumentteilen an das semantische Netz und Mechanismen zur Vereinigung mehrerer Topic Maps. Da es sich um ein von der ISO standardisiertes, SGML- oder XML-basiertes Austauschformat handelt, ist es plattformunabhängig und kann mit unterschiedlicher Software verarbeitet werden.

In Abschnitt 2 beschreiben wir, wie Wortnetze zur Zeit repräsentiert werden. Am Beispiel eines terminologischen Wortnetzes, das wir im Rahmen des DFG-geförderten Projekts „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex, siehe auch Runte et al. 2003, in diesem Band) nutzen, zeigen wir in den Abschnitten 3 bis 7, wie ein Wortnetz erstellt, gepflegt, als XML Topic Map repräsentiert, visualisiert und zum Hypertext-Linking genutzt werden kann. In den Abschnitten 8 und 9 gehen wir kurz darauf ein, welche Vorteile die Topic-Map-Modellierung auch für andere Wortnetz-Projekte haben könnte und wie die Repräsentationen von Wortnetzen als XML Topic Maps und als RDF(S) einander ergänzen könnten.

2 Bisherige Wortnetz-Repräsentationen

Es gibt verschiedene Praktiken und Vorschläge zur Repräsentation von Wortnetzen. Lemnitzer und Kunze stellen eine konzeptuelle Modellierung in Form von Entity-Relationship-Diagrammen vor, die als Ausgangsbasis für verschiedene technische Repräsentationen – Textformate, Datenbanken oder XML-Formate – dienen kann (Kunze & Lemnitzer 2002; Lemnitzer & Kunze 2003).

Eine solche technische Repräsentation kann verschiedenen Zwecken dienen, z. B. als einfach zu editierendes, menschenles- und schreibbares Format, als intermediäres Format zum Austausch zwischen verschiedenen Verarbeitungsschritten, als Ausgangsformat für die Publikation in verschiedenen Medien, als Austauschformat für die Integration verschiedener Wortnetze,

oder als Speicherformat für den schnellen Zugriff in Information-Retrieval-Anwendungen.

Rein textbasierte Formate sind am wenigsten standardisiert und erfordern die Entwicklung spezieller Parser, um sie auszulesen. Ein Beispiel ist das vom Princeton WordNet verwendete, von Menschen leicht lesbare und editierbare textbasierte Format, das als Ausgangsbasis für die automatische Erzeugung eines Datenbankformats verwendet wird (Beckwith et al. 1993). Ein solches Datenbankformat wiederum ist zwar nur schwer oder überhaupt nicht mehr menschenlesbar (in Abhängigkeit von der verwendeten Datenbank), ermöglicht dafür aber einen schnellen maschinellen Zugriff, wie er für Anwendungen im Information Retrieval unabdingbar ist.

XML-basierte Formate stellen einen Kompromiss zwischen diesen beiden Extremen dar. Sie sind einerseits menschenlesbar, andererseits aber auch für die maschinelle Verarbeitung sehr gut geeignet. Sie stellen einen Schritt in Richtung Standardisierung dar, so dass eine Reihe von Tools zur Verarbeitung bereits zur Verfügung stehen (z. B. XML-Parser, die also für ein spezielles XML-basiertes Format nicht eigens entwickelt zu werden brauchen). Für einen effizienten Zugriff ist eine Konvertierung in ein Datenbankformat möglich. Es wurden mindestens zwei solcher XML-basierten Formate mit jeweils eigenen Dokumentgrammatiken (DTDs) für Wortnetze entwickelt: für GermaNet (Kunze & Lemnitzer 2002) und für einen WordNet-Editor (Pavelek & Pala 2002). Das für GermaNet entworfene Format benutzt zudem einen standardisierten Verweismechanismus, XLink (DeRose et al. 2001), zur Repräsentation der lexikalischen und der konzeptuellen Relationen.

Zwei in unterschiedlichen XML-Formaten repräsentierte Wortnetze sind jedoch noch lange nicht kompatibel. Während auf der Ebene der Syntax eine Standardisierung erreicht ist, wird nach wie vor spezielle Software benötigt, die die *Semantik* der speziellen XML-Modellierung auswertet. Mit zwei relativ neuen WWW-Standards kann die Semantik von netzartigen Informationsstrukturen, wie sie in Wortnetzen vorliegen, zumindest teilweise erfasst werden: Resource Description Framework (RDF, Lassila & Swick 1999) mit den darauf aufbauenden Standards RDF Schema (RDFS) und Web Ontology Language (OWL) sowie Topic Maps (Pepper & Moore 2001).

Bei RDF handelt es sich um ein sehr elementar gehaltenes Modell zur Beschreibung von Metadaten, wobei das Metadaten-Set allerdings beliebig erweiterbar ist, so dass sich nahezu beliebige Sachverhalte ausdrücken lassen. Mit RDFS, der Schemasprache von RDF, ist die Validierung von erweiterten RDF-Dokumenten möglich. OWL ist eine RDF(S)-basierte, sehr mächtige Sprache zur Definition von Ontologien, die im Wesentlichen ebenfalls für die Beschreibung von Metadatenstrukturen gedacht ist. Auf Topic Maps wird im Folgenden noch genauer eingegangen.

Durch die Nutzung von Topic Maps oder RDF für die Repräsentation von Wortnetzen wird ein weiterer Schritt in Richtung Standardisierung vollzogen, so dass die Austauschbarkeit und Verarbeitung durch Standard-Software erleichtert wird. Sowohl für RDF als auch für Topic Maps gibt es – neben anderen – eine XML-Syntax.

Ein Teil des in Princeton entwickelten ursprünglichen WordNet ist bereits in Form von RDF und RDFS aufbereitet worden und unter der Netzadresse *http://www.semanticweb.org/library/* abrufbar. In der dort vorgeschlagenen Modellierung werden jedoch nur Substantive, Glosses, SimilarTo-Relationen und Hyperonymie-Relationen berücksichtigt. Ein weitergehender Vorschlag zur Abbildung von Wortnetzen auf RDF(S)-Strukturen wird in Lemnitzer & Kunze 2003 beschrieben.

3 Arbeitsschritte von der Erstellung bis zur Nutzung

Im HyTex-Projekt verwenden wir ein nach den Prinzipien von WordNet modelliertes terminologisches Netz auf eine ganz spezifische Weise, nämlich zur Nutzung in einem ontologiebasierten Hypertext (Miles-Board et al. 2001). Im gesamten Projekt, also auch bei allen Verarbeitungsschritten des Wissensnetzes, verwenden wir standardisierte, XML-basierte texttechnologische Standards. Auf diese Weise können die entstehenden Zwischenprodukte anderen Projekten zur Verfügung gestellt werden, z. B. kann die im XTM-Format vorliegende Topic Map von jeder Art von Software genutzt werden, die XTM verarbeiten kann. Verarbeitungsschritte, die die von uns eingesetzte Software nicht leistet, können wir selbst programmieren, z. B. spezielle Arten von Inferenzen auf dem Wissensnetz. Ein weiterer Vorteil liegt darin, dass die von uns verwendeten Standards alle plattformunabhängig sind. Schließlich ist es uns jederzeit möglich, einzelne Komponenten, die zur Verarbeitung notwendig sind, durch andere zu ersetzen (z. B. einen anderen Editor zu verwenden).

Die nachfolgend genannten Arbeitsschritte, welche wir im HyTex-Projekt von der Erstellung bis zur Nutzung des terminologischen Netzes durchführen, haben daher nur beispielhaften Charakter – der texttechnologische Ansatz ermöglicht es, die Vorgehensweise an jeder beliebigen Stelle der Verarbeitungskette an andere Erfordernisse anzupassen:

1. Eingabe und Wartung des Wissensnetzes mit dem Werkzeug K-Infinity und Export in eine K-Infinity-eigene XML-basierte Repräsentation
2. Konvertierung des K-Infinity-Exportformats nach XML Topic Maps (XTM)
3. Durchführung von Inferenzen und Überprüfungen auf dem Wissensnetz mittels XSLT
4. Überführung in ein grafisch aufbereitetes, strukturiertes Glossar in einen Hypertext (repräsentiert in HTML und SVG) mittels XSLT.

An verschiedenen Stellen kommt in dieser Kette die Programmiersprache XSLT zum Einsatz, eine funktionale Programmiersprache, die darauf optimiert ist, XML-Dokumente einzulesen und zu erzeugen. Nur der erste dieser Schritte geschieht manuell, alle anderen können automatisch durchgeführt werden. Die Schritte werden in den folgenden Abschnitten genauer beschrieben.

4 Erstellung mit K-Infinity

Komplexe Wissensnetze aufzubauen und zu pflegen ist eine aufwändige Aufgabe, die man am besten mit einer spezialisierten Software erledigt, die über eine Visualisierungs- und Verwaltungskomponente für Netzstrukturen und Mechanismen zur Konsistenzprüfung verfügt. Solche Mechanismen sollten z.B. sicherstellen, dass beim Löschen eines Konzeptes im Wortnetz auch alle Lexeme gelöscht werden, die dieses Konzept lexikalisieren. Wird ein Konzept gelöscht, das in einer Hyperonym-Beziehung zu anderern Konzepten steht, muss geprüft werden, wie mit den Hyponymen verfahren werden soll, ob sie gelöscht oder mit dem nächsthöheren Konzept in der Hierarchie (als Hyponyme) verbunden werden sollen. Auch die Vergabe von IDs und Verweisen auf IDs ist eine aufwändige und fehleranfällige Aufgabe, für die man am besten spezialisierte Editoren nutzen sollte.

Für die Forschungen im HyTex-Projekt, speziell für den Aufbau und die Pflege des TermNet, hat uns die Firma intelligent views das komfortable Werkzeug *K-Infinity* zur Verfügung gestellt (vgl. *http://www.i-views.de/*). K-Infinity unterstützt den Aufbau und das Editieren von Wissensnetzen mit einer grafischen Oberfläche (siehe Abb. 1), in der die Entitäten des Wissensnet-

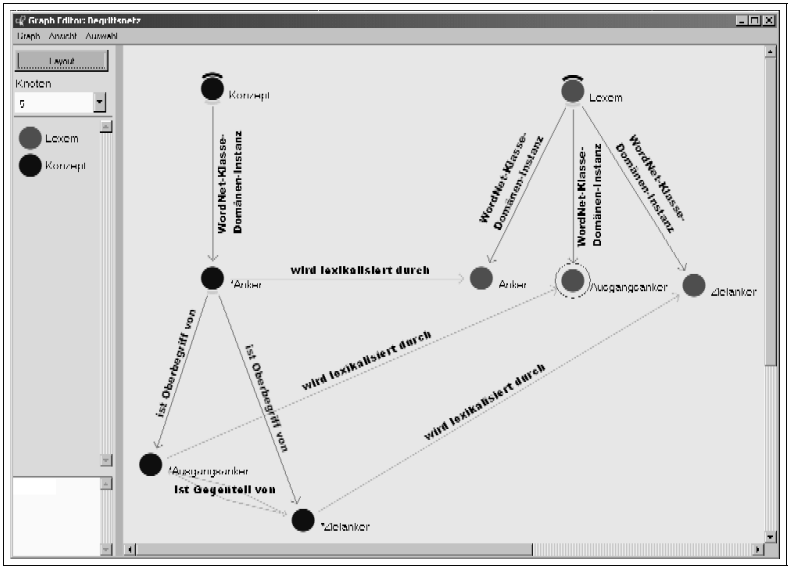

Abbildung 1: Das Werkzeug zur Verwaltung von Wissensnetzen K-Infinity ermöglicht u. a. eine grafische Eingabe von Begriffen und Relationen.

zes direkt manipuliert werden können. Es automatisiert die Verwaltung von IDs und das Umsetzen von Verweisen beim Löschen von Konzepten (in K-Infinity als „Begriffe" bezeichnet). Weil auch nachträgliche Umbenennungen von Konzepten und Relationen ohne größeren Aufwand möglich sind, kann der Aufbau des Netzes flexibel an die Anforderungen der jeweiligen Anwendung angepasst werden, was gerade in einem Forschungsprojekt von großem Vorteil ist.

Allerdings gibt es gerade bei der Modellierung von Wortnetzen (im Stil des WordNet) eine Reihe von spezifischen Anforderungen (z. B. bestimmte Inferenzen), die mit K-Infinity nicht durchgeführt werden können. Der in HyTex verfolgte texttechnologische Ansatz ermöglicht es uns aber, sie an einer späteren Stelle in der Verarbeitungskette (siehe Abschnitt 6) zu lösen. K-Infinity bietet nämlich die Möglichkeit, ein eingegebenes Wissensnetz in ein K-Infinity-eigenes XML-basiertes Format zu exportieren. Die Firma intelligent views hat uns darüber hinaus ein XSLT-Stylesheet zur Verfügung gestellt, das dieses Format automatisch in das XML Topic Map-Format überführt. Die dann vorliegende Repräsentation wird im Folgenden beschrieben.

5 Repräsentation des terminologischen Netzes als XML Topic Map

Topic Maps stellen eine standardisierte Notation zur Repräsentation von Netzwerken aus Informationseinheiten dar. Es gibt zwei syntaktische Varianten des Topic Map Standards: 1999 erschien er als ISO-Standard (ISO, 2000), der auf einer SGML-Syntax basiert. Im Jahr 2001 wurde eine XML-Syntax zur Nutzung im WWW für den ISO-Standard entwickelt: XML Topic Maps (XTM, Pepper & Moore 2001), ein Industriestandard, der inzwischen in den ISO-Standard integriert wurde. Eine sehr gute Einführung in Topic Maps findet sich bei Rath 2002. Wir beschreiben hier nur diejenigen Eigenschaften

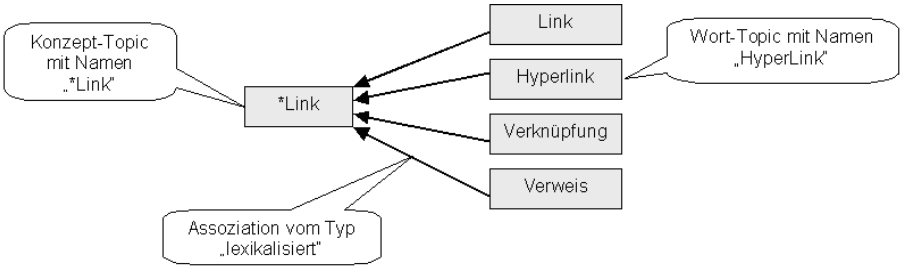

Abbildung 2: Abbildung von Konzepten und Lexemen auf Topics in der Topic Map.

von Topic Maps, die für die Repräsentation von Wortnetzen interessant sind.

Grundbausteine von Topic Maps – die Knoten des Netzwerks – sind sogenannte *Topics*. Die semantischen Beziehungen zwischen Topics – die Kanten – heißen *Assoziationen* (associations). Ein Topic kann über *Topic-Anker* (occurrences) zudem mit beliebigen adressierbaren Ressourcen, z. B. HTML-Dokumenten, verknüpft werden. In diesem Fall lassen sich Topics und Assoziationen als Metadaten zu den Dokumenten betrachten, und es ergeben sich vielfältige Anwendungsmöglichkeiten für Navigation und Suche, die wir hier aber nicht diskutieren.

Topics, Assoziationen und Topic-Anker lassen sich typisisieren. Die Typen sind in Topic Maps selbst wiederum Topics, so dass sich über Typen innerhalb desselben Formalismus Aussagen machen lassen, z. B. können Typen selbst wieder Typen haben oder durch Assoziationen mit anderen Topics verbunden werden. Dies macht Topic Maps zu einem sehr mächtigen Standard. Die Typisierung ist äquivalent mit einer Kante vom Typ *instance-of*, durch sie wird also ausgedrückt, dass zwischen einem Topic und seinem Typ eine Klasse-Instanz-Relation besteht.

Jedes Topic hat einen obligatorischen eindeutigen Bezeichner (ID), und (optional) einen oder mehrere Namen.

Entsprechend der in Wortnetzen vorhandenen grundlegenden Unterscheidung zwischen Konzepten und Lexemen enthält unsere Topic Map zwei Arten von Topics: Konzept-Topics und Wort-Topics.

Konzept-Topic: Für jedes Konzept der Domäne, für das es terminologisierte Ausdrücke gibt, führen wir ein Topic ein. Es hat einen Namen, der mit einem * gekennzeichnet ist. Jedes Konzept-Topic verbinden wir über eine Assoziation vom Typ `WordNet-Klasse-Domänen-Instanz` mit einem Topic des Namens *Konzept*.

Wort-Topic: Für jeden in der Fachdomäne terminologisierten Ausdruck – d.h. jedes Lexem, das einen Terminus darstellt – deklarieren wir ebenfalls ein Topic in der Topic Map. So erhalten wir Wort-Topics mit den Namen *Link*, *Hyperlink*, *Verknüpfung* und *Verweis*. Jedes Wort-Topic verbinden wir ebenfalls über eine Assoziation vom Typ `WordNet-Klasse-Domänen-Instanz` mit einem Topic des Namens *Lexem*.

Um nun die Zugehörigkeit der Wort-Topics zu ihrem jeweiligen Konzept-Topic auszudrücken, werden alle Wort-Topics eines Konzepts mit dem Konzept-Topic durch eine Assoziation vom Typ `lexikalisiert` verbunden (vgl. Abbildungen 1 und 2).

Lexikalische und konzeptuelle Relationen lassen sich nun auf getypte Assoziationen zwischen den Wort-Topics bzw. den Konzept-Topics abbilden. Ein Assoziationstyp für eine lexikalische Re-

lation ist `ist Abkürzung für`. Diese Relation besteht z. B. zwischen den Lexemen (bzw. Wort-Topics) *Link* und *Hyperlink*. Ein Beispiel für eine konzeptuelle Relation ist die Hyperonymie, sie besteht z. B. zwischen den Konzepten (bzw. Konzept-Topics) **Link* und **1:n-Link*. Die Assoziationstypen werden selbst wiederum als Topics repräsentiert.

Im HyTex-Projekt wird aus der Topic Map später vollautomatisch ein Glossar erzeugt. Für die hypertextuelle Verwendung verbinden wir die Wort-Topics durch Topic-Anker (occurrences) mit verschiedenen Textstellen des Korpus, z. B. mit Definitionen der Termini und mit Termverwendungsinstanzen. Diese Topic-Anker haben ebenfalls verschiedene Typen. Daraus werden später automatisch (typisierte) Hyperlinks von den Korpus-Dokumenten zum Glossar und umgekehrt erzeugt.

Einige Konzepte werden in unserer Modellierung des terminologischen Netzes durch ein Attribut gekennzeichnet, um verschiedene Arten von Ko-Hyponymie in Fachsprachen ökonomisch ausdrücken zu können, siehe dazu Runte et al. 2003, in diesem Band. Ein Attribut wird ebenfalls durch einen Topic-Anker repräsentiert, welcher auf eine in das Konzept-Topic eingebettete Ressource verweist, die den Attributwert enthält.

Ein weiteres Konstrukt von Topic Maps – neben Topics, Topic-Typen, Assoziationen, Assoziationstypen, Topic-Ankern und Topic-Anker-Typen – ist das Konstrukt des Skopus (scope). Ein Skopus ist ein Gültigkeitsbereich oder Kontext, in dem eine Aussage gültig ist. Im Semantic Web ist diese Möglichkeit von großer Bedeutung, da unterschiedliche Organisationen, Gruppen, und Menschen verschiedene Sichtweisen auf die Welt haben, die in Topic Maps nebeneinander existieren können. Zum Beispiel kann derselbe Gegenstand unterschiedlich benannt werden, oder in verschiedenen Klassifikationssystemen unterschiedlich eingeordnet werden. Auch ein Skopus ist selbst wieder ein Topic. Bei Bedarf benutzen wir die Skopen, um die Topics der Domäne von denjenigen Topics zu unterscheiden, die der WordNet-Modellierung dienen: Alle Topics, die innerhalb der Topic Map als Typen dienen (z. B. das Topic mit dem Namen *ist Abkürzung für*) bekommen den speziellen Skopus *TermNet* zugewiesen. Es wurde vorgeschlagen, eine solche Trennung zwischen „regulären“ und „deklarativen“ Topics auf andere Weise vorzunehmen (Topic Map Templates, Rath 2000). Eine Standardisierung dieses Verfahrens in Form einer Schemasprache für Topic Maps (Topic Map Constraint Language, TMCL, *http://www.isotopicmaps.org/tmcl/*) ist jedoch noch nicht abgeschlossen.

6 Inferenzen und Überprüfungen auf dem Wissensnetz

Die Datenhaltung in der Topic Map ist redundanzfrei: Redundante Information, wie z. B. die explizite Modellierung der Synonymie-Relation, ist aus Gründen der schwierigeren Lesbarkeit, Änderbarkeit und weniger effizienten Speicherung und Verarbeitung nicht erwünscht. Für einen Nutzer, dem das Wissensnetzes auszugsweise präsentiert wird, kann solche redundante Information jedoch sehr wertvoll sein. Aus diesem Grund führen wir zwei Arten von Inferenzen auf dem Wissensnetz aus, die explizite Relationen erzeugen, welche vorher nur implizit vorhanden waren:

1. Wir inferieren die Synonymie-Relation zwischen verschiedenen Lexemen aus ihrer Verbindung mit dem Konzept (über die `lexikalisiert`-Relation). Das Ergebnis wird dem Nutzer grafisch präsentiert (Abbildung 3).
2. Die Relation der Disjunktivität wird aus den in Abschnitt 5 eingeführten Attributen und Attributwerten abgeleitet: Ko-Hyponyme, die gleiche Attributwerte aufweisen, sind disjunkt.

Außerdem führen wir begrenzt Konsistenzprüfungen durch, die für unser Wissensnetz spezifisch sind und daher nicht schon von K-Infinity vorgenommen werden können. Wir überprüfen, ob jedes Konzept mit mindestens einem Lexem verbunden ist und ob jedes Lexem mindestens einem Konzept zugeordnet ist. Wenn dies nicht der Fall ist, werden Warnmeldungen ausgegeben. Von Fischer wurden eine Reihe weiterer Konsistenzprüfungen für Wortnetze vorgeschlagen und durchgeführt; für WordNet: Fischer 1997, für GermaNet: Gupta 2002. Einige dieser Überprüfungen müssen wir nicht vornehmen, da sie bereits durch den Wissensnetz-Editor K-Infinity abgefangen werden (z.B. zyklische oder „abgekürzte" Hyperonymie-Relationen) oder sich auf Relationen beziehen, die im TermNet nicht verwendet werden. Die Durchführung anderer Überprüfungen ist wünschenswert.

Die Inferenzen und Überprüfungen geschehen derzeit mit einem XSLT-Stylesheet. Prinzipiell sind natürlich beliebige andere Inferenzen auf dem Wissensnetz möglich. Das Ergebnis der Inferenzen ist wiederum eine in XTM repräsentierte Topic Map, die gegenüber der ursprünglichen Topic Map um einige Assoziationen erweitert ist. Die Programmierung in XSLT ist relativ aufwändig, da die Netzstruktur in XTM nicht direkt auf das (intrinsisch hierarchische) XML-Modell abgebildet werden kann und XTM daher stark mit Verweisen arbeitet. Von der ISO wird jedoch derzeit ein Standard geplant, der eine Anfragesprache für Topic Maps spezifiziert, vergleichbar mit SQL für relationale Datenbanken. Diese „Topic Map Query Language" (TMQL) wird es später erlauben, Inferenzen auf einer Topic Map auf einfachere Weise zu formulieren, als es derzeit mit XSLT möglich ist.

7 Nutzung als Hypertext und Visualisierung

Aus der Topic Map wird automatisch – ebenfalls durch ein XSLT-Stylesheet – ein Glossar erzeugt, aus dem der Benutzer Informationen über Fachtermini gewinnen kann, siehe auch Runte et al. 2003, in diesem Band, und Lenz et al. 2002. Aus jedem Wort-Topic (d. h. zu jedem Lexem) wird ein Glossareintrag erzeugt, der verschiedene Links in das Korpus beinhaltet, z. B. Links zu Definitionen des Terminus. Diese werden aus den Topic-Ankern erzeugt. Umgekehrt wird von Termverwendungsinstanzen im Korpus auf Glossareinträge verlinkt.

Obwohl der Ausdruck „Topic Map" eine grafische und räumliche Dimension impliziert und zur Veranschaulichung des Konzepts und der Vorteile von Topic Maps oft Beispiel-Visualisierungen herangezogen werden, beinhaltet der Standard selbst keine grafischen Komponenten. Da XTM XML-basiert ist, bietet es sich an, XSLT zu nutzen, um XTM-Daten in eine hochwertige grafische Präsentation zu transformieren, die tatsächlich das Kartenhafte an Topic Maps erkennen lässt.

HTML bietet für eine angemessene Umsetzung komplexer XTM-Strukturen nicht genügend Mittel. Abhilfe schafft der seit 2001 als W3C-Recommendation vorliegende, ebenfalls XML-basierte Standard Scalable Vector Graphics (SVG). SVG-Dokumente lassen sich über ein Plugin in (X)HTML-Dokumente einbetten und bieten eine Javascript-Schnittstelle, so dass auch anspruchvolle interaktive Karten mit verhältnismäßig wenig Aufwand automatisch aus XTM-Daten erstellt werden können, die den Anspruch von Topic Maps, Landkarten für das semantische Netz zu sein, erfüllen.

Eine relativ einfache Visualisierung jedes Lexems und seiner lexikalischen und konzeptuellen Relationen, die den jeweiligen Glossareintrag ergänzt und in diesen eingebettet wird, erzeugen wir auf diese Weise automatisch aus der Topic Map (Abbildung 3).

Abbildung 3: SVG-Visualisierung eines Lexems, wie sie einem Hypertextnutzer präsentiert wird. Lexeme und Konzepte werden verschiedenfarbig dargestellt, zu jedem Konzept werden dessen Lexeme angegeben. Durch Anklicken eines anderen Lexems kann der Nutzer zu dessen grafischer Darstellung navigieren.

8 Wortnetze als Topics Maps?

Nachdem wir am Beispiel unseres terminologischen Netzes beschrieben haben, wie sich Wortnetze als Topic Maps modellieren lassen, möchten wir an dieser Stelle einen Ausblick geben, welche Vorteile sich für die WordNet-Community ergeben könnten, wenn diese Modellierung für verschiedene Wortnetze verwendet würde.

Eine interessante Eigenschaft von Topic Maps liegt in der Möglichkeit, festzulegen, dass ein Topic mit einem anderen Topic identisch sein soll. Dies geschieht über einen URI-Verweis. Die beiden Topics müssen nicht zwangsläufig in der derselben Topic Map vorliegen, sondern es können auch „öffentliche" Topics deklariert werden, die sogenannten Published Subject Indicators (PSIs), auf die allgemein Bezug genommen werden kann und soll.

Solche öffentlichen Topics werden von Standardisierungsorganisationen, Firmen und anderen Organisationen herausgegeben. Die Organisation OASIS (Organization for the Advancement of Structured Information Standards) hat z. B. PSIs für Länder und Sprachen der Welt herausgegeben. Wenn nun zwei verschiedene Topic Maps z. B. auf das öffentliche Topic für die Sprache Deutsch referieren, dann ist klar, dass in jedem Fall dasselbe gemeint ist, auch wenn es in der einen Topic Map *allemand* und in der anderen *german* heißt.

Für das Vereinigen von Topic Maps (*Merging*) gibt der XTM-Standard explizite Regeln an, die Topic-Map-konforme Software implementieren muss. Eine dieser Regeln bewirkt, dass zwei Topics, die auf dasselbe öffentliche Topic referieren, miteinander vereinigt werden. Das so entstandene neue Topic enthält alle Eigenschaften der beiden ursprünglichen Topics. Im Beispiel würde das vereinigte Topic für die deutsche Sprache also beide Namen tragen und die Assoziationen zu anderen Topics aus beiden Topic Maps übernehmen.

Von diesem Mechanismus könnte die WordNet-Community profitieren. Man könnte PSIs

für verschiedene WordNet-Relationen veröffentlichen, ebenso für *Lexem* und *Synset*. Dann können dieselben Relationen in konkreten Wortnetzen unterschiedlich benannt werden, durch Verweis auf den PSI wird ihre Identität sichergestellt, und gleiche Relationen würden in einem möglichen Merging-Prozess miteinander vereinigt. Durch Wahl entsprechender Skopen kann sichergestellt werden, dass die Herkunftsinformationen und ursprünglichen Namen erhalten bleiben. Wenn sogar für jedes Konzept ein PSI eingeführt wird – entsprechend dem Interlingual Index, wie er in EuroWordNet (Vossen 1998) verwendet wird – dann würden auch die Konezepte beim Merging vereinigt. Der Merging-Prozess an sich kann durch standardkonforme Topic-Map-Software durchgeführt werden.

Für die Hyperonymie definiert der XTM-Standard bereits ein öffentliches Topic. Es trägt den Namen *superclass-subclass relationship* im Skopus der englischen Sprache, aber wir sind frei, ein neues öffentliches Topic mit dem Namen *Hyperonymie* festzulegen und seine Identität mit dem superclass-subclass-Topic durch einen URI-Verweis auszudrücken.

9 Ausblick

Wir haben anhand des im HyTex-Projekts verwendeten terminologischen Netzes exemplarisch aufgezeigt, wie ein Wortnetz mit XML Topic Maps repräsentiert und mit texttechnologischen Methoden weiterverarbeitet werden kann. Im letzten Abschnitt haben wir gezeigt, dass eine Topic Map-Repräsentation sehr gut dazu geeignet ist, verschiedene Wortnetze oder Teile davon miteinander zu vereinigen. Ebenso ist es möglich, eine Topic Map in ein Datenbankformat zu transformieren, das einen schnellen Zugriff ermöglicht.

Auf der anderen Seite findet RDF(S) im Zuge des voranschreitenden „Semantic Web" derzeit weite Verbreitung. Zur Repräsentation von Ontologien ist der auf RDFS aufsetzende Standard OWL weitaus ausgereifter als ein entsprechender Standard für Topic Maps, der derzeit entworfen wird.

Um diesem Dilemma zu entkommen und die Vorteile beider Repräsentationsformalismen nutzen zu können, könnte man Wortnetze zunächst in Topic Maps repräsentieren, um sie anschließend (automatisch) nach RDFS zu transformieren. Die so bereitgestellte lexikalische Ressource könnte dann im „Semantic Web" als Web-Service zur Verfügung gestellt werden (Lemnitzer & Kunze 2003).

Anmerkungen

[1] HyTex ist ein DFG-gefördertes Projekt, das seit 2002 an der Universität Dortmund unter der Leitung von Prof. Dr. Angelika Storrer durchgeführt wird. Informationen zu den Projektarbeiten finden sich unter *http://www.hytex.info/* [Zugriff April 2004].

Literatur

Beckwith, R.; Miller, G. A.; Tengi, R. (1993). "Design and implementation of the WordNet lexical database and searching software." In: Miller, G. A. et al. (1990). Five Papers on WordNet. Technical Report 43, Princeton University, Cognitive Science Laboratory, *ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf* [Zugriff April 2004, Erstveröffentlichung in: Journal of Lexicography 3(4) (1990), 235-312].

Christodoulakis, D.N.; Kunze, C.; Lemnitzer, L. (Hrsg.) (2002). Proceedings of the Workshop on Wordnet Structures and Standardizations, and how these Affect Wordnet Applications and Evaluation. 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 28th May 2002.

DeRose, S.; Maler, E.; Orchard, D. (2001). XML Linking Language (XLink) Version 1.0. World Wide Web Consortium (W3C) Recommendation, June 2001, *http://www.w3.org/TR/2000/xlink/* [Zugriff April 2004].

Fischer, D. H. (1997). "Formal Redundancy and Consistency Checking Rules for the Lexical Database WordNet." Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 22-31.

Gupta, P. (2002). "Approaches to Checking Subsumption in GermaNet." In: Christodoulakis, Kunze & Lemnitzer (2002), 8-13.

International Organization for Standardization (ISO) (2000). ISO/IEC 13250:2000 Document Description and Processing Languages – Topic Maps. Geneva, Switzerland: ISO, *http://www.y12.doe.gov/sgml/sc34/document/0129.pdf* [Zugriff April 2004].

Kunze, C.; Lemnitzer, L. (2002). "Standardizing WordNet in a web-compliant format: The case of GermaNet." In: Christodoulakis, Kunze & Lemnitzer (2002), 24-29.

Lassila, O.; Swick, R. R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium (W3C) Recommendation, February 1999, *http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/* [Zugriff April 2004].

Lemnitzer, L.; Kunze, C. (2003). Integrating Wordnets into the Resource Description Framework, 2003, Universität Tübingen, Seminar für Sprachwissenschaft, *http://www.sfs.uni-tuebingen.de/~lothar/publ/GermaNET_RDF.pdf* [Zugriff April 2004].

Lenz, E. A.; Beisswenger, M.; Storrer, A. (2002). „Hypertextualisierung mit Topic Maps – ein Ansatz zur Unterstützung des Textverständnisses bei der selektiven Rezeption von Fachtexten." In: Tolksdorf, R.; Eckstein, R. (Hrsg.) (2002). Proceedings Workshop XML-Technologien für das SemanticWeb (XSW 2002), Berlin, Juni 2003. Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-14], 151-159.

Miles-Board, T.; Kampa, S.; Carr, L.; Hall, W. (2001). "Hypertext in the Semantic Web." In: Proceedings 12th ACM Conference on Hypertext and Hypermedia (HT '01). Aarhus, Denmark, 237-238.

Pavelek, T.; Pala, K. (2002). "WordNet Standardization from a Practical Point of View." In: Christodoulakis, Kunze & Lemnitzer (2002), 30-34.

Pepper, S.; Moore, G. (2001). XML Topic Maps (XTM) 1.0. TopicMaps.Org Consortium Specification, March 2001, *http://www.topicmaps.org/xtm/1.0/* [Zugriff April 2004].

Rath, H. H. (200). "Making Topic Maps more Colourful." In: Proceedings of XML Europe 2000 Conference, Berlin, Mai 2001, *http://www.gca.org/papers/xmleurope2000/pdf/s29-01.pdf* [Zugriff April 2004].

Rath, H. H. (2002). "GPS des Web. XML Topic Maps: Themenkarten im Web." In: iX, Magazin für Professionelle Informationstechnik, Heft 6 (2002), 115-122.

Runte, M.; Beisswenger, M.; Storrer, A. (2003). „Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet." In diesem Band, 113-125.

Vossen, P. (1998). "Introduction to EuroWordNet." In: Vossen, P. (Hrsg.) (1999). EuroWordNet: A Multilingual Database with Lexical-semantic Networks. Dordrecht: Kluwer Academic Publishers, 73-89.

Simon Clematide

GermaNet und UniNet

Abstract
Relationale lexikalische Semantik in der Tradition von WordNet ist Lexikographie, welche im Spannungsfeld von Ontologie und Terminologie arbeitet. Um semantische Netze wie das GermaNet, welche orientiert sind auf den Grundwortschatz, für Anwendungen wie Informationserschliessung nutzbar zu machen, sind Erweiterung und Spezialisierung unabdingbar. Dies versucht das UniNet für das Sachgebiet „Hochschulen und ihre Administration". Die Problematik der Integration von Netzen wird allgemein sowie hinsichtlich des GermaNet/UniNet diskutiert.

1. Einleitung

Ausgehend von WordNet (Fellbaum 1998) ist in den letzten Jahren eine Vielzahl von Projekten entstanden, welche diese Art von Ressource für verschiedenste Einzelsprachen adaptieren oder WordNet direkt übersetzen (siehe *http://www.globalwordnet.org*). Für Deutsch liegt mit GermaNet (Hamp & Feldweg 1997; Kunze 2000) eine Adaption vor, welche den Grundwortschatz abdecken soll und über EuroWordNet multilingual integriert ist.

Für die direkte praktische Anwendung in sprachverarbeitenden Systemen sind diese allgemeinen semantischen Netze mit zwei Problemen behaftet:

1. Zu *hohe Allgemeinheit*: Es gibt zu viele Wortbedeutungen, welche im Anwendungsgebiet gar nicht relevant sind.
2. Zu *niedrige Abdeckung*: Es fehlen viele für das Anwendungsgebiet benötigte Wörter.

Das erste Problem lässt sich beheben, indem irrelevante Wortbedeutungen und im Anwendungsgebiet nicht anwendbare semantische Relationen eliminiert werden. Für WordNet mit dem Anwendungsbereich Aviatik haben dies z.B. (Turcato et al. 2000) stark automatisiert gemacht. Andere Ansätze versuchen die einzelnen Wortbedeutungen – wie in der Terminologiearbeit eher üblich – mit Kodes für das Sachgebiet zu ergänzen, was die Zuordnung der möglichen Lesarten eines Wortes im Kontext deutlich erleichtern kann.

Um das zweite Problem zu lösen, müssen die fehlenden Wörter bestimmt und sinnvoll integriert werden. Dies kann wie in (Buitelaar & Sacaleanu 2002) gezeigt für das Anwendungsgebiet Medizin ausgehend von GermaNet recht erfolgreich automatisch gemacht werden; insbesondere, wenn es um das Ergänzen von Unterbegriffen geht. Für das UniNet, das 1999 entstanden ist, wurde allerdings grösstenteils manuell vorgegangen.

2. UniNet

Seine Entstehung verdankt das UniNet den Bemühungen am Institut für Computerlinguistik Zürich, linguistisch fundierte Informationserschliessung im Sinne von Antwortextraktion bzw. *Passage Retrieval* (Arnold et al. 2001) zu machen. Eine Suchmaschine, welche über relativ kleinen Textmengen und einigermassen homogenem Anwendungsgebiet operiert, sollte sowohl beim Indizieren der Suchtexte wie beim Verarbeiten der Anfrage, synonyme Ausdrücke berücksichtigen und verarbeiten können; zudem sollten insbesondere für die (An-)Frage-Expansion auch einfache Hyperonymie-Relatio-

nen zugänglich sein. Die Erfahrungen aus dem englischsprachigen Antwortextraktionssystem (Dowdall et al. 2002) mit WordNet machen deutlich, dass nur eine anwendungsspezifische Ressource Vorteile einbringt.

2.1 Struktur des UniNet

Das UniNet enthält 21.974 Einträge, die aus einem Wort bestehen. Davon sind 73 Namensbezeichnungen, wobei darunter sowohl Eigennamen wie „Schweiz" als auch Bezeichnungen von Institutionen verstanden werden wie „Stiefel-Zangger-Stiftung" oder „Universitätsrat". Die beiden Wörter „Bildungsaktivität" sowie „Lernaktivität" werden als künstlich im Sinn von GermaNet verwendet. Die restlichen 21.899 Einträge sind normale Substantive – ein recht grosser Teil davon ist allerdings automatisch konstruiert aus Kombinationen von wissenschaftlichen Fachgebieten mit unterschiedlichen Kompositionsgliedern wie in „Philosophieprofessor" und „Philosophiestudentin" oder "Philosophievorlesung".

Von den 1.199 mehrteiligen Ausdrücken sind 494 komplexe Namensbezeichnungen wie etwa „schweizerische Eidgenossenschaft" oder „Maturitätsschule für Erwachsene". 12 mehrteilige Ausdrücke sind künstliche Lexikoneinträge wie etwa „sprechende Person" oder „erziehungstätige Person". Die restlichen 693 Einträge sind Nominalgruppen, die teilweise lexikalisierten Gehalt haben wie „zoologischer Garten", teilweise müssten sie eigentlich besser als künstliche Einträge gehandelt werden wie etwa „staatliche Institution".

Im Netz sind je 22.514 Unter- bzw. Oberbegriffsbeziehungen kodiert. Auf der obersten Ebene gibt es 10 Kategorien: „Entität", „Gruppe", „Handlung/Akt", „Ereignis", „Zustand", „Abstraktion", „Besitz", „Phänomen", „Lage", „Ort" sowie „Aspekt des Geistigen", welches „Kognition" und „Motivation" zusammenfasst. Wie in GermaNet kann eine Synonymklasse (*Synset*) zu mehr als einer Synonymklasse in der Oberbegriffsbeziehung stehen. Im UniNet wird davon sogar recht grosszügig Gebrauch gemacht, insgesamt gibt es 6.996 solcher Klassen. Die restlichen 5.379 der insgesamt 12.385 Synonymklassen sind direkt nur mit einem Oberbegriff verbunden. Die 2.648 Meronymie- bzw. Holonymiebeziehungen machen wie im GermaNet keine feinere Unterscheidung in Teil-Ganzes, Element-Menge oder Material-Objekt. Es existieren insgesamt 26 lexikalische Antonymie-Beziehungen.

2.2 Bildungsmuster für Wörter und Synonymklassen

Ein wichtiger Kernbereich des UniNet sind die gut 550 wissenschaftlichen Studienfächer und Disziplinen, welche in einer Taxonomie abgelegt sind. So ist beispielsweise „Rechtswissenschaft" in die Teilfächer „Privatrecht", „Staatsrecht" usw. aufgeteilt, welche wiederum untergliedert sind. Diese Fachbezeichnungen werden automatisch mit Zweitgliedern wie „-studium", „-professorin", „-professor", „-seminar" oder Erstgliedern wie „Hauptfach" verschmolzen und in die entsprechenden Synonymklassen mit den geeigneten semantischen Relationen eingefügt.

Mit dem expliziten Einbau aller Kombinationen bläht man das Netz allerdings mit vielen Wortformen auf, welche kaum je verwendet werden oder als morphologische Unfälle taxiert werden müssen wie etwa die „Hauptfachkatastrophenmedizinerin". Es wäre gerade für semantisch transparente und produktive Wortbildungen sowie für die Verwaltung von Netzen sinnvoll, solche Regularitäten als semantisches Regelwissen ablegen zu können. Allerdings würde man damit den extensionalen Charakter der traditionellen semantischen Netze sprengen.

3 Abdeckung und Lesarten

In diesem Abschnitt werden zwei kleine Experimente beschrieben, welche Umfang und Unterschiede von GermaNet und UniNet in der Anwendung zeigen sollen. Da sich die bei der Konstruktion von UniNet verwendete GermaNet-

Version von 1999 mancherorts recht deutlich unterscheidet von der Version 4 aus dem Jahr 2001, habe ich die Resultate teilweise für beide angegeben.

Grundlage für die Evaluationen war das kleine Korpus „LUIS-Texte", welches von Web-Seiten und elektronischen Broschüren aus dem Umfeld der Hochschulen auf dem Platz Zürich ursprünglich für unser *Passage-Retrieval-System* LUIS zusammengestellt wurde. Daraus sind etwa 350 Sätze mit 6.711 Token ausgewählt, getaggt (STTS-Tagset) und syntaktisch annotiert, welche letztlich noch wortsinndesambiguiert werden sollen.

3.1 Häufige Substantive

In Tabelle 1 sind die häufigsten Substantive aus dem Korpus mit der Anzahl der Synonymklassen abgebildet. Das Zeichen § wird als Wort „Paragraph" behandelt – es ist das einzige Wort, das nicht in das Kerngebiet von UniNet fällt. In 7 Fällen deckt GermaNet 1999 schlechter ab als die neuere Version. Beim Wort „Universität" darf man sich jedoch nicht täuschen lassen. Obwohl GermaNet 2001 und UniNet nur eine einzige Synonymklasse aufweisen, sind dank Mehrfachvererbung letztlich sowohl die Instituts- wie auch die Gebäudelesart präsent. Mit der Einführung mehrerer Oberbegriffsbeziehungen für eine Synonymklasse ist die Gleichung „1 Synset = 1 Lesart" aufgehoben worden.

Obwohl ein Eintrag für „Studierender" fehlt – dieser Ausdruck ist sicher genügend lexikalisiert, um eingetragen zu werden – schneidet GermaNet 2001 im Bereich der häufigsten Substantive im Vergleich zum anwendungsspezifischen UniNet sehr gut ab. Wenn man etwas genauer in die einzelnen Synonymklassen hinein schaut, treten aber einige Unterschiede auf:

So kennt GermaNet 2001 für „Fakultät" eine aus zwei Lesarten von „Fachbereich" als Synonym, UniNet hingegen nicht – dafür vermerkt es noch „Universitätsfakultät" als synonym. UniNet steckt „Prüfung", „Examen", „Test" in eine Klasse, GermaNet 2001 nimmt „Prüfung", „Überprüfung", „Kontrolle" in einer der beiden Lesarten zusammen.

Freq.	Wortform	GN 01	GN 99	UN
38	Fakultät	1	0	1
19	Prüfungen	2	1	2
18	Universität	1	2	1
16	Prüfung	2	1	2
15	§	2	0	0
14	Diplomarbeit	1	0	1
13	Universitäten	1	2	1
12	Studierende	0	0	1
12	Kandidatin	1	1	1
11	Studium	2	0	1
11	Studierenden	0	0	1
11	Latein	1	0	2

Tabelle 1: Anzahl Synonymklassen der häufigsten Substantive im Test-Korpus

3.2 Zufällig ausgewählte Substantive

In einem weiteren Experiment wurden 100 verschiedene Substantive zufällig aus dem Korpus ausgewählt und nach folgenden Kriterien beurteilt:

A Mit wievielen Synsets kommt das Wort in GermaNet 2001 bzw. UniNet vor?
B Ist die relevante Lesart darunter?
C Ist dies eindeutig oder etwas unsicher?
D Ist das Wort Teil eines Mehrwortausdrucks?

In 36 Fällen sind die Informationen von UniNet und GermaNet übereinstimmend. Bei GermaNet fehlen 27, bei UniNet 32 Wörter. In 21 Fällen gibt es das Wort nur im UniNet nicht, aber nur in 3 davon handelt es sich um Ausdrücke, die im Anwendungsgebiet relevant sind. In 16 Fällen gibt es das Wort im UniNet, aber nicht im GermaNet. Dabei sind alle Ausdrücke ausser dem Wort „Ratsuchender" relevant für das Anwendungsgebiet. Für die Wörter „Hinblick", „Literatur" und „Botschaft" kennt UniNet zwar eine Lesart, aber leider die falsche. Dies passiert

GermaNet in 2 Fällen. Im UniNet ist das Urteil, ob eine Lesart wirklich relevant ist, in 6 Fällen unsicher, im GermaNet in 13 Fällen. UniNet liefert 82 Synonymklassen für 68 Wörter (Ambiguitätsrate 1,2), GermaNet 126 für 73 (Ambiguitätsrate 1,7).

4 Verknüpfen von semantischen Netzen

Der ursprünglich rein strukturelle Ansatz der relationalen lexikalischen Semantik definiert die Bedeutung eines Wortes nur aus der Menge der semantischen Beziehungen zu den andern Wörtern. Die 1989 erfolgte Einführung definitorischer Glossen im WordNet[1] sowie die Einführung von Beispielsätzen war ein wichtiger Schritt hin bzw. zurück zur traditionelleren Lexikographie.

Die „Selbstorganisation" der Lesarten in semantischen Netzen macht ihre Integration nichttrivial – das Problem ist in unterschiedlicher Form aufgetaucht:

1. **Multilinguale Netze:** Das Verbinden von gleichbedeutenden Wörtern über mehrere europäische Sprachen hinweg war ein Ziel von EuroWordNet (Vossen et al. 1999a). Zu diesem Zweck wurde zuerst die Version 1.5 von WordNet als Referenznetz genommen und einzelsprachliche Synonymklassen mit den entsprechenden Synonymklasse des Referenznetzes indiziert (*Inter-Lingual-Index*). Verschiedene Revisionen des ILI versuchten dann, die Abdeckung und Granularität der Kernbedeutungen zu normalisieren, um ein universell anwendbares Begriffsgerüst zu erhalten.
2. **Teil-Netze**: Die Integration von themen- und anwendungsspezifischen Netzen in allgemeinere Netze ist für viele Anwendungen wichtig. Eine manuelle Integration ist nicht besonders problematisch, wenn die Lexik des anwendungsspezifischen Netzes auf einige wenige Stellen im allgemeineren Netz konzentriert ist – oder falls das allgemeinere Netz sowieso als Grundgerüst verwendet wird. (Magnini & Speranza 2002) enthält ein halbautomatisches Verfahren, um ein bereits bestehendes semantisches Netz aus dem Wirtschaftsbereich ins EuroWordNet einzuhängen.
3. **Netz-Versionen**: Der Ausbau und die Verfeinerung von semantischen Netzen ergeben verschiedene Instanzen, welche sich nicht bloss im Umfang, sondern oft auch in der internen Strukturierung mehr oder weniger stark unterscheiden. Ressourcen, welche auf Version *X* eines semantischen Netzes aufsetzen, lassen sich oft nicht einfach mit Version *Y* koppeln. Ein automatisches Abgleichen unterschiedlicher Versionen ist äusserst wünschenswert – ein erfolgreiches Verfahren für die Abbildung von WordNet 1.5 auf 1.6 stellen (Daudé et al. 2001) vor.

Selbstverständlich können obige Probleme beim Integrieren von Netzen kombiniert auftreten. Ein gutes Beispiel dafür, wie aufwändig dies für verschiedene Revisionen eines multilingualen Netzes ist, findet sich im Kontext von EuroWordNet in (Vossen et al. 1999b).

Wie steht nun das UniNet diesen Problemen gegenüber? Bei der Konstruktion war die Einbettung in das bestehende allgemeinere GermaNet (Version 1999) äusserst nützlich, d.h. ontologische Organisation, viele Überbegriffe und einzelsprachliche Gegebenheiten konnten übernommen werden. Die lexikographische Arbeit liess sich so von einer Person mit vernünftigem Aufwand erledigen. Die Integration von UniNet ins GermaNet war allerdings von Anfang an nie vollständig, d.h. UniNet kann nicht als Teilnetz von GermaNet aufgefasst werden, das als Modul per Knopfdruck die relational strukturierte Terminologie eines bestimmten Sachgebiets zur Verfügung stellt.

Mit der Weiterentwicklung von GermaNet zur Version 4 ist die Kompatibilität zum Uni-

Net vermutlich noch weiter geschrumpft. Wie stark die Abweichungen sind, soll bei uns in einem Studienprojekt abgeklärt werden.

Ein Beispiel: GermaNet 1999 kannte das Wort „Entscheid" oder „Beschluss" nicht. Im UniNet wurden sie dann als Synonym zu „Entscheidung, Festlegung, Bestimmung, Festsetzung" genommen und unter „Akt" subsumiert. In GermaNet 4 kommt sowohl „Beschluss" wie „Entscheid" vor, allerdings in leicht anderen semantischen Beziehungen: Das Wort „Entscheid" wird synonym mit „Bescheid" gesetzt und als Auskunft im Sinn von Information verstanden, die etwas als „Steuerbescheid" oder „Rentenbescheid" auftaucht. Das Wort „Beschluss" dagegen ist bei Kognition als eine Art „Entscheidung, Entschluss" aufgelistet.

Aus heutiger Sicht wäre eine Indizierung der wichtigsten Kategorien mit einem möglichst allgemeinen und stabilen Bedeutungsindex wie etwa dem ILI wünschenswert, weil dadurch eine algorithmische Verknüpfung des UniNet erheblich erleichtert werden könnte.

Ein anderer Punkt ist eine mögliche Erweiterung von UniNet: Durch den Verwendungszweck und den Entstehungsort sind viele Ausdrücke vom schweizerischen Hochschulsystem und Sprachgebrauch geprägt. Es wäre deshalb notwendig, vermehrt regionalsprachliche stilistische Markierungen in den lexikographischen Dateien zu kodieren.

Dank

Geht an Martin Volk für die UniNet-Projektleitung, an Arnold H. Bucher für die Terminologiearbeit und das Kodieren der lexikographischen Dateien, sowie an die bereitwillige Unterstützung des GermaNet-Teams.

Anmerkung

[1] George Miller schreibt im Vorwort zu Fellbaum 1998, dass mit wachsender Grösse von WordNet das Erfassen von Bedeutungen aus der Synonymie-Beziehung allein immer schwieriger wurde und konzediert selbstkritisch: "... definition by synonymy is not adequate."

Literatur

Arnold, T. et al. (2001). „LUIS – ein natürlichsprachliches, universitäres Informationssystem." In: Appelrath, H.-J. et al. (Hrsg.) (2001). Unternehmen Hochschule (UH-01). Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-6], 115-126.

Buitelaar, P.; Sacaleanu, B. (2002). "Extending Synsets with Medical Terms." In: Proceedings of the 1st International WordNet Conference, Mysore, India, January 2002.

Daudé, J.; Padró, L.; Rigau, G. (2001). "A Complete wn1.5 to wn1.6 Mapping." In: Proceedings of NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, June 2001.

Dowdall, J. et al. (2002). "Technical Terminology as a Critical Resource." In: Proceedings 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria, Spain, May/June 2002.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Kunze, C. (2000). "Extension and Use of GermaNet, a Lexical-Semantic Database." In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, May 2000, 999-1002.

Magnini, B.; Speranza, M. (2002). "Merging Global and Specialized Linguistic Ontologies." In: Proceedings of OntoLex 2002. Workshop held in conjunction with the 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria, Spain, May 2002.

Turcato, D. et al. (2000). "Adapting a Synonym Database to Specific Domains." In: Proceedings of the ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, October 2000.

Vossen, P.; Peters, W.; Gonzalo, J. (1999). „Towards a Universal Index of Meaning." In: Standardizing Lexical Resources. Proceedings of the ACL-99 / SIGLEX Workshop, College Park, MD, June 1999.

Vossen, P. et al. (1999). Extending the Inter-Lingual-Index with new concepts. EuroWordNet, Document Nr. LE-4003-2D010, *http://www.illc.uva.nl/EuroWordNet/docs/2D010RTF.zip* [Zugriff April 2004].

Petra Steiner

FrameNet und WordNet - Perspektiven für die Verknüpfung zweier lexikalisch-semantischer Netze

Abstract
Dieser Beitrag stellt das FrameNet-Projekt vor, das auf dem Ansatz der Frame-Semantik von Charles J. Fillmore basiert. Verglichen mit WordNet-Daten besitzen die Daten von FrameNet eine grundlegend andere Struktur. Die Verknüpfung der beiden semantischen Netze ist jedoch gerade aufgrund der komplementären Informationen, die sie bieten, lohnenswert. In Form eines Entity-Relationship-Diagramms wird eine Datenbankstruktur, die sämtliche Daten verbindet, dargestellt und diskutiert.

1 Einleitung

FrameNet ist ein lexikographisches Projekt, das seit 1997 unter der Leitung von Charles J. Fillmore am International Computer Science Institute (ICSI) in Berkeley durchgeführt wird. In diesem Projekt werden semantische und syntaktische Eigenschaften von Wort-Tokens in Korpora erfasst, in einer MySQL-Datenbank systematisiert und auf der FrameNet-Website der internationalen Forschungsgemeinschaft zur Verfügung gestellt (vgl. *http://www.icsi.berkeley.edu/~framenet/index.html*; Fillmore, Johnson & Petruck 2003).

FrameNet basiert auf dem Ansatz der Frame-Semantik von Charles J. Fillmore und Anderen (vgl. Fillmore 1982; Fillmore 1985; Fillmore & Atkins 1992; Fillmore & Atkins 1994).

Im nächsten Abschnitt wird eine Einführung in die Frame-Semantik gegeben, Abschnitt 3 stellt die Umsetzung dieses Ansatzes im FrameNet-Projekt dar. In Abschnitt 4 werden in kurzer Form einige Bemerkungen zu bereits bestehenden WordNet-Projekten gemacht. Abschnitt 5 zeigt den Entwurf einer Datenbank, die Informationen von WordNet bzw. GermaNet mit FrameNet verbindet. Im Ausblick werden weitere Möglichkeiten der Erweiterung einer solchen Datenbankstruktur aufgezeigt.

2 Frame-Semantik

Im Gegensatz zu anderen Ansätzen in der Semantik (wie beschrieben bei Schumacher & Steiner 2002), insbesondere der Wahrheitsbedingungen-Semantik, ist die zentrale Idee der Frame-Semantik, dass lexikalische Bedeutungen am besten vor dem Hintergrund von miteinander zusammenhängendem Wissen beschrieben werden können (vgl. Fillmore 1985).

Lexeme wie LUKEWARM (Fillmore 1982: 123) können auch im Bereich der Wortfeldtheorie beschrieben werden (vgl. Gipper 1995: 337), doch Fillmore (Fillmore 1985: 226ff.) grenzt seinen Ansatz insofern von der Wortfeldtheorie aus der Münsterschen Tradition (vgl. Trier 1931; Weisgerber 1962; Geckeler 1971; Gipper 1973; Gipper 1995; Zillig 1994) ab, als dass innerhalb der Frame-Semantik nicht *Wörter*, sondern *Konzepte* die Einheiten der Beschreibung sind, so dass – die aus der Sicht der Wortfeldtheorie postulierten – lexikalischen Lücken irrelevant für das zugrunde liegende Wissen sind. Konzepte sind unabhängig davon, ob es Bezeichnungen für bestimmte Wissenszusammenhänge gibt.

Zum Beispiel existiert im Deutschen kein Lexem für das Konzept „nicht mehr durstig sein“, während dies für „nicht mehr hungrig sein“ der Fall ist (*satt*) (vgl. Schumacher & Steiner 2002: 185). *shore* und *coast* werden im Deutschen zwar jeweils mit *Küste* übersetzt; das erste Lexem bezeichnet allerdings die Küste, die vom

Wasser aus erreicht wird, während *coast* verwendet wird, wenn die Küste von Land aus erreicht wird (vgl. Fillmore 1982: 121; Fillmore 1985: 236). Um die Konzepte SHORE, COAST und KÜSTE verstehen zu können, muss nicht das gesamte Wortfeld, wohl aber die gesamte Wissensstruktur, in die sie eingebettet sind, bekannt sein. Ähnliche Unterschiede gibt es bei *land* und *ground*: Bei *land* besteht der Hintergrund in der Kontrastierung mit dem Meer, im zweiten Fall wird die Beziehung zur Luft präsupponiert (vgl. Fillmore 1982: 121). Ein *Häretiker* präsupponiert eine Religion, von der abgewichen wird (vgl. Fillmore 1982: 123), eine *Friedensaktivistin* die Möglichkeit oder tatsächliche Existenz eines Krieges.

Fillmore bezeichnet derartiges Wissen über Hintergründe als *Frames*.[1] *Lexikalische Einheiten* im Verlauf/Gebrauch evozieren diese Frames. So evozieren die Ausdrücke *travel* und *reisen* nicht nur eine Person, die verreist, sondern auch die Konzepte *Ausgangspunkt*, *Ziel* oder *Strecke*. Eventuell wird ein Fahrzeug oder Flugzeug benutzt – hier kommt also eine Menge Wissen ins Spiel.

3 FrameNet

Im FrameNet-Projekt werden Frames und die zu ihnen gehörenden lexikalischen Einheiten beschrieben. Die definierten Frames werden innerhalb des lexikalischen Netzwerks mit anderen Frames verbunden und eine Auswahl von Sätzen aus Korpora wird mit ihnen annotiert, wobei sämtliche Daten in einer Datenbank gespeichert werden. Annotationstools helfen bei der Bewältigung der komplexen Aufgaben. Im Folgenden werden die einzelnen Arbeitsschritte für die Gewinnung der Daten beschrieben.[2]

Die Definition von Frames und Frame-Elementen

Häufig wird ein neuer Frame definiert, wenn lexikalische Einheiten sich nicht bereits bestehenden Frames zuordnen lassen. Diese Definitionen stammen vom Annotationsteam des Projekts oder aus dem *Concise Oxford Dictionary* Um beim obigen Beispiel zu bleiben, wird der Frame TRAVEL folgendermaßen definiert:

> In this frame a Traveler$_1$ goes on a journey, an activity, generally planned in advance, in which the Traveler$_1$ moves from a Source$_2$ loccation to a Goal$_3$ along a Path$_4$ or within an Area$_5$. The journey can be made with an Vehicle$_6$ and/or accompanied by Co_travelers$_7$ and Baggage$_8$. The Duration$_9$ or Distance$_{10}$ of the journey, both generally long, may also be described. Words in this frame emphasize the whole process of getting from one place to another, rather than profiling merely the beginning or the end of the journey.

Die markierten Wörter sind die *Frame-Elemente*, semantische Rollen, die zum jeweiligen Frame gehören. Im Vergleich zu den semantischen Rollen der frühen Arbeiten Fillmores zur Kasusgrammatik (vgl. Fillmore 1968 u. 1977) sind die Frame-Elemente weitaus spezifischer. Auch zu den Frame-Elementen werden Definitionen entwickelt, wie zum Beispiel:

> Traveler [Trav]$_1$
> This is the living being which travels. Normally, the Traveler$_1$ is expressed as an external argument.
>
> Vehicle [Veh]$_6$
> The Vehicle$_6$ is the means of conveyance on the journey.

Beim Prozess der Frame-Definition wird auch eine Liste mit lexikalischen Einheiten festgelegt, die prinzipiell erweiterbar ist. Die folgenden lexikalischen Einheiten evozieren den Frame TRAVEL:

> commute.v, excursion.n, expedition.n, journey.n, journey.v,

junket.n, odyssey.n, peregrination.n, pilgrimage.n, safari.n, tour.n, tour.v, travel.n, travel.v, trip.n, voyage.n, voyage.v (vgl. www.icsi.berkeley.edu/~framenet)

Homonyme evozieren unterschiedliche Frames, zum Beispiel kann *Trip* den Frame TRAVEL, andererseits aber auch den Frame INTOXICATION evozieren.

Das FrameNet-Projekt geht in seiner Konzeption nicht von Lemmata mit mehreren Bedeutungen, sondern von Frames aus, die möglichst vollständig beschrieben werden sollen. Daher sind Homonyme und polyseme Lexeme in der Datenbasis häufig nicht vollständig erfasst.[3]

Zwischen den Frames besteht eine *Framehierarchie*, derart, dass Frames der oberen Hierarchieebene Eigenschaften an die unteren vererben. Zum Beispiel ist der Frame TRAVEL ein Spezialfall des Frames SELF_MOTION, der wiederum Eigenschaften der Frames MOTION und INTENTIONALLY_ACT erbt. Dabei stehen die Frame-Elemente in definierten Beziehungen zueinander: Zum Beispiel entspricht das Frame-Element *Co_travelers* des Frames TRAVEL dem Frame-Element *Cotheme* bei SELF-MOTION (vgl. Abbildung 1).

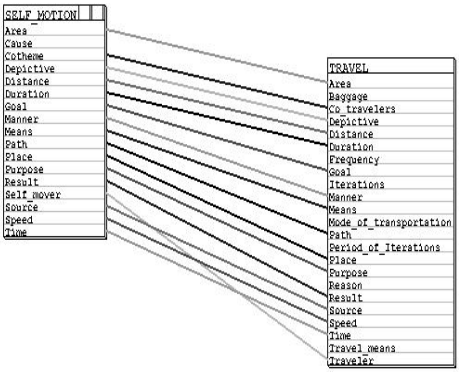

Abbildung 1: Verbindungen zwischen vererbten Frames

Wie in der Abbildung zu erkennen ist, ist der *Kind-Frame* hinsichtlich der semantischen Rollen elaborierter als der *Eltern-Frame.* Ferner können Frames die Bestandteile anderer, komplexer Frames sein. Zum Beispiel ist der Frame TRIAL ein *Subframe* von CRIMINAL_PROCESS. Ab einer gewissen Hierarchieebene werden Frames zu so genannten *Domains* zusammengefasst. CRIMINAL_PROCESS ist ein Beispiel dafür. Erfasst werden auch die Valenzmuster, also die phrasalen Kategorien und die grammatischen Funktionen (Subjekt, Objekt etc.) der Frame-Elemente.

3.2 Die Frame-Annotation in Sätzen

Nach der Definition von Frames und Frame-Elementen sowie der Festlegung der Relationen dieser Entitäten untereinander, werden Beispielsätze aus großen Textkorpora (das BNC-Korpus und das American Newswire-Korpus) extrahiert und dabei direkt unterschiedlichen Subkategorisierungsrahmen zugewiesen (so genannte *Subkorpora*[4]). Erst im nächsten Schritt werden diese Sätze annotiert.

Die annotierten Sätze sollen möglichst typische Beispiele aller möglichen Kombinationen von Frame-Elementen erfassen; sie spiegeln jedoch nicht die Verteilung dieser Muster im Text wider. Beispiele von Annotationen für den Frame TRAVEL sind:

Ellen$_1$ JOURNEYED$_{12}$ to Europe$_3$ with five suitcases$_8$.

Samantha$_1$ JOURNEYED$_{12}$ 2500 miles$_2$ with her family$_7$ by sea$_4$ to China$_3$.

The Osbournes$_1$ took a TRIP$_{12}$ from Beverly Hills$_2$ to London$_3$ on the Concorde$_6$.

Die annotierten Frame-Elemente werden nach der Relevanz für den jeweiligen Frame unterschieden. Bei *core frame elements* handelt es sich

um grundlegende konzeptuelle Bestandteile des Frames. Zum Beispiel sind Goal$_3$ und Path$_4$ Core-Frame-Elemente des TRAVEL-Frames. Häufig werden auch *periphere* Frame-Elemente annotiert, die Angaben zu Zeit, Ort, Zweck usw. sind. Für den TRAVEL-Frame sind Baggage$_8$ und Distance$_{10}$ Beispiele für periphere Frame-Elemente. *Extrathematische Frame-Elemente* gehören anderen Frames an, sie werden innerhalb der Sätze als solche annotiert. Zum Beispiel wird im folgenden Satz *in pursuit of suspected traffickers* dem Frame-Element *Depictive* zugewiesen. *Depictive* bezeichnet den Zustand des Reisenden.

Customs officers from each country$_1$ would be allowed to TRAVEL$_{12}$ to either country$_3$ in pursuit of suspected traffickers$_{11}$.

Im Gegensatz zu den anderen Frame-Elementen werden extrathematische Frame-Elemente innerhalb der Frame-Hierarchie nicht notwendigerweise vererbt.

Es gibt Belege, in denen die Core-Frame-Elemente nicht im Satz auftreten, jedoch als Teil des Frames vorhanden sind. Für solche Core-Frame-Elemente wird eine spezielle Markierung verwendet. Unterschieden wird hier zwischen *Constructional Null Instantiation* (CNI), *Definite Null Instantiation* (DNI) und *Indefinite Null Instantiation* (INI). CNI umfasst zum Beispiel weggelassene Subjekte in Imperativsätzen oder ausgelassene Agens-Formen in Passivsätzen:

Days began early and ended late so that maximum distances$_4$ could be TRAVELLED$_{12}$. (CNI$_1$)

Im Fall von DNI befindet sich das fehlende, obligatorische Element im situationalen oder linguistischen Kontext. Im folgenden Beispiel ergibt sich das Goal$_3$ aus dem Kontext.

As pilgrims the children of Israel$_1$ JOURNEYED$_{12}$, led by the guiding hand of God. (DNI$_3$)

INI bezieht sich auf Frame-Elemente, bei denen die Interpretation obligatorisch ist, die Elemente jedoch nicht im Kontext zu finden sind.
Im folgenden Beispiel ist etwa das Goal$_3$ nicht instanziiert:

I$_1$ used to TRAVEL$_{12}$ by bus$_6$ a lot$_{10}$, so I had a season ticket. INI$_3$

Ein Sonderfall der Null-Instanziierung sind *Inkorporationen*. Hier besteht eine enge morphologische Verbindung zwischen Verb und nicht instanziiertem Frame-Element, ein Beispiel hierfür ist die lexikalische Einheit *to bicycle*:

When he had leisure he$_1$ went BICYCLING$_{12}$ to Lincolnshire village churches$_3$. INC$_6$

Hier ist das Vehicle$_6$ nicht instanziiert.

Diese Markierungen erlauben es, Sätze mit Ellipsen von solchen mit tatsächlich nicht vorhandenen Frame-Elementen zu unterscheiden. Dies kann für die semantische Klassifizierung von Verben von Nutzen sein (vgl. Schulte im Walde 1998: 86f).

Für die Definition und Beschreibung der Frames, die Extraktion der Satzsammlungen aus den Korpora und die Annotation der Sätze werden die Tools für die FrameNet-Annotation verwendet (*FrameNet Desktop*, vgl. Fillmore u.a. 2003).

Abbildung 2 zeigt ein Beispiel für den Prozess einer Annotation. Mit Hilfe des Annotationstools werden neben den Frame-Elementen auch die Phrasenstrukturtypen und grammatische Funktionen annotiert, Fehler der Wortartenannotation können ebenfalls korrigiert werden.

3.3 Die Struktur der FrameNet-Datenbank

Die gesamten Daten werden in einer relationalen Datenbank abgelegt (vgl. Baker, Fillmore & Cronin 2003), die aus zwei Teilen besteht: In der so genannten lexikalischen Datenbasis sind neben den Lemmata, Lexemen, Wortformen, Wortarten auch die Frames, Frame-Elemente und die Beziehungen zwischen Frames enthalten. Die Annotationsdatenbank enthält die mit Frames annotierten Sätze aus dem BNC-Korpus und dem American Newswire Korpus. Momentan enthält die Datenbank 456 Frames mit über 7.300 lexikalischen Einheiten. Etwa 130.000 Sätze wurden seit Beginn des Projekts annotiert.

Abbildung 3 zeigt den Teilausschnitt der lexikalischen Datenbank. *WordForm* bezeichnet hier eine grammatische Ausprägung eines Lexems, damit wird jedoch nicht nach grammatischen Funktionen unterschieden. Ein oder mehrere *WordForms* sind mit einem *Lexeme*, der abstrakten lexikalischen Einheit, verknüpft. *Lexeme* umfasst lediglich lexikalische Einheiten, die keine Mehrwortlexeme, aber ggf. Bestandteile davon sind. Ein *Lemma* hingegen besteht aus einem oder mehreren Lexemen, die über *LexemeEntry* miteinander verknüpft sind. Häufig hat ein *Lemma* nur eine Verknüpfung zu genau einem *LexemeEntry*, wenn es sich nicht um einen Mehrwortausdruck wie *take off*, oder komplexe Eigennamen handelt. Für *LexemeEntry* ist vermerkt, wo es sich innerhalb des *Lemmas* befindet, ob es sich um das Head handelt und ob hier eine Trennung der Lemma-Teile vorliegen kann. Jedes *Lemma* besitzt eine Wortart (*PartOfSpeech*), ebenso wie jedes *Lexeme*. Jedes *Lemma* ist mit einem oder mehreren Einträgen in *LexUnit* verknüpft. *Lexical Units* (vgl. *lexikalische Einheiten* im Abschnitt 2) sind Verknüpfungen zwischen Lemmata und Frames.

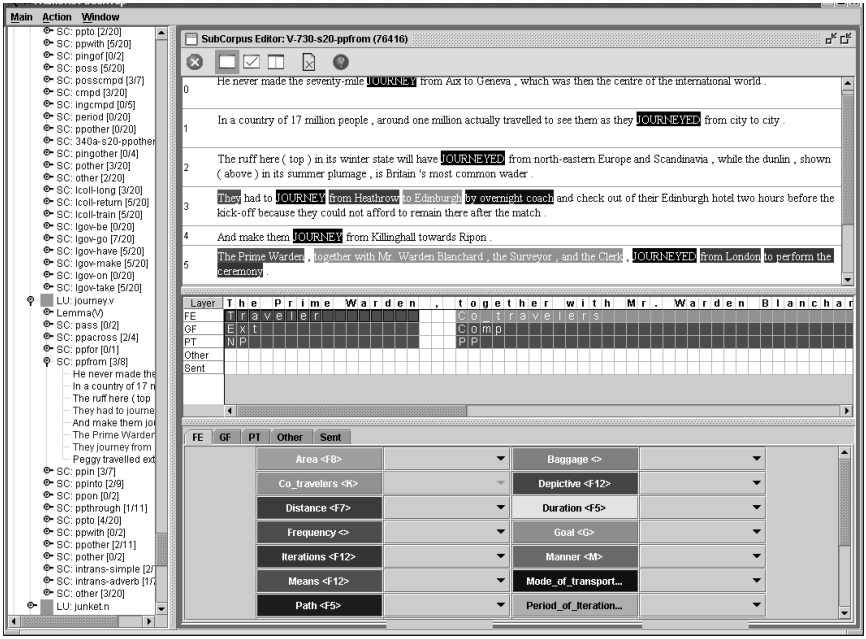

Abbildung 2: Ausschnitt aus dem FrameNet Desktop

Die Entität *Status* gibt den Bearbeitungszustand der lexikalischen Einheit an, kann aber andere Informationen enthalten, zum Beispiel, wenn die lexikalische Einheit sich analog zu anderen lexikalischen Einheiten verhält, wie es bei den Wochentagen der Fall ist. Jede lexikalische Einheit ist mit höchstens einem Frame verknüpft, umgekehrt besteht eine 1:n-Relation.

Beziehungen von Frames zu Frames und Frame-Elementen zu Frame-Elementen, wie zum Beispiel bei der Frame-Hierarchie, werden in den Tabellen *FrameRelation* und *FERelation* erfasst. Hier wird festgelegt, welche Frames bzw. Frame-Elemente in Frames in Relation zueinander stehen. Die verschiedenen Relationen befinden sich in *RelationType*.

Frame-Relationen können auch Metabeziehungen zueinander besitzen. Momentan betrifft dies die Beziehung des *Subframes* und die Metarelationstypen sind *Gleichzeitigkeit* und *Vorzeitigkeit*.

Die Entität *SemanticType* enthält zusätzliche, relevante Information, zum Beispiel zur Konnotation (*knickerig* vs. *sparsam* : negativ vs. positiv), ist aber im FrameNet-Projekt nicht systematisch erfasst. Zwischen diesen *SemanticTypes* können hierarchische *STInherit*-Beziehungen bestehen, zum Beispiel gehört eine negative Beurteilung wie *knickerig* generell zu affektiven Beurteilungen.

3.4 Andere FrameNet-Projekte

Neben einem japanischen und einem spanischen FrameNet-Projekt (vgl. Subirats & Petruck 2003; *http://gemini.uab.es/SFNI*), wird ein deutsches FrameNet-Projekt (SALSA) unter der Leitung von Manfred Pinkal in Saarbrücken durchgeführt (vgl. Erk et al. 2003). Hier soll das 1,5 Millionen Token umfassende deutsche TIGER-Korpus mit Frames annotiert werden, wobei al-

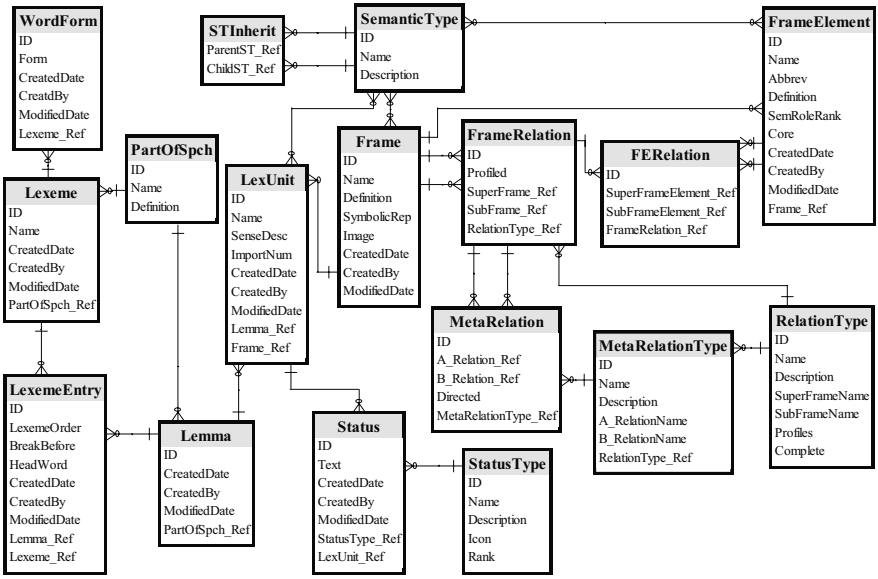

Abb. 3: Entity-Relationship-Diagramm der FrameNet-Datenbank (lexikal. Teil)

lerdings nur ein beschränktes Inventar an Frames und Frame-Elementen verwendet wird, das auf den 456 Frames des englischen FrameNet-Projekts basiert. Das bedeutet, dass die meisten Verben mit einem Tag annotiert werden, der die Existenz eines Frames und seiner Frame-Elemente lediglich markiert, diese Einheiten jedoch nicht definiert („UNKNOWN"). Informationen wie Relationen und Vererbung zwischen Frames werden nicht erfasst. Durch statistische Verfahren soll in einem späteren Schritt eine Klassifikation von Frames ermittelt werden.

4 WordNet, GermaNet, EuroWordNet

WordNet (Beckwith et al. 1991; Fellbaum 1998; Miller 1998; *http://www.cogsci.princeton.edu/~wn/*) hat den Anspruch, eine aufgrund psycholinguistischer Evidenzien aufgebaute lexikalische Datenbank zu sein. Der Ansatz von Collins u. Quillian (Collins & Quillian 1969), auf den mehrfach Bezug genommen wird, ist allerdings schon früh und zu Recht kritisiert worden (Conrad 1972).

In dieser lexikalischen Datenstruktur werden hauptsächlich semantische Relationen zwischen Konzepten beschrieben, wie Synonymie, Antonymie, Hyponymie (IS-A-relation), Hyperonymie, Meronymie und Holonymie bei Nomina. Lexikalische Einheiten mit der gleichen Bedeutung werden in so genannten *synsets* klassifiziert.

GermaNet (vgl. Hamp & Feldweg 1997; Kunze & Lemnitzer 2002; *http://www.sfs.uni-tuebingen.de/lsd/Intro.html*) und EuroWordNet (vgl. Vossen 1997) sind Parallelprojekte. EuroWordNet umfasst acht europäische Sprachen, die miteinander durch den *Inter-Lingual-Index* verknüpft sind, und integriert auch Module des GermaNet. Während EuroWordNet 1999 abgeschlossen wurde, wird das GermaNet-Projekt fortgesetzt. Hier, wie auch in EuroWordNet, sind die lexikalischen Informationen zum Teil anders strukturiert als im WordNet-Projekt. Zum Beispiel ist bei GermaNet die Struktur der Adjektive ähnlich der Hyponymierelation der Nomina konzipiert. GermaNet enthält momentan mehr als 42.000 Synsets mit über 61.000 Lesarten.

Miller selbst (Miller 1998) kritisiert den Mangel an Ausnahmeregeln für die Vererbung von Eigenschaften in der Begriffshierarchie, vor allem aber sind syntagmatische Zusammenhänge kaum repräsentiert; dies bezeichnet er als das *tennis problem*. Ball, Schläger, Tennisplatz, Schiedsrichter etc. sind in unterschiedlichen Bereichen zu finden, es gibt keine markierten Zusammenhänge in der Datenbasis: „A framelike semantics is precluded by WordNet's strict separation of its entries according to their syntactic category membership [...]." (Fellbaum 1998: 5).

GermaNet gibt für seine ca. 9.000 Verben Subkategorisierungsrahmen an (vgl. Hamp & Feldweg 1997; Kunze 1999; Kunze 2003). Auch das Konzept des Entailments wird verwendet. Die Kausalrelationen werden auch dann erfasst, wenn sie zwischen unterschiedlichen Wortarten bestehen (vgl. Vossen 1997). Ebenso werden einige Verbalternationen erfasst (vgl. Kunze 1999).

Trotz der beschriebenen Erweiterungen bietet auch GermaNet keine Informationen über Verbindungen der Argumentstrukturen von Lexemen, die durch semantische Relationen verbunden sind. Die Stärke des WordNet-Ansatzes liegt hingegen auf der Seite der Nomina und ihrer Relationen und auch bei der vorhandenen Datenmenge.

5 Entwurf einer Datenbankstruktur für die Verknüpfung von WordNet- und FrameNet-Informationen

Im Folgenden wird dargestellt, wie FrameNet- und WordNet-Daten miteinander verknüpft werden können. Dabei wird besonders berücksichtigt, welche Anforderungen an eine Datenbankstruktur für deutsche Daten erfüllt werden müssen. Außerdem wird davon ausgegangen, dass

in einer solchen Datenbank nicht nur Informationen aus zwei, sondern aus mehreren Netzen oder sonstigen Datenquellen verknüpft werden können. Auch mehrere Korpora können mit den Daten zusammenhängen.
Abbildung 4 stellt die Struktur in Form eines Entity-Relationship-Diagramms dar.

Der erste Unterschied zur FrameNet-Datenbank besteht bereits darin, dass hier nicht von der Entität *WordForm* ausgegangen wird. Stattdessen wird mit *Flexform* eine Ausprägung eines *Lexems* bezeichnet, die mit grammatikalischer Information verknüpft ist, zum Beispiel gibt es drei verschiedene Einträge zu der Wortform *spielen*, die in der Tabelle *WordformTypes* enthalten ist.

Die Entität *WordformTypes* entspricht also der Entität *WordForm* bei FrameNet. Mit der Entität *Flexform* sind über *MorphDescr* zahlreiche Tabellen zur Flexion verknüpft, die hier nicht weiter behandelt werden. Über die Entität *Token* kann auf alle *Flexform*-Tokens in einem spezifischen Korpus zugegriffen werden. In *Tokeninfo* gibt es zu jedem Token im Korpus zusätzliche Informationen, wie die Angabe zur Großschreibung von Verben, Markierung in Fett- oder Kursivdruck etc. Die Verbindung zwischen *Token* und *Corpus* erfolgt mittelbar über *Sentence*, *Paragraph* und *Document*, falls Annotationen in der Datenbank enthalten sind, oder zumindest unmittelbar für andere Korpora, etwa wenn es sich lediglich um Listen von Token handelt. Für den Fall, dass in einem Korpus die morphologischen Ausprägungen der Flexform nicht bekannt sind, besteht zumindest eine Verbindung zwischen der Entität *WordformType* und *Sentence* bzw. *Corpus*. Da mehrere Korpora untersucht werden, besteht jeweils eine 1:n-Relation von *Corpus* zu *Token* und *WordformType*.

Ein oder mehrere Flexformen sind mit einem *Lexeme*, der abstrakten lexikalischen Einheit, verknüpft.

Die Relationen zwischen *Lexeme*, *Lemma*, *LexemEntry*, *PartofSpeech* und *LexUnit* sind identisch mit denen der FrameNet-Datenbank.

Die Tabelle *Flexinfos* ist mit weiteren Datenbanktabellen verknüpft, zum Beispiel mit Informationen zu Flexionsparadigma, Umlautung und Alternanzstamm für Nomina. In der Tabelle *alternat_orthography* sind Rechtschreibvarianten und Rechtschreibfehler vermerkt. So können Korpora, die aus der Zeit vor der und nach der Rechtschreibreform stammen, gemeinsam erfasst werden.

Die Struktur zwischen *Flexform*, *Morpheme_Entry* und *Morpheme* ist analog zu der zwischen *Lexeme*, *LexemEntry* und *Lemma*. Es ist wichtig, Morpheme zu erfassen, um Frame evozierende sprachliche Einheiten, die unterhalb der Wortebene liegen, berücksichtigen zu können. Ein Beispiel hierfür ist das Kompositum *Gagenforderung*, bei dem das Determinans ein Frame-Element zum Frame *request* ist, der vom Determinandum *Forderung* evoziert wird (vgl. Erk, Kowalski & Pinkal 2003). Frame-Semantik auf der morphologischen Ebene kann aber auch für andere Sprachen als das Deutsche betrieben werden. Petruck und Boas (Petruck & Boas 2003) zeigen anhand der hebräischen, englischen und deutschen Bezeichnungen für die Wochentage, wie unikale Morpheme anhand des Frames, vor dessen Hintergrund sie stehen, interpretiert werden können.

Die Entität *LexUnit* verknüpft in diesem Entwurf nicht nur Lemmata mit Frames, sondern auch mit anderen semantischen Informationen, unter anderem mit Synsets.

Über *LexUnit* werden Informationen aus GermaNet - im Diagramm nicht vollständig dargestellt - mit Informationen zu Frames verknüpft. Dabei kann es auch Einheiten geben, die keine Informationen zu Frames bzw. zu GermaNet besitzen.

Die lexikalischen Einheiten in *LexUnit* sind ferner mit grammatischen Funktionen, Subkat-

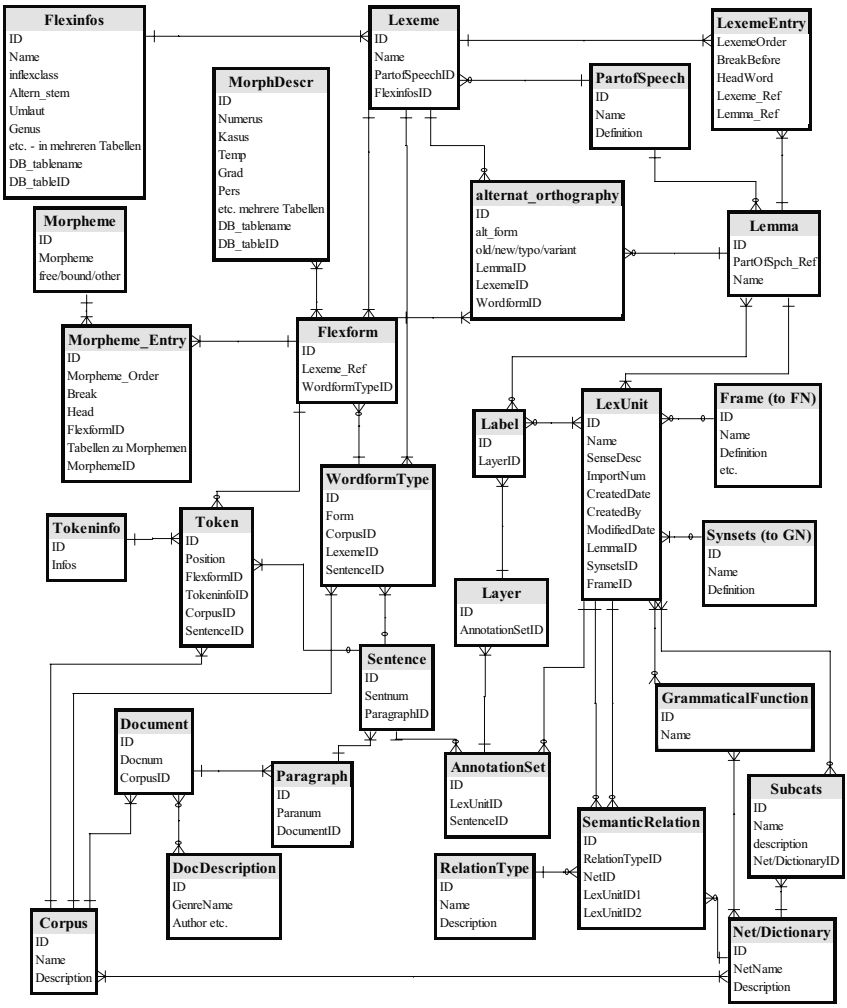

Abbildung 4: Die Verknüpfung von WordNet- und FrameNet-Daten

Informationen und semantischen Relationen verknüpft. Diese Angaben beziehen sich auf jeweils ein spezifisches Netz, Lexikon oder eine andere Datenquelle, so dass hier n:m-Beziehungen bestehen können.

Jede semantische Relation ist mit zwei lexikalischen Einheiten und einem Eintrag in der Tabelle *RelationType* verknüpft. Diese enthält alle definierten semantischen Relationen.

Die eben aufgelisteten semantischen Informationen können aus GermaNet oder einem anderen semantischen Netz stammen.

Der Entwurf enthält jedoch noch zwei weitere Verknüpfungen zwischen GermaNet und FrameNet-Daten: In den mit Frames annotierten Sätzen sind, ggf. auf mehreren Ebenen (*Layer*), die einzelnen Realisierungen der Frame-Elemente (*Label*) enthalten. Zum Beispiel ist im Satz *Peter kauft ein Buch. Peter* die Instanziierung des Frame-Elements *Buyer*. Diese Label können wiederum (hier vereinfacht dargestellt) mit Lemmata verknüpft werden. So entstehen auf einfache Art und Weise Verbindungen zwischen Frame-Elementen und weiteren semantischen Informationen, wie zu semantischen Relationen aus WordNet bzw. GermaNet oder aber zu weiteren Frames. Diese Verknüpfung liefert jedoch keine Information darüber, um welche lexikalische Einheit zu dem jeweiligen (ambigen) Lemma es sich handelt. Die Zuordnung zwischen *Label* und *LexUnit* ist vielmehr der Kern der Verknüpfung zwischen den komplementären lexikalisch-semantischen Netzen.

6 Zusammenfassung und Ausblick

Der dargestellte Datenbankentwurf verknüpft Informationen aus GermaNet oder ähnlichen lexikalischen Datenbanken und FrameNet. Dabei ist auch die Hinzunahme weiterer Informationen aus anderen Netzen oder Lexika möglich. Den Spezifika deutscher Sprachdaten wird mit Entitäten zu morphologischen und graphematischen Informationen Rechnung getragen.

FrameNet und WordNet-Projekte sind in erster Linie lexikographisch orientiert, auch wenn sie Korpora für Definitionen und Beispiele verwenden. Die annotierten Sätze geben daher kein realistisches Bild der tatsächlichen sprachlichen Situation wieder. Dies könnte sich mit Projekten wie dem SALSA-Projekt (s.o., Erk et al. 2003) ändern. Dann wäre es sinnvoll, auch die Datenbankstruktur entsprechend zu erweitern, zum Beispiel mit grundlegenden Informationen zu Häufigkeiten von Frames und Frame-Elementen in verschiedenen Korpora. Für den Fall, dass genügend Datenmaterial vorliegt, ist auch die Generierung von (statistischen) Zusammenhängen zwischen Frame-Elementen und Synsets denkbar.

Danksagungen

Ich danke Charles J. Fillmore und dem Team des FrameNet-Projekts, insbesondere Michael J. Ellsworth, Josef Ruppenhofer, Collin F. Baker und Miriam R. L. Petruck, für ihre Unterstützung. Claudia Kunze und Lothar Lemnitzer danke ich für hilfreiche Anregungen und stimulierende Fragen sowie für Informationen über GermaNet.

Die Arbeit wurde mit Unterstützung eines Stipendiums im Rahmen des Postdoc-Programms des Deutschen Akademischen Austauschdienstes (DAAD) ermöglicht.

Anmerkungen

[1] In anderen Ansätzen werden derartige zusammenhängenden Konzepte auch als Schemata, Skripte (Schank & Abelson 1977) oder Szenarien bezeichnet. Frame ist ein Oberbegriff für diese Begriffe. Zum Beispiel bezeichnet Skript eine Abfolge von Ereignissen und ist damit eine spezielle Art von Frame .

[2] [*Anmerkung der Redaktion*] Die im Original des Artikels aus dem FrameNet-Desktop (vgl. Abbildung 2) hergeleiteten farbigen Hinterlegungen der Relationstypen in den Beispielen lassen sich drucktechnisch nicht wiedergeben. Statt der farbigen Hinterlegung werden die Begriffe bzw. Mehrwortgruppen durchgehend unterstrichen. Die Relationstypen werden durch tiefgestellte Indices nach folgendem Schlüssel verwendet:

Relation	**Index**
Traveler	1
Source	2
Goal	3
Path	4

Area	5
Vehicle	6
Co_Travelers	7
Baggage	8
Duration	9
Distance	10
Purpose	11
Motion Verb	12

[3] Atkins, Rundell & Sato (2003) allerdings verwenden FrameNet für die Analyse eines polysemen Lexems.

[4] Es handelt sich hierbei eigentlich um Satzsammlungen.

Literatur

Atkins, B.T.S.; Rundell, M.; Sato, H. (2003). "The Contribution of FrameNet to Practical Lexicography." In: International Journal of Lexicography 16(3) (2003), 333-357.

Baker, C. F.; Fillmore, C. F. & Cronin, B. (2003). "The Structure of the FrameNet Database." In: International Journal of Lexicography 16(3) (2003), 281-296.

Beckwith, R. et al. (1991). "WordNet: A Lexical Database Organized on Psycholinguistic Principles. In: Zernik, U. (Hrsg.) (1991). Lexical acquisition. Exploiting On-line Resources to Build a Lexicon. Hillsdale, NJ: Lawrence Erlbaum, 211-232.

Collins, A. M.; Quillian, M. R. (1969). "Retrieval Time from Semantic Memory." In: Journal of Verbal Learning and Verbal Behavior 8 (1969), 240-247.

Conrad, C. (1972). "Cognitive Economy in Semantic Memory." In: Journal of Experimental Psychology 92(2) (1972), 149-154.

Erk, K.; Kowalski, A.; Pinkal, M. (2003). "A Corpus Resource for Lexical Semantics." In: Proceedings of The 5th International Workshop on Computational Semantics (IWCS5). Tilburg, The Netherlands, January 2003, *http://www.coli.uni-sb.de/~erk/OnlinePapers/LexProj.ps* [accessed April 2004].

Erk, K. et al. (2003). "Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In: Hinrichs, E. Roth, D. (eds.) (2003). Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 537-544.

Fellbaum, Ch. (1998). "A Semantic Network of English: The Mother of all WordNets." In: Computers and the Humanities 32 (1998), 209-220.

Fillmore, Ch. J. (1968). "The Case for Case." In: Bach, E.; Harms, R. T. (Hrsg.) (1968). Universals in Linguistic Theory. New York: Holt, Rinehard, and Winston, 1-88.

Fillmore, Ch. J. (1977). "The Case for Case Reopened." In: Cole, P.; Sadock, J. M. (Hrsg.) (1977). Syntax and semantics 8: Grammatical relations. New York: Academic Press, 59-81.

Fillmore, Ch. J. (1982). "Frame Semantics." In: The Linguistic Society of Korea (Hrsg.) (1982). Linguistics in the Morning Calm. Seoul: Hanshin Publishing Company, 111-137.

Fillmore, Ch. J. (1985). "Frames and the Semantics of Understanding." In: Quaderni di Semantica 6(2) (1985), 222-254.

Fillmore, Ch. J.; Atkins, B.T.S. (1992). "Towards a Frame-based Organization of the Lexicon: The Semantics of RISK and its Neighbors." In: Lehrer, A.; Kittay, E. (Hrsg.) (1992). Frames, Fields and Contrast: New Essays in Semantics and Lexical Organization. Hillsdale, NJ: Lawrence Erlbaum, 75-102.

Fillmore, Ch. J.; Atkins, B.T.S. (1994). "Starting where the Dictionaries Stop: The Challenge for Computational Lexicography." In: Atkins, B.T.S; Zampolli, A. (Hrsg.) (1994). Computational Approaches to the Lexicon. Oxford: Oxford University Press, 350-393.

Fillmore, Ch. J.; Johnson, C. R.; Petruck, M.R.L. (2003). "Background to FrameNet." In: International Journal of Lexicography 16(3) (2003), 235-250.

Fillmore, Ch. J. et al. (2003). "FrameNet in Action: The Case of Attaching." In: International Journal of Lexicography 16(3) (2003), 297-332.

Geckeler, H. (1971). Strukturelle Semantik und Wortfeldtheorie. München: Fink.

Gipper, H. (1973). „Das sprachliche Feld." In: Grebe, P. et al. (1973³). Duden – Grammatik der deutschen Gegenwartssprache. Mannheim: Bibliographisches Institut Dudenverlag [= Der große Duden, Bd. 4].

Gipper, H. (1995). „Jost Trier und das sprachliche Feld. Was bleibt?" In: Zeitschrift für germanistische Linguistik 23(3) (1995), 326-341.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Kunze, C. (1999). "Semantics of Verbs within GermaNet and EuroWordNet." In: Kordoni, V. (ed.) Proceedings of the ESSLI-99 Workshop on Lexical Semantics and Linking in Constraint-Based Theories, 189-200, *http://www.folli.uva.nl/CD/1999/library/coursematerial/Kordoni/kunze.pdf*, *http://citeseer.nj.nec.com/kunze99semantics.html* [accessed April 2004].

Kunze, C. (2003). „Verbsemantik in GermaNet – eine Exploration." In: Cyrus, L. et al. (Hrsg.) Sprache zwischen Theorie und Technologie. Festschrift für Wolf Paprotté zum 60. Geburtstag. Language between theory and technology. Studies in honour of Wolf Paprotté on occasion of his 60th birthday. Wiesbaden: Deutscher Universitätsverlag, 113-122.

Kunze, C.; Lemnitzer, L. (2002a). "GermaNet - Representation, Visualization, Application." In: Proceedings LREC 2002. 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain, May/June 2002, 1485-1491.

Miller, G. A. (1998). "Nouns in WordNet." In: Fellbaum, Ch. (ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA / London: MIT Press, 23-46.

Petruck, M.R.L.; Boas, H. (2003). "All in a Day's Week." Presentation at Workshop on Frame Semantics, International Congress of Linguistics. Prague, July 2003, *http://framenet.icsi.berkeley.edu/~framenet/papers/weekday.pdf* [accessed April 2004].

Schank, R. C.; Abelson, R. R. (1977). Scripts, Plans, Goals and Understanding. Hillsdale, NJ: Lawrence Erlbaum.

Schulte im Walde, S. (1998). Automatic Semantic Classification of Verbs According to their Alternation Behaviour. Diplomarbeit. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (IMS), *http://citeseer.ist.psu.edu/im98automatic.html* [Zugriff April 2004].

Schumacher, F.; Steiner, P. (2002). „Aspekte der Bedeutung: Semantik." In: Müller, H. M. (Hrsg.) (2002). Arbeitsbuch Linguistik. Paderborn: Schöningh [= UTB 2169], 170-198.

Subirats, C.; Petruck, M.R.L. (2003). "Surprise: Spanish FrameNet!" In: Proceedings of XVII International Congress of Linguistics, Prague, July 2003, 24-29.

Trier, J. (1931). Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Heidelberg: Winter.

Weisgerber, J. L. (1962). Grundzüge der inhaltsbezogenen Grammatik. Düsseldorf: Schwann.

Zillig, W. (Hrsg.) (1994). Jost Trier. Leben – Werk – Wirkung. Münster: Aa Verlag.

Autoren

Kathrin Beck
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 19-23, 72074 Tübingen
Kathrin.beck@web.de

Michael Beißwenger
Universität Dortmund
Institut für deutsche Sprache und Literatur
44221 Dortmund
beisswenger@hytex.info

Christian Biemann
Universität Leipzig
Institut für Informatik
Postfach 920, 04009 Leipzig
biem@informatik.uni-leipzig.de

Benjamin Birkenhake
Universität Bielefeld
Fakultät für Linguistik
und Literaturwissenschaft
Postfach 10 01 31, D-33501 Bielefeld
ben@vox-populi.de

Stefan Bordag
Universität Leipzig
Institut für Informatik
Postfach 920, 04009 Leipzig
sbordag@informatik.uni-leipzig.de

Kerstin Bücher
Universität Erlangen-Nürnberg
Department of Computer Science
and FORWISS
Haberstraße 2, D-91058 Erlangen
kerstin.buecher@forwiss.de

Paul Buitelaar
DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken
paulb@dfki.de

Simon Clematide
Universität Zürich
Institut für Computerlinguistik
Winterthurerstr. 190, CH-8057 Zürich
siclemat@cl.unizh.ch

Günther Görz
Universität Erlangen-Nürnberg
Department of Computer Science
and FORWISS
Haberstraße 2, D-91058 Erlangen
goerz@informatik.uni-erlangen.de

Matthias Jörg
DaimlerChrysler AG
Postfach 2360, 89013 Ulm
matthias.joerg@daimlerchrysler.com

Claudia Kunze
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 19-23, 72074 Tübingen
kunze@sfs.uni-tuebingen.de

Manuela Kunze
Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
Postfach 4120, 39016 Magdeburg
makunze@iws.cs.uni-magdeburg.de

Lothar Lemnitzer
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 19-23, 72074 Tübingen
lothar@sfs.uni-tuebingen.de

Eva Anna Lenz
Universität Dortmund
Institut für deutsche Sprache und Literatur
44221 Dortmund
lenz@hytex.info

Bernd Ludwig
Universität Erlangen-Nürnberg
Department of Computer Science
and FORWISS
Haberstraße 2, D-91058 Erlangen
bdludwig@forwiss.de

Jan Frederik Maas
Universität Dortmund
Institut für deutsche Sprache und Literatur
44221 Dortmund
maas@hytex.info

Steffen Meschkat
talk@mesch.org

Rainer Osswald
FernUniversität in Hagen
Praktische Informatik VII
Universitätstraße 1, 58084 Hagen
rainer.osswald@fernuni-hagen.de

Daniela Alina Plewe
Franz – Künstler- Str. 2
10969 Berlin
mail@danielaplewe.de

Uwe Quasthoff
Universität Leipzig
Institut für Informatik
Postfach 920, 04009 Leipzig
quasthoff@informatik.uni-leipzig.de

Dietmar Rösner
Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
Postfach 4120, 39016 Magdeburg
roesner@iws.cs.uni-magdeburg.de

Maren Runte
Universität Dortmund
Institut für deutsche Sprache und Literatur
44221 Dortmund
runte@hytex.info

Bogdan Sacaleanu
DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken
sacaleanu@dfki.de

Frank-Peter Schweinberger
Universität Erlangen-Nürnberg
Department of Computer Science
and FORWISS
Haberstraße 2, D-91058 Erlangen
goerz@informatik.uni-erlangen.de

Manfred Stede
Universität Potsdam
Institut für Linguistik
Postfach 601553, 14415 Potsdam
stede@ling.uni-potsdam.de

Diana Steffen
Consultants for Language Technology
Stuhlsatzenhausweg 69, 66123 Saarbrücken
steffen@clt-st.de

Autoren

Petra Steiner
International Computer Science Institute
1947 Center St. Suite 600
Berkeley, CA 94704-1198 , USA
petra@ICSI.berkeley.EDU

Angelika Storrer
Universität Dortmund
Institut für deutsche Sprache und Literatur
44221 Dortmund
storrer@hytex.info

Iman Thabet
Universität Erlangen-Nürnberg
Department of Computer Science
and FORWISS
Haberstraße 2, D-91058 Erlangen
iman@informatik.uni-erlangen.de

Andreas Wagner
Universität Tübingen
Sonderforschungsbereich 441
Nauklerstraße 35, D-72074 Tübingen
wagner@sfs.uni-tuebingen.de