

Using WordNets in Teaching Virtual Courses of Computational Linguistics

Lothar Lemnitzer and Claudia Kunze

Seminar für Sprachwissenschaft

Universität Tübingen

Wilhelmstr. 19, 72074 Tübingen, Germany

E-Mail: lothar@sfs.uni-tuebingen.de, kunze@sfs.uni-tuebingen.de

Abstract. This paper focuses on wordnets, especially GermaNet, as topics of teaching and learning in the field of Computational Linguistics. We are aiming at two major goals: to use wordnets for the design of tasks in core modules of the Computational Linguistics curriculum on the one hand, and, on the other hand, to enhance the wordnet structure and its accessibility by the different student projects that have been defined and accomplished. These projects, coping with various structural and content-oriented issues of wordnets, have evolved from three virtual courses taught in Tübingen and Osnabrück. They will be presented in this paper. By establishing wordnets as teaching and learning contents, advanced students should be attracted to join the international wordnet research community.

1 Introduction

In this paper, we will outline how lexical semantic wordnets like GermaNet [1] can be useful subjects of teaching and learning in the field of Computational Linguistics. GermaNet currently forms part of three virtual courses within the framework of a national E-Learning project, MiLCA¹: *Computational Lexicography* and *Applied Computational Linguistics*, held in Tübingen, and *NLP tools for Intelligent Computer Aided Language Learning*, held in Osnabrück. These are virtual courses open to students of different universities in Germany and Switzerland, which shall yield core modules of Computational Linguistics curricula. The students gather in a virtual classroom with shared work spaces, a whiteboard and communication facilities for collaboration on various exercises. In the *Computational Lexicography* course, GermaNet figure/table as a prototype of a lexical database. Within *Applied Computational Linguistics*, a course which is centered around a tool providing for intelligent access to dictionaries, the GermaNet data structure constitutes one of the underlying dictionary sources. The *I-CALL* course has developed, among other features, a vocabulary trainer on the basis of GermaNet data.

From our teaching experience, we have learned that students enjoy working with GermaNet. Lexical semantic wordnets seem to be appealing for the clarity and simplicity of

¹ The project on which this paper reports – MiLCA (media intensive learning modules for the practical training of computational linguists) - is being funded by the German Federal Ministry of Education and Research (project ID: 01NM167). The authors of this paper are, however, fully responsible for its content.

their structures, the richness of their contents and the variety of natural language processing tasks in which they may play a role.

The projects which we will describe in the following section focus on various aspects of wordnets:

- their linguistic contents,
- their data structure and presentation,
- tools for accessing and visualizing wordnet structures,
- issues of evaluation and
- wordnets as lexical resources for NLP applications.

This division may also be regarded as a proposal for an appropriate agenda in view of research topics².

Some student projects have already been completed, others are still under development and some of them are planned for teaching future courses. The outcome of the accomplished projects turned out to be quite encouraging.

2 The Student Projects in Detail

This section will present the student assignments w.r.t. the linguistic structure of wordnets (2.1), the data structure (2.2) and the development of tools (2.3). In section 2.4, evaluation tasks will be outlined, whereas section 2.5 deals with projects that are using GermaNet as a lexical resource. For each assignment, the title, a short description and the intended outcome of the work will be given.

2.1 Linguistic Aspects

Analysis of the meronymy / holonymy relation and its encoding in GermaNet. While Princeton WordNet encodes three different types of meronymy relations, and EuroWordNet realizes one generic meronymy pointer as well as five subtypes, in GermaNet only a unique pointer covering all instances of meronymy has been realized so far. Concept pairs which are encoded as meronyms should be checked under the following aspects: a) Is the application of three meronymy pointers feasible for GermaNet? b) Will a subdivision into different meronymy pointers yield transitivity for these relations or are there still examples in which transitivity is blocked? c) Are there examples for pairs of concepts where the meronymy relation is not symmetric?

The investigation is based on the subdivision of meronymy from WordNet as well as on the classification proposed by Chaffin [2, p. 274ff]. This project, which is currently under way, aims at refining the meronymy / holonymy relation in GermaNet.

Analysis of the antonymy relation. Similarly to the case of meronymy, GermaNet implements a unique pointer for encoding the antonymy relation between lexical units. Different types of opposites, like ‘man’ vs. ‘woman’, ‘busy’ vs. ‘lazy’, ‘warm’ vs. ‘cold’ or ‘arrive’ vs. ‘leave’ are, thus, subsumed and uniformly treated under the label of antonymy.

² The agenda encompasses topics which may have been discussed in the wordnet community rather than entirely new items

The student exercise consists in developing an adequate subclassification of antonymy, dividing the data into appropriate subclasses, which should account for complementary opposites as well as for scalar and gradable opposites. Furthermore, a set of relevant features should be defined, which captures opposites of, e.g., sexus or directionality for nouns and verbs. The empirical analysis of the GermaNet antonyms should consider the categories being proposed in the descriptive approaches of [3,4].

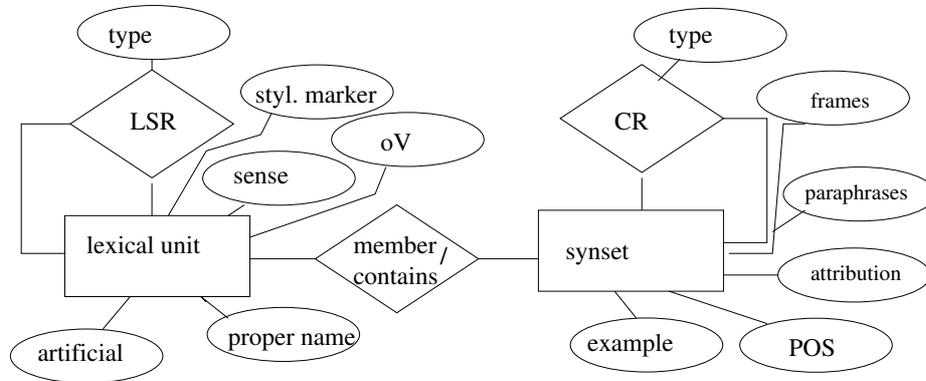
Applicability of regular polysemy in wordnets. Pustejovsky et al. criticize WordNet for ignoring existing regularities between senses [5]. It is, however, still unclear whether wordnets should implement regular sense relations or not, and which should be the appropriate hierarchical level for the application of such rules. The analysis, therefore, concentrates on lexical (sub-)fields which are in the scope of a regular sense extension, e.g. instances of the type ‘building-institution-staff’ or instances of the type ‘tree-wood-fruit’. It will be checked whether generic rules are feasible or not, and, if so, on which hierarchical level they should apply or when blocking of these rules would be necessary.

The interaction of verb concepts, verbal argument structure and Aktionsart / Aspect. Aspectual properties of verbs have recently become a major topic in semantic research. Therefore, it would be interesting to prove or disprove the necessity and feasibility of encoding further information on argument structure and Aktionsart / Aspect in the existing hierarchy of verbal concepts. Some preliminary investigation reveals that aspectual hierarchies cannot be assumed straightforwardly; otherwise, the representation of causative and inchoative variants of a verb within one synset has to be abandoned. A closer examination shall focus on a specific verb field, e.g. verbs of locomotion, and test the possibility of creating an inheritance hierarchy (with overwriting) for aspectual features in interaction with argument structure. This project will be assigned in the near future.

2.2 The Data Model and Data Structure of Wordnets

Conversion of the lexicographers’ files into an XML format. Neither the GermaNet lexicographers’ files nor the compiled database yield an ideal format for data exchange, presentation and integration into NLP tasks. XML is more convenient for these purposes. Based on the data model of GermaNet, which is captured by an Entity-Relationship graph (cf. figure 1), several students developed programs which convert the Lexicographers’ Files of GermaNet into an XML representation. The respective DTDs have been created collaboratively. The outcome of this project is documented in [6] and [7]). This task is designed to contribute to the ongoing discussion on the standardization of wordnet formats (e.g. the BalkaNet approach, cf. [8]).

Integration of the GermaNet objects and relations into the Resource Description Framework. Some work has already been done with WordNet within the framework of the Semantic Web initiative [9], but the resulting files encompass only a part of the Princeton WordNet. Before starting to convert GermaNet accordingly, and even more exhaustively, we would like to understand how well wordnet structures fit into the structures of full-fledged knowledge representation languages like DAML and OIL, which are built on top of RDF. An examination of these languages with wordnet structures in mind shall prove or disprove the usefulness of GermaNet objects and relations for the RDF and the other knowledge representation languages mentioned above. The work is under way.



CR conceptual relation; LSR lexical semantic relation; oV orthographic variant

Fig. 1. An entity relationship graph of the GermaNet data model

GermaNet representation as Scalable Vector Graphics. SVG (cf. [10]) might turn out to be a reliable standard as well as a handy tool for the visualization of wordnet objects and relations. A wordnet can be conceived as a large map where one wants to zoom in at a particular synset and see the data and relations that are associated with it. The project, which is not yet assigned, intends to explore the feasibility of data conversion into the SVG format and the functionality of existing visualization tools.

2.3 Tools

Development of tools for the extraction of the lexical and conceptual neighborhood of a lexical unit or a synset. The tools currently being developed are based on the XML representation of the data. The assignment in question addresses a user need for extracting data which are neighboring a particular synset or lexical unit. Currently, there are two projects devoted to this task: one employs a relational database for the intermediate representation of the data, the other accesses the data in their original format, using XSL Transformations to generate the output. Both methods will be evaluated in terms of their processing speed and flexibility. The GUI of one of the tools is shown in figure 2.

Visualization of the wordnet. Within another student project, a visualization tool which operates over the whole wordnet structure has been developed³. The XML representation of the wordnet is used as data base. The visualization of the data is very flexible, yet, too slow for realistic user scenarios. Results of the project have originally been presented in [7]. The outcome of the project has motivated our search for representation alternatives, e.g. Scalable Vector Graphics (see 2.2).

³ Another promising approach to the visualization of wordnet structure has been launched by the Czech wordnet group: Visdic, cf. [11]

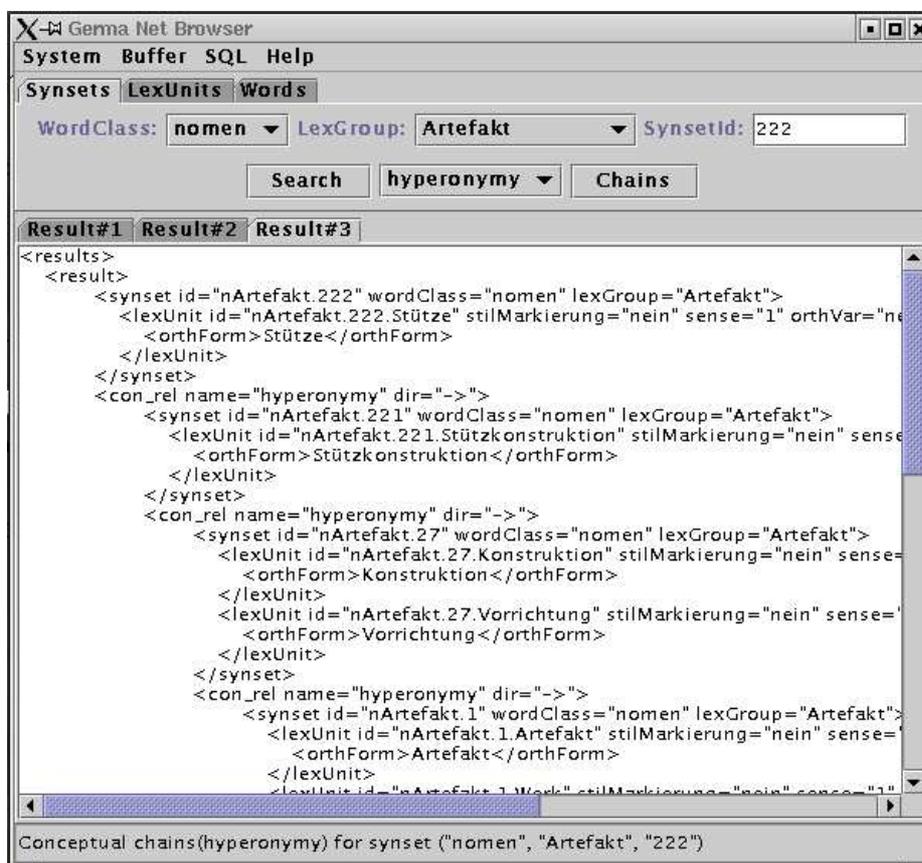


Fig. 2. GUI of a GermaNet extraction tool

2.4 Evaluation of Wordnet Data

Evaluation of the synset approach for IR and MT applications. 1. With a perspective on Machine (Aided) Translation, the EuroWordNet [12] ILI links, which are established between whole synsets (instead of lexical units), will be explored and compared with corresponding equivalence relations in a bilingual dictionary. The following questions are of interest: Does the majority of equivalence pairs between lexical units of the two languages, which are established indirectly through the relations between the synsets they are members of, really constitute pairs that can be used in substitution operations in MT? What is the relative share of mismatches for a particular language pair? 2. For IR applications, user tests will be performed on a search engine front-end, which expands query terms with their direct lexical and conceptual neighbors. The task to be developed could benefit from results of cross-language IR evaluation (cf. [13]). *Evaluation of the feasibility of the "sense clustering" approach.* On the basis of a corpus of citations, some words which are highly polysemous shall be examined and the GermaNet senses mapped onto the corpus citations. The manual encoding of these

data within different sense division scenarios serves to prepare an experiment with automatic classifiers, which will be trained and tested on the different versions of the sense encoded corpus data.

2.5 Use of GermaNet As a Lexical Resource

LSI generated lexical semantic relations compared to GermaNet relations. In a larger project, which will lead to a diploma thesis, clusters of lexical units with alleged nearness in semantic space have been extracted from a large German newspaper corpus using Latent Semantic Indexing. In this project, the conceptual and lexical relations which are used for the construction of wordnets serve to evaluate the quality of the automatically generated “sense clusters”. The aim of the evaluation is to investigate whether the lexical clusters yielded by LSI are really semantic, as the supporters of this approach claim.

GermaNet as a lexical basis for a vocabulary trainer. A group of students in Osnabrück, Edinburgh and Tübingen have developed a network-like platform for collaborative work. Within this framework, GermaNet as the central source of lexical knowledge supports a vocabulary trainer. The outcome of the project will be reported on the GLDV-Workshop on “Applications of the German Wordnet in Theory and Practice” in October 2003 (cf. [14]).

3 Conclusion

We have presented various examples of student projects that focus on GermaNet, and wordnets in general, as subjects of teaching and learning. With these examples, we have demonstrated how stimulating research and development projects can be in the teaching of advanced students. Wordnets are highly attractive for students for reasons of their simplicity and clarity and the richness of information they provide. With this paper, we want to claim the usefulness of wordnets for teaching in a broader range of subjects, including Computational Linguistics, Computational Lexicography, General Linguistics, Cognitive Science and Language Teaching. We would like to establish, within Global Wordnet Association, a repository which should encompass:

- a list of small to medium-sized research and development tasks for advanced undergraduate and graduate students (including task descriptions, methods, resources needed, possible outcomes);
- a list of results of student assignments (students might be further motivated by this prospect of publishing their work);
- a forum for discussing didactic issues.

References

1. Kunze, C., Naumann, K.: Germa Net Homepage. <http://www.sfs.uni-tuebingen.de/lsd>.
2. Chaffin, R.: The concept of a semantic relation. In Lehrer, A., Kittay, E. F., eds.: *Frames, Fields, and Contrasts. New Essays in Semantic and Lexical Organization*. Lawrence Erlbaum, Hillsdale (1992) 253–288.
3. Cruse, D. A.: *Lexical Semantics*. Cambridge University Press, Cambridge (1986).

4. Agricola, C., Agricola, E.: Wörter und Gegenwörter. Antonyme der deutschen Sprache, Leipzig: VEB Bibliographisches Institut (1987).
5. Pustejovsky, J., Boguraev, B., Verhagen, M., Buitelaar, P., Johnston, M.: Semantic Indexing and Typed Hyperlinking. In: Proceedings of the AAAI '97. (1997)
<http://www.cs.brandeis.edu/~llc/publications/aaai97.ps>.
6. Kunze, C., Lemnitzer, L.: Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet. In: Proc. LREC 2002 Workshop on Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation. (2002) 24–29.
7. Kunze, C., Lemnitzer, L.: GermaNet – representation, visualization, application. In: Proc. LREC 2002, Gran Canaria, May/June. (2002) 1485–1491.
8. Smrz, P.: Storing and retrieving Wordnet database (and other structured dictionaries) in XML lexical database management system. In: Proc. of First International WordNet Conference, Mysore, India (2002).
9. Melnik, S., Decker, S.: Wordnet RDF Representation.
<http://www.semanticweb.org/library/> (2001).
10. Ferraiolo, J., Fujisawa, J., Jackson, D.: Scalable Vector Graphics (SVG) 1.1 Specification.
<http://www.w3.org/TR/SVG11/> (2003).
11. Pavelek, T., Pala, K.: Visdic – a new tool for wordnet editing. In: Proc. of First International WordNet Conference, Mysore, India (2002).
12. Vossen, P.: Eurowordnet homepage. (URL: <http://www.illc.uva.nl/EuroWordNet/>).
13. Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3–4, 2001, Revised Papers. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: CLEF. Volume 2406 of Lecture Notes in Computer Science., Springer (2002).
14. Beck, K.: Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa (Mapping Architecture for People's Association). In Kunze, C., Lemnitzer, L., Wagner, A., eds.: Proceedings of the First GermaNet User Conference, Tübingen, 9–10 October 2003, GLDV (2003).