# Customizing GermaNet for the Use in Deep Linguistic Processing

Melanie Siegel
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
siegel@dfki.de

Feiyu Xu
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
feiyu@dfki.de

Günter Neumann
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
neumann@dfki.de

## Abstract

In this paper we show an approach to the customization of GermaNet to the German HPSG grammar lexicon developed in the Verbmobil project. GermaNet has a broad coverage of the German base vocabulary and fine-grained semantic classification; while the HPSG grammar lexicon is comparatively small und has a coarse-grained semantic classification. In our approach, we have developed a mapping algorithm to relate the synsets in GermaNet with the semantic sorts in HPSG. The evaluation result shows that this approach is useful for the lexical extension of our deep grammar development to cope with real-world text understanding.

## Introduction

The lexical-semantic information encoded in online ontologies like GermaNet (Hamp and Feldweg, 1997) is very useful for different natural language applications: information extraction, lexical acquisition and intelligent information retrieval. In this paper, we provide an approach, which customizes the GermaNet lexical semantic information to the HPSG lexicon in order to extend the lexicon for the improvement of the deep linguistic processing of real- world text.

In the DFKI project Whiteboard, we aim to integrate different natural language resources to deal with real-world text understanding. One particular goal is the integration of deep NLP (DNLP) and shallow NLP (SNLP). In recent years, a number of efforts have been spent towards the increase of parsing performance with HPSG (Flickinger et al., 2000). Especially the PET parser developed at the CL department at the University of the Saarland has demonstrated that it is now possible to use an HPSG parser for processing of real-world text using large Grammars for German and English. However, one of the bottlenecks with real-text processing is the high amount of very productive domain-specific lexical entities. The approach followed in the Whiteboard project is to let the domain-specific shallow component SPPC (Piskorski and Neumann, 2000) do the main lexical processing and integrate the lexical entities via the HPSG type system. The PET parsing system is then called with the results of the SPPC lexical processor to perform an HPSG analysis. The integration of the SPPC and PET system is based on the HPSG type system.

In our current system we have applied the German grammar developed in the Verbmobil project (Müller and Kasper 2000), which originally aimed to understand and translate dialogue language, to economic news. The result was that apart from NE's, 78.49% of the missing lexical items are nouns. Due to the integration of SPPC, NE recognition as well as coverage of nouns is now increased. However, SPPC only computes POS and morpho-syntactic information. But for the deep text analysis, a solution for retrieving nouns with their semantic sorts is essential, because the semantic sorts are useful for the semantic construction and for providing semantically based selectional restrictions, which are essential for guiding the search space defined by the HPSG grammar. GermaNet has a huge coverage of German word stems and the words are tagged with the POS information and their semantic classification. Therefore, we did experiments to automatically convert the semantic concepts in GermaNet to the semantic sorts defined in the HPSG lexicon. We have implemented an algorithm that computes the mapping relevance from a semantic concept in GermaNet to a semantic sort in the HPSG lexicon. In addition, we have also developed a *GermaNet2HPSG* tool which can not only be used for the online text analysis by assigning a word to the most adequate HPSG semantic sorts based on its GermaNet concepts, but can also be used for the offline lexicon generation. The GermaNet2HPSG tool is based on the Whiteboard Germa/WordNet ontology inference tool, which supports the search and navigation of the ontology information in Germa/WordNet.

## 2 Customization of GermaNet to the deep grammar

The semantic database (SemDb) (Bos et al. 1996) in the HPSG lexicon was set up in the Verbmobil project used in different modules. The HPSG grammar makes use of the SemDb in order to restrict and disambiguate readings via sortal restrictions on verbal arguments. It contains words and their semantic sorts as well as valence information and sortal restrictions of arguments. The semantic sorts are organized in a hierarchy. The German semantic database contains about 7800 words. Although the hierarchy is quite simple, it turned out to be very useful in the parsing process.

Let us consider the relationships between the semantic sorts and the synsets in more detail. On the one hand, there are 30 different sorts in this hierarchy as opposed to almost 20.000 synsets in the GermaNet ontology. On the other hand, each single word is annotated with one semantic sort in the SemDb and different sets of synsets in GermaNet. It is thus obvious that there cannot be a direct match from SemDb sorts to GermaNet synsets. We therefore decided to learn the relationships between the semantic sorts and the synsets.

### 2.1 Training Method

Using the nouns with semantic sort annotations from the SemDb as our training corpus, we developed a mapping algorithm from semantic sorts to synsets:

1) *Retrieve the hypernyms (synsets) in GermaNet of all nouns in the SemDb.*
2) *Count the frequency ($f_{ij}$) of each GermaNet synset$_i$ for all words in a certain HPSG semsort$_j$.*
3) *Compute the sum ($F_i$) of the frequencies of each GermaNet synset$_i$ for all HPSG semsorts in the corpus.*

$$F_i = \sum_{j=1}^{|semsorts|} f_{ij}$$

4) *Compute the mapping relevance ($R_{ij}$) of a GermaNet synset$_i$ to a certain HPSG semsort$_j$ with respect to the whole training data.*

$$R_{ij} = \frac{f_{ij}}{F_i}$$

The training results in a table of SemDb sorts and GermaNet synsets annotated with their mapping relevance; see the following example which shows the mapping from the synset 'Stelle, Ort, Stätte' (engl. place, room) and the synset 'Äußerung' (engl. uttrance) to the semantic sorts.

| Synset | Semantic Sort | Mapping Relevance (%) |
|---|---|---|
| *Stelle,Ort,Stätte* | *Symbol* | *0.60* |
| *Stelle,Ort,Stätte* | *geo_location* | *3.01* |
| *Stelle,Ort,Stätte* | *Location* | *6.02* |
| *Stelle,Ort,Stätte* | *nogeo_location* | *44.58* |
| *Äußerung* | *Field* | *2.63* |
| *Äußerung* | *abstract* | *15.79* |
| *Äußerung* | *info-content* | *21.05* |
| *Äußerung* | *communication situation* | *23.68* |

### 3.4 The Annotation of Words with SemDb sorts

Using the mapping table, words not contained in the SemDb can now be annotated with semantic sorts used in the deep grammar. The annotation algorithm works as follows:

1) *Retrieve the hypernyms (synsets) in GermaNet of a word; different senses have different sets of synsets.*
2) *For each sense,*
   *i) sum the mapping relevance weights from its GermaNet synsets to semantic*

*sorts.*

*ii) Select the best four mappings*

The result is an ordered list of semantic sorts with relevance values. A word that has more than one sense in GermaNet will also obtain more than one list of semantic sorts.

### 3.6 Evaluation

We examined a corpus of 4664 nouns extracted from economic news (Wirtschaftswoche 1992) that were not contained in the SemDb. 2312 of them are known for GermaNet. They obtain 2811 senses according to the GermaNet and were automatically annotated with semantic sorts. The evaluation of the annotation accuracy yields encouraging results:

- In 76.52% of the cases the computed sort with the highest processed probability was correct.
- In 20.70% of the cases, the correct sort was one of the next three sorts.
- In 2.74% of the cases, the first four computed sorts did not contain the correct one.

This means that the accuracy among the first four annotations is 96.52%. However, we need to improve the accuracy of the first reading. One of the reasons for errors is given the size of HPSG lexicon and therefore our mapping table is incomplete. We will consider this issue in the future work.

## 3 Implementation

The customization tool makes use of the Whiteboard Germa/WordNet inference tool. We call it Germa/WordNet inference tool because it can also be applied to retrieve the lexical semantic information in WordNet. The WordNet content has been inserted into the relational database MySQL too. Both GermaNet and WordNet share the same database design. The two tools are implemented in JAVA with JDBC access to MySQL. The GermaNet2HPSG component has already been integrated to the Whiteboard text processing server. It supports the deep text processing by assigning online the semantic sort to a word based on the GermaNet synsets. The advantage is that we do not need convert the entire GermaNet lexicon to the deep analysis lexicon. It reduces the online lexicon search and provides only the semantic sort when it is needed.

## Conclusion and Outlook

We have built a tool to automatically map GermaNet synsets to semantic sorts of the kind used in a deep HPSG grammar. The mapping result is used in a system that integrates deep and shallow processing for retrieving semantic sorts of nouns not contained in the deep lexicon. In order to extend the accuracy of the mapping table, we plan to use the evaluated annotation for the expansion of the training corpus. A next step will be the application to verbs and adjectives. We are planning to combine the information of the NEGRA treebank (Brants, 2000) with the GermaNet ontology in order to gain information about the valence and sortal restrictions of verbs. In order to extend the grammar coverage we are thinking of refining the HPSG semantic database ontology by using the GermaNet ontology.

## References

Bos, J., M. Schiehlen and M. Egg (1996). *Definition of the Abstract Semantic Classes for the Verbmobil Forschungsprototyp 1.0*. Universität des Saarlandes, IBM Heidelberg, Universität Stuttgart. Verbmobil-Report 165.

Brants, Thorsten (2000). *Inter-Annotator Agreement for a German Newspaper Corpus*. In Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.

Flickinger, D., S. Oepen, H. Uszkoreit and J. Tsuji (Eds.) (2000). *Special Issue on Efficient processing with HPSG: Methods, Systems, Evaluation*. Journal of Natural Language Engineering 6 (2000) 1. Cambridge, UK: Cambridge University Press. (in press)

Hamp, B. and H. Feldweg (1997) *GermaNet - a Lexical-Semantic Net for German*. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997

Müller, S. and W. Kasper (2000). *HPSG Analysis of German*. In “Verbmobil: Foundations of Speech-to-Speech Translation”, W. Wahlster, ed., Springer Verlag, Berlin, 238-253.

Piskorski, J. and G. Neumann (2000). *An Intelligent Text Extraction and Navigation System*. In the Proceedings of RIAO 2000 - Content-Based Multimedia Information Access, Paris, France.