# Adapting GermaNet for the Web


| **Claudia Kunze** | **Lothar Lemnitzer** |
|---|---|
| Abteilung Computerlinguistik | Abteilung Computerlinguistik |
| Seminar f. Sprachwissenschaft, Univ. Tübingen | Seminar f. Sprachwissenschaft, Univ. Tübingen |
| Wilhelmstr. 113 | Wilhelmstr. 113 |
| D-72074 Tübingen | D-72074 Tübingen |
| Germany | Germany |
| kunze@sfs.uni-tuebingen.de | lothar@sfs.uni-tuebingen.de |



**Paper ID:** GWA-2002-paper-1, short paper
**Status:** Revised version, Nov. 2001
**Keywords:** GermaNet, XML-adaption, compatibility



**Contact Author:** Claudia Kunze



**Under consideration for other conferences:** No



**Abstract:** This paper deals with the adaption of the lexical-semantic wordnet GermaNet for web-based applications. The GermaNet data have been converted into XML-conformant documents, which represent the concepts and all the basic relations defined between them, while accounting for the peculiarities of the German wordnet. We also compare GermaNet to the Princeton WordNet in order to unify the diverging representations for the use within a polylingual framework.


.

# Adapting GermaNet for the Web

**Abstract:** This paper deals with the adaption of the lexical-semantic wordnet GermaNet for web-based applications. The GermaNet data have been converted into XML-conformant documents, which represent the concepts and all the basic relations defined between them, while accounting for the peculiarities of the German wordnet. We also compare GermaNet to the Princeton WordNet in order to unify the diverging representations for the use within a polylingual framework.

## Introduction

In recent approaches to natural language processing, wordnets, which are structured along the same lines of the Princeton WordNet (see Miller et al. (1990), Fellbaum (1998))[1], have become very popular resources. These wordnets model the common and frequent words of a language and encode the basic semantic relations established between them, like synonymy, hyponymy, antonymy, and meronymy. Wordnets can support word sense disambiguation, which is crucial for information retrieval, semantic annotation, etc., and are also valuable resources for linguistic research.
Due to the increasing interest in WordNet as a pioneer resource, various language specific and interlingual initiatives to wordnet construction have been launched (GermaNet[2], French WordNet, EuroWordNet[3], BalkaNet etc.).
As web-based applications become increasingly important in the field of language engineering, it is worth making wordnets accessible to XML-based tools and services, in order to exploit the rich semantic structures encoded in them. The original WordNet has been constructed as a database, on which various tools have been built. EuroWordNet followed this strategy. A more recent version of WordNet represented as PROLOG clauses has been developed, which is therefore accessible to PROLOG-based NLP tools. It seems however quite reasonable to derive an XML representation format for wordnets. With XML being a quite powerful standard, more tools and applications will build their interfaces on this standard. For example, Chris Manning's tool for browsing and visual exploration of a structured Warlpiri dictionary (Manning et al. 2001) might be adapted. There is also a EuroWordNet-specific visualisation tool (VisDic, http://nlp.fi.muni.cz/projekty/mt/visdic/). To enable polylingual applications, it is useful to construct a common format for the various wordnets. For the German wordnet, GermaNet, XML conversion has been carried out at the Linguistics Department of the University of Tübingen.

In this paper we describe how the XML version of GermaNet has been developed. The first section outlines the German wordnet and focuses on the major differences between GermaNet and WordNet. The conversion process is briefly described in the second section. In section 3, we present the respective DTDs (Document Type Definitions). In the fourth section, compatibility issues across the English and German wordnet are raised in some detail. A few examples are given in the last section. The conclusion summarises future perspectives with regard to wordnet applications and standards.

## 1. Differences between GermaNet and WordNet

### 1.1 Some remarks on GermaNet

[1] Princeton WordNet (Miller (1990), Fellbaum (1998)) as the first in the field of wordnet construction has evolved as the quasi-standard due to various aspects: it models English, has a broad coverage, and is freely available for the research community (http://www.cogsci.princeton.edu/~wn).

[2] GermaNet has been developed within the project "SLD: Ressourcen und Methoden zur semantisch-lexikalischen Disambiguierung", which was funded by the Ministery of Research of Baden-Württemberg in 1996-1997. A second period of funding has been granted in 1999-2001. The database was built by Helmut Feldweg (first period's project leader), Claudia Kunze, Andreas Wagner, Karin Naumann, Birgit Hamp, Michael Hipp, Valérie Béchet-Tsarnou, Susanne Schüle, Christine Thielen and Rosemary Stegmann (http://www.sfs.uni-tuebingen.de/lsd).

[3] The EuroWordNet database has been built within two projects: EuroWordNet-1 (LE-4003) and EuroWordNet-2 (LE-4 8328), funded by the European Commission. The whole project was coordinated by Piek Vossen (http://www.hum.uva.nl/~ewn/).

The fundamental lack of electronic lexical-semantic resources for German (see Hamp & Feldweg (1997)) was the major motivation for constructing GermaNet a few years ago. Therefore, a first project (SLD) created an on-line thesaurus covering the German basic vocabulary. GermaNet adopted the design principles and the database technology from the Princeton WordNet. However, the German wordnet is not a mere translation of WordNet, but was built from scratch, taking into account different lexicographic resources like DUDEN 8 and DEUTSCHER WORTSCHATZ as well as corpus evidence. GermaNet includes principle-based modifications on the constructional and content-oriented level.
GermaNet currently covers some 40,000 synsets with more than 60,000 word meanings, modelling nouns, verbs, adjectives and adverbs (see Kunze (2001)). Within the EuroWordNet project, GermaNet was integrated into the polylingual EuroWordNet database (see Vossen (1999), Wagner and Kunze (1999)). We followed the merge approach, i.e., a wordnet is built independently from WordNet and the synsets are linked to the Interlingual Index (ILI) by creating the appropriate relations. The merge approach preserves language-specific patterns with differing hierarchical structures in comparison to WordNet[4].
The construction of GermaNet as well as the integration of GermaNet into EuroWordNet have been performed following independent principles, thus leading to differences with respect to the Princeton WordNet.

### 1.2 Major differences to WordNet

In spite of its general similarity with and compatibility to WordNet, we can state the following differences for GermaNet:

- we followed linguistic design principles as opposed to Princeton's psychological motivation;
- we are using artificial, i.e. non-lexicalised concepts, which have been introduced to balance the taxonomical structure more adequately;
- in GermaNet, adjectives are ordered hierarchically as opposed to Princeton's grouping by the satellite approach;
- we pursued a uniform treatment of meronymy within GermaNet, whereas WordNet has established three different pointers for *Part, Member* and *Substance*;
- within GermaNet, the causation relation can be encoded between all parts of speech, not only between verbs and adjectives;
- due to emphasising the syntax-semantics-interface for disambiguation tasks we accounted for some sixty verbal subcategorisation frames. These frames are more elaborate than the WordNet frames, and, furthermore, for each verb reading we provide a typical example.

These differences and their technical impact on compatibility for the XML conversion are outlined in more detail below.

## 2. Converting GermaNet to XML

The GermaNet lexicographer's files constitute the starting point of the conversion process[5].
The Entity-Relationship graph below visualises the structure of GermaNet and guides the conversion process (see fig. 1).
From this graph we can easily deduce:

1. the **objects**, which are synsets and lexical units[6], represented as rectangles,
2. the **attributes** of these objects, represented as circles,
3. the **relations**, represented as diamonds:
    a) **conceptual** relations which hold between instances of the synset object and
    b) **lexical-semantic** relations which hold between instances of the lexical unit object.

[4] In contrast, some language-specific wordnets were integrated via the expand approach, in which WordNet synsets were translated into the language in question. Consequently, the relational structure was adopted and therefore highly biased by WordNet.
[5] These files are the repositories in which the output of the lexicographers' work is stored.
[6] Following Cruse (1986), a lexical unit constitutes an independent form-meaning pair.

The conversion programmes are written in Perl. There are two main programmes:

1. one transforms the strings in the lexicographer's files into elements and their attributes, while checking for consistency. The transformation is structure-preserving.
2. the other checks the input files for conformancy, attaches unique IDs to all object instances and separates the object information from the relation information. From one input file, it generates two output files: one file containing the GermaNet objects, the other file containing the GermaNet relations. Furthermore, an index of all synsets and an index of all lexical units with their IDs are created.

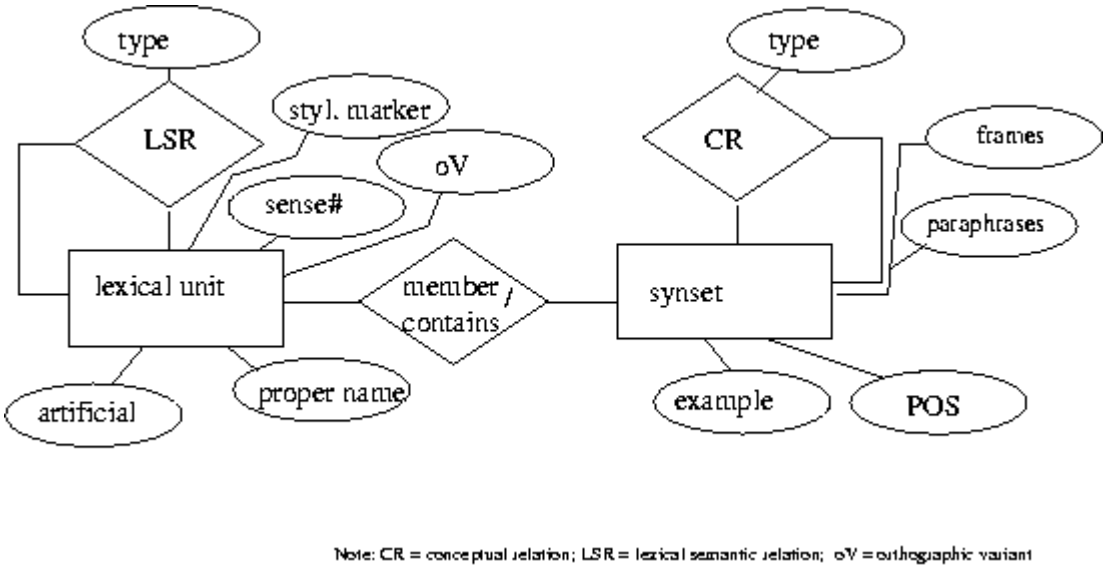


Fig.1: Entity-Relationship graph of the GermaNet structure

### 3. XML document types and their definitions

We decided to treat objects and relations with their respective attributes separately. This leads to two separate types of documents which are described by two separate document type definitions. In one document type we deal with the objects (synsets and lexical units) and their attributes, in the other document type with the relations and their attributes.
Thus the relations are easier to manage and an application-specific processing semantics can be attached once there are web-based tools, browsers etc., which are able to handle XML links properly. Second, it will be straightforward to connect one wordnet to other wordnets: the linking mechanism will be just another type of relation, now linking synsets and/or lexical units of different languages.
In the following, we present the DTDs and discuss a few design decisions. The DTDs and the documents which are generated according to them conform to the XML 1.0 and the XLink 1.0 recommendations.

### A. The object DTD

```
<!-- DTD for GermaNet objects (synsets, lexical units) -->
<!-- version 1.4, July 2001 -->
<!-- Copyright: Seminar f. Sprachwissenschaft, Universität Tübingen -->

<!ELEMENT synsets      (synset)*>
<!ELEMENT synset       ((lexUnit)+, frames?, paraphrases?, examples?)>

<!ATTLIST synset       id            ID     #REQUIRED
                       wordClass     CDATA  #IMPLIED>

<!ELEMENT lexUnit      (orthForm)>
```

```
<!ATTLIST lexUnit      id            ID     #REQUIRED
                StilMarkierung  (ja|nein)      'nein'
                sense  CDATA          #REQUIRED
                orthVar        (ja|nein)  #REQUIRED
                artificial     (ja|nein)  #REQUIRED
                Eigenname      (ja|nein) #REQUIRED >

<!ELEMENT orthForm      (#PCDATA)>

<!ELEMENT paraphrases   (paraphrase)+>
<!ELEMENT paraphrase    (#PCDATA)>
<!ELEMENT examples      (example)+>
<!ELEMENT example       (text, frame*)>
<!ELEMENT frames        (frame)*>
<!ELEMENT text          (#PCDATA)>
<!ELEMENT frame         (#PCDATA)>
```

**Description**: The document contains a set of synsets. Every synset consists of at least one lexical unit. Paraphrases may be given to characterise the meaning of the synset and examples may be added to illustrate the use of its member lexical units. For verb synsets, subcategorisation frames are given. The individual lexical units are characterised by a set of attributes, e.g. sense number and stylistic marker (*StilMarkierung*). A concept can be represented by a string which does not correspond to a lexical unit in the German vocabulary. In this case the unit is marked as artificial. The content model of most atomic elements is set to #PCDATA, therefore minimising data type restrictions. It is up to the lexicographers to fill the elements with appropriate data[7].

**B. The relation DTD**

```
<!-- DTD for GermaNet relations (conceptual and lexical-semantic) -->
<!-- version 1.0, July 2001 -->
<!-- Copyright: Seminar f. Sprachwissenschaft, Universität Tübingen -->
<ELEMENT relations (hyperonym | holonym | see | entails | causes | meronym | hyponym | antonym |
pertonym | participleOf | derivedFrom)*>
<!ELEMENT hyperonym (#PCDATA|locator|arc)*>
<!ATTLIST hyperonym xmlns:xlink CDATA #FIXED 'http://www.w3.org/1999/xlink'
      xlink:type (extended) #FIXED 'extended'
      sense CDATA #REQUIRED>
…. (the same structure for all other relations)
<!ELEMENT locator>
<!ATTLIST locator xlink:type (locator) #FIXED 'locator'
      xlink:href CDATA #REQUIRED
      xlink:label CDATA #REQUIRED>
<!ELEMENT arc>
<!ATTLIST arc xlink:type (arc) #FIXED 'arc'
      xlink:from CDATA #REQUIRED
      xlink:to CDATA #REQUIRED
      xlink:actuate (onRequest) #FIXED 'onRequest'
      xlink:show (other) #FIXED 'other'>
```

**Description**: The ‘GermaNet relations’ DTD models different types of relations (hyperonymy, antonymy etc.) which hold either between synsets or between lexical units. A link consists of two nodes (locators, specified through the IDs of the synsets or lexical units) and an arc. The attributes of

[7] The full DTDs and their documentations are available on request.

the ‘arc’ element specifies the processual behaviour whenever a link is traversed. Of course, appropriate software (browser, visualiser or else) is needed to implement these features. However, with XLink having reached the status of a W3 recommendation, its integration into standard software will be realized in the near future.

## 4. Compatibility Issues

A common access interface to the wordnets of all languages is highly desirable. NLP programmes which incorporate one or several wordnet(s) should rely on a uniform access method. Thus, it follows that all wordnets should be structurally unified, but without sacrificing the language-specific extensions and changes to the original structure of the Princeton WordNet. It is therefore necessary to consider compatibility between all wordnets. In this section, we will compare GermaNet to the Princeton WordNet, which is the original and best documented resource, raise several compatibility issues and show how they can be solved within the XML framework[8]. We distinguish six types of structural differences between WordNet and GermaNet:

1. Objects which are obviously of the same type bear different names. This can be easily solved by a mapping of these different names. The same holds for the names of relations and attributes.
2. Objects or relations have different extensions in both nets, as is the case with the CAUSE relation. In WordNet, this relation holds exclusively between verbs and adjectives. In GermaNet, synsets of all word classes are in the domain of this relation. True compatibility would require a finer granularity of the CAUSE relation in GermaNet. This could be realised by adding an attribute to it. The values of this attribute would lead to at least two subsets of items: one which is extensionally identical with the original CAUSE relation and one which characterises the GermaNet-specific extension.
3. The granularity of a relation differs. For example, WordNet divides the generic part-whole relation into three sub-relations: part (e.g. *arm,body*), member (e.g. *director, staff*), substance (e.g. *glass, glass plate*). Other values might be added to this list. GermaNet, in contrast, uniformly applies the generic relation. We recommend for WordNet to add an attribute to a truly generic part-whole-relation which divides the instances into three classes. In GermaNet, this attribute might get a value ANY, until a more fine-grained specification is implemented.
4. There are a few attributes specific to GermaNet, e.g. *StilMarkierung* (=stylistic marker) as an attribute of lexical units. For instance, the German concept *schlafen* (=sleep) has *ratzen*s, pennen*s, knacken*s, pofen*s* as hyponyms which are stylistically marked. These attributes can be INCLUDED in GermaNet and EXCLUDED elsewhere.
5. An attribute which is equivalent in both wordnets specifies a different set of values. This holds for the *verb frame* attribute. The German verb frames which are implemented in GermaNet are a closed class. For type checking, it could have been more elegant to define an attribute with a fixed set of values. For compatibility reasons, however, we voted for an element group "frames" with frames as its elements and #PCDATA as data type. We delegate the type checking to other routines.
6. The adjective domain in GermaNet differs fundamentally from that in WordNet. The domain is ordered hierarchically in GermaNet, whereas WordNet applies an associative similarity relation which groups adjectives in equivalence classes. At present, we do not see any easy solution which would preserve compatibility in this case.

## 5. Examples

In the following we will present two examples to illustrate the XML format of the data. The first example presents a verbal synset, the second one a link between a verb and its hyperonym.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE synsets SYSTEM "germanet_objects.dtd"[]>
```

[8] Alternatively, the EuroWordNet specification could serve as starting point for a uniform XML representation across languages. We decided, however, to focus on the mapping between GermaNet and WordNet in order to capture all individual properties of GermaNet and its full coverage of data.

```
....
<synset id="vKoerperfunktion.262" wordClass="verben"><lexUnit Eigenname="nein"
artificial="nein" id="vKoerperfunktion.262.pennen" orthVar="nein" sense="0" stilMarkierung="ja">
<orthForm>pennen</orthForm> </lexUnit>
<lexUnit Eigenname="nein" artificial="nein" id="vKoerperfunktion.262.knacken" orthVar="nein"
sense="0" stilMarkierung="ja"><orthForm>knacken</orthForm></lexUnit>
<lexUnit Eigenname="nein" artificial="nein" id="vKoerperfunktion.262.ratzen" orthVar="nein"
sense="0" stilMarkierung="ja"><orthForm>ratzen</orthForm></lexUnit>
<lexUnit Eigenname="nein" artificial="nein" id="vKoerperfunktion.262.pofen" orthVar="nein"
sense="0" stilMarkierung="ja"><orthForm>pofen</orthForm></lexUnit>
<frames><frame>NN</frame><frame>NN.PP</frame><frame>NN.BM</frame></frames>
<examples><example><text>Er pennt auf der Couch.</text><frame>NN.PP</frame></example>
<example><text>Sie knackt schon.</text><frame>NN</frame></example>
<example><text>Er ratzt wie ein Stein.</text><frame>NN.BM</frame></example></examples>
</synset>
```

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE relations SYSTEM "germanet_relations.dtd">
<relations>
.....
<hyperonym sense="0" xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="extended">
<locator xlink:type="locator" xlink:href="verben.Koerperfunktion.xml#vKoerperfunktion.259"
xlink:label="vKoerperfunktion.259"/> <locator xlink:type="locator"
xlink:href="verben.Koerperfunktion.xml#vKoerperfunktion.262"
xlink:label="vKoerperfunktion.262"/> <arc xlink:type="arc" xlink:from="vKoerperfunktion.262"
xlink:to="vKoerperfunktion.259" xlink:actuate="onRequest" xlink:show="other"/> schlafen
</hyperonym>
```

## Conclusion

We have developed an XML-conformant representation of GermaNet, aiming at accessibility of our large-scale semantic database on the web. This conversion, which has been described in this paper, accounts for the various peculiarities of the German wordnet while proposing a common effort of wordnet builders to standardise XML-versions of language-specific resources.
Compatibiliy issues have been explored in comparison to the Princeton WordNet, which constitutes the pioneer resource with the broadest coverage of concepts and its inherent interlingual function. A uniform representation could be very fruitful in view of various cross-lingual language processing tasks as well as polylingual research issues.

**Acknowledgements:** Our thanks go to Iris Vogel and Holger Wunsch, our students, who performed a good deal of the programming work, to Karin Naumann and Andreas Wagner for their help with structural issues and to Erhard Hinrichs who has always been an active supporter and a competent advisor.

## References

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

DEUTSCHER WORTSCHATZ. Ein Wegweiser zum treffenden Ausdruck. Eds.: Wehrle, H. and Eggers, H. Stuttgart: Ernst Klett Verlag, 1961.

DUDEN 8: Sinn-und sachverwandte Wörter. Eds.: Drosdwoski, G. and Müller, W. and Scholze-Stubenrecht, W. and Wermke, M. Mannheim et al.: Dudenverlag, $^{2}$1986.

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

Hamp, B. and Feldweg, H. (1997) GermaNet - a Lexical-Semantic Net for German. In: Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP applications. Madrid, July 7-12, 1997.

Kunze, C. (2001) Lexikalisch-semantische Wortnetze. In R. Klabunde et al. (eds.): Computerlinguistik und Sprachtechnologie - Eine Einführung. Spektrum Verlag Heidelberg, pp. 386-393.

Manning, Christopher D. / Jansz, Kevin / Indurkhya (2001) Kirrkirr - Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary. In: Literary and Linguistic Computing 16, 2, pp. 135 - 151.
Miller, G. et al. (1990) Five papers about on WordNet. CSL-Report, Vol. 43. Cognitive Science Laboratory, Princeton University.
Vossen, P., ed. (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.
Vossen, P. (1999) EuroWordNet. Building a Multilingual Database with Lexical-Semantic Networks for the European Languages. In: Proceedings of EUROLAN'99, 4th European Summer School on Human Language Technology. Iasi, Romania. July 19-31, 1999.
Wagner, A. and Kunze, C. (1999) Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database. In: Sprache und Datenverarbeitung, SDv Vol. 23.2/1999, pp. 5-20.
Xlink 1.0.: URL=http://www.w3.org/TR/2001/REC-xlink-20010627/
XML 1.0.: URL=http://www.w3.org/TR/1998/REC-xml-19980210/