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.

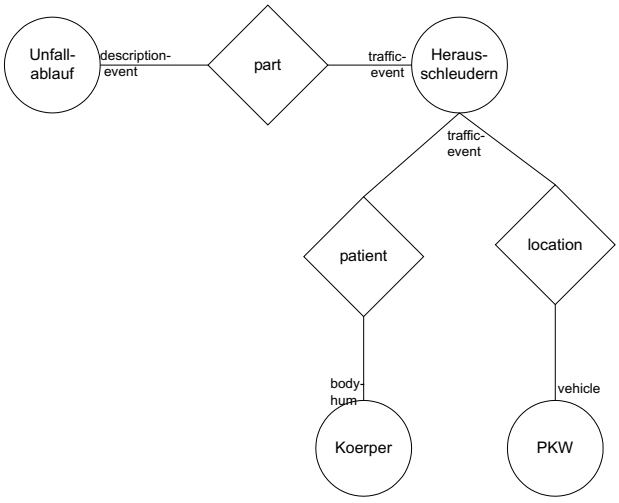

Figure 1: Results of Case Frame Analysis.

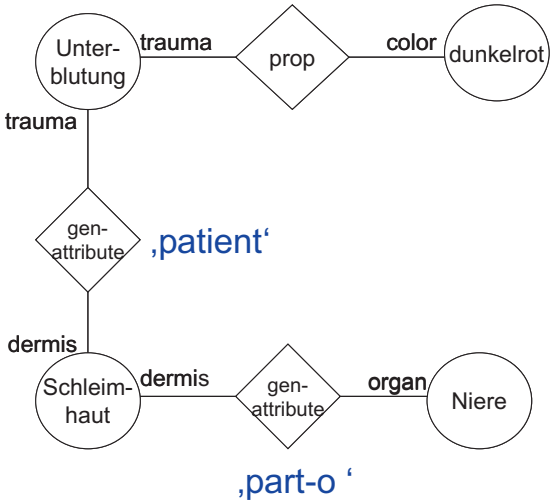

Figure 2: Results of SIsS.

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.

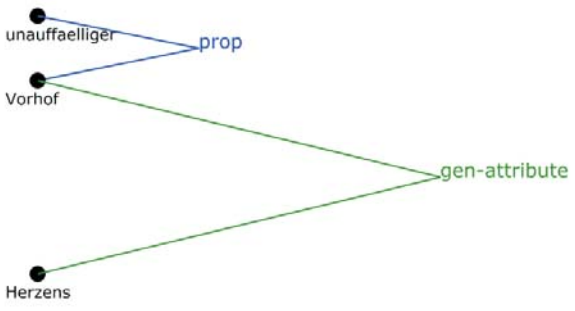

Figure 3: Result for the Phrase 'unauffaelliger Vorhof des Herzens'.

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.