# Beyond Subcategorization Acquisition – Multi-Parameter Extraction from German Text Corpora

**Kristina Spranger**
IMS, Universität Stuttgart
Azenbergstraße 12
70174 STUTTGART
GERMANY
sprangka@ims.uni-stuttgart.de

**Abstract**

In this paper, we describe a subcategorization acquisition system for German. The envisaged machine-readable lexicon is useful for both NLP tools and lexicographers. The system focuses on subcategorization extraction without being limited to this task. It also provides distributional information, selectional preferences and hints for the detection of idioms and of support-verb-constructions and other collocations. Moreover, each lexical entry is presented together with its usage contexts provided in the form of corpus examples and each subcategorization frame is presented together with its relative frequency. Thus, much additional data are given to support the lexicographer in his selection task. Furthermore, we do not only extract pairs of valency carrier and valency filler(s), but we are able to extract an almost arbitrary number of different lexicographically relevant parameters: we provide the lexicographer (and NLP tools) with quite detailed information concerning the extracted structures, such as, for example, the determiner used in the noun phrase of a verb+object collocation (definite/indefinite/possessive/null).

## 1 Motivation

We aim at constructing a large machine-readable subcategorization lexicon for German verbs and adjectives that also includes some other grammatical information, such as distributional information. It should be usable for both lexicographers and natural language processing tools.

Subcategorization information is not only important for all symbolic NLP grammars, especially for *L*exical *F*unctional *G*rammar (LFG) and *H*ead-*D*riven *P*hrase *S*tructure *G*rammar (HPSG), where it determines to a large extent the syntactic analysis of a sentence, but, as Eckle-Kohler (1999) has already shown, there is still subcategorization information missing in general dictionaries. There is, for example, so far no German dictionary available providing information concerning sentential complements subcategorized by verbs or adjectives. Another context in which subcategorization frames (and distributional information) is of interest, is the field of second language learning. Learners should be provided with the usage possibilities and contexts of a given lexeme (cf. Duden, 2001; Sommerfeldt & Schreiber, 1983).

Such a subcategorization lexicon can be acquired manually or (semi-)automatically. Since manual lexicon acquisition is very costly and time-intensive, and inevitably leads to inconsistencies, (semi-)automatic lexicon acquisition is a more promising way. Semiautomatic in this context means that the tool proposes candidates and the lexicographer selects from there.

As a knowledge source for the lexicon acquisition we use text corpora. Corpora provide us with very large amounts of data, and automatic acquisition procedures can be implemented in a generic manner. Moreover, corpora allow for determining the relative frequency of subcategorization frames and their usage contexts in the form of corpus examples.

As a prototypical example of work towards the lexicon acquisition system, we look at prenominal adjectival phrases containing an adjective or participle as head embedding one or more prepositional phrases (cf. Figure 1).

[$_{NP}$ *eine* [$_{AP}$ [$_{PP}$*für die Zukunft* $_{PP}$] *wegweisende* $_{AP}$] *Idee* $_{NP}$]
*lit.: a for the future pathbreaking idea*
*tr.: an idea pathbreaking for the future*

Figure 1: Prenominal AP embedding a PP

We opted for prenominal adjectival phrases because they represent a *secure* context for assigning the roles of valency carrier and potential valency fillers to the extracted structures (see (Kermes, 2003)). *Secure* in this context means , that such adjectival phrases represent a context where it is clear that the prepositional phrase belongs to the adjective. Prepositional phrases rather than noun phrases have been chosen because their extraction puts forth a number of interesting lexicographic and linguistic questions that have not been solved so far.

## 2 Chunked Text as a Basis for Extraction Experiments

Our extraction tool works on large German text corpora that are tokenized, part-of-speech-tagged, lemmatized, and chunked. For chunking purposes we use the *YAC*-chunker (see (Kermes & Evert, 2002)). The YAC-chunker is a fully automatic recursive chunker for unrestricted German text based on a symbolic regular expression grammar. The grammar is written in the *CQP* query language which is part of the *IMS Corpus Workbench* (see http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench). The grammar rules rely on part-of-speech and lemma annotations using the morpho-syntactic information which primarily comprises agreement information annotated using the morphological lexicon *IMSLex* (see (Lezius & Dipper & Fitschen 200)) to identify boundaries of chunks and phrases. Complex structures are built by embedding simple ones into each other using a multi-pass algorithm.
The chunker attaches feature attributes to the individual chunks and phrases: the head lemma of each chunk or phrase, such as *<ap_h wegweisend>* in Figure 2, lexical-semantic properties of the head lemma or of the chunk itself, such as, for example, temporal properties of nominal heads, and information about certain text markers (e.g. brackets or quotation marks).

[$_{NP}$ *eine* [$_{AP}$ *<ap_h wegweisend>* [$_{PP}$*für die Zukunft* $_{PP}$] *wegweisende* $_{AP}$] *Idee* $_{NP}$]

Figure 2: Prenominal AP annotated together with its lexical head

In the course of the chunking process, chunk and phrase boundaries as well as feature annotations are written back into the corpus and are thus available for query-based extraction, in the same way as lemma and part-of-speech attributes are.

Because of the additional information annotated along with the chunks and phrases - the head lemma, morpho-syntactic information, and lexical-semantic properties - YAC does not only provide a powerful basis for the extraction of subcategorization frames, but, in addition, supports the detection and extraction of selectional preferences and distributional information.

### 2.1 Extraction Method

In order to extract the relevant information, we apply queries to the corpora annotated by YAC. Complex queries or parts of queries can be stored as macros and re-used; macro calls can be nested. In Figure 3, an extraction macro for adjectival phrases like the one in Figure 1 is depicted. This macro extracts prenominal (i.e. embedded in a noun phrase (lines (1) and (6))) adjectival phrases (lines (2) and (5)), embedding one or more prepositional phrases (line (3)). The preposition lemmas must be elements of a list of 14 German prepositions (stored in the variable *$prep_sub* (cf. line (3))) that can be subcategorized by valency carriers. Since geographical adjectives (e.g. *französisch*, *afrikanisch*) are assumed to never subcategorize a prepositional phrase, a list of more than 350 geographical adjectives that is stored in the variable *$geo_adj* (cf. line (4)) is excluded from being the head of the adjectival phrase.

```
(1)  <np>
(2)          <ap>
(3)                  ([]*<pp>[]*[_.pp_h = RE($prep_sub )][]*</pp>)+
(4)                  [_.ap_h != RE($geo_adj )]
(5)          </ap>
(6)  </np>
```

Figure 3: Extraction macro for APs embedding one or more PPs

The adjectival phrases extracted by this query are postprocessed in the following way: the head of the adjectival phrase (i.e. *wegweisend* in Figure 2) is extracted together with the (potentially subcategorized) preposition (i.e. *für* in Figure 2). Moreover, the head of the noun phrase embedded in the prepositional phrase is extracted (i.e. *Zukunft* in sentence Figure 2) and morphologically analyzed (see (Schulte im Walde, 2003)). In the case of deverbal adjectival heads, the participle is matched onto the respective verb and the extracted subcategorization information is considered as belonging to the verb.

The adjective+preposition-pairs are sorted by cooccurrence frequency. The nouns occurring along with these pairs are assigned to the respective adjective+preposition-pair. For each adjective+preposition-pair, the absolute occurrence frequency and the number of different nominal heads that occur with this pair is calculated; the different nominal heads are listed together with their occurrence frequencies.

The extracted subcategorization information is compared to the information provided by (Eckle-Kohler, 1999)'s computational subcategorization lexicon that represents up to now the subcategorization information available at IMS. Eckle-Kohler's lexicon comprises 16,621 verbs and 2,399 adjectives.

We carried out our first extraction experiments on a German newspaper corpus of about 36 million words. We extracted 10,283 constructions of the above mentioned type whereby 1,884 are not covered by Eckle-Kohler. In Table 1 an excerpt of our extraction results is given (known geographical proper nouns are reduced to the class label *GEO*).

| *Adj/verb+prep* | *head nouns: frequency* | *Absolute frequency* | *number of different heads* |
|---|---|---|---|
| Befindlich in | Besitz:58, Bau:57, Aufbau:29 | 331 | 85 |
| Geraten in | Zwielicht:19, Not:17, Bedrängnis:14 | 198 | 56 |
| Tätig in | GEO:45, Bereich:14, Bau:10 | 186 | 69 |
| Einzigartig in | GEO:11, Art:5, Geschichte:3 | 33 | 12 |
| Verwickelt in | Skandal:7, Unfall:3, Konflikt:3 | 32 | 6 |

Table 1: Excerpt of the extracted results

## 2.2 Discussion

The extracted results are not a list of adjectives or verbs and their subcategorized prepositions, but they rather represent different kinds of information that have to be analyzed and further subclassified: we get instances of adjectives and verbs and their potentially subcategorized prepositions. In this context, the distinction between adjuncts and arguments is the most urgent question. We will investigate to what extent the extracted nominal heads can help automatize the distinction. To this end, we will apply heuristics. An example for such a heuristic is that German adjuncts such as *im Prinzip* (*in principle*) or *im Grunde* (*in the main*) probably occur with almost all verbs and adjectives with roughly the same frequency whereas truly subcategorized PPs are much more selective.

Moreover, the results lead to the detection of idioms, and support verb constructions and other collocations. There are verb+preposition-pairs that occur only with a very small number of different heads: the verb *treten* (*to enter*) together with *in* has an absolute cooccurrence frequency of 111, but it only appears with 8 different heads. In 92 cases *treten+in* appears together with the noun *Kraft* (*force*). And, actually, *in Kraft treten* is a German support-verb-construction with the meaning of *to take effect.* In contrast, the verb+preposition-pair *erinnern+an* (*to remind of*) has an absolute cooccurrence frequency of 56, but it appears together with 51 different nouns.

So, a next step should be to test how secure hints such as the number of different nominal heads are in order to automatically detect, filter out and collect collocational constructions and idioms.

Furthermore, some selectional preferences can be observed. The (lexicalized) participle *verwickelt* (*to be mixed up*), for example, occurs 32 times with the preposition *in.*

The preposition introduces noun phrases with 6 different heads: *Skandal* (*scandal*), *Unfall* (*accident*), *Konflikt* (*conflict*), *Affäre* (*affair*), *Betrügerei* (*swindle*) and *Mauschelei* (*underhand dealings*). All these nouns seem to belong to the same semantic field, which leads to the assumption that *verwickelt in* requires a certain semantic type of noun.

Our next step will be to thoroughly inspect the results we have extracted so far and to extract candidates from larger text corpora (up to 300 million words) in order to confirm or reject our hypotheses.

## 3 Conclusions

As the work presented in this paper is still ongoing, we have rather presented an extensive collection of different phenomena and problems we have faced when applying the extraction queries than tried to solve problems concerned with the extraction of subcategorization information. But, even the first steps have shown that we are able to face and deal with problems that go beyond mere subcategorization extraction (see (Klotz, 2000)).

The next steps are to investigate more different parameters and present more detailed hints towards the detection of idioms and collocations. In the same way, we want to provide more details concerning the question to what extent the distinction between adjunct prepositional phrases and complements can be automatized or, at least, to what extent manual checking, and hence the lexicographer's work, can be supported by our system.

## References

**Dudenredaktion Mannheim** (ed.) 2001. *DUDEN – Das Stilwörterbuch.* Mannheim: Dudenverlag.

**Eckle-Kohler, J.** 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora.* Berlin: Logos Verlag.

**Kermes, H.** 2003. *Off-line and (On-line) Text Analysis for Computational Lexicography.* PhD thesis, Stuttgart: IMS, Universität Stuttgart.

**Kermes, H. and Evert, S.** 2002.*YAC – A Recursive Chunker for Unrestricted German Text.* Las Palmas: Proceedings of LREC'02.

**Klotz, M.** 2000. *Grammatik und Lexik.* Tübingen: Stauffenberg Verlag.

**Lezius, W. and Dipper, S. and Fitschen, A.** 2000. *IMSLex – representing morphological and syntactical information in a relational database.* Stuttgart: Proceedings of the EURALEX'00.

**Schulte im Walde, S.** 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes.* PhD thesis, Stuttgart: IMS, Unversität Stuttgart.

**Sommerfeldt, K.-E. and Schreiber, H.** 1983. *Wörterbuch zur Valenz und Distribution deutscher Adjektive.* Tübingen: Max Niemeyer Verlag.