# Building a Resource for Lexical Semantics

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal
Dept. of Computational Linguistics
Saarland University, Saarbrücken, Germany
{erk,kowalski,pado,pinkal}@coli.uni-sb.de
fax +49 681 302 4351

The development of computational linguistics through the last decade has provided abundant evidence how grammar research can benefit from corpus-based methods. Computational linguistics could obviously take similar advantage from corpora on the level of semantics. However, semantic corpus annotation is currently just in its initial stages, comprising almost exclusively word sense annotation (an exception being the Prague TreeBank for Czech [4]).

We present the SCORE project, the aim of which is to create a large semantically annotated corpus and to investigate methods for its utilization. In a first step, we annotate a German 1.5 million word corpus by hand exhaustively with frame semantic roles. Additionally we will selectively annotate word senses and anaphoric links. For the semantic role annotation, we use the FrameNet [1] database of frames, extending it to a light version of a German FrameNet. In the next step, we will train statistical systems on the annotated corpus to further extend the corpus (semi)-automatically. Similar tools already exist, for the FrameNet paradigm [3] as well as the Prague Treebank [6].

The SCORE corpus can be used in a number of interesting ways, e.g. for the automatic acquisition of lexical semantic information, the training of statistical parsers on a combination of syntactic and semantic role information and the improvement of linguistically guided techniques for information access and extraction.

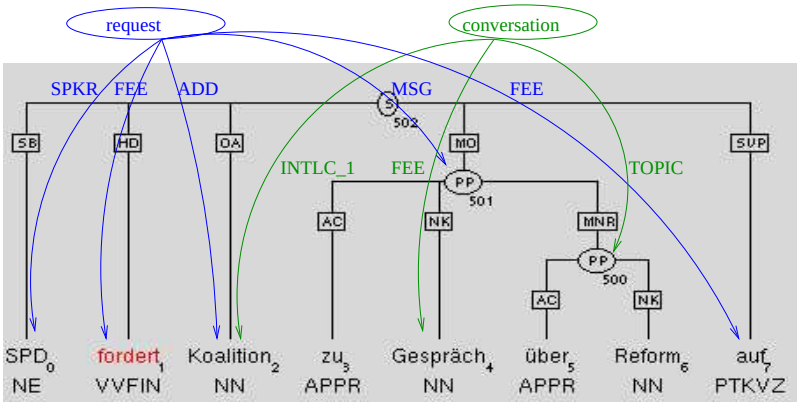


Figure 1: Frame tree and syntactic structure

**The Annotation Scheme.** As a basis for the semantic annotation in SCORE we use the TIGER corpus [2], a German newspaper corpus annotated for syntactic structure. In this corpus we tag all frame evoking elements with their appropriate frames, and specify their frame elements. In the annotation we represent frame structures as flat trees of depth 1. The root node of a frame tree is labelled by the frame name. The edges are labeled by abbreviated frame

element names or as (parts of) the frame evoking element (FEE). The terminal nodes of the frame trees are sets of nodes of the syntactic annotation trees. Figure 1 shows an example from the TIGER corpus: the tree with straight edges describes the syntactic structure, and the two trees with arched edges describe the frames REQUEST and CONVERSATION introduced by the verb *fordert auf* (demand) with separable verb prefix, and the noun *Gespräch* (conversation), respectively.

The decision to use flat trees allows us to keep frame annotation as close as possible to the syntactic structure, and thus to build on decisions that have been made in the syntactic annotation. Additionally the annotation scheme lets us organize the annotation in a more modular and flexible way, as the annotation of one frame is never dependent on the previous annotation of another frame. Furthermore, the availability of small units of semantic information locally related to syntactic units is crucial to the extraction of lexical semantic information and the training of analyzers. Whenever the overall semantic structure is relevant, it can be easily recovered from the flat representation.

**Additional Frames.** In an exhaustive corpus annotation with frame semantic roles we will need many frames not yet covered by the FrameNet database. In these cases we will construct suitable "proto-frames" on the basis of the available corpus examples.

**Underspecification.** Apart from "real" disagreement, or annotation mistakes, disagreement between two annotators may be due to the fact that more than one tag applies at the same time, or that several tags seem equally possible judging from the available context, or that the distinction between two tags is systematically vague or unclear.

In cases like that, we allow the annotators to assign more than one tag, leaving the decision between different readings open. A similar design decision for *underspecified* tags has been made e.g. in the manual word sense annotation done for SENSEVAL [5].

In SCORE, we allow underspecification on the level of frames and on the level of frame elements. On the level of frames, it may be left open which of two frames a word introduces. For example the verb *verlangen* can introduce the frame REQUEST, but in the pilot study that we describe below we found that for some corpus examples it is hard to decide between REQUEST and TRANSACTION_COMMERCE:

(1) Gleichwohl versuchen offenbar Assekuranzen, [das Gesetz] zu umgehen, indem sie von Nichtdeutschen mehr Geld verlangen.

Nonetheless insurance companies obviously try to circumvent [the law] by asking/demanding more money from non-Germans.

On the level of frame elements, the annotators may leave it open whether or not a single tag applies, which of two tags applies, and how many words a frame element encompasses. For example, in the pilot study that we describe below we often found it problematic to assign the MEDIUM (for which an uncontroversial example would be *in a letter*) in the REQUEST frame: is *in der Fortsetzung der Diskussion in Computerdenken (in the sequel of the discussion in Computer Thought)* a MEDIUM or not?

Allowing for "underspecified tags" can be useful in several respects. During annotation, it avoids (sometimes dubious) decisions for a unique tag. Furthermore it may be valuable to know that annotators systematically found

| annotators | A-B | A-C | B-C |
|---|---|---|---|
| agreement on frame | 97% | 97% | 97% |
| agreement on frame elements (given same frame) | 80% | 78% | 80% |
| complete agreement | 78% | 76% | 78% |

Table 1: Pilot study: agreement between pairs of annotators

it hard to distinguish between frames A and B. It may allow us to find find relations between frames that go beyond an inheritance hierarchy, horizontal rather than vertical connections, e.g. between the frames REQUEST and TRANSACTION_COMMERCE and their respective frame elements.

**Results of a Pilot Study.** In a first pilot study we examined lemmas that may introduce the REQUEST frame, 3 verbs (*auffordern*, *fordern*, *zurückfordern*), one noun (*Forderung*), and compound nouns ending in *-forderung*. We analyzed all 441 instances of these lemmas present in the TIGER corpus. Each sentence was tagged by three different annotators. Additionally all 118 instances of another REQUEST verb (*verlangen*) were tagged by two of the annotators. Annotation was done using an XML formulation of the flat frame trees described above; the syntactic structure assigned to the sentences could only be viewed in a separate viewer. (A graphical tool that makes the syntactic structure directly available during frame annotation will be available from early 2003.)

Table 1 shows the agreement between pairs of annotators: pairs of annotators agreed in the frame they assigned in 97% of the corpus instances. Of these, they also agreed in the assignment of frame elements in 78% to 80%. So we found complete agreement in between 76% and 78% of the corpus instances, an encouraging result, given the strict definition of agreement – annotators had to assign the same frame, and determine all frame elements of the frame exactly in the same way –, the unavailability of the syntactic structure, and the lack of a strict annotation guideline at that time.

# References


[1] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada, 1998.

[2] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.

[3] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[4] E. Hajičová. Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In *Proceedings of TSD'98*, pages 45–50, Brno, Czech Republic, 1998.

[5] A. Kilgarriff and J. Rosenzweig. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2), 2000.

[6] Z. Žabokrtský. Automatic functor assignment in the Prague Dependency Treebank. In *Proceedings of TSD'00*, 2000.