Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation

Katrin Erk and Andrea Kowalski and Sebastian Padó and Manfred Pinkal
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{erk, kowalski, pado, pinkal}@coli.uni-sb.de

Abstract

We describe the ongoing construction of a large, semantically annotated corpus resource as reliable basis for the large-scale acquisition of word-semantic information, e.g. the construction of domain-independent lexica. The backbone of the annotation are semantic roles in the frame semantics paradigm. We report experiences and evaluate the annotated data from the first project stage. On this basis, we discuss the problems of vagueness and ambiguity in semantic annotation.

1 Introduction

Corpus-based methods for syntactic learning and processing are well-established in computational linguistics. There are comprehensive and carefully worked-out corpus resources available for a number of languages, e.g. the Penn Treebank (Marcus et al., 1994) for English or the NEGRA corpus (Skut et al., 1998) for German. In semantics, the situation is different: Semantic corpus annotation is only in its initial stages, and currently only a few, mostly small, corpora are available. Semantic annotation has predominantly concentrated on word senses, e.g. in the SENSEVAL initiative (Kilgarriff, 2001), a notable exception being the Prague Treebank (Hajičová, 1998) . As a consequence, most recent work in corpus-based semantics has taken an unsupervised approach, relying on statistical methods to extract semantic regularities from raw corpora, often using information from ontologies like WordNet (Miller et al., 1990).

Meanwhile, the lack of large, domain-independent lexica providing word-semantic information is one of the most serious bottlenecks for language technology. To train tools for the acquisition of semantic information for such lexica, large, extensively annotated resources are necessary.

In this paper, we present current work of the SALSA (SAarbrücken Lexical Semantics Annotation and analysis) project, whose aim is to provide such a resource and to investigate efficient methods for its utilisation. In the current project phase, the focus of our research and the backbone of the annotation are semantic role relations. More specifically, our role annotation is based on the Berkeley FrameNet project (Baker et al., 1998; Johnson et al., 2002). In addition, we selectively annotate word senses and anaphoric links. The TIGER corpus (Brants et al., 2002), a 1.5M word German newspaper corpus, serves as sound syntactic basis.

Besides the sparse data problem, the most serious problem for corpus-based lexical semantics is the lack of specificity of the data: Word meaning is notoriously ambiguous, vague, and subject to contextual variance. The problem has been recognised and discussed in connection with the SENSEVAL task (Kilgarriff and Rosenzweig, 2000). Annotation of frame semantic roles compounds the problem as it combines word sense assignment with the assignment of semantic roles, a task that introduces vagueness and ambiguity problems of its own.

The problem can be alleviated by choosing a suitable resource as annotation basis. FrameNet roles, which are local to particular *frames* (abstract situations), may be better suited for the annotation task than the "classical" thematic roles concept with a small, universal and exhaustive set of roles like *agent, patient, theme*: The exact extension of the role concepts has never been agreed upon (Fillmore, 1968). Furthermore, the more concrete frame semantic roles may make the annotators' task easier. The FrameNet database itself, however, cannot be taken as evidence that reliable annotation is possible: The aim of the FrameNet project is essentially lexicographic and its annotation not exhaustive; it comprises representative examples for the use of each frame and its frame elements in the BNC.

While the vagueness and ambiguity problem may be mitigated by the using of a "good" resource, it will not disappear entirely, and an annotation format is needed that can cope with the inherent vagueness of word sense and semantic role assignment.

Plan of the paper. In Section 2 we briefly introduce FrameNet and the TIGER corpus that we use as a basis for semantic annotation. Section 3 gives an overview of the aims of the SALSA project, and Section 4 describes the annotation with frame semantic roles. Section 5 evaluates the first annotation results and the suitability of FrameNet as an annotation resource, and Section 6 discusses the effects of vagueness and ambiguity on frame semantic role annotation. Although the current amount of annotated data does not allow for definitive judgements, we can discuss tendencies.

2 Resources

SALSA currently extends the TIGER corpus by semantic role annotation, using FrameNet as a resource. In the following, we will give a short overview of both resources.

FrameNet. The FrameNet project (Johnson et al., 2002) is based on Fillmore's Frame Semantics. A *frame* is a conceptual structure that describes a situation. It is introduced by a *target* or *frame-evoking element* (*FEE*). The roles, called *frame elements* (FEs), are local to particular frames and are the participants and props of the described situations.

The aim of FrameNet is to provide a comprehensive frame-semantic description of the core lexicon of English. A *database of frames* contains the frames' basic conceptual structure, and names and descriptions for the available frame elements. A *lexicon database* associates lemmas with the frames they evoke, lists possible syntactic realizations of FEs and provides annotated examples from the BNC. The current on-line version of the frame database (Johnson et al., 2002) consists of almost 400 frames, and covers about 6,900 lexical entries.

Frame: REQUEST	
FE	**Example**
SPEAKER	**Pat** *urged* me to apply for the job.
ADDRESSEE	Pat *urged* **me** to apply for the job.
MESSAGE	Pat *urged* me **to apply for the job**.
TOPIC	Kim made a *request* **about changing the title**.
MEDIUM	Kim made a *request* **in her letter**.
Frame: COMMERCIAL_TRANSACTION (C_T)	
BUYER	**Jess** *bought* a coat.
GOODS	Jess *bought* **a coat**.
SELLER	**Kim** *sold* the sweater.
MONEY	Kim *paid* **14 dollars** for the ticket.
PURPOSE	Kim *bought* peppers **to cook them**.
REASON	Bob *bought* peppers **because he was hungry**.

Figure 1: Example frame descriptions.

Figure 1 shows two frames. The frame REQUEST involves a FE SPEAKER who voices the request, an ADDRESSEE who is asked to do something, the MESSAGE, the request that is made, the TOPIC that the request is about, and the MEDIUM that is used to convey the request. Among the FEEs for this frame are the verb *ask* and the noun *request*. In the frame COMMERCIAL_TRANSACTION (henceforth C_T), a BUYER gives MONEY to a SELLER and receives GOODS in exchange. This frame is evoked e.g. by the verb *pay* and the noun *money*.

The TIGER Corpus. We are using the TIGER Corpus (Brants et al., 2002), a manually syntactically annotated German corpus, as a basis for our annotation. It is the largest available such corpus (80,000 sentences in its final release compared to 20,000 sentences in its predecessor NEGRA) and uses a rich annotation format. The annotation scheme is surface oriented and comparably theory-neutral. Individual words are labelled with POS information. The syntactic structures of sentences are described by relatively flat trees providing information about *grammatical functions* (on edge labels), *syntactic categories* (on node labels), and *argument structure of syntactic heads* (through the use of dependency-oriented constituent structures, which are close to the syntactic surface). An example for a syntactic structure is given in Figure 2.

3 Project overview

The aim of the SALSA project is to construct a large semantically annotated corpus and to provide methods for its utilisation.

Corpus construction. In the first phase of the project, we annotate the TIGER corpus in part man-

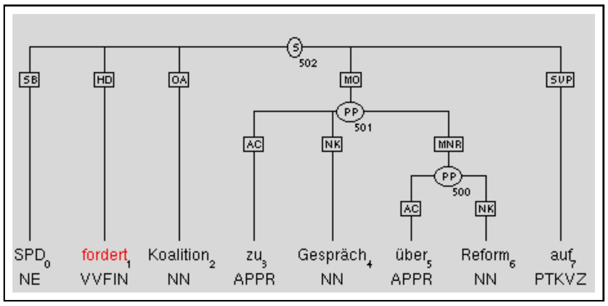

Figure 2: A sentence and its syntactic structure.

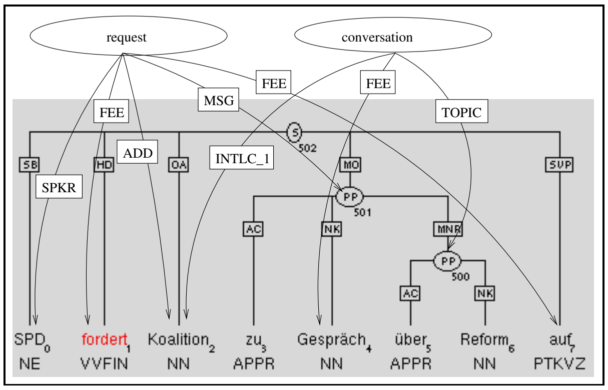

Figure 3: Frame annotation.

ually, in part semi-automatically, having tools propose tags which are verified by human annotators. In the second phase, we will extend these tools for the weakly supervised annotation of a much larger corpus, using the TIGER corpus as training data.

Utilisation. The SALSA corpus is designed to be utilisable for many purposes, like improving statistical parsers, and extending methods for information extraction and access. The focus in the SALSA project itself is on lexical semantics, and our first use of the corpus will be to extract selectional preferences for frame elements.

The SALSA corpus will be tagged with the following types of semantic information:

FrameNet frames. We tag all FEEs that occur in the corpus with their appropriate frames, and specify their frame elements. Thus, our focus is different from the lexicographic orientation of the FrameNet project mentioned above. As we tag all corpus instances of each FEE, we expect to encounter a wider range of phenomena. which Currently, FrameNet only exists for English and is still under development. We will produce a "light version" of a FrameNet for German as a by-product of the annotation, reusing as many as possible of the semantic frame descriptions from the English FrameNet database. Our first results indicate that the frame structure assumed for the description of the English lexicon can be reused for German, with minor changes and extensions.

Word sense. The additional value of word sense disambiguation in a corpus is obvious. However, exhaustive word sense annotation is a highly time-consuming task. Therefore we decided for a selective annotation policy, annotating only the heads of frame elements. GermaNet, the German WordNet version, will be used as a basis for the annotation.

Coreference. Similarly, we will selectively annotate coreference. If a lexical head of a frame element is an anaphor, we specify the antecedent to make the meaning of the frame element accessible.

4 Frame Annotation

Annotation schema. To give a first impression of frame annotation, we turn to the sentence in Fig. 2:

(1) SPD fordert Koalition zu Gespräch über Reform auf.
(SPD requests that coalition talk about reform.)

Fig. 3 shows the frame annotation associated with (1). Frames are drawn as flat trees. The root node is labelled with the frame name. The edges are labelled with abbreviated FE names, like SPKR for SPEAKER, plus the tag FEE for the frame-evoking element. The terminal nodes of the frame trees are always nodes of the syntactic tree. Cases where a semantic unit (FE or FEE) does not form one syntactic constituent, like *fordert ... auf* in the example, are represented by assignment of the same label to several edges.

Sentence (1), a newspaper headline, contains at least two FEEs: *auffordern* and *Gespräch*. *auffordern* belongs to the frame REQUEST (see Fig. 1). In our example the SPEAKER is the subject NP *SPD*, the ADDRESSEE is the direct object NP *Koalition*, and the MESSAGE is the complex PP *zu Gespräch über Reform*. So far, the frame structure follows the syntactic structure, except for that fact that the FEE, as a separable prefix verb, is realized by two syntactic nodes. However, it is not always the case that frame structure parallels syntactic structure. The second FEE *Gespräch* introduces the frame CONVERSATION. In this frame two (or more) groups talk to one another and no participant is construed as only a SPEAKER or only an ADDRESSEE. In our example the only NP-internal frame element is the TOPIC ("what the message is about") *über Reform*, whereas the INTERLOCUTOR-1 ("the prominent participant in the conversation") is realized by the direct object of *auffordern*.

As shown in Fig. 3, frames are annotated as trees of depth one. Although it might seem semantically more adequate to admit deeper frame trees, e.g. to allow the MSG edge of the REQUEST frame in Fig. 3 to be the root node of the CONVERSATION tree, as its "real" semantic argument, the representation of frame structure in terms of flat and independent semantic trees seems to be preferable for a number of practical reasons: It makes the annotation process more modular and flexible – this way, no frame annotation relies on previous frame annotation. The closeness to the syntactic structure makes the annotators' task easier. Finally, it facilitates statistical evaluation by providing small units of semantic information that are locally related to syntax.

Difficult cases. Because frame elements may span more than one sentence, like in the case of direct speech, we cannot restrict ourselves to annotation at sentence level. Also, compound nouns require annotation below word level. For example, the word "Gagenforderung" (demand for wages) consists of "-forderung" (demand), which invokes the frame REQUEST, and a MESSAGE element "Gagen-". Another interesting point is that one word may introduce more than one frame in cases of coordination and ellipsis. An example is shown in (2). In the elliptical clause *only one fifth for daughters*, the elided *bought* introduces a C_T frame. So we let the *bought* in the antecedent introduce two frames, one for the antecedent and one for the ellipsis.

(2) Ein Viertel aller Spielwaren würden für Söhne erworben, nur ein Fünftel für Töchter.
(One quarter of all toys are bought for sons, only one fifth for daughters.)

Annotation process. Frame annotation proceeds one frame-evoking lemma at a time, using subcorpora containing all instances of the lemma with some surrounding context. Since most FEEs are polysemous, there will usually be several frames relevant to a subcorpus. Annotators first select a frame for an instance of the target lemma. Then they assign frame elements.

At the moment the annotation uses XML tags on bare text. The syntactic structure of the TIGER-sentences can be accessed in a separate viewer. An annotation tool is being implemented that will provide a graphical interface for the annotation. It will display the syntactic structure and allow for a graphical manipulation of semantic frame trees, in a similar way as shown in Fig. 3.

Extending FrameNet. Since FrameNet is far from being complete, there are many word senses not yet covered. For example the verb *fordern*, which belongs to the REQUEST frame, additionally has the reading *challenge*, for which the current version of FrameNet does not supply a frame.

5 Evaluation of Annotated Data

Materials. Compared to the pilot study we previously reported (Erk et al., 2003), in which 3 annotators tagged 440 corpus instances of a single frame, resulting in 1,320 annotation instances, we now dispose of a considerably larger body of data. It consists of 703 corpus instances for the two frames shown in Figure 1, making up a total of 4,653 annotation instances. For the frame REQUEST, we obtained 421 instances with 8-fold and 114 with 7-fold annotation. The annotated lemmas comprise *auffordern* (*to request*), *fordern*, *verlangen* (*to demand*), *zurückfordern* (*demand back*), the noun *Forderung* (*demand*), and compound nouns ending with *-forderung*. For the frame C_T we have 30, 40 and 98 instances with 5-, 3-, and 2-fold annotation respectively. The annotated lemmas are *kaufen* (*to buy*), *erwerben* (*to acquire*), verbrauchen (*to consume*), and *verkaufen* (*to sell*).

Note that the corpora we are evaluating do not constitute a random sample: At the moment, we cover only two frames, and REQUEST seems to be relatively easy to annotate. Also, the annotation results may not be entirely predictive for larger sample sizes: While the annotation guidelines were being developed, we used REQUEST as a "calibration" frame to be annotated by everybody. As a result, in some cases reliability may be too low because detailed guidelines were not available, and in others it may be too high because controversial instances were discussed in project meetings.

Results. The results in this section refer solely to the assignment of fully specified frames and frame elements. Underspecification is discussed at length

frames	average	best	worst
REQUEST	96.83%	100%	90.73%
COMM.	97.11%	98.96%	88.71%

elements	average	best	worst
REQUEST	88.86%	95.69%	66.57%
COMM.	74.25%	90.30%	69.33%

Table 1: Inter-annotator agreement on frames (top) and frame elements (below).

in Section 6. Due to the limited space in this paper, we only address the question of *inter-annotator agreement* or *annotation reliability*, since a reliable annotation is necessary for all further corpus uses.

Table 1 shows the inter-annotator agreement on frame assignment and on frame element assignment, computed for pairs of annotators. The "average" column shows the total agreement for all annotation instances, while "best" and "worst" show the figures for the (lemma-specific) subcorpora with highest and lowest agreement, respectively. The upper half of the table shows agreement on the assignment of frames to FEEs, for which we performed 14,410 pairwise comparisons, and the lower half shows agreement on assigned frame elements (29,889 pairwise comparisons). Agreement on frame elements is "exact match": both annotators have to tag exactly the same sequence of words. In sum, we found that annotators agreed very well on frames. Disagreement on frame elements was higher, in the range of 12-25%. Generally, the numbers indicated considerable differences between the subcorpora.

To investigate this matter further, we computed the Alpha statistic (Krippendorff, 1980) for our annotation. Like the widely used Kappa, α is a chance-corrected measure of reliability. It is defined as

$$\alpha = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}}$$

We chose Alpha over Kappa because it also indicates unreliabilities due to unequal coder preference for categories. With an α value of 1 signifying total agreement and 0 chance agreement, α values above 0.8 are usually interpreted as reliable annotation.

Figure 4 shows single category reliabilities for the assignment of frame elements. The graphs shows that not only did target lemmas vary in their difficulty, but that reliability of frame element assignment was also a matter of high variation. Firstly, frames introduced by nouns (*Forderung* and *-forderung*) were more difficult to annotate than verbs. Secondly, frame elements could be assigned to three groups: frame elements which were always annotated reliably, those whose reliability was highly dependent on the FEE, and the third group whose members were impossible to annotate reliably (these are not shown in the graphs). In the REQUEST frames, SPEAKER, MESSAGE and ADDRESSEE belong to the first group, at least for verbal FEEs. MEDIUM is a member of the second group, and TOPIC was annotated at chance level ($\alpha \approx 0$). In the COMMERCE frame, only BUYER and GOODS always show high reliability. SELLER can only be reliably annotated for the target *verkaufen*. PURPOSE and REASON fall into the third group.

5.1 Discussion

Interpretation of the data. Inter-annotator agreement on the frames shown in Table 1 is very high. However, the lemmas we considered so far were only moderately ambiguous, and we might see lower figures for frame agreement for highly polysemous FEEs like *laufen* (to run).

For frame elements, inter-annotator agreement is not that high. Can we expect improvement? The Prague Treebank reported a disagreement of about 10% for manual thematic role assignment (Žabokrtský, 2000). However, in contrast to our study, they also annotated temporal and local modifiers, which are easier to mark than other roles.

One factor that may improve frame element agreement in the future is the display of syntactic structure directly in the annotation tool. Annotators were instructed to assign each frame element to a single syntactic constituent whenever possible, but could only access syntactic structure in a separate viewer. We found that in 35% of pairwise frame element disagreements, one annotator assigned a single syntactic constituent and the other did not. Since a total of 95.6% of frame elements were assigned to single constituents, we expect an increase in agreement when a dedicated annotation tool is available.

As to the pronounced differences in reliability between frame elements, we found that while most central frame elements like SPEAKER or BUYER were easy to identify, annotators found it harder to agree on less frequent frame elements like MEDIUM, PURPOSE and REASON. The latter two with their

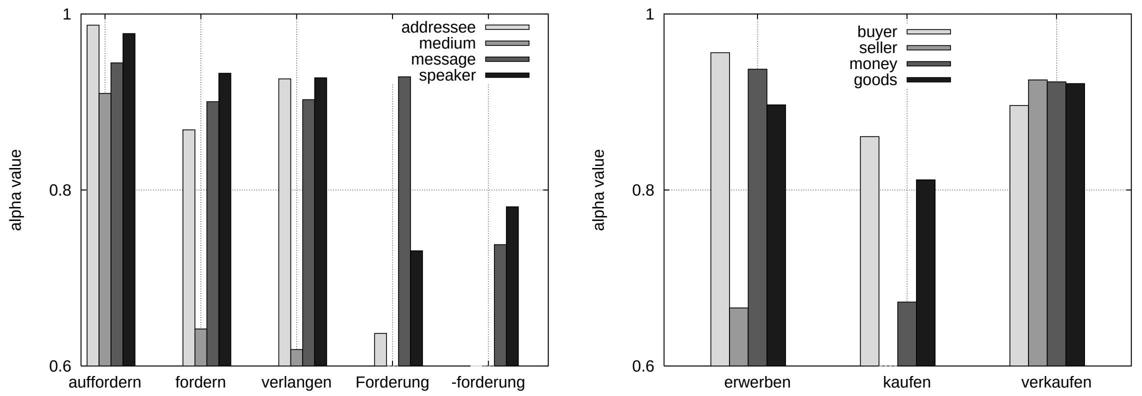

Figure 4: Alpha values for frame elements. Left: REQUEST. Right: COMMERCIAL_TRANSACTION.

particularly low agreement ($\alpha < 0.8$) contribute towards the low overall inter-annotator agreement of the C_T frame. We suspect that annotators saw too few instances of these elements to build up a reliable intuition. However, the elements may also be inherently difficult to distinguish.

How can we interpret the differences in frame element agreement across target lemmas, especially between verb and noun targets? While frame elements for verbal targets are usually easy to identify based on syntactic factors, this is not the case for nouns. Figure 3 shows an example: Should *SPD* be tagged as INTERLOCUTOR-2 in the CONVERSATION frame? This appears to be a question of pragmatics. Here it seems that clearer annotation guidelines would be desirable.

FrameNet as a resource for semantic role annotation. Above, we have asked about the suitability of FrameNet for semantic role annotation, and our data allow a first, though tentative, assessment.

Concerning the portability of FrameNet to other languages than English, the English frames worked well for the German lemmas we have seen so far. For C_T a number of frame elements seem to be missing, but these are not language-specific, like CREDIT (for *on commission* and *in installments*).

The FrameNet frame database is not yet complete. How often do annotators encounter missing frames? The frame UNKNOWN was assigned in 6.3% of the instances of REQUEST, and in 17.6% of the C_T instances. The last figure is due to the overwhelming number of UNKNOWN cases in *verbrauchen*, for which the main sense we encountered is "to use up a resource", which FrameNet does not offer.

Is the choice of frame always clear? And can frame elements always be assigned unambiguously? Above we have already seen that frame element assignment is problematic for nouns. In the next section we will discuss problematic cases of frame assignment as well as frame element assignment.

6 Vagueness, Ambiguity and Underspecification

Annotation Challenges. It is a well-known problem from word sense annotation that it is often impossible to make a safe choice among the set of possible semantic correlates for a linguistic item. In frame annotation, this problem appears on two levels: The choice of a frame for a target is a choice of word sense. The assignment of frame elements to phrases poses a second disambiguation problem.

An example of the first problem is the German verb *verlangen*, which associates with both the frame REQUEST and the frame C_T. We found several cases where both readings seem to be equally present, e.g. sentence (3). Sentences (4) and (5) exemplify the second problem. The italicised phrase in (4) may be either a SPEAKER or a MEDIUM and the one in (5) either a MEDIUM or not a frame element at all. In our exhaustive annotation, these problems are much more virulent than in the FrameNet corpus, which consists mostly of prototypical examples.

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

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

(4) Die nachhaltigste Korrektur der Programmatik fordert *ein Antrag...*

(The most fundamental policy correction is requested by *a motion...*)

(5) Der Parteitag billigte *ein Wirtschaftskonzept*, in dem der Umbau gefordert wird.

(The party congress approved of *an economic concept* in which a change is demanded.)

Following Kilgarriff and Rosenzweig (2000), we distinguish three cases where the assignment of a single semantic tag is problematic: (1), cases in which, judging from the available context information, several tags are equally possible for an ambiguous utterance; (2), cases in which more than one tag applies at the same time, because the sense distinction is neutralised in the context; and (3), cases in which the distinction between two tags is systematically vague or unclear.

In SALSA, we use the concept of *underspecification* to handle all three cases: Annotators may assign *underspecified frame and frame element tags*. While the cases have different semantic-pragmatic status, we tag all three of them as underspecified. This is in accordance with the general view on underspecification in semantic theory (Pinkal, 1996). Furthermore, Kilgarriff and Rosenzweig (2000) argue that it is impossible to distinguish those cases

Allowing *underspecified tags* has several advantages. First, it avoids (sometimes dubious) decisions for a unique tag during annotation. Second, it is useful to know if annotators systematically found it hard to distinguish between two frames or two frame elements. This diagnostic information can be used for improving the annotation scheme (e.g. by removing vague distinctions). Third, underspecified tags may indicate frame relations beyond an inheritance hierarchy, horizontal rather than vertical connections. In (3), the use of underspecification can indicate that the frames REQUEST and C_T are used in the same situation, which in turn can serve to infer relations between their respective frame elements.

Evaluating underspecified annotation. In the previous section, we disregarded annotation cases involving underspecification. In order to evaluate underspecified tags, we present a method of computing inter-annotator agreement in the presence of underspecified annotations. Representing frames and frame elements as predicates that each take a sequence of word indices as their argument, a frame annotation can be seen as a pair (CF, CE) of two formulae, describing the frame and the frame elements, respectively. Without underspecification, CF is a single predicate and CE is a conjunction of predicates. For the CONVERSATION frame of sentence (1), CF has the form CONVERSATION(Gespräch)[1], and CE is INTLC_1(Koalition) $\wedge$ TOPIC(über Reform). Underspecification is expressed by conjuncts that are disjunctions instead of single predicates. Table 2 shows the admissible cases. For example, the CE of (4) contains the conjunct SPKR(ein Antrag) $\vee$ MEDIUM(ein Antrag). Our annotation scheme guarantees that every FE name appears in *at most one* conjunct of CE. *Exact* agreement means that every conjunct of annotator A must correspond to a conjunct by annotator B, and vice versa. For *partial* agreement, it suffices that for each conjunct of A, one disjunct matches a disjunct in a conjunct of B, and conversely.

frame annotation	
F(t)	single frame: F is assigned to t
$(\mathrm{F}_1(t) \vee \mathrm{F}_2(t))$	frame disjunction: F_1 or F_2 is assigned to t

frame element annotation	
E(s)	single frame element: E is assigned to s
$(\mathrm{E}_1(s) \vee \mathrm{E}_2(s))$	frame element disjunction: E_1 or E_2 is assigned to s
$(\mathrm{E}(s) \vee \mathrm{NOFE}(s))$	optional element: E_1 or no frame element is assigned to s
$(\mathrm{E}(s) \vee \mathrm{E}(s_1 s s_2))$	underspecified length: frame element E is assigned to s or the longer sequence $s_1 s s_2$, which includes s

Table 2: Types of conjuncts. F is a frame name, E a frame element name, and t and s are sequences of word indices (t is for the target (FEE))

Using this measure of partial agreement, we now evaluate underspecified annotation. The most striking result is that annotators made little use of underspecification. Frame underspecification was used in 0.4% of all frames, and frame element underspecification for 0.9% of all frame elements. The frame element MEDIUM, which was rarely assigned outside

[1] We use words instead of indices for readability.

underspecification, accounted for roughly half of all underspecification in the REQUEST frame. 63% of the frame element underspecifications are cases of optional elements, the third class in the lower half of Table 2. (Partial) agreement on underspecified tags was considerably lower than on non-underspecified tags, both in the case of frames (86%) and in the case of frame elements (54%). This was to be expected, since the cases with underspecified tags are the more difficult and controversial ones. Since underspecified annotation is so rare, overall frame and frame element agreement including underspecified annotation is virtually the same as in Table 1.

It is unfortunate that annotators use underspecification only infrequently, since it can indicate interesting cases of relatedness between different frames and frame elements. However, underspecification may well find its main use during the merging of independent annotations of the same corpus. Not only underspecified annotation, also disagreement between annotators can point out vague and ambiguous cases. If, for example, one annotator has assigned SPEAKER and the other MEDIUM in sentence (4), the best course is probably to use an underspecified tag in the merged corpus.

7 Conclusion

We presented the SALSA project, the aim of which is to construct and utilize a large corpus reliably annotated with semantic information. While the SALSA corpus is designed to be utilizable for many purposes, our focus is on lexical semantics, in order to address one of the most serious bottlenecks for language technology today: the lack of large, domain-independent lexica.

In this paper we have focused on the annotation with frame semantic roles. We have presented the annotation scheme, and we have evaluated first annotation results, which show encouraging figures for inter-annotator agreement. We have discussed the problem of vagueness and ambiguity of the data and proposed a representation for underspecified tags, which are to be used both for the annotation and the merging of individual annotations.

Important next steps are: the design of a tool for semi-automatic annotation, and the extraction of selectional preferences from the annotated data.

Acknowledgments. We would like to thank the following people, who helped us with their suggestions and discussions: Sue Atkins, Collin Baker, Ulrike Baldewein, Hans Boas, Daniel Bobbert, Sabine Brants, Paul Buitelaar, Ann Copestake, Christiane Fellbaum, Charles Fillmore, Gerd Fliedner, Silvia Hansen, Ulrich Heid, Katja Markert and Oliver Plaehn. We are especially indebted to Maria Lapata, whose suggestions have contributed to the current shape of the project in an essential way. Any errors are, of course, entirely our own.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Katrin Erk, Andrea Kowalski, and Manfred Pinkal. 2003. A corpus resource for lexical semantics. In *Proceedings of IWCS5*, pages 106–121, Tilburg, The Netherlands.

Charles J. Fillmore. 1968. The case for case. In Bach and Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York.

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

C. R. Johnson, C. J. Fillmore, M. R. L. Petruck, C. F. Baker, M. Ellsworth, J. Ruppenhofer, and E. J. Wood. 2002. FrameNet: Theory and Practice. `http://www.icsi.berkeley.edu/~framenet/book/book.html`.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2).

Adam Kilgarriff, editor. 2001. *SENSEVAL-2*, Toulouse.

Klaus Krippendorff. 1980. *Content Analysis*. Sage.

M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Gerguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA HLT Workshop*.

G. Miller, R. Beckwith, C. Fellbaum, D. Gros, and K. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–44.

Manfred Pinkal. 1996. Vagueness, ambiguity, and underspecification. In *Proceedings of SALT'96*, pages 185–201.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of LREC'98*, Granada.

Zdeněk Žabokrtský. 2000. Automatic functor assignment in the Prague Dependency Treebank. In *Proceedings of TSD'00*, Brno, Czech Republic.