Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs

Ines Rehbein
SFB Information Structure
German Department
Potsdam University
irehbein@uni-potsdam.de

Josef Ruppenhofer
Information Science and
Natural Language Processing
Hildesheim University
ruppenho@uni-hildesheim.de

Caroline Sporleder **Manfred Pinkal**
Computational Linguistics
Cluster of Excellence MMCI
Saarland University
csporled,pinkal@coli.uni-sb.de

Abstract

This paper presents Release 2.0 of the SALSA corpus, a German resource for lexical semantics. The new corpus release provides new annotations for German nouns, complementing the existing annotations of German verbs in Release 1.0. The corpus now includes around 24,000 sentences with more than 36,000 annotated instances. It was designed with an eye towards NLP applications such as semantic role labeling but will also be a useful resource for linguistic studies in lexical semantics.

1 Introduction

We present SALSA Release 2.0, a lexical-semantic resource for German. SALSA provides annotations of word senses, expressed through the frame-semantic classification of predicates, their semantic roles and syntactic realization patterns. These frame-semantic annotations in the flavor of the Berkeley FrameNet (Baker et al., 1998) are added as a complementary annotation layer to the TIGER treebank (Brants et al., 2002), a syntactically annotated corpus of German newspaper text.

Now that the SALSA project has concluded we do not only want to present the resulting resource but also take the opportunity to revisit some central methodological and analytical issues which came up during the frame development and annotation process. Since SALSA is so centrally related to FrameNet, we will typically cast our discussion as a comparison between SALSA and FrameNet, highlighting key differences. In particular, we discuss the differences in the workflow; differences in the organization and representation of lexical items; the use of underspecification; and the treatment of metaphor.

The remainder of this paper is structured as follows. In Section 2, we briefly recap the central ideas of Frame Semantics. Section 3 then gives an overview of the size and composition of SALSA. In Section 4, we describe our efforts at quality control in the creation of the resource, focusing on inter-annotator agreement. Section 5 discusses some central methodological and analytical issues that came up in the work of SALSA, situating the discussion against the background of how FrameNet handles the same issues. Finally, we offer conclusions in Section 7.

2 Frame Semantics

SALSA provides annotations in the framework of Frame Semantics (Fillmore, 1982). Frames are representations of prototypical events or states and their participants. In the FrameNet database (Baker et al., 1998), a lexical database of English which implements the ideas of Frame Semantics in the sense of Fillmore (1982), both frames and their participant roles are arranged in various hierarchical relations (most prominently, the is-a relation). FrameNet links these descriptions of frames with the words and multi-words (lexical units, LUs) that evoke these conceptual structures. It also documents all the ways in which the semantic roles (frame elements, FEs) can be realized as syntactic arguments of each frame-evoking word by labeling corpus attestations.

By way of example, consider the *Change position on a scale* frame (Figure 1), evoked in English by lexical units across different word classes such as accelerated.a, advance.v, climb.v, contraction.n, and others. In the German SALSA corpus, the frame is licensed by frame-evoking elements like, e.g., *abstürzen.v, erhöhen.v, gewinnen.v*, and *klettern.v* (crash, increase, win, climb). The core set of frame-specific roles that apply includes ATTRIBUTE, DIFFERENCE, FINAL_STATE, FINAL_VALUE, INITIAL_STATE, INITIAL_VALUE, ITEM and VALUE_RANGE, out of which only ITEM and ATTRIBUTE are realized in our example.

3 Overview of the SALSA corpus

The SALSA project used the frames of FrameNet Releases 1.2 and 1.3 to perform German annotation on top of the TIGER corpus. Since the English frames were not always available or appropriate, SALSA additionally developed a number of "proto-frames", i.e., predicate-specific frames, to provide coverage for predicate instances that were not covered by FrameNet at the time SALSA analyzed relevant German vocabulary. Figure 1 shows a sentence from the TIGER corpus that is annotated with one original FrameNet frame, *Change position on a scale* with frame elements ITEM and ATTRIBUTE, and with one SALSA proto-frame, *Zahl1-salsa* with its frame element INDIVIDUALS, which is assigned to the same NP as the frame element ITEM of *Change position on a scale*.

SALSA Release 1.0 (Burchardt et al., 2006) was published in October 2007. The total size of the annotations in SALSA Release 1.0 includes

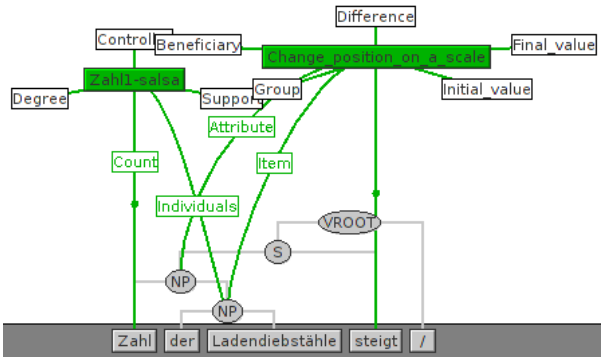

Figure 1: "Number of shoplifting cases increases."

20.380 annotated instances for 493 verb lemmas and 348 annotated instances for 14 noun lemmas (Table 1).

SALSA	Release 1.0		Release 2.0	
	token	type	token	type
verb	20,380	493	20,380	493
noun	348	14	15,871	155
total	20,728	507	36,251	648

Table 1: Annotated instances in SALSA 1.0 and 2.0

SALSA Release 2.0 complements the verb annotations in Release 1.0 with the annotation of more than 15 000 noun instances (Table 1). The selection of nouns that were annotated mostly includes nominalizations (e.g. *Eroberung* (conquest), *Freude* (delight)) and relational nouns like *Bruder* (brother) or *Seite* (side). The annotation scheme follows the one for verbs as closely as possible. There are, however, differences due to

- multi-word expressions, e.g. *guter Hoffnung sein* (expect a baby, be expecting)
- constructions, e.g. *wählen* (choose) – *X n-ter Wahl* (as in e.g. *second quality socks*)
- named entities, e.g. *Der Heilige Krieg* (holy war)
- lack in parallelism of verb and noun senses, e.g. *gut ankommen* ('be well received' – *Arrival)

The goal in selecting the nouns chosen for analysis was to achieve a balanced distribution with lexical units across all frequency bands (Table 2).

An interesting point of comparison is the number of realized frame elements for nouns and verbs in the SALSA corpus. As shown by Table 3, the average number of FEs is higher for verbs at 1.91 than for nouns at 1.45. Note that this is the case despite the fact that during the annotation of

SALSA	Release 1.0	Release 2.0
>500	2	4
301-500	6	17
101-300	41	81
51-100	68	93
31-50	75	89
11-30	99	122
<=10	217	242

Table 2: Frequency distribution of lexical units in SALSA (annotated frame instances)

SALSA	verbs	nouns
0 FE	0	2
1 FE	68	91
2 FEs	376	57
3 FEs	49	4
4 FEs	0	1
avg. # FEs	1.91	1.45

Table 3: Avg. number of realized frame elements (FEs) per lemma for verbs and nouns in SALSA

nouns the annotators were allowed to also assign non-core frame elements when suitable, while the vast number of verbal annotations do not include any non-core elements.[1]

In the next section we describe our efforts at quality control during the annotation process.

4 Inter-Annotator Agreement

The annotation process in SALSA involved a thorough quality control, including double annotation by two independent annotators for *all* instances in the corpus. Disagreements in the annotations were resolved by two expert annotators in an adjudication phase. Table 4 shows inter-annotator agreement[2] for frames and frame elements. Despite showing approximately the same degree of polysemy (Table 4), agreement on noun frames is higher than for verbal targets.

For frame elements, however, percentage agreement for nominal instances is far below the one for verbs. This is mostly due to the annotation of SUPPORT and CONTROLLER, which rank highest among the frame elements on which the annotators disagreed. Percentage agreement for SUPPORT is 43.2%, and the agreement for CONTROL is even lower with 23.6%.[3] Another reason for the low coder agreement on frame elements is caused by relaxations in the annotation guidelines which, unlike in Release 1.0, allowed the annotators to assign non-core frame elements for nouns, when appropriate. Apparently, some annotators made more use of this than others. Amongst the 15 highest-ranking frame elements on which annotators disagreed most frequently, we found DESCRIPTOR, DEGREE, PERSISTENT_CHARACTERISTIC, PLACE and TIME, all of which are non-core frame elements. In most cases, these FEs had only been annotated by one annotator, while the other one had assigned the core frame elements only.

SALSA	*frames*	*frame elements*	*# word senses*
verbs	84.6	81.0	2.6
nouns	92.9	73.3	2.7
all	87.1	78.2	2.6

Table 4: Percentage agreement for human annotators in SALSA 2.0 on frames and frame elements and avg. number of word senses (frames) per verb/noun lemma

Amongst the frames which proved to be most difficult for the annotators are the ones expressing fine-grained distinctions between related meanings, such as the distinction between *Statement* and *Telling*, between *Being_born* and *Birth*, or between *Judgment* and *Judgment_communication*. Other difficult cases involved the annotation of more abstract concepts like *Causation*, which was often mixed up with the *Reason* frame.

5 Comparison of SALSA and FrameNet

Now we want to focus on some central methodological and analytical issues which came to our attention during the annotation process.

5.1 Frame development and workflow

A key difference between SALSA and FrameNet lies in how the vocabulary to be analyzed and annotated is chosen. FrameNet has two modes of working, a lexicographic and a full text one. In full-text analysis mode the goal is to cover running text as densely as possible with frame semantic annotations. When frames are found to be

[1] In the first phase of the project when only verbs were annotated, the policy was to usually forgo annotating non-core FEs in favor of annotating a greater number of targets with core FEs.

[2] We do not report chance-corrected agreement but percentage agreement for reasons discussed in (Burchardt et al., 2006). In addition, chance-corrected agreement measures like the kappa statistics are often misleading when applied to unbalanced data sets. Consider, e.g., the lemma *Bruder* (brother), where we have 29 annotated instances out of which 28 are assigned the frame *Kinship* by both annotators. However, due to the skewed distribution of word senses the chance-corrected agreement (Fleiss κ) is only 0.491, while the percentage agreement of 96.6% better reflects that the annotators agreed on all but one instance for the lemma *Bruder*.

[3] Efforts to overcome this difficulty by replacing support and control predicates with fully flashed out frames and proto-frames did not succed.

missing, they have to be created "on the fly". By contrast, in lexicographic work, frames are developed and lemmas identified that can evoke them. In this mode, patterns of polysemy of lemmas typically lead the analysts from the description of the current frame to the description of a related but different one. The database as a whole grows organically.

Correlated with the differences in how the frames are chosen are differences in the annotators' task. In the lexicographic mode, annotators proceed by focusing sequentially on the different lexical units of the frame. For each LU, they label a set of instances that were extracted from a large corpus based on syntactic pattern or collocates. The annotators deal with only one LU per sentence and they get to select cases where the target lemma at issue clearly evokes the frame of interest.

SALSA's way of working combined aspects of FrameNet's lexicographic and full-text modes, as it aimed to achieve full coverage for all the uses of a set of lemmas in the TIGER corpus. The set of lemmas to be analyzed was chosen mainly with reference to the lemmas' frequency in the corpus. SALSA annotators, like FrameNet full-text annotators, had to classify every instance of a predicate, not being able to only mark up the clearest examples. In terms of the frame inventory, SALSA re-used as many FrameNet frames as possible. In some cases, minor modifications were made to the FrameNet frames, either to accommodate peculiarities of German or to allow for a more consistent annotation when the English FrameNet seemed to make too fine-grained distinctions. Still in other cases, SALSA had to create so-called proto-frames for specific word senses not covered by FrameNet. Since the set of lemmas to be analyzed was mainly frequency-based, the average SALSA lexical unit has fewer known frame-mates than the average FN lexical unit.

SALSA, having 1826 lexical units (1349 verbal ones, 477 nominal ones) and 36,251 annotated frame instances, has defined more than 1,000 different frames. By comparison, FrameNet's release 1.5, which has more than 10,000 lexical units and includes about 150,000 manually annotated frame instances has only 1019 frames. Table 5 shows that many of SALSA's frames are proto-frames, defined for a specific sense of a specific lemma.

	Verbs	**Nouns**	**LUs**
FN frame	865	163	1,028
modified FN frame	33	35	68
proto-frame	451	279	730
Total (LUs)	1,349	477	1,826

Table 5: Distribution of frames across lexical units

However, the numbers for proto-frames are somewhat misleading. Since SALSA kept the FrameNet frame inventory of Releases 1.2 and 1.3 as its reference point, some of SALSA's verb-sense specific proto-frames have already been covered by later FrameNet Releases. For instance, the proto-frame *Abnehmer1-salsa* for the lemma *Abnehmer.n* has been superseded by FrameNet's frame *Commerce_scenario*.

As a measure of the degree to which SALSA could re-use FrameNet's analyses, consider that of the 1023 frame types used by SALSA, 730 are proto-frames, 256 original FrameNet frames and 37 modified FrameNet frames (these latter recognizable by a frame name ending in "fnsalsa"). In other words, about 1 in 8 of the FrameNet frames used was adapted in some way. These adaptations typically concern Frame Elements: they either receive broader definitions, pairs of them are merged, or, more rarely, new ones are introduced into the frame.

Conversely, we can also ask whether the proto-frames that SALSA developed might be of use to FrameNet, too. Since no completely up-to-date record exists for which SALSA proto-frames have been made redundant by new frames created by the FrameNet project, we will consider a 50-item random sample out of the 730 proto-frames. In 15 cases, the proto-frame could be replaced directly by a now existing FrameNet frame, indicating compatible analyses. 15 more proto-frames represent cases where there are very clear English translation equivalents, suggesting that a FrameNet frame would be needed anyway. An example of this is *Attentat* 'attempt (on sb's life)'.

For the remaining 20 proto-frames, the question of integration depends on policy decisions and preferences for generality or specificity. As an example, consider the German verb *aufschla-*

gen, which in one of its meanings is typically translatable as *add*. However, it has very narrow selectional and contextual restrictions, being typically used to talk about adding a surcharge to the cost of a good or service. German also has a morphologically related noun *Aufschlag* (and also *Zuschlag*), which translate English *surcharge*. If one decides to have a specific frame in English for this narrow concept, then the German proto-frame may serve as a seed. If English FrameNet decided to only use a more general frame for adding, *aufschlagen* could evoke that, too, but the original definition of the proto-frame would need to be given up. As another example consider the German reflexive verb *sich durchsetzen*, which covers greater semantic ground than its possible translations, among which are *win out, get one's way*, and *become accepted*. English FrameNet may not have need for a frame general enough to host *sich durchsetzen*, if it chooses to relate e.g. *get one's way* more specifically to contexts of human competition or argument. On the other hand, the FrameNet hierarchy would probably not be harmed by having a general frame, to which more specific ones can be related. Generally, the challenges in aligning frames and lexical units across the two languages do not seem to be fundamentally different from what is involved in aligning frames and lexical units in one language.

5.2 Organization and representation of lexical units

A basic difference between SALSA and FrameNet is that the coverage of SALSA is limited to the frames (word senses) needed for the TIGER corpus. Senses/frames of a lemma that were not attested are not accounted for. For instance, while the lemma *einräumen* has an attested sense of 'conceding' in the TIGER corpus, its sense of 'filling up, stocking' is not attested and, consequently, is missing from SALSA. FrameNet, by comparison, does list lexical units in frames anyway even when it cannot provide annotations for lack of attestations in the corpus or lack of resources for performing the annotations.

In SALSA, unlike FrameNet, lemmas have a more prominent role. For one, the annotations are distributed in one file per lemma. More importantly, multi-word expressions are not represented outright but are treated as senses of one of their component lemmas. For instance, *ins Auge gehen* 'go wrong, backfire' is simply a sense of the lemma *Auge*. Note that this treatment of multi-words as part of the treatment of a component lemma is motivated by purely pragmatic considerations. Multiwords in Salsa were included from a decoding perspective, that is, because one of their component lemmas (e.g. *Auge* 'eye') was being analyzed and the multi-word uses needed to be covered so as to guarantee the exhaustive treatment of all tokens of the lemma in question. In FrameNet, by contrast, idiomatic multi-word expressions are treated as lexical units and they are included in the resource from an encoding perspective, that is, because they have the right semantics to fit a particular frame. In FrameNet, multi-word lexical units may consist of lemmas that have no other sense in the resource.

Because of the lack of explicit representation, we estimated the percentage of multi-words in SALSA in a 100 item random sample of lexical units: 8 percent of the lexical units are multi-word items.

Another area where SALSA differs subtly from FrameNet is in the handling of support and control predicates, which are recorded in the annotation of frame elements outside of the maximal projection of the frame evoking element (Figures 2 and 3). Table 6 displays the distribution of the special governors that are recognized by SALSA.

	Contr.	**Supp. N**	**Supp. V**	**Total**
instances	692	1644	752	3088
types	451	249	52	663

Table 6: Distribution of Support and Control in SALSA

In SALSA, support predicates can receive two types of treatment. Within frames evoked by nouns, an honorary frame element SUPPORT is used to label support verbs and prepositions (cf. column *Supp. N* in Table 6). An example can be seen in Figure 2. Copular verbs are also treated as instances of SUPPORT predicates, unlike in FrameNet. Unlike Support predicates, Controllers are said to introduce a distinct event from that of the target. They do, however, share at least some frame element with the event of the target. The constituent expressing that shared par-

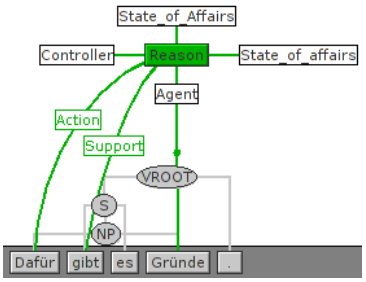

Figure 2: Support verb annotated as honorary FE
"For this, there are reasons."

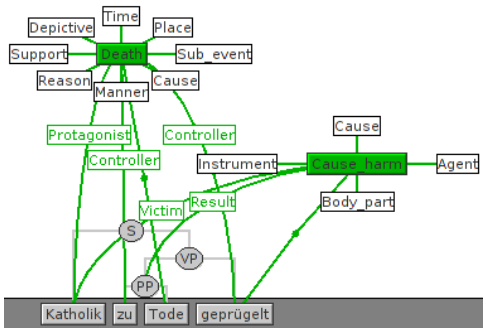

Figure 3: Annotation of Controllers as honorary FEs
"Catholic beaten to death"

ticipant is labeled with a frame element relative to the noun target.

A second treatment that support uses can receive is that they are represented by a support sense of a verb lemma as shown in Figure 4 (cf. column *Supp. V* in Table 6). However, such support-frames have only one frame element, SUPPORTED, for the supported noun. The other arguments of the support verb, let alone those of the noun, are not annotated. Note that this second treatment is motivated by SALSA's self-imposed goal of analyzing all instances of the lemmas it works with. If support verbs could only be marked from within noun frames, frames for all the nouns that the verb lemmas can support would have to be created, too.

Overall, neither FrameNet nor SALSA meet Heid (1998)'s desiderata of which types of information to collect about noun-verb collocates. Most importantly, although both resources acknowledge the importance of support verbs, neither resource treats support verb-noun combinations as first-class citizens, that is, as separate lexical entries, as noted above. Neither FrameNet nor SALSA give an explicit characterization of the semantics of support verbs, for instance, in terms of (Mel'čuk, 1996) lexical functions. Implicitly, all occurring support verbs appear to be synonyms and pragmatically (including registrally) equivalent. For an attempt to semi-automatically categorise FrameNet support verbs in terms of lexical functions see (Ramos et al., 2008). Morphosyntactic constraints on the support predicate or the supported predicate are not stated and may only be observed from the annotations. For instance, the noun *Opfer* 'victim' can occur as part of the structure *zum Opfer fallen* 'fall victim'. In this collocation, the contraction *zum* cannot be uncontracted to the combination consisting of the preposition *zu* and the dative determiner form *dem*, which is possible in productively formed combinations.

5.3 Underspecification

One problem for annotating in full-text mode is the fact that all instances of the target lemma have to be resolved. There are many cases where the context does not fully disambiguate the word sense of the lemma of interest. When annotating in the lexicographic mode, it is possible to simply dismiss those cases and focus on the prototypical uses of a particular word meaning. In SALSA, ways had to be found to deal with this issue.

Therefore, underspecification was introduced to increase inter-annotator agreement for cases where the annotators were confronted with a decision where two or more solutions seem equally adequate. Underspecification also accommodates the fact that word meanings are often by no means clear-cut but rather seem to reflect gradience by showing a certain degree of sense overlap (Erk and Padó, 2007).

Underspecification is used in either of two cases. Firstly, two frames (or frame elements) can be underspecified when they both cover part of

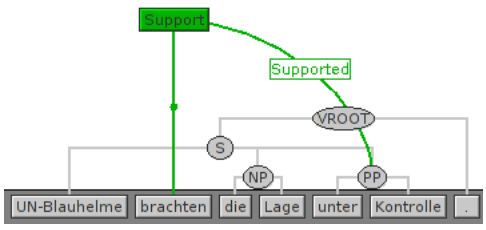

Figure 4: Support verb annotated as verb sense
"UN soldiers brought the situation under control."

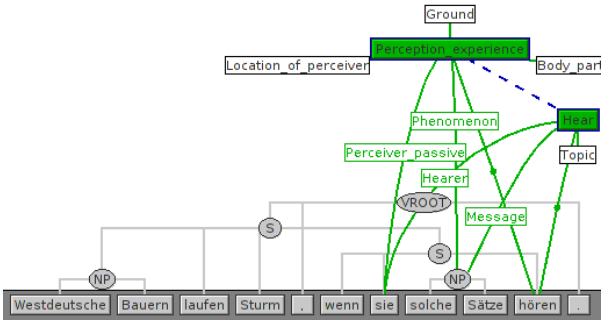

Figure 5: Underspecification in SALSA

the meaning of a predicate (frame element), but fail to represent the whole meaning. The second case applies when it is clear that only one of the two frames can apply, but – due to ambiguity in context – it is not clear which one does represent the intended meaning.

Figure 5 gives an example of frame underspecification where both frames represent a possible meaning of the sentence in (1). It is not clear whether the event of listening was intentional (*Hear*) or whether the farmers are passive listeners who cannot help but undergo the perception of the speech signal (*Perception_experience*).

(1) Westdeutsche Bauern laufen Sturm , wenn sie solche Sätze hören.
West-German farmers run storm , when they such sentences hear
"West-german farmers are up in arms when listening to such sentences."

Table 7 shows the number of underspecified frames and frame elements for both noun and verb lemmas in SALSA. As expected, the numbers for underspecification are much higher for the verb senses. What comes as a surprise is the large gap between the numbers for verb frames and noun frames (7.34 versus 0.64%). Comparing this with the lower inter-annotator agreement for verbal frames (Table 4), it seems as if the annotators use underspecification as a means to deal with the hard cases, regardless of the fact that it was never intended as such.

SALSA	verbs	nouns
frames	7.34%	0.64%
FEs	1.67%	0.16%

Table 7: Percentage of underspecified frames and frame elements in SALSA 2.0

In the next section we describe how SALSA deals with metaphoric expressions.

5.4 Metaphors

In SALSA, idioms and entrenched metaphorical uses are assigned a frame of their own, as are support uses of many verbal predicates. This is in contrast to WordNet (Miller, 1995) and results in a seemingly higher polysemy when comparing numbers for word senses for both resources (Burchardt et al., 2006).

The entrenchment of metaphors is not always easily ascertained but there are some criteria that are used in combination. One is that entrenched metaphors are not perceived as creative. Another is that entrenched metaphors, especially when they are basic conceptual metaphors, often apply to several frame-mates. For instance, many predicates in the *Change position on a scale* frame such as *rise, fall, plummet, climb* etc. also have uses in the *Motion directional* frame.

Metaphors that do not involve completely entrenched metaphorical meanings are described by a combination of two frames in SALSA: a source frame, expressing the literal meaning of the multi-word, and a target frame describing the understood meaning. This follows the ideas of Lakoff and Johnson (1980) on metaphorical transfer, where a mapping can be defined from the conceptual domain from which the metaphorical expression is drawn (source domain) to the target domain which represents the figurative meaning of the expression. Figure 6 gives an example where the source frame (*Request*) captures the literal meaning of *fordern* (demand, require) and the target frame (*Causation*) expresses the meaning understood by most listeners, namely that there was a CAUSE (the riot) which had an EFFECT (the death of at least four people).[4]

In cases where the target meaning was not easy to capture, only the source frame was annotated. This practice allowed for the annotation to proceed swiftly while, at the same time, assuring that metaphors can be retrieved from the corpus.

[4]Interestingly, this use of German *fordern* is not treated as an established word sense by the Duden online dictionary or by GermaNet (Hamp and Feldweg, 1997), the German counterpart of WordNet, while its usual English translation *claim* has a separate sense in WordNet.

SALSA 2.0	verbs	nouns
	266 (1.3%)	4 (0.02%)

Table 8: Metaphorical expressions annotated in SALSA (numbers in brackets give the percentage of all annotated verbs/nouns annotated as metaphors)

Table 8 gives the number of annotated metaphorical expressions in SALSA. The vast majority of them are evoked by verb lemmas. This, similar to the case of underspecification, might reflect a personal bias of the annotators who created the frames. While the annotator who created the verbal frames tended to flag existing frames as metaphorical, the annotator who created the noun frames had a bias for defining new, more fine-grained frames that captured the metaphorical meaning of the expression. It is not clear to us whether either of the approaches has a significant advantage over the other.

Finally, the above discussion of metaphor does not capture another class of cases. Some frames in SALSA, though inspired by existing FrameNet frames, were created as proto-frames. Here, the differences between SALSA and FrameNet do not concern the Frame elements but typically involve some notion of 'generalization'. For instance, the verb lemma *ablegen* has a proto-frame *ablegen1-salsa*, which is described as follows:

> An Agent causes a Theme to leave a location, the Source. Unlike in the non-generalized version of this frame, neither the location nor the Theme need be a location in the literal sense. The Source is profiled by the words in this frame, just as the Goal is profiled in the Placing frame.

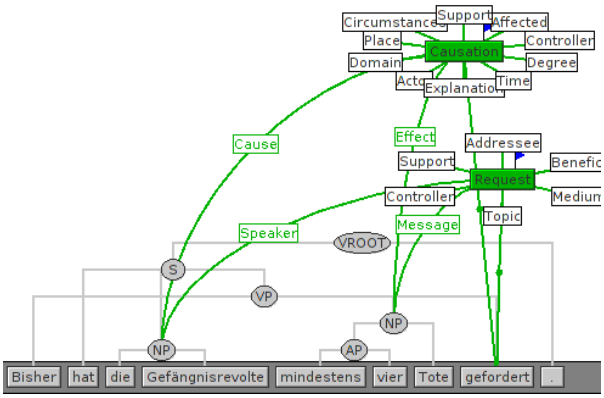

Figure 6: Annotation of metaphors in SALSA
"Up to now, the prison riot has claimed at least 4 lifes."

As the definition suggests this proto-frame targets uses of the lemma that talk about things like "getting rid of [lit. doffing, taking off] an image or a habit". A possible analysis of these cases is that they involve metaphor rather than merely more general meanings. These frames are not explicitly marked but their definition typically refers to the 'generalization' of meaning in the frame relative to an existing FrameNet frame or to another proto-frame.

6 Discussion

Having outlined the different approaches of FrameNet and SALSA, a question that naturally comes to mind is which impact these differences have on practical applications. Regarding coverage, SALSA seems to be more domain-specific, as only vocabulary from the newspaper domain has been dealt with, while FrameNet provides more general frames but is to be expected to have coverage gaps on newstext. Underspecification, as applied in SALSA to deal with ambiguous instances, might result in 'harder' training and test sets for machine learning applications, while the prototypical instances in FrameNet might be easier to classify. It is still unclear which effect these different training sets will have when used as training data for Semantic Role Labelling systems that are to be applied to new text domains.

7 Conclusions

We presented Release 2.0 of the SALSA corpus, which provides frame-semantic annotations for German nouns and verbs. The corpus now contains more than 36,000 annotated instances from the newspaper domain. In the paper we described the workflow in SALSA, discussed our efforts to ensure annotation quality and reported inter-annotator agreement for frame and frame element annotations. The core of our discussion then focused on methodological choices made in SALSA and compared them to the approach taken by FrameNet. The SALSA corpus is freely available[5] and can be used as training data for semantics-like NLP applications as well as for linguistic studies in lexical semantics.

[5] http://www.coli.uni-saarland.de/projects/salsa/

Acknowledgments

This work was funded by the German Research Foundation DFG (grant PI 154/9-3). We gratefully acknowledge the work of our annotators, Lisa Fuchs, Andreas Rebmann, Gerald Schoch and Corinna Schorr. We would also like to thank the anonymous reviewers for helpful comments.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

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

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Katrin Erk and Sebastian Padó. 2007. Towards a computational model of gradience in word sense. In *Proceedings of IWCS-7*, Tilburg, The Netherlands.

Charles J. Fillmore, 1982. *Linguistics in the Morning Calm*, chapter Frame Semantics, pages 111–137. Hanshin Publishing, Seoul.

B. Hamp and H. Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Ulrich Heid. 1998. Towards a corpus-based dictionary of german noun-verb collocations. In *Proceedings of the EURALEX International Congress*, Liège, Belgium.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Igor Mel'čuk. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical functions in lexicography and natural language processing*, volume 31, pages 37–102. John Benjamins, Amsterdam/Philadelphia.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Margarita Alonso Ramos, Owen Rambow, and Leo Wanner. 2008. Using semantically annotated corpora to build collocation resources. In *Proceedings of Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.