Published in: Language Resources and Evaluation vol. 46 (2012) no. 1, pp. 1-23.

Is it worth the effort? Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation

Ines Rehbein · Josef Ruppenhofer · Caroline Sporleder

Abstract Corpora with high-quality linguistic annotations are an essential component in many NLP applications and a valuable resource for linguistic research. For obtaining these annotations, a large amount of manual effort is needed, making the creation of these resources time-consuming and costly. One attempt to speed up the annotation process is to use supervised machine-learning systems to automatically assign (possibly erroneous) labels to the data and ask human annotators to correct them where necessary. However, it is not clear to what extent these automatic pre-annotations are successful in reducing human annotation effort, and what impact they have on the quality of the resulting resource. In this article, we present the results of an experiment in which we assess the usefulness of partial semi-automatic annotation for frame labeling. We investigate the impact of automatic pre-annotation of differing quality on annotation time, consistency and accuracy. While we found no conclusive evidence that it can speed up human annotation, we found that automatic pre-annotation does increase its overall quality.

Keywords Linguistic annotation · Semantic role labelling · Frame semantics · Semi-automatic annotation

1 Introduction

Linguistically annotated resources play a crucial role in natural language processing. Many recent advances in areas such as part-of-speech tagging, parsing,

I. Rehbein (✉) · J. Ruppenhofer · C. Sporleder
Saarland University, PO Box 15 11 50, 66041 Saarbrücken, Germany
e-mail: rehbein@coli.uni-saarland.de

J. Ruppenhofer
e-mail: josefr@coli.uni-saarland.de

C. Sporleder
e-mail: csporled@coli.uni-saarland.de

co-reference resolution, and semantic role labeling have only been possible because of the creation of manually annotated corpora, which then serve as training data for machine-learning based NLP tools. However, human annotation of linguistic categories is time-consuming and expensive. While this is already a problem for major languages like English, it is an even bigger problem for less-used languages.

This data acquisition bottleneck is a well-known problem and there have been numerous efforts to address it on the algorithmic side. Examples include the development of weakly supervised learning methods such as co-training and active learning. However, addressing only the algorithmic side is not possible or desirable in all situations. First, some machine learning solutions are not as generally applicable or widely re-usable as one might think. It has been shown, for example, that co-training does not work well for problems which cannot easily be factorized into two independent views (Mueller et al. 2002; Ng and Cardie 2003). Some active learning studies suggest both that the utility of the selected examples strongly depends on the model used for classification and that the example pool selected for one model can turn out to be sub-optimal when another model is trained on it at a later stage (Baldridge and Osborne 2004). Also, Rehbein et al. (2010) applied active learning to a frame assignment task and showed that annotation noise caused by biased annotators as well as erroneous annotations mislead the classifier and result in skewed data sets, and that for the task of frame assignment for highly ambiguous words no time savings are to be expected when applied to a realistic scenario. Furthermore, there are a number of scenarios for which there is simply no alternative to high-quality, manually annotated data; for example, if the annotated corpus is used for empirical research in linguistics (Meurers and Müller 2007; Meurers 2005).

In this paper, we look at the data acquisition problem from the data creation side. Specifically, we investigate whether a semi-automatic annotation set-up in which a human expert corrects the output of an automatic system can help speed up the annotation process without sacrificing annotation quality.

For our study, we explore the task of frame-semantic argument structure annotation (Baker et al. 1998; Fillmore et al. 2003). We chose this particular task because it is a rather complex—and therefore time-consuming—undertaking, and it involves making a number of different but interdependent annotation decisions for each instance to be labeled (e.g. frame assignment and labeling of frame elements, see Sect. 3.1). Semi-automatic support would thus be of real benefit.

More specifically, we explore the usefulness of automatic pre-annotation for the first step in the annotation process, namely frame assignment (which can be viewed as a word sense disambiguation task). Since the available inventory of frame elements is dependent on the chosen frame, this step is crucial for the whole annotation process. Furthermore, semi-automatic annotation is more feasible for the frame labeling sub-task. Most automatic semantic role labeling systems (ASRL), including ours, tend to perform much better on frame assignment than on frame role labeling, and correcting an erroneously chosen frame typically also requires fewer physical operations from the annotator than correcting a number of wrongly assigned frame elements.

We aim to answer three research questions in our study: First, we explore whether pre-annotation of frame labels can indeed speed up the annotation process. This question is important because frame assignment, in terms of physical operations of the annotator, is a relatively minor effort compared to frame role assignment and because checking a pre-annotated frame still involves all the usual mental operations that annotation from scratch does. Our second major question is whether annotation quality would remain acceptably high. Here the concern is that annotators might tend to simply go along with the pre-annotation, which would lead to an overall lower annotation quality than they could produce by annotating from scratch.[1] Depending on the purpose for which the annotations are to be used, trading off accuracy for speed may or may not be acceptable. Our third research question concerns the required quality of pre-annotation for it to have any positive effect. If the quality is too low, the annotation process might actually be slowed down because annotations by the automatic system would have to be deleted before the new correct ones could be made. In fact, annotators might ignore the pre-annotations completely. To determine the effect of the pre-annotation quality, we not only compared a null condition of providing no prior annotation to one where we did, but we in fact compared the null condition to two different quality levels of pre-annotation, one that reflects the performance of a state-of-the-art ASRL system and an enhanced one that we artificially produced from the gold standard.

2 Related work

While semi-automatic annotation is frequently employed to create labeled data more quickly (see, e.g., Brants and Plaehn 2000), there are comparatively few studies which systematically look at the benefits or limitations of this approach. One of the earliest studies that investigated the advantages of manually correcting automatic annotations for linguistic data was carried out by Marcus et al. (1993) in the context of the construction of the Penn Treebank. Marcus et al. (1993) employed a post-correction set-up for both part-of-speech and syntactic structure annotation. For pos-tagging they compared the semi-automatic approach to a fully manual annotation. They found that the semi-automatic method resulted both in a significant reduction of annotation time, effectively doubling the word annotation rate, and in increased inter-annotator agreement and accuracy.

Chiou et al. (2001) explored the effect of automatic pre-annotation for treebank construction. For the automatic step, they experimented with two different parsers and found that both reduced overall annotation time significantly while preserving accuracy. Later experiments by Xue et al. (2002) confirmed these findings.

Ganchev et al. (2007) looked at semi-automatic gene identification in the biomedical domain. They, too, experimented with correcting the output of an

[1] This problem is also known in the context of resources that are collaboratively constructed via the web (Kruschwitz et al. 2009).

automatic annotation system. However, rather than employing an off-the-shelf named entity tagger, they trained a tagger maximized for recall. The human annotators were then instructed to filter the annotation, rejecting falsely labeled expressions. Ganchev et al. (2007) report a noticeable increase in speed compared to a fully manual set-up.

The approach that is closest to ours is that of Chou et al. (2006) who investigate the effect of automatic pre-annotation for Propbank-style semantic argument structure labeling. However that study only looked into the properties of the semi-automatic set-up; the authors did not carry out a control study with a fully manual approach. Nevertheless Chou et al. (2006) provide an upper bound of the savings obtained by the semi-automatic process in terms of annotator operations. They report a reduction in annotation effort of up to 46%.

Another annotation experiment similar in spirit to ours is the one by Dandapat et al. (2009), who present a case study of part-of-speech annotation in Bangla and Hindi. They compare the time requirements needed for fine-grained part-of-speech annotation done by two groups of annotators (all of them trained linguists), where the first group has been subject to extensive training and in-house supervision, while the second group was self-trained and did not get any feedback during the annotation process. Dandapat et al. (2009) systematically tested the impact of automatic pre-annotation (a) by a part-of-speech tagger trained on a small data set, producing low quality pre-annotations, and (b) by a part-of-speech tagger trained on a larger data set, providing high accuracy pre-annotations. Unfortunately, Dandapat et al. (2009) did not have access to a manually annotated gold standard for evaluation and therefore had to assess the accuracy of the annotation indirectly by means of inter-annotator agreement.

Dandapat et al. (2009) report higher inter-annotator agreement when annotating text with automatically assigned low quality part-of-speech tags, compared to inter-annotator agreement on text without pre-annotation. For the high-quality pre-annotation, inter-annotator agreement further increased. Their most important finding concerns the impact of training and supervision: while the use of an appropriate annotation tool in combination with automatic pre-annotation reduced annotation time for the untrained annotator group to the same level as needed by the trained annotators, inter-annotator agreement for the first group (and hence the reliability of the annotated data) was under all conditions lower than for the trained annotators. The authors conclude from this that training and supervision are worthwhile the effort and are indispensable for obtaining high-quality linguistic annotations.

While the studies mentioned above focus on the same research question, namely to what extent automatic pre-annotation can lower the costs for human annotation, it should be noted that automatic systems for tasks like part-of-speech tagging and parsing are far more advanced than e.g. systems for automatic semantic role labeling. As a result, the quality of the automatically produced pre-annotations is much higher and therefore less effort is needed for manual correction during the annotation process. It is not yet clear if the results gained from these studies carry over to highly complex annotation tasks such as frame-semantic annotation.

3 Experimental set-up

3.1 Frame-semantic annotation

The annotation scheme we use is that of FrameNet, a lexicographic project that produces a database of frame-semantic descriptions of English vocabulary. Frames are representations of prototypical events or states and their participants in the sense of Fillmore (1982), Fillmore and Baker (2010). In the FrameNet database, both frames and their participant roles are arranged in various hierarchical relations (most prominently, the is-a relation).[2] FrameNet links these descriptions of frames with the words and multi-words (lexical units) that evoke these conceptual structures. It also documents all the ways in which the semantic roles (frame elements) can be realized as syntactic arguments of each frame-evoking word by labeling corpus attestations. As a small example, consider the Collaboration frame, evoked in English by lexical units such as *collaborate*.v, *conspire*.v, *collaborator*.n and others. The core set of frame-specific roles that apply include Partner_1, Partner_2, Partners and Undertaking. A labeled example sentence is

(1) [The two researchers $^{\textit{Partners}}$] **collaborated** [on many papers $^{\textit{Undertaking}}$].

FrameNet uses two modes of annotation: full-text, where the goal is to exhaustively annotate the running text of a document with all the different frames and roles that occur, and lexicographic, where only instances of particular target words used in particular frames are labeled (Fillmore et al. 2003).

In lexicographic mode, the job of a FrameNet annotator is to look over sentences containing a particular target word. The sentences are extracted from a balanced corpus (mostly, the British National Corpus[3]) and pre-grouped into subcorpora based on syntactic contexts or collocates. From each subcorpus, the annotator is to label a small number of prototypical instances where the appropriate frame is clearly evoked by the target word. Specifically, with the target being pre-marked, annotators select the phrases that identify particular semantic roles in the sentences, and tag them with the names of these roles (Frame Elements). Since FrameNet does not work with a pre-parsed corpus, a chunk parser then provides grammatical information, in terms of grammatical function and phrase type, about the tagged phrases, which the annotators hand-correct as needed.

[2] In FrameNet, the participant roles are called *frame elements*, while in a more general context the term *semantic roles* is commonly used. Also, in the FrameNet lexicon, the lexical entries are called *lexical units*. A lexical unit (LU) is a pairing of a lemma and a frame that it evokes. Most of FrameNet's lemmas consist of a single morphological lexeme but multi-word expressions consist of several. In this paper, we will sometimes allow ourselves to use the term *word senses* to refer to the frames a lemma evokes because, as noted by Erk (2005), the process of frame assignment can be treated as a word sense disambiguation task.

[3] http://www.natcorp.ox.ac.uk/.

3.2 Pilot study

Prior to the present study we carried out a pilot experiment comparing manual and semi-automatic annotation of different segments of running text. In this experiment we saw no significant effect from pre-annotation. Instead we found that the annotation speed and accuracy depended largely on the order in which the texts were annotated and on the difficulty of the segments. The influence of order is due to the fact that FrameNet has more than 825 frames and each frame has around two to five core frame elements plus a number of non-core elements. Therefore even experienced annotators can benefit from the recurrence of frames during the ongoing annotation process.

Drawing on our experiences with the first experiment, we chose a different experimental set-up for the present study. To reduce the training effect, we opted for annotation in lexicographic mode, restricting the number of lemmas (and thereby frames) to annotate, and we started the experiment with a training phase (see Sect. 3.5). Annotating in lexicographic mode also gave us better control over the difficulty of the different batches of data. Since these now consist of unrelated sentences, we can control the distribution of lemmas across the segments (see Sect. 3.4).

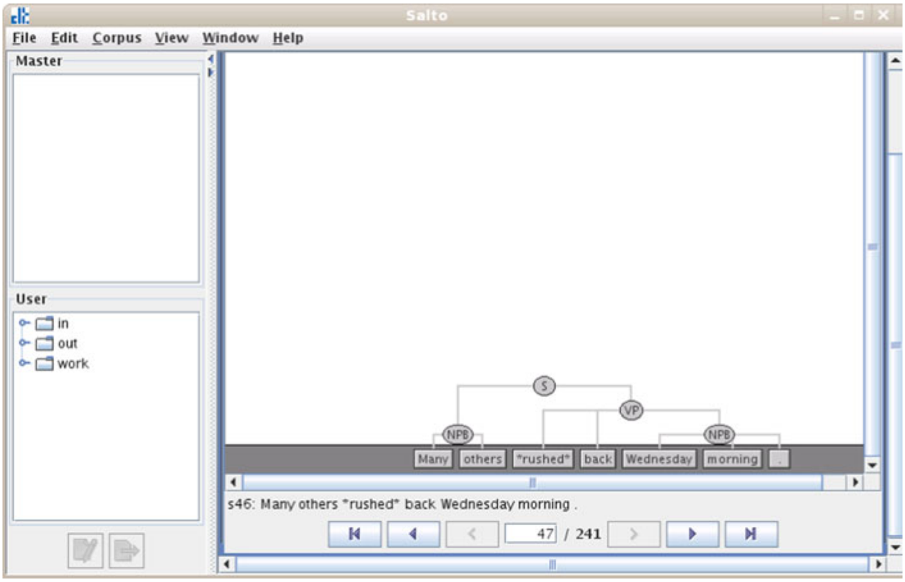
Furthermore, since the annotators in our pilot study had often ignored the error-prone pre-annotation, in particular for frame elements, we decided not to pre-annotate frame elements and to experiment with an enhanced level of pre-annotation to explore the effect of pre-annotation quality.

3.3 Annotation set-up

The annotators included the authors and three computational linguistics undergraduates who have been performing frame-semantic annotation for at least 1 year. While we use FrameNet data, our annotation set-up is different. The annotation consists of decorating automatically derived syntactic phrase structure trees with semantic role labels using the Salto tool (Burchardt et al. 2006). As input, the annotators are shown a phrase structure syntax tree (Fig. 1). Parsing was done using a state-of-the-art statistical parser. Due to time restrictions, the annotators were instructed to ignore syntax errors produced by the parser. By contrast, in FrameNet annotation a chunk parser is used to provide phrase type and grammatical relations for the arguments of the target words. Further, FrameNet annotators need to correct mistakes of the automatic grammatical analysis, unlike in our experiment.

In our experiment, the first annotation step, frame assignment, involves choosing the correct frame for the target lemma from a pull down menu (Fig. 2); the second step, role assignment, requires the annotators to draw the available frame element links to the appropriate syntactic constituent(s) (Fig. 3). Figure 4 shows the completed annotation with frame and frame elements assigned.

Note that we took as our input sentences contained in the FrameNet data release but stripped of their annotations (see Sect. 3.4). Thus, our annotators, unlike the FrameNet annotators, did not have to decide whether an instance of the target word



s46: Many others *rushed* back Wednesday morning .

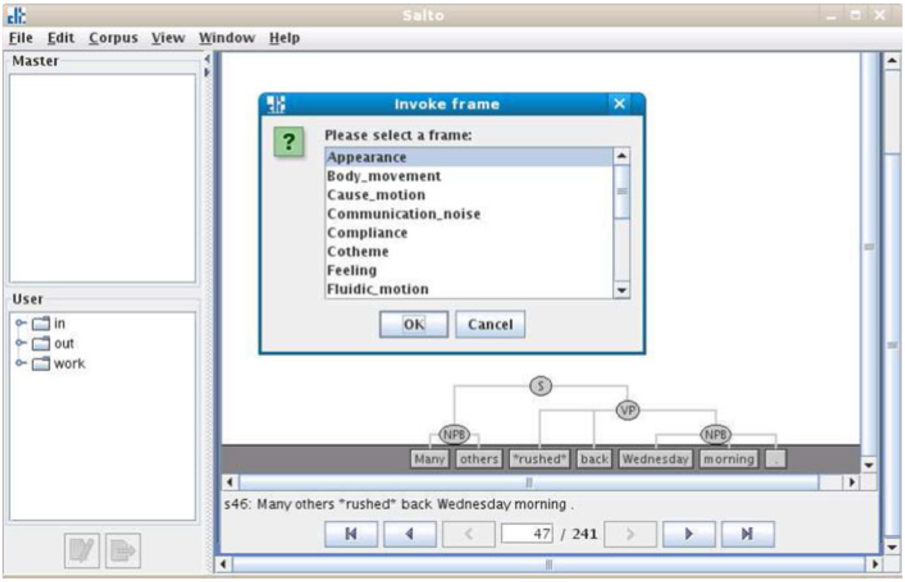
Fig. 1 The Salto Annotation Tool—input for annotation: phrase structure trees

exhibited a known word sense or one not yet treated by FrameNet. They could simply assume the former.[4] Further, our annotators did not have to choose clear, prototypical examples but were instructed to label all instances.[5] Given that our instances come out of FrameNet's annotated corpus, we assumed that the FrameNet annotators had already identified these instances as being good, prototypical examples of the relevant word senses. Thus, overall our annotation set-up ought to have been somewhat easier than that of FrameNet. Of course, not all of FrameNet's annotations may in fact be correct or clear, which would reintroduce some of the difficulty faced by the original annotators. We will come back to this issue in Sect. 4.2.

The annotators performed their annotation on computers where access to the FrameNet website, where gold annotations could have been found, was blocked. They did, however, have access to local copies of the frame descriptions needed for the lexical units in our experiment. As the overall time needed for the annotation was too long to do in one sitting, the annotators did it over several days. They were instructed to record the time (in minutes) that they took for the annotation of each annotation session.

[4] FrameNet analyzes the English lexicon from an encoding point-of-view: given a frame, it finds words that evoke that frame. FrameNet proceeds from frame to frame, rather than analyzing all senses of a given lemma. This means that as long as FrameNet is not complete, polysemous words may not have all their senses covered by FrameNet.

[5] One exception concerns metaphorical usages which may not be covered well by any of the frames provided by FrameNet for the lemma. In those cases, our annotators occasionally left the target word unannotated (see Sect. 4.2).



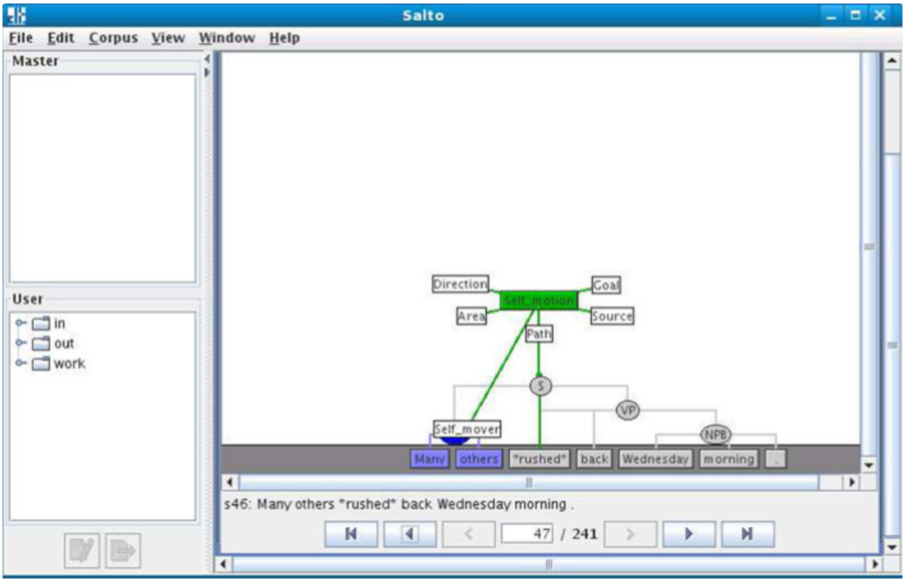
Please select a frame:

Appearance
Body_movement
Cause_motion
Communication_noise
Compliance
Cotheme
Feeling
Fluidic_motion

Many others *rushed* back Wednesday morning .

s46: Many others *rushed* back Wednesday morning .

Fig. 2 Frame assignment: choosing the appropriate frame from a list of frame candidates

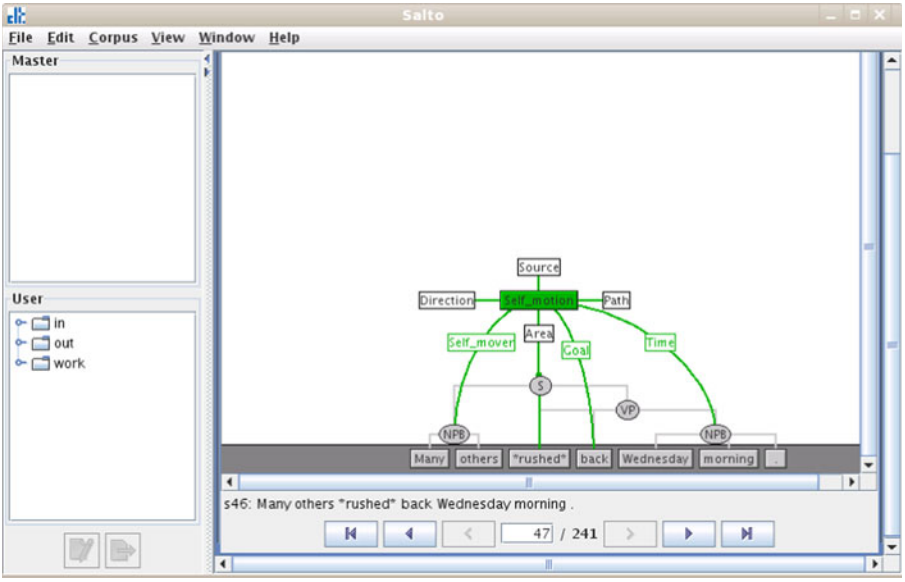


Many others *rushed* back Wednesday morning .

s46: Many others *rushed* back Wednesday morning .

Fig. 3 Linking the frame elements of the Self_motion frame to appropriate syntactic constituent(s)

Our ASRL system for state-of-the-art pre-annotation was Shalmaneser (Erk and Pado 2006). The enhanced pre-annotation was created by manually inserting substitution errors with uniform distribution into the gold standard.



s46: Many others *rushed* back Wednesday morning .

Fig. 4 The input sentence with completed frame-semantic annotation

3.4 Data

We annotated 360 sentences exemplifying all the senses that were defined for six different lemmas in FrameNet release 1.3.[6] The lemmas were the verbs *rush, look, follow, throw, feel* and *scream*. These verbs were chosen for three reasons. First, they have enough annotated instances in the FrameNet release that we could use some instances for testing and still be left with a set of instances sufficiently large to train our ASRL system. Second, we knew from prior work with our automatic role labeler that it had a reasonably good performance on these lemmas. Third, these lexical units exhibit a range of difficulty in terms of the number of senses they have in FrameNet (see Table 1) and the subtlety of the sense distinctions—e.g. the FrameNet senses of *look* are harder to distinguish than those of *rush*, due to the higher number of different word senses for *look*. However, the number of different frames for one particular target word is not the only indicator to predict how difficult the frame assignment task might be. Some frames are relatively easy for humans to disambiguate while others encode more subtle distinctions and therefore are more likely to be mixed up during annotation.

See, for example, the difference between Fluidic_motion and Self_motion (Table 2), the two frames for *rush*. Examples involving liquids or animate agents like (2) and (3) do not pose any challenge to the human annotator. However, less prototypical examples like the one in (4), often were incorrectly labeled as

[6] FrameNet release 1.3 was released in June 2006. It contained 795 frames and listed 10,195 lexical units. The successor version 1.5 with 1,019 frames and 11,829 lexical units became available only as this paper went to press.

Table 1 Lemmas used in our experiment

	Instances	Senses
Feel	134	6
Follow	113	3
Look	185	4
Rush	168	2
Scream	148	2
Throw	155	2

Table 2 Definitions of the Fluidic_motion and Self_motion frames

Fluidic_motion

In this frame a *Fluid* moves from a *Source* to a *Goal* along a *Path* or within an *Area*

Self_motion

The *Self_mover*, a living being, moves under its own power in a directed fashion, i.e. along what could be described as a *Path*, with no separate vehicle

Frame elements are shown in italics

Table 3 Definitions of the Make_noise and Communication_noise frames

Make_noise

A physical entity, construed as a point-*Sound_source*, emits a *Sound*. This includes animals and people making noise with their vocal tracts. Sometimes the sound itself is referred to with a nominal expression, in which case it is called the *Sound*. *Manner* expressions may also be relevant in this frame, if they describe properties of the sound as such. A pathschema can be overlaid on the simple noise-making scene, adding a *Location_of source* and/or a *Path*.

Communication_noise

This frame contains words for types of noise which can be used to characterize verbal communication. It inherits from Communication (possibly more specifically Communication manner) and the Sound emission frame (which simply characterizes basic sounds of whatever source, including those made by animals and inanimate objects). As such, it involves a*Speaker* who produces noise and thus communicates a *Message* to an *Addressee*

Frame elements are shown in italics

Self_motion by our annotators, while the correct frame for these instances is the Fluidic_motion frame.

(2) Tons of water **rushed** over the falls. (Fluidic_motion)
(3) Fred **rushed** into the basement. (Self_motion)
(4) A buffeting wind **rushed** over the land. (Fluidic_motion)

Make_noise and Communication_noise (Table 3) also seemed to be rather hard to distinguish, so that in spite of its low number of frames, *scream* also proved to be one of the harder cases for frame assignment. Again, there are prototypical examples in which it is relatively clear whether a verbal communication takes place (5) or not (6). But in examples like (7), choosing the correct frame is more difficult. In this case an annotator might reason that the event of somebody screaming at somebody else typically involves verbal communication, which might lead to the (wrong) assignment of the Communication_noise frame.

(5) He **screamed** some incoherent threat. (Communication_noise)
(6) Then a woman **screamed**, and all the lights came on again. (Make_noise)
(7) The bald man **screamed** loudly at Ivan. (Make_noise)

We randomly grouped our sentences into three batches of equal size and for each batch we produced three versions corresponding to our three levels of annotation.

3.5 Study design

In line with the research questions that we want to address and the annotators that we have available, we choose an experimental design that is amenable to an analysis of variance. Specifically, we randomly assign our 6 annotators (1–6) to three groups of two (Groups I–III). Each annotator experiences all three annotation conditions, namely no pre-annotation (N), state-of-the-art pre-annotation (S), and enhanced pre-annotation (E). This is the within-subjects factor in our design, all other factors are between subjects. Namely, each group was randomly matched to one of three different orders in which the conditions can be experienced (see Table 4).[7] The orderings are designed to control for the effects that increasing experience may have on speed and quality. While all annotators end up labeling all the same data, the groups also differ as to which batch of data is presented in which condition. This is intended as a check on any inherent differences in annotation difficulty that might exist between the data sets. In practice this means that all annotators had to annotate exactly the same sentences in the first, the second and the third batch, but the annotation condition varied over the three groups of annotators: while the first group had to annotate these sentences with no pre-annotation, for the second group the same batch was provided with state-of-the-art pre-annotations, and group three had to assign labels to the same batch of sentences with enhanced pre-annotations.

The three batches were presented to the annotators as three distinct subcorpora. While all annotators had to annotate in all three conditions, they were not told which of the batches belonged to the state-of-the-art condition and which one provided the enhanced pre-annotations. Finally, to rule out difficulties with unfamiliar frames and frame elements needed for the lexical units used in this study, we provided some training to the annotators. In the week prior to the experiment, they were given 240 sentences from the FrameNet database which exemplified all 6 verbs in all their senses, stripped off all their annotations. After annotating these sentences we met to discuss any questions the annotators might have about frame or frame element distinctions etc. The annotated instances of these 240 sentences contained in the FrameNet corpus were used as the gold standard to train the ASRL system.

[7] We avoided the order NSE because in that order the pre-annotation quality would have improved between all adjacent batches (from ’no annotation’ to ’state-of-the-art annotation’ to ’enhanced annotation’), in which case we might have had a confounding effect between pre-annotation quality and a possible ongoing training effect. From the remaining five theoretically possible orders we randomly selected three subject to the constraint that each annotation condition came first for exactly one group.

Table 4 Annotation condition by order and group

	1st	2nd	3rd	Annotators
Group I	E	S	N	5, 6
Group II	S	N	E	2, 4
Group III	N	E	S	1, 3

4 Results

In addition to time, we measured precision, recall and F1 score for frame assignment and semantic role assignment for each annotator. We then performed an analysis of variance (ANOVA) on the outcomes of our experiment. Our basic results are presented in Table 5. As can be seen and as we expected, our annotators differed in their performance both with regard to annotation quality and speed. Differences concerning the overall number of targets annotated in each batch are due to the fact that we did not mark the target words which had to be annotated. Some sentences by chance included more than one possible target. The annotators, who expected to encounter one target per sentence, sometimes missed these. Below we discuss our results with respect to the research questions named above.

4.1 Can pre-annotation of frame assignment speed up the annotation process?

Not surprisingly, there are considerable differences in speed between the six annotators (Table 6), which are statistically significant with $p \leq 0.05$. Focussing on the order in which the three batches were given to the annotators, we observe a significant difference ($p \leq 0.05$) in annotation time needed for each of the batches. With one exception, all annotators took the most time on the batch given to them first, which hints at an ongoing training effect. This was somewhat surprising, as all our annotators had at least one year of experience in frame-semantic annotation, and as we also had tried to rule out any training effects by means of providing a training trial, where all annotators had to assign frame and role labels to 240 training instances. However, annotation times speak for themselves. On average, the annotators needed 125 min for annotating the first batch, while for the second and the third batch the average time requirements were 91 and 90 min, respectively. This shows that for highly complex annotation tasks like frame-semantic annotation we have to account for training effects even for experienced annotators, and that a training trial might not be enough to control these effects.

While the overall time needed for annotating all three batches was too long to do it in one sitting, most of the annotators processed their batches on subsequent days, which explains the observed speed-up between the first two batches. Even experienced annotators are not expected to be able to memorize all the different frames for each lemma and all the different frame elements for each frame. So for batch 1 the annotators have to look up the frame and frame element candidates for each new instance, while after annotating enough instances they are likely to remember the more frequent labels, which crucially reduces the look-up costs and results in a considerably lower amount of time needed for annotation. The training

Table 5 Results for frame assignment: precision, recall, F1 score (F1), time in minutes (t) (frame and role assignment), pre-annotation (p): Non (N), Enhanced (E), Shalmaneser (S)

Precision		Recall		F1 (%)	t	p
Annotator 1						
(94/119)	79.0	(94/112)	83.9	81.4	75	N
(99/113)	87.6	(99/113)	87.6	87.6	61	E
(105/113)	92.9	(105/113)	92.9	92.9	65	S
Annotator 2						
(93/112)	83.0	(93/112)	83.0	83.0	135	S
(86/116)	74.1	(86/113)	76.1	75.1	103	N
(98/114)	86.0	(98/113)	86.7	86.3	69	E
Annotator 3						
(95/117)	81.2	(95/112)	84.8	83.0	168	N
(103/113)	91.2	(103/113)	91.2	91.2	94	E
(99/113)	87.6	(99/113)	87.6	87.6	117	S
Annotator 4						
(106/112)	94.6	(106/112)	94.6	94.6	80	S
(99/114)	86.8	(99/113)	87.6	87.2	59	N
(105/113)	92.9	(105/113)	92.9	92.9	52	E
Annotator 5						
(104/116)	89.7	(104/112)	92.9	91.3	170	E
(91/115)	79.1	(91/113)	80.5	79.8	105	S
(96/120)	80.0	(96/113)	85.0	82.4	105	N
Annotator 6						
(102/112)	91.1	(102/112)	91.1	91.1	124	E
(94/113)	83.2	(94/113)	83.2	83.2	125	S
(93/116)	80.2	(93/113)	82.3	81.2	135	N

Table 6 Average annotation time for frame and role assignment (in minutes) for the 6 annotators

Anot1	Anot2	Anot3	Anot4	Anot5	Anot6
67.0	102.3	126.3	63.7	126.7	128.0

trial helped to familiarize the annotators with the set of frames and frame elements, however, the time between training and the actual experiment (1 week) was too long to keep all the different frames and frame elements in memory and therefore was not able to control the training effect we observed in our experiment. We will come back to this issue and propose a way to alleviate the impact of this ongoing training effect in Sect. 6.

The different conditions of pre-annotation (none, state-of-the-art, enhanced) did not have a significant effect on annotation time. However, all annotators except one were in fact faster under the enhanced condition than under the unannotated condition. The one annotator who was not faster annotated the segment with the enhanced pre-annotation before the other two segments and thus lacked a training effect at that point. This interaction between training effect and degree of pre-annotation might be one reason why we do not find a significant effect between annotation time and pre-annotation condition. Another reason might be that the pre-annotation only reduces the physical effort needed to *annotate* the correct frame which is relatively minor compared to the cognitive effort of *determining* (or verifying) the right frame, which is required for all degrees of pre-annotation. Furthermore, the amount of time needed for frame assignment is only part of the overall annotation time, which also includes the annotation of semantic roles. Therefore it might be hard to obtain a statistically significant result for any speed-up taking place during frame assignment.

4.2 Is annotation quality influenced by automatic pre-annotation?

To answer the second question, we looked at the relation between pre-annotation condition and F1 score. Even though the results in F1 score for the different annotators vary in extent (Table 7), there is no significant difference between annotation quality for the six annotators.

Next we performed a two-way ANOVA (Within-Subjects design), and crossed the dependent variable (F1 score) with the two independent variables (order of text segments, condition of pre-annotation). Here we found a significant effect ($p \leq 0.05$) for the impact of pre-annotation on annotation quality. All annotators achieved higher F1 scores for frame assignment on the enhanced pre-annotated batches than on the ones with no pre-annotation. However, most annotators introduced some new errors by manually changing the correct, high-quality enhanced pre-annotations so that F1 scores for manual annotation for 4 out of 6 annotators are lower than the ones of the enhanced pre-annotations itself (Table 8).

Table 7 Annotation quality for frame assignment (average F1 score) for the 6 annotators

Anot1	Anot2	Anot3	Anot4	Anot5	Anot6
87.3	81.5	87.3	91.6	84.5	85.2

Table 8 Precision, recall and F1 scores for automatic pre-annotation for frames by Shalmaneser and for the enhanced pre-annotation

Batch	Precision		Recall		F1 (%)
Shalmaneser					
1	(70/112)	62.5	(70/112)	62.5	62.5
2	(75/113)	66.4	(75/113)	66.4	66.4
3	(66/113)	58.4	(66/113)	58.4	58.4
Enhanced pre-annotation					
1	(104/112)	92.9	(104/112)	92.9	92.9
2	(103/113)	91.2	(103/113)	91.2	91.2
3	(99/113)	87.6	(99/113)	87.6	87.6

This is somewhat frustrating but also shows that the human annotators do not tend to simply accept the annotations provided to them, and it also shows that the concept of a gold standard for highly ambiguous word senses is more under debate than for a task like part-of-speech tagging or syntactic parsing.

A look at the data shows that around one third (33.3%) of the errors manually inserted into the batches with enhanced pre-annotation are in fact cases where the annotators replaced the pre-annotated label with the label *Unannotated*, indicating that they considered none of the existing FrameNet frames as adequate in those particular contexts. Most of the unannotated frames are metaphoric uses of the target lemma like the one in (8), showing that not all data in FrameNet are in fact clear, prototypical examples.

(8) The tales rushed back into her mind.

Other miscorrections which occurred with high frequency concern the frame pairs Make_noise/Communication_noise, Fluidic_motion/Self_motion, Perception_active/Perception_experience, and Seeking/Scrutiny. Most interestingly, for some frame pairs these miscorrections have been symmetrical, e.g. for Make_noise/Communication_noise, where in 21 cases the annotators substituted the first frame for the second, while in 18 cases they manually substituted in the other direction (Table 9). This hints at a real ambiguity between the two frames. For the frame pair Fluidic_motion/Self_motion, on the other hand, manual changes have been strictly asymmetric. With one exception, annotators replaced Fluidic_motion with Self_motion (Table 7). This type of error is probably caused by an imprecise frame description in the annotation guidelines, making it hard for the annotators to correctly assign some of the instances.

The next issue concerns the question whether annotators make different types of errors when provided with the different styles of pre-annotation. We would like to know if erroneous frame assignment, as done by a state-of-the-art ASRL will tempt annotators to accept errors they would not make in the first place.

Table 9 Confusion matrix for frame pairs for which annotators manually introduced errors compared to the (correct) pre-annotation in the enhanced batches

Manual	Enhanced	
	Make_n.	Com_n.
Make_noise	0	21
Com_noise	18	0

Manual	Enhanced	
	Fluidic_m.	Self_m.
Fluidic_motion	0	1
Self_motion	30	0

Please note that we only look at those instances where the annotators introduced an error into the (correct) enhanced pre-annotation, therefore counts for correct frame pairs in the confusion matrices are 0

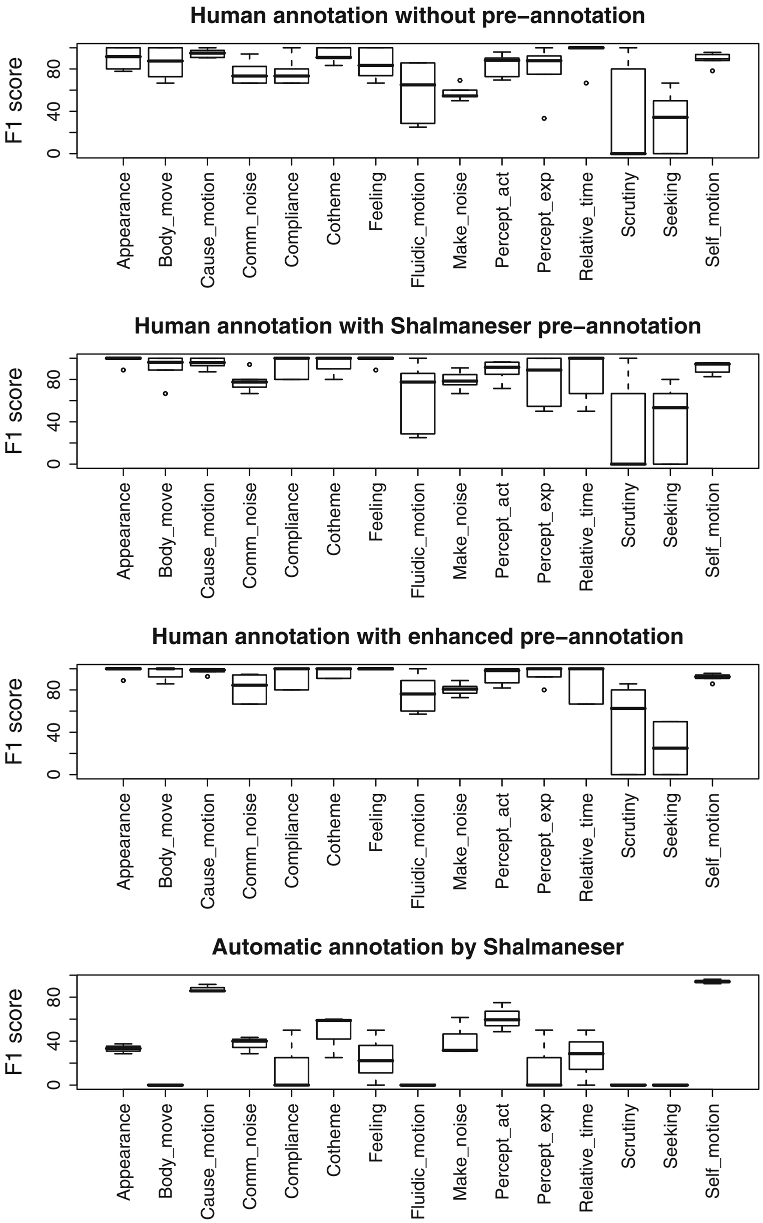


Fig. 5 F1 scores per frame for human annotators on different levels of pre-annotation and for Shalmaneser. Frames: Appearance, Body_movement, Cause_motion, Communication_noise, Compliance, Cotheme, Feeling, Fluidic_motion, Make_noise, Perception_active, Perception_experience, Relative_time, Scrutiny, Seeking, Self_motion

To investigate this issue, we compared F1 scores for each of the frames for all three pre-annotation conditions with F1 scores for frame assignment achieved by Shalmaneser. The boxplots in Fig. 5 shows the distribution of F1 scores for each frame for the different pre-annotation styles and for annotating the same sentences with Shalmaneser. We can see that the same error types are made by human annotators throughout all three annotation trials, and that these errors are different from the ones made by the ASRL.

As indicated by the F1 score for frame assignment, the most difficult frames in our data set are Scrutiny, Fluidic_motion, Seeking, Make_noise and Communication_noise. This shows that automatic pre-annotation, even if noisy and of low quality, does not corrupt human annotators on a grand scale. This is in line with previous studies for other annotation tasks (Marcus et al. 1993).

4.3 How good does pre-annotation need to be to have a positive effect?

Comparing annotation quality on the automatically pre-annotated texts using Shalmaneser, four out of six annotators achieved a higher F1 score than on the non-annotated sentences. The effect is statistically significant with $p \leq 0.05$. This means that pre-annotation produced by a state-of-the-art ASRL system is not yet good enough to significantly speed up the annotation process, but is able to improve the quality of the annotation itself.

Most interestingly, the two annotators who showed a lower F1 score on the batches pre-annotated by Shalmaneser (compared to the batch with no pre-annotation provided) had been assigned to the same group (Group I, A5 and A6). Both had first annotated the enhanced, high-quality pre-annotation, in the second trial the sentences pre-annotated by Shalmaneser, and finally the texts with no pre-annotation. It might be possible that they benefitted from the ongoing training, resulting in a higher F1 score for the third batch (no pre-annotation).

Figure 6 (left) illustrates a noticeable trend for the interaction between pre-annotation and annotation quality: the four annotators who did benefit from automatic pre-annotation all show a lower annotation quality on the batches without pre-annotation, while both types of pre-annotation (Shalmaneser, Enhanced) yield higher F1 scores for human annotation. There are, however, differences between the impact of the two pre-annotation types on human annotation quality: two annotators show better results on the enhanced, high-quality pre-annotation, the other two perform better on the texts pre-annotated by the state-of-the-art ASRL. This observation is somewhat unexpected. Looking at the data, it turned out that the two annotators who achieved a higher F1 score on the text with Shalmaneser pre-annotations were in fact the two annotators who had the most experience in frame-semantic annotation. It might be possible that therefore they relied more on their own intuition and were less likely to be influenced by the difference in quality for automatic pre-annotation.

Next we investigated the influence of pre-annotation style on annotation time. Again we can see an interesting pattern: The two annotators (A5, A6) who first annotated the batches with enhanced pre-annotation do take the highest amount of time for these (Fig. 6, right). Annotators (A1, A3) who annotated in the order N-E-S, both take most time for the texts without pre-annotation, getting faster on the text pre-processed by Shalmaneser, while the least amount of time was needed

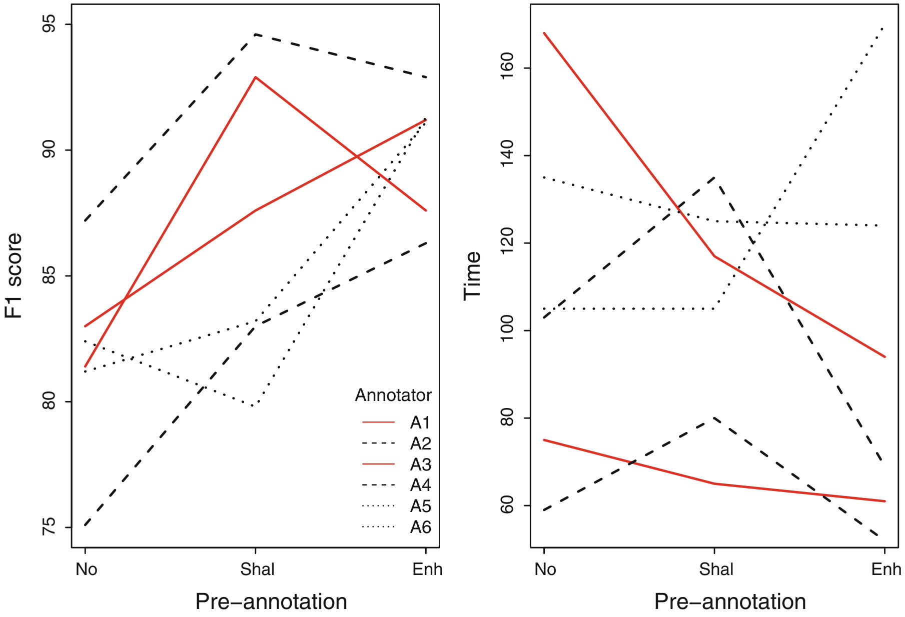


Fig. 6 Interaction between pre-annotation and F1 score for frame assignment (*left*) and between frame pre-annotation and time for full task (*right*)

for the enhanced pre-annotated texts. The two annotators (A2, A4) who processed the texts in the order S-N-E, showed a continuous reduction in annotation time, probably caused by the interaction of training and data quality. These observations, however, should be taken with a grain of salt, as they outline trends, but due to the low number of annotators, could not be substantiated by statistical tests.

4.4 Does automatic pre-annotation have an impact on inter-annotator agreement?

In the last section we showed that high-quality automatic pre-annotation can improve the accuracy of human annotation, and that even noisy and error-prone pre-annotation does not corrupt the quality of human annotation. Based on this, we would expect that automatic pre-annotation also has a positive effect on the consistency of human annotation, namely on inter-annotator agreement (IAA).

To test this hypothesis, we computed inter-annotator agreement for each pairing of human annotators a_{ij} like this: We considered the annotations by annotator a_i as a gold standard and evaluated the annotations by all the other annotators a_j .. a_n against the ones by a_i. This results in two F1 scores for each pair of annotators, a_{ij} and a_{ji}. For each annotator pair we took the mean of the two F1 scores as inter-annotator agreement. We then performed another ANOVA on the results.

As for annotation time, the differences between the pairs of annotators with regard to inter-annotator agreement are highly significant ($p \leq 0.001$). No effect was found for the impact of the order of batches on inter-annotator agreement, while the condition of pre-annotation proved to be significant with $p \leq 0.01$. This provides more evidence for the claim that automatic pre-annotation can improve the quality of human annotation. However, one may wonder whether these results are only due to the correctly annotated instances in the pre-annotated data, meaning that automatic pre-annotation is useful only if the quality of the automatic pre-annotation is high enough.

To test this suspicion we removed all those instances from the data set where frames had been assigned correctly by Shalmaneser, and computed inter-annotator agreement on the remaining instances only. To allow for a meaningful comparison, we removed the same sentences from the batches without pre-annotation and computed inter-annotator agreement on the same subset of sentences for both, batches without pre-annotation and texts with automatic pre-annotation assigned by the state-of-the-art ASRL system. For the sake of completeness we also computed inter-annotator agreement on the sentences with enhanced pre-annotation. Please note that the latter also include correct pre-annotations and therefore are expected to show higher F1 scores for inter-annotator agreement.

Figure 7 shows the distribution of inter-annotator agreement for the different conditions of pre-annotation, NN (no pre-annotation), SS (both batches have been pre-annotated by Shalmaneser), and EE (both batches provide enhanced pre-annotation). Results for inter-annotator agreement on the automatically pre-annotated batches (SS) are clearly higher than the ones for the batches without pre-annotation (NN). Please note that these results are for instances which have been annotated incorrectly by the ASRL system. As expected, inter-annotator agreement on the sentences with enhanced pre-annotation is much higher, most probably due to the correct annotations included in the data.[8]

While the impact of automatic pre-annotation on inter-annotator agreement results computed on batches including all, correct *and* incorrect automatic pre-annotations, was statistically significant, we could not obtain a significant effect for the impact of pre-annotation on inter-annotator agreement computed on the subset of incorrectly pre-annotated instances. This means that we could not show that even falsely assigned automatic pre-annotations have a positive effect on annotation quality. What we could show, however, is that even incorrect pre-annotations do not corrupt the human annotators, and that the reliance of human annotation on these instances is demonstrably not worse than on unannotated text.

4.5 Semantic role assignment

As described in Sect. 3.5, we provided pre-annotation for frame assignment only, therefore we did not expect any significant effects of the different conditions of pre-annotation on the task of semantic role labeling. To allow for a meaningful comparison, the evaluation of semantic role assignment was done on the subset of frames annotated correctly by all annotators.

[8] The small number of instances for the enhanced pre-annotation (batch 1:8 sentences, batch 2:10 sentences, batch 3:14 sentences) did not allow for a reliable analysis of the incorrectly annotated sentences with enhanced pre-annotation.

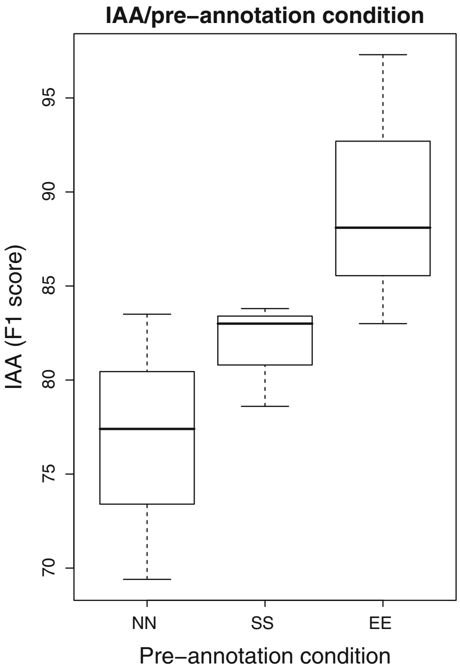


Fig. 7 Inter-annotator agreement (F1 score) on batches with different conditions of pre-annotation (NN, SS, EE), considering incorrect Shalmaneser pre-annotation instances only

As with frame assignment, there are considerable differences in annotation quality between the annotators. In contrast to frame assignment, here the differences are statistically significant ($p \leq 0.05$). Table 10 shows the average F1 score for each annotator on the semantic role assignment task.

As expected, neither the condition of pre-annotation nor the order of batches had any significant effect on the quality of semantic role assignment.[9]

5 Discussion

Our experiments showed that partial automatic pre-annotation of frames (word senses) can have a positive impact on the quality and consistency of frame-semantic annotation. We obtained a statistically significant effect for a real-life scenario where we used a state-of-the-art ASRL system for pre-annotation. We suspect that the strong interaction between the order in which the batches are given to the annotators and the annotation conditions lessen the observed effect, resulting in lower f-scores for the group of annotators who processed the ASRL pre-annotations in the first trial, where they could not yet profit from the same amount of training as the other two groups. This problem is even more severe for annotation time. While

[9] The annotation of frame and role assignment was done as a combined task, therefore we do not report separate results for annotation time for semantic role assignment.

Table 10 Annotation quality (average F1 scores) for semantic role assignment for the 6 annotators

Anot1	Anot2	Anot3	Anot4	Anot5	Anot6
85.2	80.1	87.7	89.2	82.5	84.3

for most annotators the annotation times decreased when annotating batches that had been pre-annotated with a state-of-the-art semantic role labeller, this speed-up was not statistically significant. However, we suspect that here, too, the interaction between training effect and annotation condition made it difficult to reach a significant improvement. This confirms the findings by Dandapat et al. (2009) that training often has a larger effect on both annotation time and quality than other factors such as the use of automatic pre-annotation. Nonetheless, pre-annotation can still have a noticeable positive effect if its quality is good enough.

Another possible reason why we might not observe a significant reduction of annotation time was suggested in Sect. 4.1. Pre-annotation of frames only reduces the physical effort needed to annotate the correct frame, while the cognitive effort of verifying (or determining) the right frame remains the same. This is a major difference between our annotation task and the one by Chou et al. (2006) who produced Propbank-style semantic annotations on bio-medical data (see Sect. 2). In constrast to our experiment, Chou et al. (2006) only annotate predicates with exactly one word sense, which means that the effort of disambiguating between different frames does not appear. While we assess the benefits from automatic state-of-the-art pre-annotation of frames, Chou et al. (2006) provide the correct frames and try to estimate time savings for state-of-the-art semantic role labelling on *gold* frames.

When estimating an upper bound for reducing the annotation effort, Chou et al. (2006) consider only the last step in the annotation process, namely the manual validation or correction of pre-annotated semantic roles. In our experiment, semantic role labelling is done completely by hand, and we measure annotation time for the whole task, namely frame assignment and semantic role labelling. As a result, the annotation savings reported by Chou et al. (2006) are extremely optimistic and cannot be compared to our findings.

6 Conclusion and future work

In the paper we presented experiments to assess the benefits of partial automatic pre-annotation on a frame assignment (word sense disambiguation) task. We compared the impact of (a) pre-annotations provided by a state-of-the-art ASRL, and (b) enhanced, high-quality pre-annotation on the annotation process. We showed that pre-annotation has a positive effect on the quality and consistency of human annotation: the enhanced pre-annotation clearly increased f-scores for all annotators, and even the noisy, error-prone pre-annotations provided by the ASRL system were able to improve the quality of human annotation.

In the last section we pointed out the interactions between different variables in our experimental design. Of particular concern to us is the interaction between the order of batches and the pre-annotation condition. Here a strong training effect that exists only in our experimental set-up may over-shadow the benefit from automatic pre-annotation especially with regard to reducing annotation time that we could use in the real world.

One way to address this problem with our set-up would be a further split of the test data, so that the different types of pre-annotation could be presented to the annotators at different stages of the annotation process. This would allow us to control for the strong bias through incremental training, which we cannot avoid if one group of annotators is assigned data of a given pre-annotation type in the first trial, while another group encounters the same type of data in the last trial. Due to the limited number of annotators we had at our disposal as well as the amount of time needed for the experiments we could not sort out the interaction between order and annotation conditions. We will take this issue up in future work.

Acknowledgments We would like to thank Berry Claus for extensive discussion and comments on our design. We are also grateful to our annotators Markus Dräger, Lisa Fuchs, and Corinna Schorr and to the anonymous reviewers for their insightful comments and useful feedback. Ines Rehbein and Josef Ruppenhofer are supported by the German Research Foundation DFG under grant PI 154/9-3 and Caroline Sporleder as part of the Cluster of Excellence Multimodal Computing and Interaction (MMCI).

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th international conference on computational linguistics* (pp. 86–90). Morristown, NJ, USA: Association for Computational Linguistics.

Baldridge, J., & Osborne, M. (2004). Active learning and the total cost of annotation. In *Proceedings of EMNLP.*

Brants, T., & Plaehn, O. (2000). Interactive corpus annotation. In *Proceedings of LREC-2000.*

Burchardt, A., Erk, K., Frank, A., Kowalski, A., & Padó, S. (2006). SALTO—a versatile multi-level annotation tool. In *Proceedings of LREC.*

Chiou, F. D., Chiang, D., & Palmer, M. (2001). Facilitating treebank annotation using a statistical parser. In *Proceedings of HLT-2001.*

Chou, W. C., Tsai, R. T. H., Su, Y. S., Ku, W., Sung, T. Y., & Hsu, W. L. (2006). A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of FLAC-2006.*

Dandapat, S., Biswas, P., Choudhury, M., & Bali, K. (2009). Complex linguistic annotation—no easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the third linguistic annotation workshop* (pp. 10–18). Suntec, Singapore: Association for Computational Linguistics.

Erk, K. (2005). Frame assignment as word sense disambiguation. In *Proceedings of the 6th international workshop on computational semantics (IWCS-6).* The Netherlands: Tilburg.

Erk, K., & Pado, S. (2006). Shalmaneser—a flexible toolbox for semantic role assignment. In *Proceedings of LREC*, Genoa, Italy.

Fillmore, Charles J. (1982). Frame semantics. In The Linguistic Society of Korea (Eds.), *Linguistics in the morning calm* (pp. 111–137). Seoul: Hanshin.

Fillmore, C. J., & Baker, C. (2010). A frame approach to semantic analysis. In B. Heine & H. Narrog (Eds.), *Oxford handbook of linguistic analysis.* Oxford: Oxford University Press.

Fillmore, C. J., Petruck, M. R., Ruppenhofer, J., & Wright, A. (2003). FrameNet in action: The case of attaching. *International Journal of Lexicography, 16*(3), 297–332.

Ganchev, K., Pereira, F., Mandel, M., Carroll, S., & White, P. (2007). Semi-automated named entity annotation. In *Proceedings of the linguistic annotation workshop* (pp. 53–56). Prague, Czech Republic: Association for Computational Linguistics.

Kruschwitz, U., Chamberlain, J., & Poesio, M. (2009). (Linguistic) science through web collaboration in the ANAWIKI project. In *Proceedings of WebSci'09*.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua, 115*(11), 1619–1639.

Meurers, W. D., & Müller, S. (2007). Corpora and syntax (article 44). In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics*. Berlin: Mouton de Gruyter.

Mueller, C., Rapp, S., & Strube, M. (2002). Applying co-training to reference resolution. In *Proceedings of 40th annual meeting of the association for computational linguistics* (pp. 352–359). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Ng, V., & Cardie, C. (2003). Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 conference on empirical methods in natural language processing (EMNLP-2003).*

Rehbein, I., Ruppenhofer, J., & Palmer, A. (2010). Bringing active learning to life. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*, Beijing, China.

Xue, N., Chiou, F. D., & Palmer, M. (2002). Building a large-scale annotated Chinese corpus. In *Proceedings of the 19th international conference on computational linguistics (COLING 2002).*