Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7132–7139
Marseille, 11–16 May 2020

Do you Feel Certain about your Annotation? A Web-based Semantic Frame Annotation Tool Considering Annotators' Concerns and Behaviors

Regina Stodden & Behrang QasemiZadeh & Laura Kallmeyer
Heinrich Heine University,
Düsseldorf, Germany
{stodden, zadeh, kallmeyer}@phil.hhu.de

Abstract

In this paper, we present an open-source web-based application with a responsive design for modular semantic frame annotation (SFA). Besides letting experienced and inexperienced users perform suggestion-based and slightly-controlled annotations, the system keeps track of the time and changes annotators made during the annotation process and logs certain metadata. This collected metadata can be used to get new insights regarding the difficulty of annotating specific types of frames, and as an input of an annotation cost measurement for an active learning algorithm. The tool was already used to build a manually annotated corpus with semantic frames and its arguments for task 2 of SemEval 2019 regarding unsupervised lexical frame induction (QasemiZadeh et al., 2019). Although English sentences from the Wall Street Journal corpus of the Penn Treebank (Marcus et al., 1999) are annotated for this task, it is also possible to use the proposed tool for the annotation of sentences in any other languages.

Keywords: annotation tool, semantic frames, multilingual semantic annotation tool

1. Introduction

In computational linguistics, manually annotated corpora are in high demand. In machine-learning-based natural language processing tasks, corpora with manual annotations are necessary to train, evaluate, and compare systems and algorithms in terms of quantitative measures. However, the development of manually annotated corpora is resource-intensive and complex; and, often experienced annotators and annotation tools specialized for the purpose of the annotation task are required. This paper addresses these problems in the context of semantic frame annotation by introducing a web-based open source software.

Simply put, semantic frames as used in this paper are event representations that are assigned to lexical units. A frame consists of a frame type (also known as event type) and a set of semantic roles (slots/elements), each of which links the item evoking the event to the lexical filler of the role (see Fig. 1 for an example)[1]. Frame information is useful for various natural language understanding tasks, e.g., question answering, information extraction, or text summarization. The basic idea is that the understanding of an utterance requires knowledge about the denoted events and the roles of their participants (Fillmore and Baker, 2010). The sentence (1) for instance expresses an event of "something originating from somewhere". The verbal multi-word expression (VMWE) *come_from* evokes the ORIGIN frame with the semantic roles or frame elements ENTITY and ORIGIN. The two role labels are assigned to the lexical units *Criticism* and *Wall Street* (see Fig. 1).

(1) Criticism of futures COMES FROM Wall Street.

FrameNet (Ruppenhofer et al., 2016), the most well-known frame resource, lists nearly 800 frames with the definition of the event type, definitions of suitable frame elements, and possible lexical units that can evoke the frame. The mapping between lexical units and corresponding frame types is a many-to-many relation. For example, the ORIGIN frame can also be evoked by the verbs *originate* or *date* and the verb *come* can also evoke the ARRIVING or MOTION frame (Ruppenhofer et al., 2016).

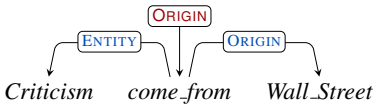

Figure 1: A graph-based representation of the frame for (1)

In order to encourage the development of corpora annotated with the rather complex FrameNet frames, we are presenting an annotation tool with a responsive and modular design. Our system aims at

(I) simplifying the annotation task to involve not only experienced annotators but also inexperienced annotators,

(II) reducing the time used for the annotation, and

(III) measuring the performance and concerns of the annotators to get insights in the annotation process.

The paper is structured as follows. Section 2 briefly reviews other (semantic frame) annotation tools. Section 3 describes our system, including its architecture, instructions to reuse and its workflow. Afterwards, Section 4, a description of an already realized use case, exemplifies the usage of the system. Limitations of the tool and points for future work are discussed in Section 5, and Section 6 concludes the paper.

[1] Throughout the paper, frame types are highlighted in red, and frame elements, i.e., semantic roles, are highlighted in blue. Lexical units are notated in italic.

2. Related Work

Annotation tools are mostly realized with a graphical or a command-line user interface. For example, Vossen et al. (2018) propose a typing-based command-line tool which might be challenging to use for users with low media literacy or persons using a mobile device.

In contrast, the SALTO annotation tool (Burchardt et al., 2006) has a graphical user interface. It was initially developed to annotate semantic roles in the context of semantic frames, but now it is more focused on annotating syntactic structure in a graphical environment.

FrameNet itself also offers a frame annotation tool with a graphical user interface[2], but it is only an online demo version for which the code is not freely available. The system provides demo annotation records with a target verb and a pre-annotated frame. On a click-based user interface a user can annotate the core and non-core frame elements. It is similar to the system proposed here in the click-based user interface as well as in the supplement of links to frame definitions and examples.

The most similar annotation tool to our system is WebAnno (Yimam et al., 2013; Eckart de Castilho et al., 2016), a web-based tool based on BRAT (Stenetorp et al., 2012) for the annotation of semantic frames and morphological or syntactical data. Further development of BRAT and WebAnno has resulted in APLenty, a system that can be used for annotating with various types of labels, in particular for sequence labeling. In order to reduce the annotation effort, the system uses active learning methods to predict the most relevant annotation records. In comparison to the annotation tool presented here, these tools focus more on the emerging annotations than on the metadata of the annotation process.

Another group of work relevant to SFA is concerned with studying the usability[3] of an annotation tool. The evaluation of the ease of use of a web based annotation tool is important because if a web interface is too difficult to use, users will leave the website (Nielsen, 2012) and are presumably not willing to annotate many data. Furthermore, a good usability helps also experienced users because it supports a efficient usage. In order to reach usability, Burghardt (2012) extends and adjusts the ten well-known heuristics of Nielsen (1994) regarding linguistic annotation tools. Most of the recommendations, which are equally relevant for semantic frame annotation, are considered in this paper and will be elaborated and referred to during the detailed system description in the following (see Section 3). Another approach to simplify the ease of use and navigation is to allow the users to restrict or filter their list of annotations (Abend et al., 2017).

Furthermore, the design criteria regarding the journaling system of the annotation are important. While other tools only save the end version of the annotation, Zeldes (2016) and Marcu et al. (1999) propose to automatically log all states of the annotation during the process (including all additions and revisions). Additionally, in the annotation tools of, e.g. Ringger et al. (2008) and Tomanek and Hahn (2009), the time per annotation step is also stored.

3. System Description

The semantic frame annotation tool presented in this paper (SFA) is an open-source, web-based application with a responsive design[4] for a modular semantic frame annotation following the guidelines of FrameNet version 1.7. The annotation process is modular in the sense that it is separated into smaller subtasks or steps, the results of which are separately stored.

In the following, Section 3.1 introduces the conceptual design of the tool. In Section 3.2, the architecture and use of the annotation system are described. Finally, the workflow of the annotation process is introduced from the users' and admins' perspective, starting with the preparation and ending with the evaluation of the generated annotated corpus (see Section 3.3).

3.1. Design

SFA is designed as a click-based system so that the system is easy to use for inexperienced users because of its conventionalized interaction mechanisms (Burghardt, 2012, Recommendation 18). Because of the clicks, the annotation is quick and easy; for each subtask of the annotation process, only a few clicks are needed (select annotation body, mark annotation, apply annotation) (Burghardt, 2012, R22), which make the usage effective. The click-based user interface and the division in smaller subtasks also have the advantage of a potential usage on a portable device with a smaller screen.

The visualization design is consistent during all steps. This consistent step-separated design guides the users through the complete annotation process and hampers to forget an intermediate step. A screenshot of the tool at the annotation step of frame annotation is provided in Figure 2.

Burghardt (2012, R21;25;27) proposes that a linguistic annotation tool should be able to distinguish and easily switch between different annotation layers visually. In SFA, this is addressed with sliders on the right of the sentence box (see part III in Figure 2) and color divided highlighting of the layers, e.g., *Wall Street* in Figure 2 is highlighted as a frame element (blue) and multi-word unit (yellow). Additionally, the frame element annotation layer is also shown in the frame structure (see part IV in Figure 2).

Furthermore, the annotation tool automatically keeps track of any changes (including all additions and revisions per step) to the annotation records instead of simply maintaining their current state– also referred to as journaling system –, as proposed in Zeldes (2016) or Marcu et al. (1999). This procedure facilitates getting insights in the annotation process, e.g., keeping track of revised and finally chosen frames of a user per annotation record.

In addition to the logged changes, the time per step and the

[2] `https://framenet.icsi.berkeley.edu/fndrupal/annotation_tool`

[3] The usability is the extent to which a system can be used to achieve a goal in a specified context of use with consideration of effectiveness, efficiency, and satisfaction (ISO, 2010).

[4] If a webpage can be used without any problems on both large screens, e.g., a desktop used with a mouse, as well as on small screens, e.g., tablets with touchscreens, its design is called responsive.

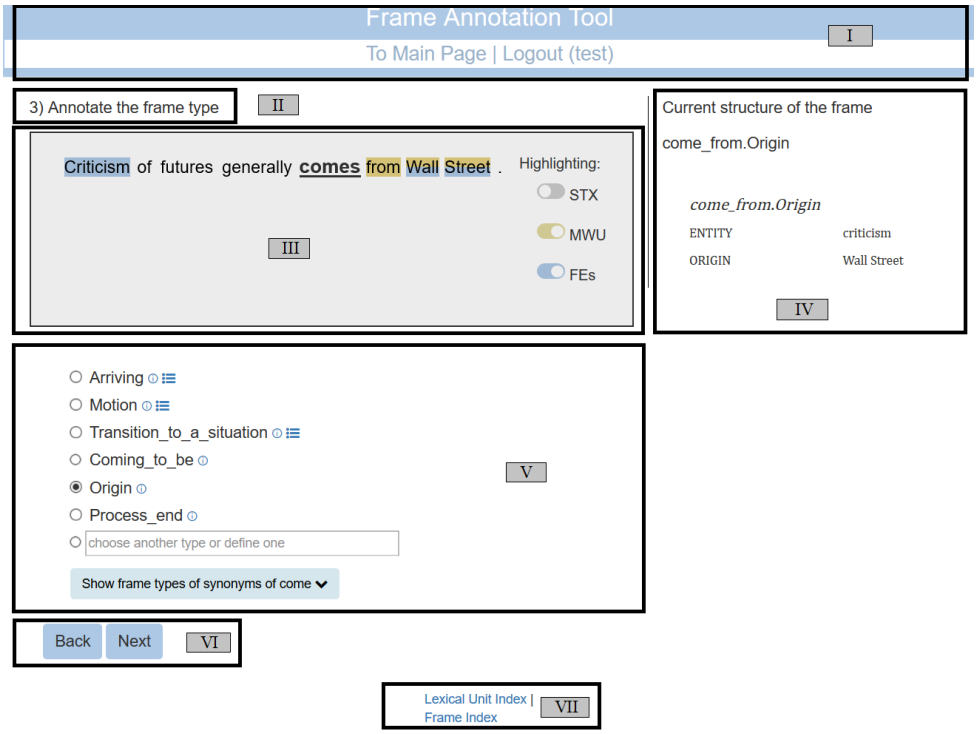

Figure 2: Screenshot of the annotation tool at the step of frame annotation. The window is split into seven parts: I) header, II) instruction, III) sentence with highlighted annotation layers, IV) frame structure, V) step specific actions (here choosing a frame), VI) navigation, and VII) additional resources.

number of changes per annotation are saved[5]. According to Ringger et al. (2008), the data could be helpful to compare 'inter-annotators performance' or could be used to calculate annotation costs for a (pro-)active learning algorithm[6]. Based on the annotation cost, the most costly or hardest annotation records would be assigned next to the annotators, so always the most informative annotation records would be picked. The annotation cost is mostly measured based on sentence length or word length (Arora et al., 2009). Tomanek and Hahn (2010) show that the usage of an annotation cost based on timestamps can help to create a high-quality corpus of annotated named entities with less amount of annotation records and less time spent on annotation than with randomly picked annotation records. Further investigation is needed to test if the number of changes and the number of annotations for a record can also enhance the development of a corpus with less effort.

Furthermore, the time per annotation record can yield information regarding the difficulty of an annotation. To enhance this, we also ask the annotators about their concerns regarding each annotation record after finishing an annotation. This measurement facilitates a high quality of annotations. Furthermore, if the records are sorted by their concern labels, the difficulty of the annotation of different frames can be compared.

3.2. System Architecture, Download and Demo Site

The annotation system proposed here is a web-based application, a current online version of it is accessible at `http://sfa.phil.hhu.de:8080/`. The tool works best with the Firefox web browser[7]. On other web browsers some visual components working with HTML5 mark-ups may not be adequately rendered. Furthermore, JavaScript is needed, which is normally enabled in web browsers by default. With the user account *test* (password: *guest123*), one is able to navigate through the annotation records[8], annotate a few of them and play with the functionality of the tool.

SFA is based on a Python web framework called Django (Django-Software-Foundation, 2019) and implemented with MySQL, HTML, CSS and JavaScript. Due to the design of Django, no expert knowledge of databases, server, or HTML are required to reuse the tool.

The code of the tool is licensed under the MIT license and is available at `https://github.com/rstodden/semantic_frame_annotation_tool`. The required steps to use the tool after a download, e.g., filling the database, are described in Section 3.3.1.

[5] All data which are connected to a user account can always be requested, the tool respects The EU's regulations concerning privacy (DSGVO in Germany; i.e., in English, General Data Protection Regulation, GDPR).

[6] A pro-active learning algorithm, is catered for experienced as well as inexperienced users (Nghiem and Ananiadou, 2018) so that a fault tolerance is included. The implementation of both features is planned for the next version of SFA.

[7] `https://www.mozilla.org/de/firefox/`

[8] In the following an *annotation record* describes a sentence with a highlighted target verb which evoking frame should be annotated. If a sentence contains more than one target verb, each combination of a verb with its sentence corresponds to one annotation record.

3.3. Workflow

The workflow of SFA is split into four phases,

(I) preparation phase: first steps to (re-)use the tool,

(II) annotation phase: users annotate the annotation records,

(III) reviewing phase: reviewers or users review and potentially revise the annotated records,

(IV) evaluation phase: exporting the completed annotated corpus and analyzing the metrics regarding the users' behavior.

A detailed description of each phase follows.

3.3.1. Preparation Phase

In the preparation phase, the administrator has to make everything ready for the usage of the tool. In the download package of the tool, a database with two user accounts (*admin* and *test*, both with the password *guest123*), an example annotation record and its referred tokenized sentence are already included.

The admin can add more records by uploading files in the admin interface. The upload is split into (a) the annotation records in a specified format[9] as recommended in Burghardt (2012, R12-14), and (b) the referred sentences in the CONLL-U format[10].

The records file can either contain raw annotation records, which indicate only the sentence and the annotating verb, or pre-annotated records, which also include a name of a frame and optionally pre-annotated frame elements. Following the assumption of Haertel et al. (2008) and Ringger et al. (2008), it is costlier to annotate data from scratch than revising pre-annotated data, so we recommend to use pre-annotated data even if priming is possible.

After the upload, the admin assigns the annotation records to one or more users using the provided feature in the admin interface. In an extended version of SFA, new records will be automatically assigned to the users using an active learning algorithm.

3.3.2. Annotation Phase

As mentioned before, in the annotation phase, an annotation record will be annotated following distinct annotation steps:

0. *reading the annotation guidelines* (QasemiZadeh and Petruck, 2018 2019), which contain instructions on the annotation tool, and an introduction to frame semantics,
1. *picking an annotation record* by optionally filtering the records by their frame, verb, etc. (similar to Abend et al. (2017)) start, continue or review the annotation of a record,
2. *reading* the sentence and mental *comprehension* of the frame,
3. *choosing a frame* (e.g., ORIGIN in Figure 1), which fits best to the usage of the target verb, out of a list of most likely frames. Each frame is provided with a link to its definition and proved examples by FrameNet (see part V in Figure 2),
4. *annotating frame elements* or arguments (e.g., Criticism – ENTITY in Figure 1) by choosing suitable labels for the lexical units[11] and annotate whether they are a reference, and if yes, which kind of,
5. *rating and commenting* on the annotation to make the annotation more transparent and comprehensible for reviewers and the users themselves.

In each of the five steps, instructions for the users are provided. Furthermore, each step is slightly controlled so that the users cannot lose the thread or forget an important step. Even if the user interrupts the annotation for any reason, the tool will resume at the breakpoint due to the underlying journaling system. A more detailed description of each step follows.

Guidelines & Records In the beginning, all users are provided with annotation guidelines (QasemiZadeh and Petruck, 2018 2019), which contain instructions on the annotation tool, as well as an introduction to frame semantics. In addition to the introduction for beginners, the guidelines can support congruence in the annotation between beginners as well as experts.

Picking a Record Before starting with the annotation, the user has to pick an annotation record. The users can decide if they want to start the annotation of a new annotation record from the beginning (unannotated frames), continue their annotation (annotations in progress), review an already completed record (annotated records), or try to understand a previously skipped record (skipped frames). Similar to the annotation tool proposed in Abend et al. (2017), the users can restrict or filter their list of annotations to a verb, (pre-)annotated frame, or concern.

Reading and Comprehension In the first step of the annotation, only the sentence of the annotation record with the highlighted target verb is shown and all annotation options are disabled to ensure the focus on the comprehension of the sentence. The annotators decide if they understand the sentence, especially the meaning of the verb, and its semantic function. A confirmation redirects to the next annotation step, whereas the contrary terminates the annotation of this record and moves it to the list of skipped records.

Choosing a Frame Next, the annotators identify and annotate the frame type whose definition fits best to the usage of the target verb in the annotation record, e.g., ORIGIN in Example 1. In this context, the users might annotate the target verb as a verbal multi-word expression, e.g., *come from* like in Example (1), or annotate the slot fillers as multi-word units (MWUs), e.g., *Wall Street* like in Example (1) or

[9] The input format of the annotation records is described in detail in QasemiZadeh et al. (2019).

[10] `https://universaldependencies.org/format.html`

[11] The list of suitable labels is limited by the previous chosen frame. If a frame is changed afterwards, only the suitable frame elements are kept to minimize errors in the annotation. Additionally, the annotation tools warns if a label is illicitly used twice for different lexical units.

$ 100[12]. Because picking the frame out of the list of all 800 frames defined by FrameNet is inefficient, the annotation tool assists in this step in different ways: First, similar to FrameNet's demo annotation tool, the frame definition and annotated examples of each frame (in combination with the verb)[13] are provided besides their listing point in the list of options (see Figure 2 box V).
Second, the choices of frames are restricted to the list of the most likely frames regarding the target verb. This list contains all frames which are evoked by the target verb following FrameNet or following previously made annotations. Nevertheless, the user still has the possibility to choose another frame, based on a provided list of synonyms and their suitable frames or based on a search through all frames supported with an auto-complete function.
SFA "learns" from the made annotations, if a chosen frame is not evoked by the target verb (following FrameNet), but the user feels confident that his/her/their choice is correct, the system will remember this decision and will suggest this frame for all other records with the same target verb.

Annotating Frame Arguments In the third step, the arguments or frame elements (FE) are annotated, e.g., the slot fillers *Criticism* and *Wall Street* with the labels ORIGIN and ENTITY (see Figure 1). Concretely, the annotators click on a token or a previously annotated MWU in the sentence box (see Figure 2 box III) and can either add a FE or adapt a FE by changing its role label, or remove a FE (see Figure 3). The easy modification or deletion of annotations that is possible here is in line with the recommendations for an easy to use user interface by Burghardt (2012, R20).
Besides the annotation of the slot filler's label, the users also annotate whether and which kind of reference is expressed by a slot filler. A selection out of labels (R) for a pronoun whose reference is resolved in the sentence, (D) for a pronoun whose reference is not resolved in the sentence, i.e., has to be resolved in the discourse, and (C) for an expression whose reference requires coercion in order to be resolved, i.e., semantic reinterpretation, is possible (see Figure 3). An example of coercion is "Wall Street" of Example 1, since one needs to understand that the criticism comes from one or more leading American financial institutions, perhaps located at the Wall Street.
An annotated FE is highlighted in the sentence box and added to the current frame structure (see Figure 2 box IV). The administrator can previously decide whether all FEs should be annotated or only all core FEs. Core FEs are all FEs which are essential for the meaning, non-core FEs contribute additional information (Ruppenhofer et al., 2016).
Furthermore, the annotation tool keeps track of whether the annotation of the user is valid, e.g., the system emits a warning if a FE is assigned more than once, to different role fillers, which is quite unusual. SFA also automatically deletes FEs on a frame change if the previous FEs are not suitable for the new frame type. Potential careless mistakes could be minimized through this procedure.

Rating, Commenting and Revising The last step of the annotation is rating, commenting and revising. In addition to the time measured during the annotation, the users can give a self-measured value regarding their annotation in this step. The users give feedback for each record of how confident they are about their annotation and how well the annotation fits the frame definition. In an optional comment, they explain their concerns or doubts to make the annotation more transparent and comprehensible (also for themselves). In an extended version of SFA, the self-measured concern value could be used in combination with the automatically derived metadata, e.g., time spent on an annotation record, to get more insights into the difficulty of the annotation task and to calculate a more precise annotation cost.
This idea regarding difficulty measurement is derived from usability tests. In post-task tests, participants rate the difficulty of a task to get insights into problems of a system, conveyed to the annotation task it could indicate intricate frames or target verbs. In an extended version of SFA, these metadata could be enriched with more *psychometric data*, such as derived by eye trackers. This would produce more insights into the process, which could help for further interpretation of the annotated linguistic data or language processing.

3.3.3. Reviewing Phase

In the reviewing phase, a user can verify the annotated records of other users or the admin can calculate an inter-annotator agreement. Therefore, all completely annotated frames of all users are listed in a human-readable way[14]. If the users want to change the annotation of a record, they can either change their own annotation, or they can create a copy of the wrong record into their own user account and change the annotation there. Afterwards, the admin can choose the annotation records with a double agreement or a high inter-annotator agreement for the final corpus.

3.3.4. Evaluation Phase

In the evaluation phase, the admin user can download and analyze the resulting corpus. In the analysis, the performance and self-measured metrics could be compared regarding different frames, verbs or users. For example, the users' performance could be compared with each other and could be presented on a ranking board. This gamification approach might motivate the users to annotate more.
In addition, correlation tests between human behavior during the annotation of a frame and a semantic frame parser's prediction of the same annotation records or frames could be interesting.

4. Case Study

The proposed annotation tool was used to build a manually annotated corpus of semantic frames for the SemEval 2019

[12]Detailed information regarding the choice of a VMWE or MWU are provided in the annotation guidelines (QasemiZadeh and Petruck, 2018 2019).

[13]For example the definition of the frame Activity_start is accessible at `http://corpora.phil.hhu.de/framenet/fndata-1.7/frame/Activity_start.xml` and the annotation report in combination with the verb begin at `http://corpora.phil.hhu.de/framenet/fndata-1.7/lu/lu4448.xml?mode=annotation`.

[14]See here for an example of the human-readable annotations `http://corpora.phil.hhu.de/fi/frames.html`.

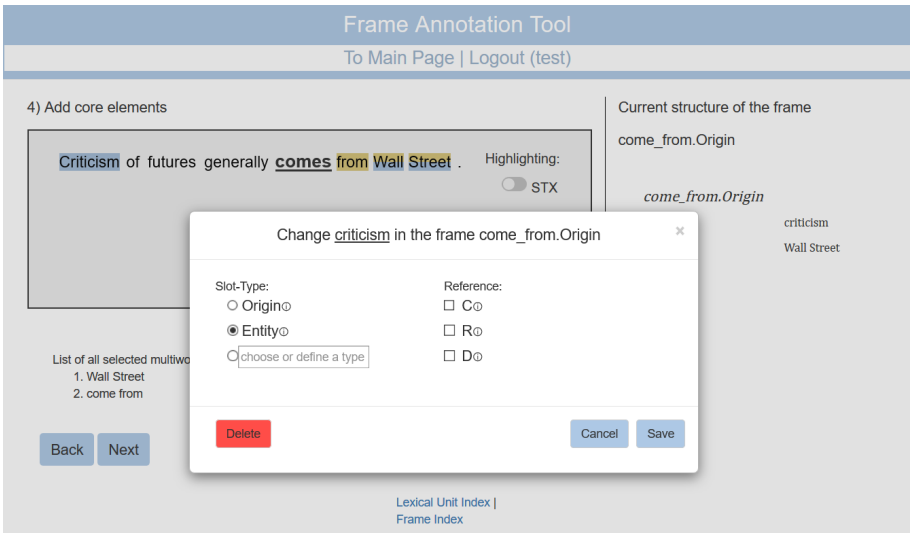

Figure 3: Screenshot of the annotation tool at the step of frame element annotation.

task 2 (QasemiZadeh et al., 2019). In this context, 4,620 double-checked annotation records and 9,510 frame elements of 3,800 English sentences from the Wall Street Journal corpus of the Penn Treebank (Marcus et al., 1999) were annotated and revised by different users, some of whom were experienced as well as some inexperienced with semantic frames. Overall, this process took roughly 539h and span 68.784 annotation steps, which are all logged by the system.

Annotator Activity	**Time**	**Moves**
Reading and Comprehension	78	4,795
Choosing a Frame	177	9,737
Annotating Frame Elements	81	19,510
Rating, Revising, Commenting	115	25,793
Multi-word Unit Annotation	89	8,949
Total	539	68,784

Table 1: Total hours and number of moves for each annotation step for the 4,620 record dataset. Move is the total number of logical moves for each annotation step between all annotators and the annotation system, i.e., logged changes in the process of frame and core FE annotation.

Table 1 shows the amount of effort to develop the SemEval dataset in terms of time and moves that the annotation system has recorded. Overall, the choice of an adequate frame was the most time consuming step (177 h), which can be explained by the fact that it involved comparing similar sentences or reading frame definitions. The highest numbers of moves or changes are recorded for the rating and commenting step because this step also includes the clicks for checking or revising an annotation.
In addition to the time and moves used per annotation step, the confidence per frame or annotation record can help to estimate the difficulty of the annotation task and quality of the created data. As shown in Table 2, the confidence regarding a frame is unrelated to its number of annotation records or the number of distinct verb forms evoked by the frame.

Frame Type	**#VF**	**#Rec**	**Conf**
DECIDING	1	13	4.31
AGREE_OR_REFUSE_TO_ACT	1	15	4.13
TAKE_PLACE_OF	1	11	4
BEING_EMPLOYED	1	6	4
STATEMENT	8	149	3.97
TAKING_SIDES	3	16	3.88
ACTIVITY_STOP	4	16	3.88
COMMERCE_SELL	6	168	3.82
BRINGING	1	5	3.8
GIVE_IMPRESSION	4	39	3.79

(a) Frames with Highest Average Confidence

Frame Type	**#VF**	**#Rec**	**Conf**
BEING_IN_CONTROL	2	5	1.6
COMING_TO_BE	2	5	1.8
OPERATING_A_SYSTEM	2	10	1.8
AWARENESS	1	6	1.83
REMOVING	3	8	1.88
INTENTIONALLY_CREATE	6	19	1.95
CERTAINTY	1	68	2.03
OPINION	2	91	2.1
THWARTING	2	22	2.32
FIRST_RANK	1	21	2.38

(b) Frames with Lowest Average Confidence

Table 2: Frame types with the highest (2a) and the lowest (2b) confidence (**Conf**) by the number of records (**#Rec**) with double annotator agreement. **#VF** reports the number of distinct verb forms that evoke a frame.

Furthermore, the confidence value facilitates the interpretation of the performances of the frame induction systems from the shared task on the created data. In a primary analysis, we observed a strong uphill positive correlation using Spearman's rank correlation (Spearman's Rho; $r = 0.75$, $p <= 0.05$, $n = 3$) between the confidence value of the human annotators and the automatic annotation of the proposed systems of the SemEval task (see Table 3). More concretely, if an annotation record was easy to annotate for a human, the annotation systems also achieved better results (QasemiZadeh et al., 2019).

Cnf	#I	(Arefyev et al., 2019)	(Anwar et al., 2019)	(Ribeiro et al., 2019)
1	286	73.79	70.57	67.70
2	677	66.45	63.80	60.46
3	1,115	76.71	75.98	70.01
4	2,458	76.65	74.05	73.45
5	84	86.14	84.65	85.13

Table 3: Results of the systems of the SemEval task 2 (columns) subtask A grouped by the annotators' confidence (rows) regarding the annotation records. The number of the records per confidence level is presented in the second column.

5. Limitations and Future Work

Even though the SFA tool has been successfully employed in production, still it has certain limitations. So far, the system has been limited to frames with the depth of one; the annotation of recursive/nested frames would be interesting (or even necessary) addition to the tool. Furthermore, lexical units such as verbs (i.e. per annotation record) are allowed to be annotated only with one frame, which may not be sufficient for preserving ambiguity. When annotating a lexical item in a sentence, for a clearer interpretation of the sentence meaning, the context of the sentence could be displayed next to it.

Another use case would be to test the SFA tool with FrameNet editions in languages other than English, e.g., French, Japanese, etc.[15]; similarly, other inventories of semantic role labels, e.g., VerbNet, can be loaded into the tool. Furthermore, in place of the fine-grained FrameNet frames, the annotation of more coarse-grained or universal frames would be interesting. Additionally, the metadata collected during the annotation process can be extended with eye and/or mouse tracking to get information on the incidents during the time spent on the task, e.g., shift of opinion regarding frames or frame elements.

We plan to conduct a usability study to verify the implementations of the usability recommendations regarding linguistic annotation tools, e.g. as in (Burghardt, 2012). The result of a usability study can also help to identify usability concerns of SFA by measuring the users' satisfaction and perception regarding the annotation tool.

6. Conclusion

In this paper, a modular, suggestion-oriented, web-based, open-source annotation tool with a responsive design is presented. The annotation tool can be predominantly used for semi-controlled semantic frame annotation by both experienced as well as inexperienced annotators. It is shown that an assistive and suggestion-based user interface considerably helps to simplify the complex task of semantic frame annotation.

Furthermore, the advantages of a journaling system, which keeps track of any changes to the annotation records instead of simply recording their current state, and the advantages of the metadata of the annotations are presented, e.g., time spent and changes made during annotations as well as concerns of the annotation. This metadata leads to promising insights regarding comparisons between the difficulty of annotating different frames and their correlations to machine learning systems performance, which are using the resulting annotated data.

[15]The French FrameNet is accessible at `http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml` and the Japanese FrameNet at `http://sato.fm.senshu-u.ac.jp/frameSQL/jfn23/notes/index2.html`.

Acknowledgments

This work was supported by the CRC 991 "The Structure of Representations in Language, Cognition, and Science" funded by the German Research Foundation (DFG). This research was also part of the PhD-program "Online Participation", financed by the North Rhine-Westphalian funding scheme "Forschungskolleg".

7. References

Abend, O., Yerushalmi, S., and Rappoport, A. (2017). UCCAApp: Web-application for syntactic and semantic phrase-based annotation. In *Proceedings of ACL 2017, System Demonstrations*, pages 109–114, Vancouver, Canada, July. Association for Computational Linguistics.

Anwar, S., Ustalov, D., Arefyev, N., Ponzetto, S. P., Biemann, C., and Panchenko, A. (2019). HHMM at SemEval-2019 task 2: Unsupervised frame induction using contextualized word embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 125–129, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Arefyev, N., Sheludko, B., Davletov, A., Kharchev, D., Nevidomsky, A., and Panchenko, A. (2019). Neural GRANNy at SemEval-2019 task 2: A combined approach for better modeling of semantic relationships in semantic frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 31–38, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Arora, S., Nyberg, E., and Rosé, C. P. (2009). Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 18–26, Boulder, Colorado, June. Association for Computational Linguistics.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., and Pado, S. (2006). Salto - a versatile multi-level annotation tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Burghardt, M. (2012). Usability recommendations for annotation tools. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 104–112, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Django-Software-Foundation. (2019). Django.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated

annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Fillmore, C. J. and Baker, C. (2010). A frame approach to semantic analysis. In Bernd Heine et al., editors, *Oxford Handbook of Linguistic Analysis*, chapter 13, pages 313–341. OUP.

Haertel, R., Ringger, E., Seppi, K., Carroll, J., and McClanahan, P. (2008). Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 65–68, Columbus, Ohio, June. Association for Computational Linguistics.

ISO. (2010). ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.

Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Towards Standards and Tools for Discourse Tagging*.

Nghiem, M.-Q. and Ananiadou, S. (2018). Aplenty: annotation tool for creating high-quality datasets using active and proactive learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 108–113, Brussels, Belgium, November. Association for Computational Linguistics.

Nielsen, J. (1994). Heuristic evaluation. In *Usability Inspection Methods*, pages 25–62, New York. Wiley.

Nielsen, J. (2012). Usability 101: Introduction to usability. Accessed: 2019-01-29.

QasemiZadeh, B. and Petruck, M. R. L. (2018–2019). Guidelines for frame and frame element identification: Hhud semantic frame annotation system. corpus annotation guidelines TR.9.2018, SFB991 - ICSI. https://user.phil.hhu.de/ stodden/Frame_Annotation_Instruction_v1.pdf.

QasemiZadeh, B., Petruck, M. R. L., Stodden, R., Kallmeyer, L., and Candito, M. (2019). SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Ribeiro, E., Mendonça, V., Ribeiro, R., Martins de Matos, D., Sardinha, A., Santos, A. L., and Coheur, L. (2019). L2F/INESC-ID at SemEval-2019 task 2: Unsupervised lexical semantic frame induction using contextualized word representations. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 130–136, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Ringger, E., Carmen, M., Haertel, R., Seppi, K., Lonsdale, D., McClanahan, P., Carroll, J., and Ellison, N. (2008). Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Tomanek, K. and Hahn, U. (2009). Timed annotations — enhancing MUC7 metadata by the time it takes to annotate named entities. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 112–115, Suntec, Singapore, August. Association for Computational Linguistics.

Tomanek, K. and Hahn, U. (2010). Annotation time stamps — temporal metadata from the linguistic annotation process. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Languages Resources Association (ELRA).

Vossen, P., Fokkens, A., Maks, I., and van Son, C. (2018). Towards an open dutch framenet lexicon and corpus. In *Proceedings of LREC 2018 Workshop International FrameNet Workshop 2018. Multilingual Framenets and Constructicons*, pages 75–80, Miyazaki, Japan, 5.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zeldes, A. (2016). rstWeb - a browser-based annotation interface for rhetorical structure theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, San Diego, California, June. Association for Computational Linguistics.

8. Language Resource References

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Treebank-3 LDC99T42*. Linguistic Data Consortium, Philadelphia. Web Download.

QasemiZadeh, B., Petruck, M. R. L., Stodden, R., Kallmeyer, L., and Candito, M. (2019). SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.