Treebanking Finnish

Katri Haverinen,[1,3] Timo Viljanen,[1] Veronika Laippala,[2]
Samuel Kohonen,[1] Filip Ginter[1] and Tapio Salakoski[1,3]

[1]Department of Information Technology,
[2]Department of French studies
[3]Turku Centre for Computer Science (TUCS)
20014 University of Turku, Finland
first.last@utu.fi

Abstract

In this paper, we present the current version of a syntactically annotated corpus for Finnish, the Turku Dependency Treebank (TDT). This is the first publicly available Finnish treebank of practical size, currently consisting of 4,838 sentences (66,042 tokens). The treebank includes both morphological and syntactic analyses, the morphological information being produced using the FinCG analyzer, and the syntax being human-annotated in the Stanford Dependency scheme. Additionally, we conduct an experiment in automatic pre-annotation and find the overall effect positive. In particular, pre-annotation may be tremendously helpful in terms of both speed and accuracy for an annotator still in training, although for more experienced annotators such obvious benefit was not observed.

In addition to the treebank itself, we have constructed a custom annotation software, as well as a web-based interface with advanced search functions. Both the treebank, including the full edit-history with exact timings, and its associated software are publicly available under an open license at the address http://bionlp.utu.fi.

1 Introduction

The applications of treebanks and their benefits for natural language processing (NLP) are numerous and well-known. Many languages, regardless of how widely spoken, already have a treebank, and for many others one is currently being developed. Finnish is among the less fortunate languages in the sense that it previously long lacked a publicly available treebank entirely. Even now, prior to this work, the only such treebank is our previously published small-scale treebank [3], which does not yet truly enable NLP research.

In this work, we aim to address the serious lack of NLP resources for Finnish, by extending our previous work into a freely available, practically sized treebank

for Finnish, the Turku Dependency Treebank (TDT). The current, extended version of the treebank presented in this paper includes 4,838 sentences. The whole treebank has manually created syntax annotations in the well-known Stanford Dependency (SD) scheme [1, 9] and automatically created morphological analyses. The text of the treebank is drawn from four sources: the Finnish Wikipedia and Wikinews, popular blogs and a university web-magazine.

As a second contribution, we also conduct an experiment on the effect of automated pre-annotation on annotation speed and quality, by using a preliminary statistical parser induced from the treebank to produce an initial analysis.

The linguistic aspects of the work, such as the choice of the annotation scheme and the modifications needed to accommodate the specific features of the Finnish language have been thoroughly discussed in our previous paper on the first release of the treebank [3]. In particular, we have found the Stanford Dependency scheme suitable for the Finnish language, with only minor modifications needed. Thus this paper will rather focus on the annotation process point of view of the work.

2 Related Work

The only publicly available treebank of general Finnish is our previously released treebank version [3]. This version only consists of 711 sentences, and, unlike the extended treebank release presented here, lacks morphological information. Also, no inter-annotator agreement figures were presented for this previous release. In addition to the general Finnish treebank, there exists a recently published small-scale treebank and PropBank of clinical Finnish [4]. The size of this corpus is 2,081 sentences (15,335 tokens), and it includes morphological, syntactic and semantic annotation.

Due to the lack of a large, publicly available treebank, also Finnish NLP tools are scarce. Tools targeted at Finnish morphology include FinTWOL and FinCG, a commercial morphological analyzer and a constraint grammar parser that resolves morphological ambiguity [5, 6]. These tools are used in this work to provide morphological analyses for the treebank. The only previously available broad-coverage syntactic parser for Finnish is Machinese Syntax,[1] which is a closed-source commercial parser.

The syntactic representation scheme used in this work, the Stanford Dependency (SD) scheme [1, 9], is relatively widely used in NLP applications. Both the above mentioned treebanks of Finnish use this scheme and additionally, there is a third treebank that has native SD annotation. The BioInfer [13] treebank is an English language corpus of scientific abstracts in the biomedical domain. In addition to these native corpora, also any English language treebank that uses the Penn Treebank scheme [8] can be automatically converted into the SD scheme using existing tools[2].

[1] http://www.connexor.eu

[2] http://nlp.stanford.edu/software/lex-parser.shtml

3 Treebank Text

In the current version of the treebank, there are four distinct sources of text: the Finnish Wikipedia and Wikinews, popular blogs and a university web-magazine. These sources are selected on the basis of two criteria.

First, it is our fundamental belief that the treebank should be freely available under an open license, which restricts our choice of texts to those which either are published under an open license originally or for which we can, with reasonable effort, negotiate such a license. Three of the current sources, Wikipedia, Wikinews and the university web-magazine, were originally published under an open license, and for the blog texts, we have obtained the permission to re-publish the text from individual authors.

Second, we have strived for linguistic variety in the texts. We have specifically limited the amount of text chosen about the same topic, and by the same author. In Wikipedia, this is naturally achieved by choosing the articles randomly, as both the amount of articles and the amount of authors are large. In the Finnish Wikinews, the number of authors is substantially smaller, and thus we have, when choosing an article from this source, first randomly selected an author, and only after that randomly selected one individual article by them. This selection process was repeated until a sufficient amount of articles had been chosen.

When selecting the blog texts, we have used several lists of most popular blogs and only selected blogs where the entries appeared to be of sufficient grammatical quality to allow proper annotation of syntax. In addition, the blogs were divided into categories, based on which topic the majority of the entries were about, and the amount of blogs selected from each category was limited. The current selection consists of two blogs from the category *personal and general*, one from the category *style and fashion* and one from the category *relationships and sex*. Naturally, this selection was affected by the permissions given by the authors. The individual texts were selected starting from the newest entries, discarding entries containing certain problematic properties, such as long quotes which could cause copyright issues. We limited the amount of text to be selected from one blog author to be approximately 200 sentences, so that individual entries from a blog were selected in order until the total amount of sentences surpassed 200. Thus the amount of entries chosen from each blog varies according to the length of the entries in that blog.

In the case of the university web magazine, articles were selected starting from the newest writings. Given the relatively limited topics, this section was restricted to a total of 50 articles, which results in a total of 942 sentences.

The breakdown of articles and sentences in different sections of the treebank is shown in Table 1. The table shows that the largest section of the treebank is currently the Wikipedia section, followed by the Wikinews section and the university web-magazine section. The smallest section is at the moment the blog section, which is due to the difficulty of gaining re-publication permissions from individual authors. Altogether the current version of the treebank consists of 4,838 sentences

Section	articles	sentences	tokens
Wikipedia	199	2,260	32,111
Wikinews	67	760	9,724
Blogs	32	876	10,918
Web-magazine	50	942	13,289
Total	348	4,838	66,042

Table 1: Breakdown of the treebank sections. As the annotation work still continues, it should be noted that the current breakdown of sections does not reflect the final composition of the treebank.

(66,042 tokens). Out of all sentences, 5.8% are non-projective. For comparison, the clinical Finnish treebank [4] is reported to have a non-projectivity rate of 2.9% of sentences.

In all sections of the treebank, we have annotated each selected text that is shorter than 75 sentences in its entirety. As for instance some Wikipedia articles may be as long as 300 sentences, longer texts have been truncated after the first 75 sentences to avoid biasing the treebank towards the topics of long texts. This strategy was also used in the construction of the first treebank release.

4 Syntax and Morphology Annotation in the Treebank

4.1 Syntax in the SD Scheme

Our choice for the syntactic representation scheme, the established Stanford Dependency (SD) scheme of de Marneffe and Manning [1, 9], is naturally the same as that used in our previous work. It is a dependency scheme, where syntax is represented as a graph of directed, labeled dependencies between words. The scheme has four different representation variants, which include a different subset of dependency types each. In effect, these variants are layers of dependencies that can be added on top of the basic dependency tree, to offer deeper information on the structure of the sentence. Therefore, the structures in all variants are not necessarily trees. The reader is referred to the original work of de Marneffe and Manning for further details on the SD scheme.

The current annotation of the treebank is based on the so called *basic* variant, where the analyses are trees and the dependencies are for the most part syntactic. The original *basic* variant of the SD scheme includes 55 dependency types, and our modified version 44 types. An example of a syntactic analysis of a Finnish sentence in the SD scheme is given in Figure 1.

In our previous work [3], we have shown that although originally designed for English, the SD scheme is well-suited for Finnish as well, with some minor changes. The reader is referred to this work for details of the modifications made to the original SD scheme, as the version of the scheme used in the current work is identical.

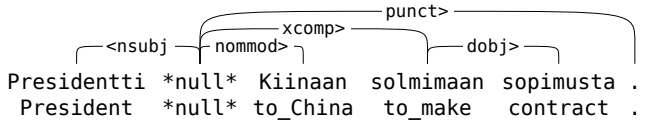

Figure 1: An example of a Finnish sentence annotated in the SD scheme. The sentence can be translated as *The president to China to make a contract*. The *null token* present in the analysis stands for the main verb which this fragmentary sentence lacks but which is necessary for the construction of a dependency analysis in the SD scheme.

word	transl.	lemma	POS	comp.	case	tense	voice	num.	person
Ryöstäjä	Burglar	**ryöstäjä**	**N**		**NOM**			**SG**	
poistui	leave	**poistua**	**V**			**PAST**	**ACT**		**SG3**
pimeän	darkness	**pimeä**	**N**		**GEN**			**SG**	
	dark	pimeä	A	POS	GEN			SG	
turvin	chub	turpa	N		INS			PL	
	safety	**turva**	**N**		**INS**			**PL**	

Figure 2: FinTWOL and FinCG analyses. The words of the sentence and their translations are given in the two leftmost columns and the lemma in the third column, followed by all tags given to the word by FinTWOL. The readings selected by FinCG are shown in bold. The example sentence as read from the leftmost column can be translated as *The burglar left in the safety of the darkness*.

4.2 Morphology with FinTWOL and FinCG

We also add to the whole treebank, including our previously released subcorpus, morphological analyses created using two Finnish morphology tools: FinTWOL and FinCG[3]. For each word, FinTWOL gives all possible readings, each of which includes a detailed morphological analysis. Given the analysis by FinTWOL, FinCG aims to disambiguate which of the readings is correct in the current context. When unable to fully disambiguate a word, FinCG may select multiple readings. In the treebank, each token is given all of its FinTWOL readings, and those selected as correct by FinCG are marked. An illustration of the morphological information present in the treebank is given in Figure 2.

Manually annotated morphological analyses are currently left as future work, pending further investigation of the various issues involved, such as morphological analyzer licensing, defining a suitable annotation scheme and, naturally, funding.

[3] http://www.lingsoft.fi

Section	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Overall
Wikipedia	95.1	84.0	90.4	-	89.5
Wikinews	96.3	87.7	-	-	92.0
Blogs	94.6	86.4	-	-	90.5
Web-magazine	96.6	89.5	92.0	70.6	88.6
Overall	95.5	86.2	90.8	70.6	**89.9**

Table 2: The inter-annotator agreement of different annotators across the sections of the treebank. All agreement figures are given in labeled attachment scores (%). Note that the LAS is calculated across all tokens and the averages in the table are thus implicitly weighted by the size of the various sections and annotator contributions. Therefore the overall figures are not the same as the averages of the individual annotator or section figures.

5 Annotating the Treebank

5.1 Annotation Process and Quality

Our annotation method for all sections of the treebank is the so called *full double annotation*. Each sentence is first independently annotated by two different annotators, and the resulting annotations are automatically merged into a single analysis, where all disagreements are marked. Disagreements are then jointly resolved, typically by all annotators in the group, and this results in the *merged annotation*. These annotations are further subjected to consistency checks, the purpose of which is to ensure that even old annotations conform to the newest annotation decisions. The result of these consistency checks is called the *final annotation*.

This annotation procedure allows us to measure the quality of the annotation and the suitability of the SD scheme for its purpose, using *inter-annotator agreement*. Rather than the *final annotation*, the agreement is measured for each annotator against the *merged annotation*, so as to avoid unfairly penalizing an annotator on decisions that were correct at annotation time but have later become outdated due to changes in the annotation scheme. Additionally, the *final annotation* may differ from the individual annotations in terms of numbers of tokens and sentences, as sentence boundaries and tokenization are corrected at this level where necessary. We use as the measure of inter-annotator agreement *labeled attachment score (LAS)*, which is the percentage of tokens that receive the correct head and dependency label. On average, our annotators achieved an inter-annotator agreement of 89.9% over the entire treebank. Figure 3 illustrates the development of inter-annotator agreement over time and Table 2 lists the agreements of individual annotators across the different sections of the treebank.

5.2 The Effect of Pre-annotation

On a 45 article subset of the web-magazine section, we have performed an experiment regarding annotation speed and quality, in order to find whether automatic

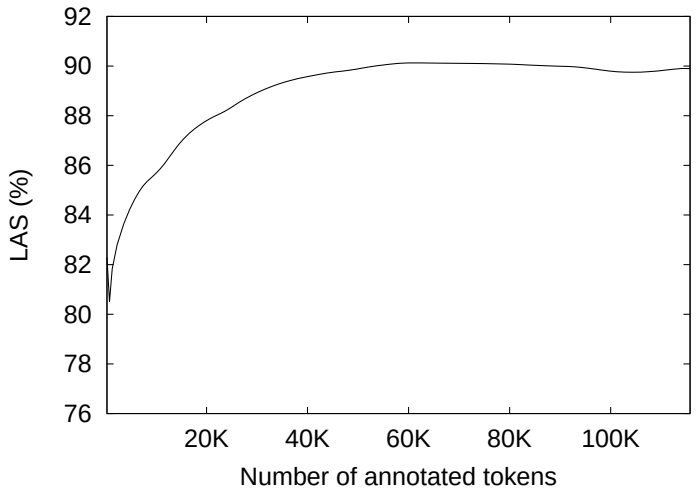

Figure 3: Development of inter-annotator agreement as the amount of annotated tokens grows. Note that the total number of tokens is twice the size of the treebank due to the double-annotation protocol.

Speed [sec/token]				**LAS [%]**			
Annotator	**plain**	**pre.**	***p***	**Annotator**	**plain**	**pre.**	***p***
Annotator 1	3.63	3.01	0.01	Annotator 1	97.4	95.3	<0.001
Annotator 2	4.61	4.45	0.60	Annotator 2	89.0	89.6	0.64
Annotator 3	5.60	5.39	0.65	Annotator 3	92.3	91.7	0.63
Annotator 4	6.92	4.59	0.001	Annotator 4	64.0	78.7	<0.001
Δ		-0.76	<0.001	Δ		2.29	0.034

Table 3: Results of the pre-annotation experiment. The left-hand side shows annotation speed and the right-hand side the LAS. For each annotator are given their base speed, averaged across all plain documents, their pre-annotated speed, and the p-value for the difference. Similarly for LAS. The Δ values are the change of speed or LAS across annotators, corrected for each annotator's base speed or labeled attachment score.

pre-annotation would be helpful (or possibly harmful) for our annotation process. Previously, beneficial effects have been reported by for instance Rehbein et al. [14] and Fort and Sagot [2], on different linguistic annotation tasks.

For the purposes of this experiment, we have produced the first baseline statistical parser of Finnish. Our parser was built using the MaltParser system of Nivre et al. [11], which can be used to automatically induce a parser for a new language, given a treebank. The parser was developed using the body of annotated data available before commencing the experiment discussed in this section, in total 3,648 sentences (48,950 tokens), gathered from all sections of the treebank, including

also a small portion of the web-magazine section. From this data, 80% was used for parser training, 10% for parameter estimation, and 10% for testing. Substantial effort was invested into parameter and feature selection in inducing the parser; the LAS of 70% achieved by the parser thus forms a non-trivial parsing baseline for Finnish. The parser was used to provide automatically pre-annotated versions of each of the documents in our experiment set. The division of documents among annotators was then performed so that for each document, one annotator was assigned the pre-annotated version and the other annotator was to start from an unannotated one (referred to as *plain* hereafter) exactly as in our regular annotation setting. The automatically produced dependencies were visually marked so that the annotator could easily distinguish between dependencies already considered and those still awaiting confirmation or correction.

We calculate the annotation speed in seconds per token and annotation accuracy in terms of LAS. To evaluate the effect of pre-annotation for individual annotators, we compare their speed and LAS on *pre-annotated* vs. *plain* documents and establish statistical significance using the unpaired, two-tailed t-test. The results are shown in Table 3. Our Annotator 4, who has only recently started annotation training and consequently receives the lowest base speed and LAS by far, benefited from the pre-annotation by a tremendous amount, with regard to both speed and LAS. In fact, this annotator's LAS on the plain documents is worse than that of the baseline parser, but given a pre-annotated text, the annotator's LAS clearly exceeds the parser performance. Our most experienced annotator, Annotator 1, gained a small benefit in speed, but suffered a small but statistically significant decrease in LAS. Annotators 2 and 3 did not have a statistically significant difference in speed or LAS.

Since each document was annotated by two different annotators, once as plain and once as pre-annotated, we can further establish the overall effect of pre-annotation across all documents regardless the annotator, using the paired, two-tailed t-test to test for statistical significance. This, however, involves comparing speeds and accuracies between different annotators, which are not directly comparable since individual annotators differ notably in their typical annotation speed and LAS. To take this into account, we establish the base speed and LAS of each annotator across all plain documents (columns *plain* in Table 3) and subtract these from the per-document values before performing the comparison. We are thus comparing changes in speed and LAS, rather than directly their values. We find that pre-annotated documents were on average annotated 0.76 seconds/token faster than plain documents (significant with $p<0.001$) and their LAS was on average 2.29 percentage points higher (significant with $p=0.034$).

Therefore, we conclude that whether pre-annotation is beneficial or harmful depends strongly on the annotator. It would seem that an inexperienced annotator can greatly benefit from a starting point for their work, but for more experienced annotators there was no similar benefit. The risk of overlooking mistakes in a pre-annotated text may contribute to this, and additionally at least Annotator 2 reported difficulties in adapting to the new style and technique of annotation. Naturally, it

could also be suggested that a parser with a better performance could potentially be more helpful for even more experienced annotators. This matter is worth investigating further.

6 Released Data and Software

When releasing the treebank, we do not merely release the text together with its final annotation, but rather the full history leading to the final data. That is, in addition to the final data, we release the independent annotations of both annotators on each document, as well as the *merged* annotation, which is the result of discussing the disagreements between annotators.

Our most important reason for releasing this intermediate data is that each annotated document contains the full edit history of that document, including the exact times (at the resolution of a millisecond) of each edit action performed by an annotator. We believe that this kind of data could potentially be very useful for research, especially for studies on the difficulty of different phenomena encountered in an annotation task, such as the recent work by Tomanek et al. [15]. To our knowledge such detailed data included in a treebank is unique, and it may be a useful resource for future research. In addition, the data makes our own work more transparent. For instance, it allows the replication of the results presented in this paper.

We note that a fraction of approximately 10% of the treebank data in this as well as future releases will be held private, for the purposes of possible future shared tasks on Finnish parsing and parser comparison in general.

Finally, we release a web-based interface for the treebank. This interface allows the user to browse the treebank, as well as make advanced searches. It is possible to search in the text of the treebank, in the morphological analyses, and in the syntactic trees. Morphological and syntactic searches can also be combined, by for instance searching for present tense third singular form verbs that have as their subject a noun that is in partitive. Also searches with a more complex dependency structure are possible, using a syntax akin to TRegex [7] and Tgrep[4]. Detailed documentation of the search features is beyond the scope of this paper and can be found on the project web-page.

7 Conclusion and Future Work

In this work, we have presented an extended version of a freely available treebank for Finnish, the Turku Dependency Treebank (TDT). The size of the current treebank is 4,838 sentences (66,042 tokens), and it consists of four sections, with text from different sources: the Finnish Wikipedia and Wikinews, assorted blogs and a university web-magazine. These sources were selected with the aim to keep the

[4] http://crl.ucsd.edu/software/

treebank freely available under an open license, as well as to ensure a sufficient variation of topics and authors.

The treebank has two levels of analysis: morphological and syntactic. The morphological analyses are created automatically, using existing tools for Finnish. In the manually created syntax annotation, we have used the well established Stanford Dependency scheme, and in order to ensure high quality of the annotation, we have used the *full double annotation* protocol. The average inter-annotator agreement across the treebank was 89.9%. Such a high agreement suggests that annotator training is sufficient and that the annotation scheme is well-defined.

We have also performed an experiment on the effect of pre-annotation on annotation speed and quality and observed greatly improved performance, in terms of both speed and accuracy, for an annotator still in training. An expert annotator achieved a statistically significant gain in speed, although at the cost of a decrease in accuracy.

The treebank, our custom annotation software, detailed data on the annotation process, and a web-based interface of the treebank are available at the address `http://bionlp.utu.fi`.

This work has several important future work directions. The first and most obvious one is to further increase the size of the treebank, adding also new text sources. For instance, fiction text would be a valuable addition, and we are searching for fiction published under an open license. Our current goal is to annotate approximately 10,000 sentences, which appears to generally be enough to produce a robust statistical parser, as for example the results of the multiple language parsing study by Nivre [10] indicate. The second future work direction is to investigate the possibilities to improve the performance of the current parser and to release a fast and robust statistical parser for Finnish.

Thirdly, our goal is to enhance the treebank with additional annotation. Such annotation could, for instance, include human-validated morphological analyses. Also additional dependencies on top of the *basic* SD variant, following one of the extended variants of SD, could be a useful extension of the treebank. With such further annotation in place, it would be possible to add semantic information, for example more detailed analysis of the highly common *nominal modifiers*, with labels such as *temporal* and *cause*. Ultimately, these annotations would enable the development of the treebank into a fully fledged PropBank according to the model set by Palmer et al. [12]. The interaction between the SD and PropBank schemes has already been investigated in connection with the clinical Finnish PropBank [4], and the schemes were found compatible.

Acknowledgements

We are grateful to Lingsoft Ltd. for making FinTWOL and FinCG available to us, as well as the permission to publish their analyses together with the treebank. We would also like to thank all the blog authors who kindly gave us the permission to

include their work in the treebank. This work has been supported by the Academy of Finland.

References

[1] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies manual. Technical report, Stanford University, September 2008.

[2] Karën Fort and Benoît Sagot. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, 2010.

[3] Katri Haverinen, Filip Ginter, Veronika Laippala, Timo Viljanen, and Tapio Salakoski. Dependency annotation of Wikipedia: First steps towards a Finnish treebank. In *Proceedings of The Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, pages 95–105, 2009.

[4] Katri Haverinen, Filip Ginter, Veronika Laippala, Timo Viljanen, and Tapio Salakoski. Dependency-based propbanking of clinical Finnish. In *Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV)*, pages 137–141, 2010.

[5] Fred Karlsson. Constraint Grammar as a framework for parsing unrestricted text. In *Proceedings of COLING'90*, pages 168–173, 1990.

[6] Kimmo Koskenniemi. Two-level model for morphological analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 683–685, 1983.

[7] Roger Levy and Galen Andrew. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC'06*, pages 2231–2234, 2006.

[8] Mitchell Marcus, Mary Ann Marcinkiwicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[9] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies representation. In *Proceedings of COLING'08, Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, 2008.

[10] Joakim Nivre. Deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

[11] Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.

[12] Martha Palmer, Dan Gildea, and Paul Kingsbury. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, 2005.

[13] Sampo Pyysalo, Filip Ginter, Veronika Laippala, Katri Haverinen, Juho Heimonen, and Tapio Salakoski. On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. In *Proceedings of BioNLP'07*, pages 25–32, 2007.

[14] Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 19–26, 2009.

[15] Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167, 2010.