TüBa-D/W: a large dependency treebank for German

Daniël de Kok

Seminar für Sprachwissenschaft
University of Tübingen
E-mail: daniel.de-kok@uni-tuebingen.de

Abstract

We introduce a large, automatically annotated treebank, based on the German Wikipedia. The treebank contains part-of-speech, lemma, morphological, and dependency annotations for the German Wikipedia (615 million tokens). The treebank follows common annotation standards for the annotation of German text, such as the STTS part-of-speech tag set, TIGER morphology and TüBa-D/Z dependency structure.

1 Introduction

In this paper we introduce the automatically annotated TüBa-D/W dependency treebank. Our goal with TüBa-D/W is to provide a large treebank of modern written German, that follows common annotation standards and is freely available under a permissive license. The TüBa-D/W is based on Wikipedia text, consists of 36.1 million sentences (615 million tokens), and is distributed under the same license as Wikipedia.[1] After discussing related work, we will describe how the material for this treebank was collected. Then we will discuss the annotation layers in the treebank and how they are constructed. Finally, we will discuss the treebank format and future work.

2 Related work

In the past two decades three major manually corrected treebanks have been developed for German: NEGRA [6], TIGER [5], and TüBa-D/Z [20]. Although these treebanks are in principle phrase structure treebanks, edges are labeled with grammatical roles. The presence of grammatical roles makes them amenable for

[1] https://creativecommons.org/licenses/by-sa/3.0/

conversion to dependency structure. Such conversions exist for both TIGER [19] and TüBa-D/Z [21].

Recent research has shown that larger automatically annotated treebanks can be a useful resource to gauge the distribution of lexical or syntactic phenomena in a language [4, 16, 10]. Although the use of automatic annotation usually implies a loss of annotation accuracy compared to manually corrected treebanks, their size makes it possible to get more fine-grained statistics and discover low-frequency phenomena. For instance, the largest of the aforementioned treebanks (TüBa-D/Z) has annotations for 1.6 million tokens, while the automatically annotated corpus used in [10] is more than two orders of magnitude larger.

Given that vast computational resources and fast parsers are now readily available, it is perhaps surprising that the number of large automatically annotated treebanks for German is small. The TüPP-D/Z [14] corpus contains partial parses for 204 million tokens from the German newspaper taz. The VISL Corpuseye provides a public search interface and syntactic analyses for Europarl (15 million tokens), Wikipedia (28.7 million tokens), and the Leipzig internetcorpus (47 million tokens). Unfortunately, the annotations do not follow common annotation standards for German and the Wikipedia material is older and substantially smaller than that in the present work. The German reference corpus (DeReKo) contains a recent version of Wikipedia, including discussion pages [7]. However, this corpus does not contain syntactic annotations.

Our contribution is a dependency treebank that is larger than the aforementioned treebanks, using annotation standards that are broadly used for German resources, using a pipeline that can be reproduced and applied to new material easily.

3 Material

For the construction of the treebank, we use a dump of the German Wikipedia that was downloaded on May 6, 2014. Since Wikipedia dumps contain MediaWiki markup, we use the Wikipedia Extractor[2] to convert the Wikipedia dump to plain text. We then convert the plain-text files to the Text Corpus Format (TCF) [9]. The conversion to TCF allows us to process Wikipedia using WebLicht[11], an environment for automatic annotation of corpora. In WebLicht users can compose annotation pipelines of annotation tools that are hosted by CLARIN centers. After composing the pipeline in WebLicht, the corpus was processed using WebLicht as a Service [8], which is a non-interactive version of WebLicht that is tailored to processing of large corpora.

Another preprocessing step that was required, was the replacement of 78 unicode characters that are problematic for many for off-the-shelf natural language processing tools. This set of characters mainly consists of quotation characters, dashes/underscores, and arithmetic operators. To this end, we developed and added

[2] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

a small annotation tool to WebLicht that performs replaced these characters with ASCII equivalents.

4 Annotations

In this section, we give an overview of the annotation layers in the treebank. For each layer we discuss the annotation standard and the tools we use for the automatic annotation.

Tokenization The tokenization and sentence splitting of the corpus is performed using the OpenNLP[3] tokenizer. We retrained the tokenizer on a detokenized version of the TüBa-D/Z treebank [20], release 9. Detokenization reverses tokenization using a set of rules, inserting special markers where the splits occurred. For instance, the tokenized sentence:

> " Gut für sie , gut für Europa " steht klein darunter .

is detokenized to:

> "<SPLIT>Gut für sie<SPLIT>, gut für Europa<SPLIT>" steht klein darunter<SPLIT>.

We used the detokenization rules that were provided by OpenNLP. However, we found that we had to add a rule to handle forward slash (/) characters. The OpenNLP tokenizer obtained an average f-score of 0.9986 in ten-fold cross-validation. For the sentence splitter, we use the model provided for German by OpenNLP.

One problem that we found in the sentence splitter is that it merges headlines with the first first sentence, because headlines usually do not end with end-of-sentence punctuation in Wikipedia. Fortunately, since new lines do not occur in running text in the plain text dump, we could segment the text by using newline characters as boundaries. We then apply the sentence splitter and tokenizer per segment.

Part-of-speech tags The treebank is tagged using the OpenNLP POS tagger, trained on TüBa-D/Z release 9. TüBa-D/Z uses the Stuttgart-Tübingen-TagSet (STTS) [17]. Two changes were made to the tag set in TüBa-D/Z before training the model to make it compatible with the tag set of the TIGER treebank [1]: (1) the pronominal adverb tag was changed from *PROP* to *PROAV* and (2) TIGER does not make the distinction between between attributive indefinite pronouns with and without determiner (*PIDAT* and *PIAT*), so we replaced all *PIDAT* tags by *PIAT*.

The OpenNLP tagger has an accuracy of 96.93% when performing ten-fold cross-validation on TüBa-D/Z with these modifications.

[3] https://opennlp.apache.org/

Morphology Morphological annotations are added using RFTagger [18], with the model for German included in RFTagger, which was trained on the TIGER treebank [5]. Morphological information that is added include gender, case, number, person, tense, and degree. We add morphology annotations because it improves the output of the dependency parser and is useful in some types of treebank queries.

The morphology and part-of-speech tag layers provide overlapping annotations. For instance, the OpenNLP tagger marks a finite verb as *VVFIN*, while RFTagger assigns the category *verb* and attributes such as the tense, person and number. Sometimes the analyses of the part-of-speech tagger and the morphological tagger diverge. In such cases, we do not perform any filtering or post-processing. The parser, which is discussed below, uses both part-of-speech tags and morphological information as features. We expect the training procedure to reduce the weights of features in cases of systematic errors.

Dependency structures The sentences are dependency parsed using the MaltParser [15]. We constructed a model that uses tokens, part-of-speech tags, and morphology as features. The feature templates were constructed using MaltOptimizer [2], using 17072 dependency structures from TüBa-D/Z release 9 as training data with cross-validation on 17071 dependency structures from TüBa-D/Z. We then trained the model using the aforementioned training instances and evaluate it on a third, held-out set of another 17070 dependency structures. In these sets, we used gold standard part-of-speech tags and the output of RFTagger for creating morphological features. The resulting model has a labeled attachment score of 89.0% (88.2% without morphology features).

Lemmatization For lemmatization, we use the SepVerb lemmatizer. This is a lemmatizer that was developed in-house to produce lemmatizations that follow TüBa-D/Z [22]. It first uses the MATE lemmatizer [3], trained on a simplified version of the TüBa-D/Z and then applies post-processing rules to obtain the canonical TüBa-D/Z lemmatization.

TüBa-D/Z lemmatization differs from standard lemmatization in the following ways: (1) the suffix *%passiv* is added to *werden* in passive constructions; (2) the suffix *%aux* is added to auxiliary and modal verbs; (3) particles are added to and marked in separable verbs, for instance *gehen* in *geht davon aus* 'to assume' is lemmatized as *aus#gehen*; (4) reflexives get the lemma *#refl*; and (5) *zu* is removed from infinitives that contain *zu*, for instance *einzufordern* becomes *ein#fordern* 'to demand'. Furthermore, (6) SepVerb uses the lemma *d* and *ein* respectively for definite and indefinite articles.

Transformations for 4-6 can be performed using rules that use the lemma and part-of-speech. However, transformations for 1-3 require syntactic information. For this reason, the SepVerb lemmatizer requires input from a parser. The transformations in SepVerb operated on constituency trees. For the construction of dependency treebanks, we extended SepVerb with rules that work on dependency

structures. The rules for 1-3 for a lemma *l* are:

1. If lemma *l* is the head of lemma *m* with dependency relation *AVZ* (separable verb prefix) and *m* is marked with part-of-speech tag *PTKVZ* (verb particle), then *l* is replaced by *l#m* and *m* is replaced by the empty lemma.

2. Else if $l = werden$ is the head of a token with the tag *VVPP* (perfect participle) with dependency relation *AUX*, *l* is replaced by *l*%passiv.

3. Else if *l* dominates a token with the dependency relation *AUX*, *l* is replaced by *l*%aux.

The lemmatizer uses a model that was trained on TüBa-D/Z release 8 and applies verb processing rules after lemmatization. The lemmatizer achieves 97.66% accuracy in 10-fold cross-validation on the TüBa-D/Z.

5 Availability and future work

TüBa-D/W is provided in the CONLL-X dependency format. Moreover, we added the treebank to the TüNDRA [12] visualization and search tool. To this end, we optimized TüNDRA to work efficiently with treebanks of this size [8].

This paper only describes the first version of TüBa-D/W. We plan to provide updates of the treebank. The initial changes will focus on making the annotations as close to TüBa-D/Z as possible. For instance, we plan to use the morphological information from RFTagger and the dependency information from MaltParser to use gender-specific lemmas (e.g. *der*, *die*, *das*) as in TüBa-D/Z. We would also like to extend the morphology layer such that it provides features in TüBa-D/Z-style in addition to the current TIGER morphology.

Statistical dependency parsing is an active field of work and state-of-the-art parsers such as TurboParser [13] provide an improvement over the MaltParser in our initial experiments with German. If performance and computing facilities permit, we might parse a future version with a parser such as TurboParser to improve the dependency annotations.

Acknowledgments

The development of this treebank was supported by CLARIN-D. We would like to thank the CLARIN-D center IMS Stuttgart for making the RFTagger available as a WebLicht web service.

References

[1] Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pußel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith, and Hans Uszkoreit. TIGER Annotationsschema. Universität des Saarlandes, Universität Stuttgart and Universität Potsdam, 2003.

[2] Miguel Ballesteros and Joakim Nivre. MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–62. Association for Computational Linguistics, 2012.

[3] Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics, 2010.

[4] Gosse Bouma and Jennifer Spenader. The distribution of weak and strong object reflexives in Dutch. In F van Eynde, A Frank, K D Smedt, and G van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, pages 103–114, 2009.

[5] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, Sozopol, Bulgaria, 2002.

[6] Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. Syntactic annotation of a german newspaper corpus. In Anne Abeillé, editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 73–87. Springer Netherlands, 2003.

[7] Noah Bubenhofer, Stefanie Haupt, and Horst Schwinn. A comparable Wikipedia corpus: From Wiki syntax to POS tagged XML. *Multilingual Resources and Multilingual Applications*, 96B:141–144, 2011.

[8] Daniël de Kok, Dörte de Kok, and Marie Hinrichs. Build your own treebank. In *Proceedings of the CLARIN Annual Conference 2014*, 2014.

[9] Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard W Hinrichs. A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In *Proceedings of LREC 2010, Malta*, 2010.

[10] Erhard Hinrichs and Kathrin Beck. Auxiliary fronting in German: A walk in the woods. In *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, page 61, 2013.

[11] Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. Weblicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics, 2010.

[12] Scott Martens. TüNDRA: A web application for treebank search and visualization. In *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, page 133, 2013.

[13] Andre Martins, Miguel Almeida, and Noah A. Smith. Turning on the Turbo: Fast third-order non-projective Turbo Parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[14] Frank Henrik Müller. Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). 2004.

[15] Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135, 2007.

[16] Tanja Samardžić and Paola Merlo. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology*, 7:1–14, 2012.

[17] Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, 1995.

[18] Helmut Schmid and Florian Laws. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 777–784. Association for Computational Linguistics, 2008.

[19] Wolfgang Seeker and Jonas Kuhn. Making ellipses explicit in dependency conversion for a german treebank. In *Proceedings of LREC 2012, Istanbul*, pages 3132–3139, 2012.

[20] Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the tübingen treebank of written German (TüBa-D/Z). In *Seminar für Sprachwissenschaft, Universität Tübingen, Germany*, 2003.

[21] Yannick Versley. Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, 2005.

[22] Yannick Versley, Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. A syntax-first approach to high-quality morphological analysis and lemma disambiguation for the TüBa-D/Z treebank. In *Ninth International Workshop on Treebanks and Linguistic Theories*, page 233, 2010.