Detecting Annotation Scheme Variation in Out-of-Domain Treebanks

Yannick Versley, Julius Steen
ICL Univ. Heidelberg / Leibniz ScienceCampus
(versley|steen)@cl.uni-heidelberg.de

Abstract

To ensure portability of NLP systems across multiple domains, existing treebanks are often extended by adding trees from interesting domains that were not part of the initial annotation effort. In this paper, we will argue that it is both useful from an application viewpoint and enlightening from a linguistic viewpoint to detect and reduce divergence in annotation schemes between extant and new parts in a set of treebanks that is to be used in evaluation experiments. The results of our correction and harmonization efforts will be made available to the public as a test suite for the evaluation of constituent parsing.

Keywords: treebank, variation, annotation schemes

1. Introduction

In treebank creation, there is a growing trend to complement existing in-domain (newspaper text) treebanks by out-of-domain treebanks that allow researchers to study syntactic variation and train statistical parsing models on texts of other domains.

Most often, the creation of these out-of-domain treebanks follows the existing annotation guidelines created by the creators of the newspaper treebank, which would – at least in principle – make it possible to evaluate the generalization performance both of purely-supervised and domain-adapted parsers on these new datasets. However, some caution is in order, as in many cases, annotation guidelines or even informal practices guiding the annotation deviate.In the case of Dredze et al. (2007), researchers report that the divergence between the annotation schemes of Penn Treebank and the Penn Biomedical treebank was greater than the influence of divergence between domains, and that manual annotation conversion of the training dataset without any domain adaptation yielded better results than those of many other participants'.

It has long been argued that detecting and correcting errors in treebanks is a sensible thing to do, as these become a more prominent source of errors as the performance of automatic taggers and parsers gets better (Blaheta, 2002; Manning, 2011). In a study on part-of-speech tag usage in treebanking (Telljohann et al., 2013), it was found that even after the first public release of a treebank, about 0.8% of tags would be changed in routine revisions (which would correspond about to one every eight sentences). In the case of *out-of-domain* treebanks, it is often the case that a new group of people is involved in the construction, and that the annotations between old and new resource differ in tacit, and sometimes also explicit, aspects of the annotation scheme. Dredze et al. (2007) argue that, in adapting parsers from the Penn Treebank to the Penn Biomedical Treebank, *the primary cause of loss from adaptation is from differences in the annotation guidelines themselves*.

While the OntoNotes version of the Penn Treebank (Weischedel et al., 2008) now allows to carry out this kind of domain adaptation experiment, it is to be expected that any such experiment with domain adaptation will yield evaluation results that depend both on the actual performance of the model and a difference between an "ideal" annotation (or at least something as close as possible to the original treebank) and the actual annotation in the out-of-domain treebank.

This difference is usually a compound of several factors, including a different group of people involved in the annotation (with implicit understanding sometimes varying), sometimes a different annotation process (which often involves semi-automatic tools – such as the Annot@te tool (Brants and Plaehn, 2000) that can make suggestions for annotated structures, or in the case of (Pado and Lapata, 2009), the revision of parses from one specific parser), as well as a different distribution of linguistic phenomena – in particular, some phenomena may be marginal in newspaper text but necessitate a principled treatment in domains where they occur more often.

1.1. Related Work

Different processes have been used for quality assurance in the creation of existing treebanks, which each include several ways of checking for errors. In one example, the *Hamburg Dependency Treebank* (Foth et al., 2014), the annotation tool uses handwritten constraints on trees to flag likely errors during annotation Foth et al. (2004). In the second and third step, an approach based on consistency assumptions is used: the part-of-speech tags are checked using the *n-gram*

variation method of (Dickinson and Meurers, 2005), but also pairs of governor and dependent are checked for having a consistent label across the treebank, and are manually inspected if they do not.

The annotation process of the TüBa-D/Z treebank is also based on a combination of intelligent tools, explicit rules, and ad-hoc checking (Telljohann et al., 2013): after initial annotation using the Annot@te tool, explicit queries are used for checking consistency between syntactic structure and part-of-speech or morphological tags, and ad-hoc queries regarding specific phenomena are used for ensuring consistency in certain classes of ambiguous phenomena.

Linguistic knowledge thus plays an important role in the creation and qulity assurance of treebanks: The weighted constraint grammar used in the annotation tool for the Hamburg Dependency Treebank is detailed enough for parsing – indeed use in a parser was the motivation for creating the set of heuristic rules – and represents a significant investment. In some cases, including that of the HDT, the tools used are accessible to the public, but in many cases they either just used informally or simply not disclosed to the public.

Creating such a large and universal set of checking rules anew is also not for the faint of heart: On one hand, such a set of rules is independent from the corpus and may carry over to out-of-domain annotation, on the other hand such an investment stops being useful whenever the annotation scheme changes.

But let us look at the procedures that promise to find candidates for inconsistencies without explicit linguistic knowledge. In finding subtrees that diverge from the normal patterns of annotation, three approaches have been used in the literature:

n-gram-based approaches look for n-grams that occur more than once in a corpus while carrying different annotation, in particular a differen parent category (Dickinson and Meurers, 2005), or different subtrees, or being a non-constituent in the treebank annotation. Because the syntactic structure of an n-gram without context is often ambiguous, Dickinson and Meurers use a *non-fringe* heuristic to filter out n-grams that may be part of an ambiguous construction.

Approaches that aim to find **ad-hoc rules** (Dickinson, 2011) first extract the CFG rules from the treebank and then try to find rules that do not meet a certain minimum support – either by being rare themselves, having rule bigrams that are rare, or by not being part of a cluster of similar rules that is frequent enough.

Finally, approaches based on **predictive modeling** use classifiers to spot trees or nodes that are likely errors. In particular, Ule and Simov (2004) go beyond single CFG expansion and look at the correlation between grandparents and daughter lists; Volokh and Neumann (2011) use an ensemble of two off-the-shelf parsers that generate a re-annotation of the training set. Volokh and Neumann treat sentences as erroneous when the two parsers agree with each other but not with the human annotation. Haverinen et al. (2011) pursue a supervised approach based on the initial version of a treebank and the finished, which they use to train a classifier that detects erroneous edges.

In work on German, Seeker and Kuhn (2014) present a conversion of the Smultron and EuroParl700 corpora to dependencies, in which they updating their converter to cover the structures occurring in Smultron, and manually adjusting instances where the converter fails. Seeker and Kuhn's explicit goal is a compatible dependency conversion rather than a (re-)annotation of the original corpora.

2. Materials

2.1. Treebanks

The Negra/Tiger scheme is a constituency scheme that sees dependencies between a phrase head and its arguments and adjuncts as the fundamental motivation, yielding trees that have discontinuous phrases. It also uses flat adjunction for modifiers (i.e. adding modifiers to an existing projection node rather than adding an additional projection) and elides any unary node projections. The treebank treats specially both coordination, which introduces an additional node governing the conjoined phrases, and prepositional phrases, where the NP that is the preposition's argument is elided to ease the annotation.

The Tiger scheme (Brants et al., 2002) is based on the Negra scheme, but contains a few simplifications; in particular, the categories `QL` (quasi-language, for fixed phrases) and `PIDAT` (a noun modifier that can play the role of an article but does not always do so) are not used in the Tiger scheme.

The **Negra treebank** (Skut et al., 1997) is a treebank that predates Tiger. The Negra treebank uses texts from the same newspaper (*Frankfurter Rundschau*), and its annotation scheme served as the starting point for the definition of the Tiger scheme, however these two differ in several points. In the study of Maier and Sogaard (2008), it is shown that the distribution of block degrees (roughly: number of discontinuities within one LCFRS expansion) is very similar in Negra and in Tiger.

The **Smultron** corpus (Volk and Samuelsson, 2004) contains treebanks in a variety of domains ranging from the manual of a DVD player over company reports and mountaineering articles to a novel about the history of Philosophy (Sophie's world).

Treebank	Text source	number of sentences	number of words	Avgerage words/s.	nodes/word (orig)	nodes/word (revised)
Tiger	FR Newspaper	50472	88238	17.60	0.42	0.52
Negra	FR Newspaper	10027	176371	17.59	0.41	0.50
Smultron Alpine	Mountaineering reports	1060	19467	18.37	0.66	0.48
Smultron DVD	DVD manuals	547	8988	16.43	0.65	0.48
Smultron Economy	Company reports	518	10987	21.21	0.65	0.55
Smultron Sophie	Youth novel	529	7416	14.02	0.65	0.44
EuroParl700	Parliament debates	707	16437	23.25	0.46	0.49
PCC v2	Opinion pieces	2193	33222	15.15	0.42	0.50

Table 1: In-domain and out-of-domain treebanks for the German Tiger scheme

The **EuroParl 700** dataset was created for the study of Pado and Lapata (2009) on crosslingual projection of semantic information (frame roles), and was used by Fraser et al. (2013) as an out-of-domain dataset for parsing evaluation.

The **Potsdam Commentary Corpus** (Stede and Neumann, 2014) is a corpus of news commentary from the *Märkische Allgemeine Zeitung* which has been annotated with syntax and discourse information, with one of the annotators for the syntax part having been part of the effort of creating the original Tiger treebank.

We harmonized several obvious differences as a first step before undertaking comparisons:

- In the *Smultron* treebanks, unary projection nodes are added to aid the crosslingual tree alignment with parts in other languages. For our tests, we simply remove these nodes.
- In both the *Negra* corpus and the *Smultron* treebanks, the label `MPN` is used for multiword proper names, whereas the Tiger scheme uses `PN` for this role. We therefore rename any `MPN` to `PN`.
- Only Smultron contains `NP` nodes inside `PP` nodes, all while these are often added when, e.g. converting the treebanks to dependencies. We therefore add `NP` nodes into `PP`s in all the non-Smultron treebanks.
- In the case of *EuroParl700*, we remove additional `TOP` nodes and attach all punctuation to the root.

As can be seen from Table 1, the unary nodes in the Smultron treebanks in Smultron, as well as the absence of PP-internal NP nodes, make a large difference in the number of nonterminals governing each word. Our normalization approach produces much more similar counts.

We also add automatically generated lemmas to Smultron and the other treebanks using Lemming (Müller et al., 2015), as well as morphology predicted using the Marmot tagger (Müller et al., 2013) using case hints from the annotated trees to the EuroParl700, PCC2 and Negra data, in order to be able to use parsers that rely on them.[1]

3. Manual investigation

In all long-term treebanking projects, manual checking is part of the process of treebank creation besides any rule-based or statistical tools.

For an idea of what errors are present in a treebank at all, we undertook a careful investigation of the first 50 annotated sentences of the EuroParl 700 treebank Pado and Lapata (2009). EuroParl 700 was built by manually checking and correcting parses from the Sleepy Student parser Dubey (2005), which was trained on the Tiger treebank. As such, the process for creating this treebank was perhaps more lightweight than with long-running treebanking projects, but not atypical for the initial effort in an out-of-domain treebank.[2].

Based on manual inspection of the trees, we found about 30 likely errors, of which some are due to peculiarities of the domain: In example (1), it is not clear whether "*Frau Präsidentin*" should be a fragment of its own or is integrated into the sentence as it is in the English version; all cases in the newspaper text of the Tiger treebank have such adresses in integrated form and none use sentence punctuation.

(1) Frau Präsidentin! Ich bin Herrn Wurtz in der Tat eine Erklärung schuldig, die ich nun dem ganzen Parlament geben will.
Madam President, I clearly owe Mr Wurtz an explanation, and I am going to give this explanation before the whole House.

[1] Negra contains morphology for the first 3 000 sentences only, but not for the rest of the treebank.

[2] The EuroParl 700 dataset's main selling point is that it contains frame annotation

Many cases of errors concern grammatical function labels, which are out of scope for the other methods considered here; in some cases, such as subjects of sentences with modals/auxiliaries, however, a change in grammatical function also means a change of attachment.

Grammatical function errors involving subjects (and hence the need to reattach nodes) were most frequently found in passive constructions, where the patient is then mis-tagged as a direct object despite passive sentences not taking direct objects.

4. Criteria based on Keyness

Going past criteria for detecting suspicious trees or parts of trees by manual inspection, we have techniques that decompose each tree into a set of parts (such as rules, or parts of them), and then look at the frequency of these parts to detect heuristically which of them are anomalies.

Previous approaches for finding deviating rules such as Dickinson (2011) use rarity of a rule. However, we can clearly distinguish between the rules in the Tiger treebank (which we assume to conform to the Tiger scheme) and the rules from a given out-of-domain treebank (which may be suspect of being a deviation from the Tiger scheme).

In our case, deviations from an annotation scheme may be relatively frequent and also relatively systematic, so that Dickinson's way of using frequency (and similarity to other rules) as criteria would not yield a good indication of deviant behaviour.

The problem of finding linguistic structures that occur in one dataset but would be atypical in the general distribution of data is not new: In corpus linguistics, **Keyness** refers to quantitative metrics that can be use to find items that are characteristic for a specific profile (i.e., author, genre, or dialect) when compared with another set of texts. In our case we want to compare the frequency profiles of syntactic parts (in particular CFG rules and bigrams of child categories in a rule) to detect rules or parts of rules that are uncharistically frequent or infrequent in the out-of-domain corpus with respect to the Tiger treebank.

On the positive side, this allows to find rules that are frequently used in the out-of-domain treebank (which a purely frequency-based account would tell us to ignore); As a downside, our keyness criterion also captures rules that are more or less frequent because the associated syntactic constructions are rare in one genre but frequent in the other.

We implemented different variations on the basic idea of keyness: from a simple ratio of the smoothed frequencies to Pearson's X^2 statistic, to an interval estimate of the odds (Johnson, 2001).

However, we found that a more intuitive way than comparing all rules or all rule bigrams found in the treebank was to compare just the rule bigrams for one particular category – Tiger has less than 25 categories of nonterminals, so it is definitely feasible to look at these individually – and to display these visually.

Figure 1 shows a parallel-axes visualization of an earlier version of the Tiger and Smultron conversions – in this case, we see a raising diagonal line from 0 (in this case 0.3) to 52 for PPs with a coordinated noun phrase argument that get an NP insertion in one case but not the other.

In our case of the EuroParl corpus, for example, we detected rule bigrams where the children included a finite verb within verb phrases as a parent category, which normally does not occur (projections of finite verbs are `S` nodes, not `VP` nodes, so it is most likely that these finite verbs are mistagged).

5. Explicit Modeling

Similar to the work of Volokh and Neumann (2011), we use off-the-shelf parsing models to detect irregularities by training them on a dataset including the part that is of interest to us (in our case, the Tiger treebank in addition to the out-of-domain treebank) and testing them on the out-of-domain treebank, in the hope that any regularities that are present in the Tiger corpus can still influence the behaviour of the parser on the out-of-domain treebank while keeping the latter's structures for constructions atypical to Tiger.

While Volokh and Neumann used dependency parsers for their work, we use two approaches that can produce discontinuous constituents, selected to produce near-state-of-the-art results with tolerable train and test times. Since both of the parsers produce discontinuous constituents, we can hope to be able to detect errors that would be hidden by either a dependency conversion or the 'standard' conversion to projective constituents.

5.1. Parsers

The first parser in our ensemble is based on the **BLLIP parser** of Charniak (2000), which we train on trees transformed to contain additional information e.g. on the case of noun phrases (see e.g. Dubey, 2005, Versley and Rehbein, 2009).

To get trees similar to those in the actual treebank, we undo these transformations in the parser output, and use heuristic reattachment to reproduce any discontinuous phrases.

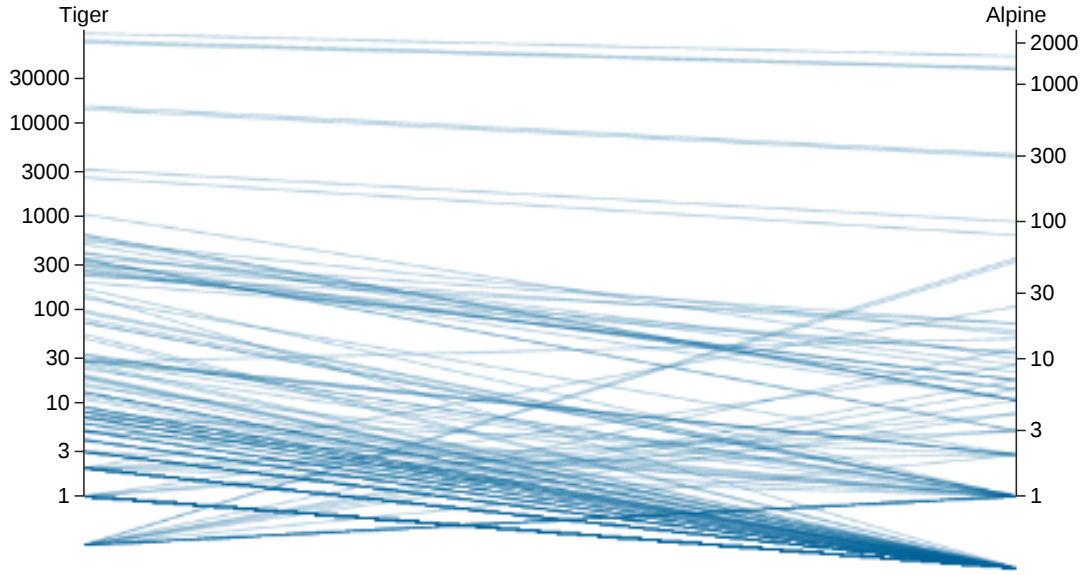

Figure 1: Visual comparison of rule bigram frequencies (for PP parents) between the Tiger treebank and the Alpine portion of Smultron.

As the second part of the ensemble, we use the **Parsing as Reduction** approach of Fernández-González and Martins (2015), which decomposes discontinuous constituent trees into labeled dependency trees that contain enough information for reconstructing the constituents from a dependency backbone. Fernandes-Gomez and Martins' approach has the best published results for this task.

We pair our reimplementation of the constituents-to-dependencies reduction with a second-order model trained using the **TurboParser** (Martins et al., 2013), which uses a dual decomposition approach for the decoding of dependency trees with higher-order factors. We only use the second-order model instead of the third-order model that is also available, in the hope of limiting overfitting.

6. Experiments

Using the keyness statistic, we found several genre-dependent effects within the Smultron treebanks:

- The DVD manuals contain a large number of imperative sentences instructing the user to perform certain actions.
- The Economy texts contain many numeric expressions.
- The *Sophie's World* novel contains more questions (in accordance with the inquisitive nature of the philosophical investigations)

However, we also see patterns that indicate likely annotation errors, in this case a `PP` → `APPR NP ADV` rule that indicates a postmodifying adverbial that should (in all likelihood) be either postmodifying the noun phrase or be attached at sentence level.

In the PCC2 and Negra corpora, we find atypical patterns involving a sentence node with only an infinitive verb, as well as prepositions with a sentence conjunction (`PP` → `KOUI NP`) which are most likely mistagged.

We can see in Table 2 that parsing the data using a model trained on data including our treebank gives quite high exact match scores around 75% for the BLLIP parser and currently around 50-65% for the conversion-to-dependencies approach, which yields 50-55% of sentences where both parsers agree.

Of these, a small number (around 2-3% of the whole corpus) do not agree with existing annotations, and cursory inspection yielded a mixture of clear errors (a prepositional phrase with the `NP` node label), likely errors (an adverb that post-modifies a PP), and cases where only experts in the Tiger annotation scheme could make a firm prediction.

In contrast, the modeling-based approach would find parses where an extraposed constituent was attached to a `VP` according to both parsers but to the `S` node in the treebank. While such attachment decisions are relatively tedious to make for humans, we think that may represent a pattern.

	Alpine	DVD	Economy	Sophie	PCC v2	EuroParl700
BLLIP Exact	79.2	75.3	74.9	85.8	74.9	60.4
Turbo Exact	59.9	50.5	57.5	66.7	62.0	33.7
BLLIP ∩ Turbo	56.9	46.8	54.8	64.1	55.9	30.6
$\lvert(B \cap T) \setminus G\rvert$	19 (1.7%)	19 (3.4%)	15 (2.9%)	11 (2.0%)	29 (1.3%)	15 (2.1%)

Table 2: Exact matches and suspicious sentences on the Smultron subcorpora

7. Summary

With this paper, we hope to make several contributions: Firstly, to present an approach for the consistency checking of out-of-domain treebanks that can also be applied for treebanks in other annotation schemes or in other languages. Secondly, we hope that by checking and harmonizing all treebanks that are currently available in the Negra/Tiger annotation scheme, we can create a high-quality testsuite for out-of-domain parsing of constituents in that scheme. In part, identified annotation errors will have to be fixed manually when they do not obey a systematic pattern. As we have shown, our efforts could find at least some deviant constructs in the treebanks, some of which would not receive differing dependency structures, but also some of which would. In the future, we hope to be able to manually check the added morphology and lemma annotations.

8. Bibliographical References

Blaheta, D. (2002). Handling noisy training and testing data. In *Proceedings of the 7th conference on Empirical Methods in Natural Language Processing*.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. TLT 2002*.

Brants, T. and Plaehn, O. (2000). Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Sixth Applied Natural Language Processing Conference (ANLP-NAACL 2000)*.

Dickinson, M. (2011). Detecting ad-hoc rules for treebank development. *Linguistic Issues in Language Technology (LiLT)*, 4(3).

Dickinson, M. and Meurers, W. D. (2005). Prune diseased branches to get healthy trees! How to find erroneous local trees in a treebank and why it matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain.

Dredze, M., Blitzer, J., Pratim Talukdar, P., Ganchev, K., Graca, J. a., and Pereira, F. (2007). Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1051–1055, Prague, Czech Republic. Association for Computational Linguistics.

Dubey, A. (2005). What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *ACL-2005*.

Fernández-González, D. and Martins, A. F. T. (2015). Parsing as reduction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1523–1533, Beijing, China. Association for Computational Linguistics.

Foth, K., Köhn, A., Beuck, N., and Menzel, W. (2014). Because size does matter: The Hamburg dependency treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*.

Foth, K. A., Daum, M., and Menzel, W. (2004). Interactive grammar development with WCDG. In *Proceedings of the ACL Interactive Poster and Demonstrations Session (ACL 2004)*.

Fraser, A., Schmid, H., Farkas, R., Wang, R., and Schuetze, H. (2013). Knowledge sources for constituent parsing of German, a morphologically rich and less-configurational language. *Computational Linguistics*, 39(1):57–85.

Haverinen, K., Ginter, F., Laippala, V., Kohonen, S., Viljanen, T., Nyblom, J., and Salakoski, T. (2011). A dependency-based analysis of treebank annotation errors. In *Proceedings of International Conference on Dependency Linguistics (Depling 2011)*.

Johnson, M. (2001). Trading recall for precision with confidence sets. Technical report, Brown University.

Maier, W. and Sogaard, A. (2008). Treebanks and mild context-sensitivity. In *Proceedings of Formal Grammar 2008*.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of CICLing 2011*.

Martins, A., Almeida, M., and Smith, N. (2013). Turning on the turbo: Fast third-oreder non-projective

turbo parsers. In *ACL 2013*.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of EMNLP 2015*.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings fo EMNLP 2013*.

Pado, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*.

Seeker, W. and Kuhn, J. (2014). An out-of-domain test suite for dependency parsing of German. In *Proceedings of LREC 2014*.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*.

Stede, M. and Neumann, A. (2014). Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of LREC 2014*.

Telljohann, H., Versley, Y., Beck, K., Hinrichs, E., and Zastrow, T. (2013). Stts als part-of-speech-tagset in tübinger baumbanken. *Journal for Language Technology and Computational Linguistics*.

Ule, T. and Simov, K. (2004). Unexpected productions may well be errors. In *Proceedings of LREC 2004*.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

Volk, M. and Samuelsson, Y. (2004). Bootstrapping parallel treebanks. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC) at Coling 2004*.

Volokh, A. and Neumann, G. (2011). Automatic detection and correction of errors in dependency treebanks. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., and Houston, A. (2008). Ontonotes release 2.0. LDC2008T04, Philadelphia, Penn.: Linguistic Data Consortium.