Originally published in: Henrich, Verena/Hinrichs, Erhard/de Kok, Daniël/Osenova, Petya/ Przepiórkowski, Adam (eds.): Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13). December 12-13, 2014, Tübingen, Germany. - Tübingen: University of Tübingen, 2014. Pp. 172-183

# POS Tagset Refinement for Linguistic Analysis and the Impact on Statistical Parsing

Ines Rehbein and Hagen Hirschmann

German Department
Potsdam University
E-mail: rehbein@uni-potsdam.de

German Studies and Linguistics
Humboldt University Berlin
hirschhx@hu-berlin.de

## 1 Introduction

The annotation of parts of speech (POS) in linguistically annotated corpora is a fundamental annotation layer which provides the basis for further syntactic analyses, and many NLP tools rely on POS information as input. However, most POS annotation schemes have been developed with written (newspaper) text in mind and thus do not carry over well to text from other domains and genres. Recent discussions have concentrated on the shortcomings of present POS annotation schemes with regard to their applicability to data from domains other than newspaper text.

For German, ongoing efforts [18, 14, 17, 5, 20, 19] discuss the restructuring of the Stuttgart-Tübingen Tagset (STTS) [15], which is a quasi standard for German POS tagging. While the discussion so far has focussed on the extension of the STTS to non-canonical data such as spoken language or user-generated content from the web, we put our attention on a different, but related matter, namely the restructuring of the tagset in order to provide us with a more detailed linguistic analysis of modification. At the same time, we are interested in the impact of the tagset refinements on the accuracy of NLP tools which rely on POS as input, in particular of syntactic parsers.

Recent work investigating the impact of POS annotation schemes on parsing has yielded mixed results [8, 3, 7, 9, 13]. While a preliminary study on providing the parser with more fine-grained *gold* annotations gave proof-of-concept that, at least for German dependency parsing, a linguistically motivated distinction of modifier tags can indeed improve parsing results [13], it still has to be shown that these results carry over to larger data sets and to a real-world scenario where the parser is provided with automatically predicted POS.

In the paper, we fill this gap and present experiments on dependency and constituency parsing of German with a more fine-grained and syntactically motivated tagset for modifier relations. We explore whether the new modifier distinctions can be automatically predicted with an accuracy that is good enough to increase pars-

ing accuracy. Our results show a modest, but statictically significant improvement when training the parsers on the modified tagset.

The paper is structured as follows. In Section 3, we briefly describe the new tag distinction and report on our efforts to improve POS tagging results on the new POS tags. In Section 4, we present parsing experiments investigating the impact of the different POS distinctions on the accuracy of statistical parsers. We discuss our results and put them into context in Section 5, and finally conclude in Section 6.

## 2 Tagset Refinements

The classification in the standard part of speech tagset for German, the STTS, is based on very heterogeneous criteria – some definitions refer to the word's inflectional status, some to its syntactic status, some to semantic or to purely lexical classes. The open word class ADV (adverb) can be described as a residual category where adverbs are defined as modifiers of verbs, adjectives, adverbs, or clauses, which are not derived from adjectives (STTS guidelines, p. 56). Since there are other parts of speech that can also modify each of these heads (e.g. modal particles, regular particles, pronominal adverbs, and ordinals), this definition is not sufficient.

We thus propose a more fine-grained subcategorisation of the residual class ADV in the STTS tagset which distinguishes between a) "real" adverbs (ADV), b) modal particles (MODP), c) focus particles (PTKFO), d) intensifiers (PTKINT), and e) lexical particles (PTKLEX). These classes are defined from a *functional syntactic* perspective, which does not include semantic classes like temporal or manner adverbs that are specific semantic subcategories of the class ADV. Furthermore, we redefine the dissociation of adverbs (ADV) and adjectives (ADJD), which –according to the STTS– is based on the criterium of *inflectibility*, in favour of a syntactically motivated notion of lexical modifiers (for a more detailed discussion of the new tag distinctions see [6, 13]).

As an example, consider Sentence 145 from the TIGER treebank [2] (Figure 1). In the original Tiger POS annotation (ORIG) which follows the STTS, the four lexical modifiers "etwa" (for instance), "so" (as), "stark" (strong), "allgemein" (generally) are described by the tags ADV (adverb: "etwa", "so") and ADJD (predicative adjective: "stark", "allgemein"). This distinction is motivated morphologically – "etwa" and "so" cannot be inflected in German, whereas "stark" and "allgemein" can. The POS tags for "etwa", "so", "stark", and "allgemein" do not express any syntactic differences between the words. Furthermore, most grammarians will question the analysis of "etwa" and "so" as adverbs in this particular context.

Compare the original tags to the new tag distinctions (NEW) in Figure 1 which show a (more) syntactically motivated POS analysis of the same lexical items. In the case of "etwa" and "so", new POS tags (PTKFO and PTKINT) have been introduced which reflect the syntactic status of the respective words. The POS tag

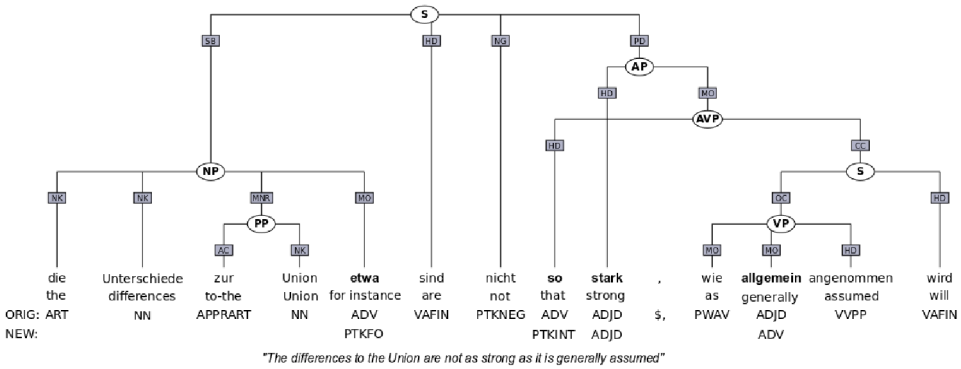


Figure 1: Example tree from TIGER illustrating four different modifers (PTKFO, PTKINT, ADJD, ADV) in the original STTS and in the new classification scheme

ADJD for "stark" is the same as in the original STTS anlysis, because "stark" is used as a predicative adjective in the given context. The POS tag for "allgemein", however, deviates from the original STTS analysis because of the adverbial function of "allgemein" in this context, which is reflected by assigning it the ADV tag in the new classification scheme.

In conclusion, our new classification provides (four) new POS tags and redefines (two) existing categories of the original STTS in order to restructure the part of speech analysis of modifying words in the STTS in a principled, syntactically motivated way. Table 1 gives an overview over the categories that are conceptually altered or newly implemented in the new tagset.

# 3 POS tagging experiments

## 3.1 Data

The data we use in our experiments are the first 10.000 sentences from the TIGER treebank [2], reannotated according to the new modifier classification. As these distinctions are sometimes ambiguous, we cannot expect the same inter-annotator agreement as for the original STTS modifier tags.[1] However, as shown in [13], even POS annotations with a lower IAA are able to provide a statictical parser with useful information. The question which remains to be answered, is whether these tag distinctions can also be learned reliably by NLP tools to achieve the same effect.

After annotating a goldstandard with 1.000 sentences randomly extracted from TIGER, the first 10.000 sentences of the treebank have been relabelled semi-automatically, using manually defined patters which make use of the syntactic information in the treebank as well as the STTS tags and the lexical forms. The automatically predicted new tags were then checked manually by one expert annotator.

[1] In an annotation study with two human coders, we obtained an inter-annotator agreement of 0.838 (Fleiss' $\kappa$) on newspaper text for the new modifier distinctions [13].

| Tag | Meaning | Description | Restriction/Test |
|---|---|---|---|
| ADV | Adverb in a syntactic sense | Modifier of verbs or clauses | Can appear in the prefield |
| *Example:* | | *Sie läuft* **schnell** – *Sie läuft* **glücklicherweise**<br>she runs quickly – she runs fortunately<br>"She runs quickly" – "Fortunately, she runs" | |
| ADJD | Predicative adjective | Complement of a copula verb | Can appear in the prefield |
| *Example:* | | *Die Läuferin ist* **schnell**<br>the runner is quick<br>"The runner is quick" | |
| MODP | Modal particle | Clausal modifier | Cannot appear in the prefield |
| *Example:* | | *Sie läuft* **ja** *bereits*<br>she runs PTC already<br>"She is already running" **[as it is well known]** | |
| PTKFO | Focus particle | Associated with a focus element, modifying a set of alternatives | Cannot appear on their own in the prefield |
| *Example:* | | **Auch** *sie läuft schnell*<br>also she runs quickly<br>"She runs quickly, too" | |
| PTKINT | Intensifier | Intensifying or quantifying a gradable expression | Cannot appear on their own in the prefield |
| *Example:* | | *Sie läuft* **sehr** *schnell*<br>she runs very quickly<br>"She runs very quickly" | |
| PTKLEX | Part of a multiword expression | Particle which cannot be analysed compositionally | cannot appear on their own in the prefield |
| *Example:* | | *Sie läuft* **immer** *noch*<br>she runs always still<br>"She is still running" | |

Table 1: Overview of the new tag distinctions and examples for each tag

Please note that we evaluate POS accuracy on the 1.000 sentences which have been reannotated from scratch for POS by two annotators, while we use the whole 10.000 sentences in a 10-fold cross-validation setting for evaluating parsing accuracy. As we do not include the POS in the evaluation of parsing results, potential POS errors in the semi-automatically annotated data should not influence the parser evaluation.

### 3.2 Setup

Our POS tagger is similar to the FLORS tagger [16] and makes use of 4 different feature types: a) shape features, b) prefix/suffix features, c) context features, and d) distributional features. We also use the linear L2-regularized L2-loss SVM implementation provided by LIBLINEAR [4] to train a one-vs-all classifier for each POS in the training set.[2] In contrast to [16], we also include POS context features from POS tags predicted by the Hunpos tagger trained on the original STTS tags.[3]

### 3.3 POS tagging baseline

As our baseline, we train the Hunpos tagger on sentences 1-9.500 from TIGER (excluding the sentences which are part of the goldstandard; the last 500 sentences have been held out as development data) and evaluate on the randomly extracted, manually annotated goldstandard (1.000 sentences). Table 2 shows results for different taggers on the original STTS (orig) and on the new classification (new).

| tagger | setting | acc. |
|---|---|---|
| baseline1 (Hunpos) | orig tags | 96.11 |
| baseline2 (Hunpos) | new tags | 94.78 |
| own tagger (w/o POS context) | new tags | 94.91 |
| own tagger (with POS context) | new tags | **96.68** |

Table 2: POS tagging accuracy for the Hunpos tagger and for our own tagger (without and with POS context features) on the gold standard (500 sentences)

Not surprisingly, the baseline tagging accuracy on the new tagset is lower than the one on the more coarse-grained STTS distinctions. The Hunpos tagger, trained on the original tags, achieves an accuracy of 96.11% while the accuracy of the same tagger on the new tagset is more than 1% lower at 94.78%. Our own tagger achieves a slightly higher accuracy of 94.91% on the new tags when using word-form context only, and a considerably higher accuracy of 96.68% when also using context features based on the original STTS tags predicted by Hunpos.

Table 3 shows results for individual tags, evaluated on the larger data set (10.000 sentences) in a 10-fold cross-validation setting. We can see that precision for most

[2]This amounts to 52 POS for the original STTS and to 56 POS for the modified annotation scheme.

[3]The Hunpos tagger is an open source reimplementation of the TnT tagger (`https://code.google.com/p/Hunpos`)

of the tags is considerably higher than recall. Exceptions are the two most frequent classes ADV and PTKFO. This is not surprising as ML methods are known to have a bias towards the more frequent classes and tend to overuse them.

| TAG | prec. | rec. | f-score | row counts |
|---|---|---|---|---|
| ADV | 87.25 | 89.76 | 88.49 | (6276/6992) |
| PTKFO | 82.32 | 84.73 | 83.51 | (1243/1467) |
| ADJD | 81.10 | 75.46 | 78.18 | (824/1092) |
| PTKINT | 80.81 | 72.87 | 76.63 | (779/1069) |
| MODP | 85.37 | 69.31 | 76.51 | (70/101) |
| PTKLEX | 81.87 | 67.62 | 74.07 | (307/454) |

Table 3: POS tagging accuracy on the larger data set (10.000 sentences)

The overall accuracy of our tagger on all 58 tag distinctions on the larger data set is 97.0%. This is slightly higher than the accuracy of the Hunpos tagger on the same data (96.4%), using the original STTS tag distinctions.

Whether or not a precision in the range of 80-87% and f-scores between 74% and 88% are good enough to be used in linguistic analyses is hard to answer and certainly depends on the research question. We would like to argue that the additional information provided by the new tag distinctions is useful, and that the new tags, even if not perfect, can at least be used to extract candidates for linguistic analysis. Furthermore, the new classes MODP, PTKFO, PTKINT and PTKLEX can easily be subsumed under the ADV class, so no information is lost.

# 4 Impact of modifier distinctions on statistical parsing

To find out whether the more fine-grained modifier distinctions are able to improve parsing accuracy when predicted automatically, we use the new tags as input for training two statictical parsers. The first parser is the Berkeley parser [10], a PCFG-LA constituency parser, and the second one the MATE parser [1], a transition-based dependency parser. Both systems are language-agnostic.

## 4.1 Impact on constituency parsing

Table 4 gives results for constituency parsing when training the parser on the original STTS tags (orig) and on the new tags (new). We can see a modest improvement of 0.3% f-score for the new tag distinctions over all folds. When including the grammatical function labels in the evaluation, the average improvement is 0.2.

These results are for letting the parser assign its own POS tags. When providing it with the POS tags assigned by our tagger, results are similar with an average f-score of 75.26 for the original STTS tags and a slightly higher f-score of 75.54 (excluding grammatical functions) for the new tag distinctions. The difference in f-scores between the original STTS POS tags and the new tags is statictically

significant with $p = 0.025$.[4] While the improvements are small, they do show that our new tag distinctions do not hurt parsing accuracy and might even have the potential to improve it.

| | *original STTS* | | | *new tags* | | |
|---|---|---|---|---|---|---|
| | **rec** | **prec** | **fscore** | **rec** | **prec** | **fscore** |
| fold 1-10 (avg.) | 75.38 | 75.12 | 75.25 | **75.45** | **75.64** | **75.55** |

Table 4: Parseval results (10fold cross-validation excluding grammatical functions)

One drawback of using the Berkeley parser in our experiments is that even when provided with "gold" POS tags, the parser, when it cannot find a good analysis for the prelabelled tags, takes the liberty to reject them and reassing its own POS. Also, the Berkeley parser does not take the tags as they are but, during training, refines the annotations by applying merging and splitting operations to the nodes in the tree, and only keeps those labels which have been shown to be useful during training. By just looking at the parsing results, we do not know what the internal representation used by the parser after the training cycles looked like.

In the next section, we turn to dependency parsing, which provides us with a more straight-forward way to compare the influence of different POS tagset distinctions on syntactic parsing.

## 4.2 Impact on dependency parsing

In the next experiments, we use the CoNLL conversion of the same 10.000 TIGER sentences to train the MATE parser. First, we replicate the experiment of [13] on a larger data set and use the gold tags (original STTS and new classification) for training. We find a small improvement of around 0.3 (UAS) and around 0.4 (LAS) when providing the parser with the new tags (Table 5). The results are consistent with the ones of [13] obtained on a smaller data set.

| fold 1-10 (avg.) | *UAS orig.* | *new* | *LAS orig.* | *new* |
|---|---|---|---|---|
| gold POS | 91.88 | **92.23** | 90.02 | **90.46** |
| pred POS | 89.68 | **89.81** | 86.94 | **87.13** |

Table 5: Unlabelled and labelled attachment scores for gold / predicted POS

Next, we test the parser on automatically predicted POS tags. The training data was annotated using 10-fold jackknifing. For the original STTS we used POS tags predicted by the MATE tagger, for the new classification we provided the parser with the tags predicted by the POS tagger described in Section 3.2.

While the improvements are smaller than for the gold tags, the difference is still statictically significant with $p = 0.002$ (LAS). When looking at f-scores for

[4]For significance testing, we used Dan Bikel's Randomized Parsing Evaluation Comparator with 10.000 iterations.

identifying specific dependencies and their attachments,[5] we observe improved results for 23 (out of 40) dependencies, the same results on 3 dependencies, and lower scores on the remaining 14 dependency relations (Table 6). Amongst the ones where we obtain improved results are not only the most frequent modifier relations (MO, MNR), but also the core grammatical functions (SB: subject, +0.7%; OA: direct object, +0.5%; DA: indirect object, +2.5%). We thus argue that, despite the improvements in f-score being small, the parse trees we obtain when training the parser on the new tag distinctions are of a higher quality as we achieve higher f-scores for identifying the arguments of a sentence.

As expected, we can see that the parser benefits from the new tag distinctions when parsing modifier relations. Figure 2 shows the parser output tree for the MATE parser when provided with the original STTS tags, and the parse tree triggered by the new tags. The POS tagger correctly predicted the two more fine-grained new tags for "auch" (also) (PTKFO) and "gegenwärtig" (at present) (ADV), which helped the parser trained on the new tags to correctly identify the low attachment for PTKFO, while the original STTS tag ADV incorrectly triggers high attachment for "auch" (dotted-red arrow). For "gegenwärtig", the redefined ADV tag in the new scheme again results in the correct attachment decision, while the same parser trained on the original STTS is only provided with the underspecified ADV tag and thus again produces the wrong analysis.

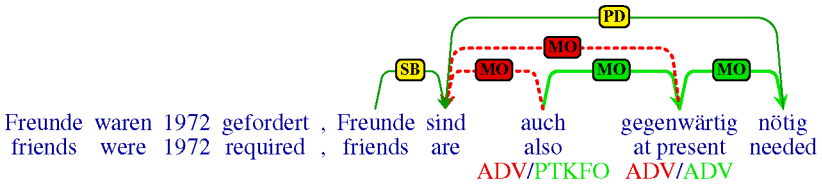


Figure 2: Parser output tree for orig (dotted-red) and new tags (green)

For predicate (PD) dependencies, however, f-scores for identifying the correct dependency label and attachment site are below the ones obtained by a parser trained on the original STTS tags. This is due to the low accuracy of the POS tagger on the ADJD part of speech tag. The distinction between adverbially used adjectives and predicative adjectives in the new tagset is difficult for the tagger which has only access to local information. For the STTS distinctions, the tagger can rely on word form information which, in many cases, in enough to identify the "correct" tag.

When providing the parser with gold POS tags, we observe an improvement in f-score of 1.7% for the PD dependency, from 76.7% for the original STTS tags up to 78.4% f-score when training the parser on the new modifier tags. In future work, we will try to improve the tagger by adding linguistically motivated features which might help to increase the tagging accuracy especially for predicative adjectives.

[5]For the evaluation we used a slightly modified version of the CoNLL07 evaluation script provided by `http://pauillac.inria.fr/~seddah/eval07.pl`.

| DEPREL | freq. | orig f-score | new f-score | + (%) | = (%) | - (%) |
|---|---|---|---|---|---|---|
| AC | 172 | 78.3 | 79.0 | 0.7 | | |
| ADC | 4 | 75.0 | 85.7 | 10.7 | | |
| AG | 4929 | 92.9 | 93.0 | 0.1 | | |
| AMS | 113 | 76.2 | 73.6 | | | -2.6 |
| APP | 796 | 55.9 | 56.5 | 0.6 | | |
| AVC | 11 | 50.0 | 42.9 | | | -7.1 |
| CC | 446 | 61.0 | 59.3 | | | -1.7 |
| CD | 3991 | 85.1 | 85.1 | | 0.0 | |
| CJ | 5668 | 80.6 | 80.6 | | 0.0 | |
| CM | 480 | 77.1 | 75.2 | | | -1.9 |
| CP | 1759 | 90.8 | 92.9 | 2.1 | | |
| CVC | 172 | 52.8 | 48.5 | | | -4.3 |
| DA | 1045 | 57.1 | 59.6 | 2.5 | | |
| DM | 16 | 55.2 | 66.7 | 11.5 | | |
| EP | 377 | 77.9 | 78.6 | 0.7 | | |
| JU | 320 | 86.4 | 87.0 | 0.6 | | |
| MNR | 5227 | 67.0 | 67.5 | 0.5 | | |
| MO | 22378 | 74.9 | 75.2 | 0.3 | | |
| NG | 1097 | 76.0 | 76.1 | 0.1 | | |
| NK | 55439 | 97.4 | 97.5 | 0.1 | | |
| NMC | 552 | 96.4 | 96.1 | | | -0.3 |
| OA | 6678 | 79.7 | 80.2 | 0.5 | | |
| OA2 | 5 | 0.0 | 0.0 | | 0.0 | |
| OC | 8133 | 86.9 | 88.1 | 1.2 | | |
| OG | 29 | 5.0 | 0.0 | | | -5.0 |
| OP | 1597 | 51.5 | 51.7 | 0.2 | | |
| PAR | 456 | 44.1 | 43.6 | | | -0.5 |
| PD | 1888 | 73.8 | 73.2 | | | -0.6 |
| PG | 613 | 80.1 | 79.3 | | | -0.8 |
| PH | 21 | 38.7 | 36.4 | | | -2.3 |
| PM | 989 | 98.9 | 98.8 | | | -0.1 |
| PNC | 2217 | 89.8 | 89.7 | | | -0.1 |
| RC | 1395 | 66.4 | 68.4 | 2.0 | | |
| RE | 597 | 68.1 | 68.8 | 0.7 | | |
| RS | 78 | 21.8 | 29.1 | 7.3 | | |
| SB | 13065 | 86.4 | 87.1 | 0.7 | | |
| SBP | 359 | 76.5 | 74.7 | | | -1.8 |
| SVP | 1045 | 91.7 | 90.6 | | | -1.1 |
| UC | 71 | 11.0 | 12.9 | 1.9 | | |
| VO | 15 | 0.0 | 11.1 | 11.1 | | |

Table 6: Results for specific dependency relations + attachment

## 5 Discussion

Our experiments have shown that more fine-grained, linguistically motivated POS distinctions can, at least to a small extent, improve results of a data-driven statistical parser.

There is related work by Maier et al. [9] who compare the impact of three different POS tagsets on constituency parsing results. The tagsets they use are the coarse-grained Universal Tagset (UTS) [11] which distinguishes 12 tags, the STTS (54 tags), and a fine-grained tagset enriched with morphological information (> 700 tags). They also use the Berkeley parser in their experiments, which always obtained best results when trained on the STTS tagset, no matter if the POS tags were i) gold tags, ii) predicted by a HMM tagger, or iii) assigned by the parser itself. Surprisingly, the results for using the coarse-grained UTS were only slightly lower when provided as gold tags or learned by the parser. The fine-grained morphological tagset, however, proved to be too sparse and resulted in a substantial decrease in f-score. Maier et al. [9] did not modify the STTS, and only report results for constituency parsing. It would be interesting to see the impact of the UTS on dependency parsing, as it might be the case that the Berkeley parser can cope with the underspecified tags only because it applies its own refinement techniques to the annotation scheme during training.

Another relevant study is the work by Plank et al. [12] who discuss the problem of ambiguity caused by unreliable POS annotations by human coders. They show that incorporating annotator disagreements into the loss function of the POS tagger can improve results for POS tagging as well as increase the accuracy of a syntactic chunker that uses the POS tags as input. Their study can be described as complementary to ours. While we try to reduce the ambiguity in the data by refining the tagset and augmenting it with new information, their approach is to incorporate the ambiguity directly in the tagging model. In future work, it might be interesting to combine both approaches.

## 6 Conclusions

In the paper, we argued for a new classification of modifier distinctions in the STTS, the standard German POS tagset, which overcomes the drawbacks of the residual category ADV in the original tagset by providing more fine-grained and syntactically well motivated tag distinctions. The new tagset not only supports a more detailed linguistic analysis, it also has the potential to improve the accuracy of statistical parsers. We showed that even for automatically predicted POS tags we obtained a small, but significant improvement over the original STTS. As these improvements concern the core grammatical functions, we argue that the new modifier classification not only leads to modest improvements in parsing accuracy but, more importantly, also to a qualitatively improved syntactic analysis.

## References


[1] Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, 2010.

[2] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42, 2002.

[3] Markus Dickinson. An investigation into improving part-of-speech tagging. In *Proceedings of the Third Midwest Computational Linguistics Colloquium (MCLC-06)*, Urbana-Champaign, IL, 2006.

[4] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[5] Ulrich Heid und Kathrin Beck Heike Zinsmeister. Das Stuttgart-Tübingen Tagset – Stand und Perspektiven. 28(1), 2013.

[6] Hagen Hirschmann. Richtlinien zur Wortartenannotation von Adverb- und Partikelklassen – eine Granularisierung des STTS im Bereich von Modifikatoren. Technical report, Humboldt-Universität zu Berlin, 2014.

[7] Sandra Kübler and Wolfgang Maier. Über den Einfluss von Part-of-Speech-Tags auf Parsing-Ergebnisse. *Journal for Language Technology and Computational Linguistics*, 1(28):17–44, 2014.

[8] Andrew MacKinlay and Timothy Baldwin. Pos tagging with a more informative tagset. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 40–48, Sydney, Australia, 2005.

[9] Wolfgang Maier, Sandra Kübler, Daniel Dakota, and Daniel Whyatt. Parsing German: How much morphology do we need? In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, Dublin, Ireland, 2014.

[10] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING)*, 2006.

[11] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *The 8th International Conference on Language Resources and Evaluation (LREC-12)*, May 2012.

[12] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014.

[13] Ines Rehbein and Hagen Hirschmann. Towards a syntactically motivated analysis of modifiers in german. In *Conference on Natural Language Processing (KONVENS)*, Hildesheim, Germany, 2014.

[14] Ines Rehbein and Sören Schalowski. STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language. *Journal for Language Technology and Computational Linguistics*, 1(28):199–227, 2014.

[15] Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen, 1999.

[16] Tobias Schnabel and Hinrich Schütze. FLORS: Fast and Simple Domain Adaptation for Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics (TACL)*, 2:15–26, February 2014.

[17] Thomas Schmidt Swantje Westpfahl. POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *JLCL*, 28(1):139–153, 2013.

[18] Angelika Storrer Thomas Bartz, Michael Beißwenger. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28(1):157–198, 2013.

[19] Swantje Westpfahl. STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 1–10, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[20] Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. Adapting a part-of-speech tagset to non-standard text: The case of STTS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 4097–4104, 2014.