Towards a syntactically motivated analysis of modifiers in German

Ines Rehbein
Universität Potsdam
German Department
SFB 632 "Information Structure"
irehbein@uni-potsdam.de

Hagen Hirschmann
Humboldt-Universität zu Berlin
Department of German Studies
and Linguistics
hirschhx@hu-berlin.de

Abstract

The Stuttgart-Tübingen Tagset (STTS) is a widely used POS annotation scheme for German which provides 54 different tags for the analysis on the part of speech level. The tagset, however, does not distinguish between adverbs and different types of particles used for expressing modality, intensity, graduation, or to mark the focus of the sentence. In the paper, we present an extension to the STTS which provides tags for a more fine-grained analysis of modification, based on a syntactic perspective on parts of speech. We argue that the new classification not only enables us to do corpus-based linguistic studies on modification, but also improves statistical parsing. We give proof of concept by training a data-driven dependency parser on data from the TiGer treebank, providing the parser a) with the original STTS tags and b) with the new tags. Results show an improved labelled accuracy for the new, syntactically motivated classification.

1 Introduction

The Stuttgart-Tübingen Tagset (STTS) (Schiller et al., 1999) is a widely used POS annotation scheme for German. It provides 54 different tags for the analysis of German, partly based on morphological and distributional properties, partly also taking semantics into account. The tagset, however, does not distinguish between adverbs and different types of particles used for expressing modality, intensity, graduation, or to mark the focus of the sentence. This is understandable, as these distinctions are often hard to make and thus might decrease the consistency of the annotations as well as make the annotation process more time-consuming.

Nonetheless, there are many tasks where one would wish for a more fine-grained analysis, especially when analysing spoken language or user-generated content from the web, but also for newspaper text where we can find a high variety of different modifiers. Consider, e.g., examples (1)-(3) below.

(1) Russland ist doch aber auch noch da.
Russia is however but also still there.
"But after all, Russia is also still there."
[spoken language utterance]

(2) [...] , im Roman heißt sie ja ohnehin
[...] , in the novel is called she PTC anyway
zumindest fast immer nur Caro.
at least nearly always only Caro.
"[...], in the novel, she is nearly always only called Caro, anyways." [from Twitter]

(3) [...] , jetzt vielleicht sogar noch mehr.
[...] , now maybe even still more.
"[...], but now maybe even more so."
[newspaper text (TiGer)]

According to the STTS, the modifier sequences in (1)-(3) would be annotated as shown in (4)-(6).

(4) doch aber auch noch da
ADV ADV ADV ADV ADV

(5) ja ohnehin zumindest fast immer nur
ADV ADV ADV ADV ADV ADV

(6) jetzt vielleicht sogar noch mehr
ADV ADV ADV ADV ADV

Long sequences of adverbs and particles are particularly frequent in spoken dialogues and in conceptually spoken registers[1] but are also common in newspaper text. Thus, an analysis telling us that *ja* in (2) is a modal particle, *fast* (nearly) modifies *immer* (always) on a gradual scale, and that *nur* (only) is associated with the focus would be far more informative than the analysis given above. The question is, is such an analysis feasible with respect to annotation consistency and time, and how hard is it for automatic methods to learn these distinctions.

In the paper, we follow up on these questions and present a new classification for the analysis of modifiers in German, based on a syntactic perspective on part of speech categories (for details see Section 3).

Section 2 starts with a brief review of related work, then we describe the new tagset and motivate the linguistic basis of the distinctions between the tags (Section 3). In Section 4 we present an annotation study where we report on inter-annotator agreement and discuss the difficulties we encounter when applying the new classification to data from the TiGer treebank (Brants et al., 2002). In Section 5 we investigate the impact of the syntactically motivated annotations on the accuracy of a syntactic parser. We train a data-driven dependency parser on a subset of the TiGer treebank which we re-annotated using the new tags. Results show that the new classification improves labelled accuracy scores (LAS) especially for modifier relations. In Section 6 we discuss our results and outline future work.

2 Related Work

There is some previous work on improving natural language processing by refining POS tagsets. However, most of these studies have been conducted on English (with the exception of Kübler and Maier (2014)) and have reported negative results.

MacKinlay and Baldwin (2005) investigate the impact of different POS tagsets on automatic tagging accuracy by introducing finer distinctions between the tags. Their refined tagsets did not succeed in improving tagging accuracy. The authors attribute this to data sparseness.

Dickinson (2006) also tries to improve the results of automatic POS tagging by redefining ambiguous tags in the tagset. His approach is to add complex tags to the tagset which reflect the ambiguity of certain word forms. This approach gave slight improvements on the test set but proved to be less robust than the same tagger trained on the original tagset.

In contrast to the studies mentioned above, our main motivation for refining the STTS is not to improve tagging accuracy but to investigate whether taking a syntactically motivated perspective on POS tagset distinctions is reflected in the outcome of a syntactic parser, where (manually or automatically assigned) POS tags are crucial information to build up the syntax tree.

There is some evidence against our hypothesis. Kübler and Maier (2014) compare the influence of different POS tagsets, the German STTS, the coarse-grained universal tagset of (Petrov et al., 2012), and a fine-grained German tagset including morphological information, on constituency parsing results. They use the Berkeley parser (Petrov et al., 2006), a PCFG-LA parser, and show that in some settings, the coarse-grained universal tags are more useful to the parser than the more fine-grained STTS tags, while the morphologically enriched tags seem to be too sparse for the parser to benefit from the information. However, it is hard to draw conclusions from this, as 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.

We argue that a more straight-forward way to compare the influence of different POS tagset distinctions on syntactic parsing consists in using a

[1] Here we refer to the model of Koch and Oesterreicher (1985) who describe texts from written registers which display many features of spoken language as *conceptually oral*. A case in point are texts from computer-mediated communication (CMC) such as chat, facebook comments or Twitter messages.

dependency parser where the POS tags are provided as features, thus making it easier to directly compare their impact on the parsing results. In contrast to Kübler and Maier (2014), we do not compare the STTS with a general version of the tagset where all tags have been modified. Our tagset only applies linguistically motivated changes to specific tags, namely to those dealing with modification. As these are fairly frequent, we hypothesise that data sparseness will not be a big issue and that a theoretically well-funded analysis will have a positive impact on parsing results.

Relevant to our work is also the study by Plank et al. (2014), who discuss the problems of unreliable POS annotations. They show that incorporating annotator disagreements into the loss function of the POS tagger does yield better results not only on different POS tagsets but also in an extrinsic evaluation where these POS tags are used as input to a syntactic chunker.

This study is of interest to us as it gives some evidence that providing the parser with more specific information on ambiguous word forms might improve parsing. Our approach, however, is different from the one in Plank et al. (2014) who do incorporate the ambiguity in the tagging model. Instead, we aim at reducing the ambiguity in the data by refining the tagset and thus by providing the parser with more useful information.

Dalrymple (2006) follows the question how much POS tagging can help for reducing ambiguity during parsing. She presents a thorough study assessing the impact of POS tagging on parse disambiguation, applied to the output of a large-scale English LFG parser. Her findings show that presenting the parser with perfect tags would resolve ambiguitiy for around 50% of the parse trees, but that for 30% of the sentences in the test corpus even perfect POS tags would not help to disambiguate the parser output. In contrast to our work, Dalrymple does not investigate in how far modifications to the tagset might help.

3 The annotation scheme

In the standard part of speech tagset for German, the STTS, about 54 tags were defined which can be categorised into eleven major classes on a less fine-grained level (Schiller et al. 1999, pp. 4f). 48 of the tags represent word classes as such, six tags refer to punctuation marks, special characters, truncated word parts, and non-German words. The classification is based on very heterogeneous criteria – some definitions refer to the word's inflectional status (as for subclasses of verbs there are distinct categories for finite and infinite verb forms, past participles, and imperatives), to its syntactic status (as for predicative/adverbial vs. prenominal adjectives or attributive vs. substitutive pronouns), to semantic classes (e.g. different kinds of pronouns like demonstrative, indefinite or possessive pronouns), or to pure lexical classes (the word class PTKNEG (negated adverb) is represented by exactly one lexical form *nicht* (not); the same is true for all subclasses of the major class "particle" apart from the morphological class PTKVZ (verb particle)).

While all the major parts of speech contain at least two subclasses, the open word class ADV (adverb) is the only one which has not been subdivided any further. The STTS, in fact, does provide a part of speech tag PAV (pronominal adverb). This class is a purely morphologically or lexically defined class, which contains words with a prepositional and a pronominal component (words like *darauf* (literally: on that)). These words, however, are, similarly to prepositional phrases, syntactically extremely heterogeneous: they can occur as prepositional objects (*Ich warte* **darauf** (I am waiting for that)) or as adverbials (**Darauf** *solltest du nicht treten* (You should not step on that)). From a syntactic or functional perspective, only in the second case they can be regarded as adverbs. For that reason we, like most grammars, treat pronominal adverbs strictly as a morphological class which hierarchically stands above all syntactically motivated word classes and should not be mixed up with them.

According to the STTS, adverbs are defined as modifiers of verbs, adjectives, adverbs, or clauses, which are not derived from adjectives (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. As a matter of fact, the category ADV in the STTS tagset can be described as a residual category. This situation is unsatisfactory for the annotation of cor-

pora which are intended for the study of adverbs, particles, or one of the other parts of speech mentioned above. Therefore, we would like to propose a more fine-grained subcategorisation of the residual class ADV in the STTS tagset.

With regard to the fact that the part of speech category ADV in the STTS contains different word classes, we have divided the class ADV into "real" adverbs (ADV), modal particles (MODP), and other particles (PTK). The PTK category is further subdivided into focus particles (PTKFO), intensifiers (PTKINT), and lexical particles (PTKLEX). These classes are defined from a purely *functional syntactic* perspective, which does not incude semantic classes like temporal or manner adverbs which are specific semantic subcategories of the class ADV. Furthermore, we redefine the dissociation of adverbs (ADV) and adjectives (ADJD) in favour of a syntactically motivated notion of lexical modifiers. In the following section, we will first describe the newly defined classes which are already present in Schiller et al. (1999). Then we will discuss the new part of speech categories.

3.1 ADV vs ADJD

The distinction between the STTS categories ADV and ADJD is motivated inflectionally: Words that cannot be inflected and modify heads of any kind are, according to Schiller et al. (1999), p. 56, classified as adverbs (ADV). Words that can be inflected but are used as adverbials or predicatives are categorised as adjectives (ADJD) (see Schiller et al. 1999, p. 23). We argue, however, that this distinction is syntactically irrelevant and also hard to operationalise. Consider the following examples (7-12).

(7) Sie hat **behände**/ADV (?) den Baum
She has skilfully the tree
beklettert.
climbed.
"She has skilfully climbed the tree."

(8) Sie hat **elegant**/ADJD den Baum beklettert.
She has elegantly the tree climbed.
"She has elegantly climbed the tree."

(9) Sie hat **oft**/ADV den Baum beklettert.
She has often the tree climbed.
"She has often climbed the tree."

(10) Sie hat **häufig**/ADJD (?) den Baum
She has frequently the tree
beklettert.
climbed.
"She has frequently climbed the tree."

(11) Sie hat **wahrscheinlich**/ADJD (?) den
She has probably the
Baum beklettert.
tree climbed.
"She has probably climbed the tree."

(12) Sie hat **vielleicht**/ADV den Baum beklettert.
She has perhaps the tree climbed.
"Perhaps she has climbed the tree."

According to the STTS, the words in bold are assigned the tags shown above (examples (7)-(12)). However, from a syntactic perspective it is hard to justify that the different modifiers in (7)-(12) belong to fundamentally different categories; they have the same inflectional status, their distribution is exactly the same, and they have similar syntactic functions insofar as they are all modifying the main verb or are attached at a higher level in the respective sentence.[2] Since we assume that part of speech categories are often the basis for further syntactic analysis, this is our main argument against an inflectional morphological approach for distinguishing adverbs and adjectives. Furthermore, there are conceptional problems for the operationalisation offered in Schiller et al. (1999) and in many German grammars.

The different tags shown in (7)-(12) result from one particular feature of the modifier in question, namely from its *inflectibility* (+*infl.*→ADJD, -*infl.*→ADV). This means that if a given modifier can be used adverbially and at the same time prenominally, it has to be classified as ADJD. Since the feature *inflectibility* cannot be tested properly (there is, for instance, no general agreement on the question whether *hoffentlich* (hopefully) is inflectible or not), another syntactic test is given in the guidelines (Schiller et al. 1999, p. 57): If the word in question can be used as a predicative adjective, it has to be annotated as ADJD

[2]The different semantic classes have a different scope which has provable distinct syntactic effects. This is why different kinds of adverbials are not only discussed from a semantic, but also from a syntactic point of view. Here we subsume all different kinds of adverbs (like adverbial versus adsentential adverbs) under one category 'adverb' (ADV).

(*sie ist elegant*/ADJD (she is elegant); **das ist oft* (this is often) →oft/ADV).

Inflectibility and the ability to function as a predicate, however, are independent features; words can be uninflectible but, at the same time, be used as a predicate (*er ist pleite* (he is broke) – *ein *pleiter Mensch* (a broke guy)), and there also are inflectible forms which cannot be used as predicates (*der eigentliche Termin* (the actual date) – *der Termin ist *eigentlich* (the date is actual)).

Not only can the tests for distinguishing adjectives from adverbs provide contradictory outcomes, in many cases they simply fail. For instance, acceptability judgments by German native speakers do not give a clear picture on whether examples (13)-(15) are grammatical or not.

(13) Der Sprung war behände.
The jump was agile

(14) Der Vorfall war häufig.
The incident was frequent.

(15) eine wahrscheinliche Baumbesteigung
a probable tree climb

To get rid of the inflectibility criterion, we propose that all adverbial or adsentential modifiers (like the ones in 7-12) are analysed as adverbs, whereas uninflected adjectives have to be used as a syntactic predicate in order to be tagged as ADJD. This means that only complements of copula verbs are tagged as predicative adjectives.[3]

3.2 Particles

Since the residual category ADV in the STTS guidelines (Schiller et al., 1999) includes different kinds of particles (a fact not discussed in the guidelines themselves), we move these to the main class PTK of the STTS which, so far, includes the tags PTKA (particle with adjective or adverb), PTKANT (answer particle), PTKZU (*zu* (to) with infinitive), and PTKVZ (separated verb particle). Particles are modifiers which can not, on their own, stand in the German pre-field (Vorfeld) and which, in general, can not be moved around freely in the sentence but which are restricted to appearing adjacent to a specific lexical head. This can be tested easily by human annotators with the help of permutation tests – if a given modifier cannot be placed (alone) in the pre-field position, it will be analysed as a particle. We distinguish between three different types of particles.

3.2.1 Focus particles – PTKFO

Focus particles are associated with a given focus element and modify the set of alternatives which is connected with the focus itself. Consider examples (16) and (17) below.

(16) Petra ist **nur** zum KLETTERN
Petra is only for rock climbing
gekommen.
went.
"Petra only came for rock climbing"

In (16), the focus is on *klettern* (rock climbing). The particle *nur* (only) is associated with the focus and opens up a set of alternatives (any other activity). However, the modifier *nur* tells us that none of the other activities besides rock climbing should be considered in this context.

(17) Petra hat **sogar** UNTER dem Tisch
Petra has even under the table
nachgeschaut.
looked.
"Petra has even looked under the table."

In (17), the focus is *unter* (under), the set of alternatives includes any other positions in relation to the table, and the focus particle *sogar* (even) tells us that all the other possible alternatives are valid options as well (on the table, next to the table, ...).

3.2.2 Intensifiers – PTKINT

Intensifiers are expressions of graduation, intensification, or quantification. In most cases, they are modifying (gradable) adjectives or adverbs. In (18), *sehr* (very) is intensifying the adverb *kurz* (shortly) while in (19), *überaus* (extremely) strengthens the adjective *groß* (great).

[3]Please refer to Hirschmann (2014) for more detailed information on the distinction between adverbs and adjectives: https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/hirschmann-adv-stts.pdf.

(18) Petra ist **sehr** kurz zum
Petra is very shortly to the
Klettern gegangen.
rock climbing went.
"Petra went rock climbing for a very short time."

(19) Petra hat **überaus** großen Hunger.
Petra has extremely great hunger.
"Petra is extremely hungry."

3.2.3 Lexical particles – PTKLEX

Lexical particles are associated with a lexical head element with which they form a complex lexeme. In (20), for example, the complex lexeme *nicht mehr* (not any more) is composed of the head *nicht* and the lexical particle *mehr*, while in (21), we have a complex lexeme *immer noch* (still) with *noch* as the head. The meaning of the complex lexeme can not be derived by a compositional analysis of its individual components.

(20) Petra gefällt das [nicht **mehr**]
Petra pleases this not more
"Petra doesn't like that any more"

(21) Petra gefällt das [**immer** noch]
Petra pleases this always still
"Petra still likes that"

3.2.4 Modal particles – MODP

Modal particles (like particles in general) are also not *vorfeldfähig*, meaning they can not on their own fill the pre-field position in a Standard German sentence. They can, however, be placed relatively freely within the German middle field (Mittelfeld), a crucial feature which does not apply to any other type of particle. Because of this – and also for other semantic-syntactical reasons (modal particles modify the sentential level of a given clause) – we consider modal particles as a distinct major class. Modal particles can be treated as a closed word class. Please refer to the tagging guidelines by Hirschmann (2014) for a comprehensive list of candidates.

4 Annotation experiment

To test the new classification, we applied it to 1000 sentences randomly selected from the TiGer treebank and reassigned labels to all tokens where the original tag was one of either ADJD (adverbially used or predicative adjective), ADV (adverb), or a past participle[4] (VAPP, VVPP). In the beginning, the annotators were presented with the original POS tags. As we had the impression that this influenced the annotators' decision, we replaced all instances of the modifier tags with the same dummy tag.

POS	# orig	# new	# agr.	Fleiss' κ
ADJD	191	74	63	0.891
ADV	445	378	343	0.800
MODP	-	12	6	0.515
PTKFO	-	80	67	0.797
PTKINT	-	63	49	0.788
PTKLEX	-	33	17	0.594
VAPP	21	21	21	1.000
VVPP	173	172	172	0.989
total	**830**	**833**	**88.3%**	**0.838**

Table 1: Distribution (orig, new) and agreement (percentage agreement and Fleiss' κ) for the different tags

We started off with annotating samples of 100 sentences, then discussed the mismatches and updated the annotation guidelines. After having finished the first 400 sentences (samples 1-4), we annotated a larger batch including the remaining 600 sentences of our goldstandard. As we still made changes to the guidelines at this stage, we report inter-annotator agreement on an additional test set of 500 sentences from Tiger (sentence 9501-10000).

Our test set includes 830 instances of modifiers which had to be re-annotated (Table 1).[5] The annotators could assign one of the tags ADV, ADJD, MODP, PTKFO, PKTINT, PKTLEX, VAPP, VVPP. We achieved an inter-annotator agreement of 0.838 (Fleiss' κ), and an overall percentage agreement for all modifier tags of 88.3%.

Table 1 also shows that modal particles (MODP) and lexical particles (PTKLEX) are the most difficult ones to annotate, maybe partly due to their low frequency in the corpus.

[4] We included past participles in the annotation as some of them had to be reannotated as ADJD → ADV.

[5] The numbers for the original data set and the reannotated set vary slightly, as also some other instances not labelled as ADV or ADJD in TiGer have been assigned a new label, e.g. *"um/KOUI/PKTLEX so scheinheiliger"* (so much more sanctimonious).

	ADJD	ADV	PFO	PINT	PLEX	MODP
ADJD	63	6	0	0	0	0
ADV	6	343	15	6	6	5
PFO	0	12	67	2	1	0
PINT	0	9	0	49	2	0
PLEX	0	9	0	1	17	0
MODP	0	5	0	0	1	6

Table 2: Confusion matrix for adverbs (ADV), predicative adjectives (ADJD), focus-associated particles (PFO), intensifiers (PINT), lexicalised particles (PLEX) and modal particles (MODP)

4.1 Ambiguous cases

Below we show some examples where the annotators disagreed. The confusion of ADV and ADJD mostly concerned cases like (22) where the lexeme in question was interpretated as a verb modifier (ADV) by one annotator and as a predicative adjective by the other. These cases can be handled by providing more specific instructions in the annotation guidelines, e.g. by providing a list of potential copula verbs which link the subject to the adjectival predicate.

(22) ADV vs ADJD

Wer sich weigere, werde durch Drogen
Who himself refuses, is by drugs
gefügig gemacht
compliant made

"Who refuses is made compliant by drugs"

For the distinction between adverbs (ADV) and focus particles (PTKFO), many cases were indeed ambiguous (see example 23). It is not clear how much context should be taken into account in order to resolve the ambiguity in the sentence. In our experiments, we decided to only use the sentence context in order to speed up the annotation process, and to use the combined label ADV:PTKFO for those cases which could not be resolved during adjudication. However, often the annotators were only aware of one of the possible readings, which resulted in many disagreements for these tags.

(23) ADV vs PTKFO

Hennemann hatte seinen Rückzug **bereits**
Hennemann had his withdrawal already
im September angeboten.
in September offered.

"Hennemann had already offered his withdrawal in September."

Better agreement can be achieved especially for the lexicalised particles (24), which mostly consist of frequent, co-occurring lexemes. Many disagreements concerned new instances which had not been seen before. Listing the most frequent instances in the guidelines might improve inter-annotator agreement for PTKLEX.

(24) ADV vs PTKLEX

Diese werden **immer wieder** missbraucht
These become always again abused

"Again and again, these become abused"

5 Parsing experiments

This section presents a parsing experiment where we test the learnability of our new classification using a statistical dependency parser.

5.1 Data expansion

To obtain more training data than the manually annotated 1000 sentences, we extracted patterns from the goldstandard capturing the syntactic context in which each of the new tags might occur, and applied them to the whole TiGer treebank.

Example (25) shows such a pattern. It extracts all tokens #p which have a lemma form from a predefined list (*rund* (around), *etwa* (about), *kaum* (hardly), ...), which are assigned the grammatical function MO (modifier), and which are directly followed by a cardinal number which has the same mother node as #p. We use TiGerSearch for pattern extraction, identify the terminal ids of the #p nodes and assign the new tag PTKINT (intensifier) to all #p.

(25) #p:[lemma=("rund"|"etwa"|...|"kaum")] &
#p . #card:[pos="CARD"] &
#mother >MO #p &
#mother > ∗ #card

Another example is shown in (26). Here we look for a token with the POS tag ADV (adverb) which is the leftmost child of an NP and which has one of the following lemma forms: *allein*

(only), *auch* (also), ..., *zwar* (indeed). These instances are then relabelled as PTKFO (focus particles).

(26) #cat:[cat="NP"] >@l #p:[pos="ADV"] &
#p:[lemma=("allein"|"auch"|...|"zwar")]

Overall, we defined 49 different patterns, which assigned tags to 90.9% of the modifiers in the sample. Sometimes, these patterns overgeneralise. We manually checked potential errors in the first 5000 sentences of the treebank and manually annotated the remaining 478 cases which were not captured by our pattern approach. After the manual clean-up we had an additional data set with 4922 new sentences (86,517 tokens).[6] This dataset is not as "high-quality" as the 1000 sentences of the goldstandard which have been individually annotated from scratch by the authors, and where all disagreements have been resolved in discussion. However, as we do not evaluate the accuracy of the POS tags themselves but the impact of the new classification on parsing accuracy where we only evaluate the dependency labels and relations, this is not a problem for our experimental setup. Table 3 shows the distribution of our new tags in the goldstandard and in the expanded dataset.

Tag	**gold**	**expanded**
ADJD	142	478
ADV	686	3,289
MODP	18	36
PTKFO	161	675
PTKINT	135	516
PKTLEX	54	201
ambiguous tags		
ADJD:ADV	1	-
ADV:MODP	1	-
ADV:PTKFO	22	-
ADV:PTKINT	2	-
ADV:PTKLEX	1	-
PTKFO:PTKINT	1	-
Total	1,224	5,195

Table 3: Distribution of the different modifier classes in the goldstandard

[6] 78 of the 5000 sentences were already included in the goldstandard.

5.2 Setup

The parsers we use in our experiments are the Malt parser (Nivre et al., 2007) and the MATE parser (Bohnet, 2010), both language-independent systems for data-driven dependency parsing. We trained the parsers on the first 5000 sentences from the TiGer treebank and evaluated them in a 10-fold crossvalidation setting. The parsers have been trained on two different versions of the data, a) on the original treebank trees, and b) on the same trees, but replacing the original POS tags with our new POS classification.

For each version of the data, we separately optimised the parameters for the Malt parser, using MaltOptimizer (Ballesteros and Nivre, 2012), and then trained the parser with the parameter and feature settings optimised for each dataset.

5.3 Results

Table 4 shows labelled attachment scores (LAS) for the 10 folds and averaged scores for the whole dataset. For both, Malt and MATE parser, we observe a small, but highly significant difference between the two datasets.[7]

	Malt		MATE	
fold	**orig**	**new**	**orig**	**new**
1	84.0	84.3	85.4	86.3
2	84.2	84.7	87.1	87.6
3	89.0	89.3	91.7	91.7
4	85.3	85.9	88.5	89.1
5	89.0	88.9	91.2	91.5
6	86.0	85.5	88.0	88.4
7	86.0	86.2	88.7	89.2
8	89.1	89.2	91.6	91.9
9	89.7	89.8	92.0	92.1
10	85.0	85.9	87.4	88.1
avg.	**86.7**	**87.0**	**89.2**	**89.6**

Table 4: Parsing results (Malt and MATE parsers, LAS) for original and new tags

This difference becomes more substantial when only looking at the modifier (MO) dependency relation. Table 5 shows precision, recall and f-score for the 10 folds and results averaged over all folds for the combined evaluation of dependency relation and attachment for the label

[7] For significance testing we used Dan Bikel's Randomized Parsing Evaluation Comparator with $n = 10000$.

		orig			new		
fold	freq.	prec.	rec.	f1	prec	rec.	f1
1	1301	72.2	70.4	71.3	76.2	74.5	75.3
2	1261	73.9	71.7	72.8	76.5	73.8	75.2
3	916	78.4	76.3	77.3	81.1	77.5	79.2
4	1159	74.2	73.5	73.8	77.9	77.0	77.5
5	1031	76.4	75.7	76.1	79.7	79.1	79.4
6	1125	75.1	74.9	75.0	76.7	77.0	76.8
7	1151	75.2	73.6	74.4	77.8	76.7	77.3
8	978	76.9	78.2	77.6	80.0	79.6	79.8
9	867	81.8	79.2	80.5	82.2	80.5	81.3
10	1081	73.6	73.4	73.5	77.2	78.5	77.8
avg.	1087	75.8	74.7	**75.2**	78.5	77.4	**78.0**

Table 5: Precision, recall and f-score for dependency relation and attachment for MO (MATE parser)

MO.[8] Here the gap is nearly 3 percentage points (MATE parser), giving evidence that our syntactically motivated classification of modifiers supports the parser in analysing these structures.

Table 6 shows that our new tag distinctions not only help when analysing MO dependencies but also improve results for other dependencies.

		orig			new		
DEP	freq.	prec.	rec.	f1	prec	rec.	f1
CJ	2497	84.5	83.1	83.8	85.0	83.4	84.2
DA	533	86.1	78.0	81.9	87.8	78.4	82.8
MNR	2618	64.9	67.5	66.2	65.3	68.6	66.9
NG	496	75.1	75.6	75.4	76.3	76.4	76.3
OP	846	57.8	33.0	42.0	57.7	33.6	42.4
PD	879	77.2	70.2	73.5	81.5	71.3	76.1
RE	272	58.5	50.7	54.3	64.0	53.7	58.4
SBP	182	71.5	78.6	74.9	76.0	80.2	78.1

Table 6: Precision, recall and f-score for other dependency relations (and attachment) where the new tags improved results (MATE parser; CJ: conjunct, DA: dative object, MNR: postnominal modifier, NG: negation, OP: prepositional object, PD: predicate, RE: repeated element, SBP: passivised subject)

6 Conclusions and future work

The results presented in the paper are interesting in many ways. First of all, we proposed an extension to the STTS which gives a more detailed, as well as linguistically well-founded analysis of modifiers in German. This is of interest especially for spoken and conceptually spoken language such as CMC data, where modifiers are extremely frequent and an analysis based on the core STTS tags is not very informative. Second, we presented an annotation study where we tested the applicability of the new classification to newspaper text. We discussed the problems arising during annotation, which are mostly based on real ambiguities in the data. The new annotations are available to the research community.[9]

Last, and most important, we gave proof of concept that a more detailed analysis of modification on the POS level which is linguistically motivated can indeed support data-driven syntactic parsing.

So far, we have only shown that our new classification scheme does improve data-driven syntactic parsing of modification relations when providing the parser with gold (or, as for our extended dataset, with nearly gold standard) tags. It remains to be shown that the new tags can be learned by a POS tagger (or parser) with sufficient accuracy to be useful to the parser. Also, the parsing results are based on a small testset only and thus need to be validated on a larger dataset. Additional annotations are under way, and we plan to address both issues in future work.

Acknowledgments

This work was supported by a grant from the German Research Association (DFG) awarded to SFB 632 "Information Structure" of Universität Potsdam, Humboldt Universität Berlin and Freie Universität Berlin, Project B6: "The Kiezdeutsch Korpus (KiDKo)".

References

Miguel Ballesteros and Joakim Nivre. 2012. 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*, EACL '12.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER

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

[9] Download from `https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/tiger_adv.tgz`

treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42.

Mary Dalrymple. 2006. How much can part-of-speech tagging help parsing? *Natural Language Engineering*, 12(4):373–389.

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

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

Peter Koch and Wulf Oesterreicher. 1985. Sprache der nhe – sprache der distanz. mündlichkeit und schriftlichkeit im spannungsfeld von sprachtheorie und sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

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

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

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

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. 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)*.

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

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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.

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