Optimizing a PoS Tag Set for Dependency Parsing

Anonymous ACL submission

Abstract

Abstraction ensues.

1 Introduction

Part-of-speech (PoS) tagging is an important preprocessing step for many NLP tasks, such as dependency parsing (Nivre et al., 2007; Hajič et al., 2009), named entity recognition (Tjong Kim Sang and De Meulder, 2003) and sentiment analysis (Wilson et al., 2009). Whereas much effort has gone into the development of PoS-taggers – to the effect that this task is often considered more or less a solved task – considerably less effort has been devoted to the empirical evaluation of the PoS tag sets themselves. Error analysis of PoS taggers indicate that, whereas tagging improvement through means of learning algorithm or feature engineering seems to have reached something of a plateau, linguistic and empirical assessment of the distinctions made in the PoS tag sets may be a venue worth investigating further (Manning, 2011). Clearly, the utility of a PoS tag set is tightly coupled with the downstream task for which it is performed. Even so, PoS tag sets are usually employed in a "one size fits all" fashion, regardless of the requirements posed by the task which makes use of this information.

It is well known that syntactic parsing often benefits from quite fine-grained morphological distinctions (Zhang and Nivre, 2011; Seeker and Kuhn, 2013). Morphology interacts with syntax though phenomena such as agreement and case marking, and incorporating information on morphological properties of words can therefore often improved parsing performance. However, in a realistic setting where the aim is to automatically parse raw text, the generation of morphological information will often require a separate step of morphological analysis that can be quite costly.

In this paper, we optimize a PoS tag set for the task of data-driven dependency parsing of Norwegian. We report on a set of experiments where various morphological distinctions are introduced to the PoS annotations and evaluated both intrinsically, i.e., in terms of PoS tagging accuracy, and extrinsically, in terms of parsing accuracy. Our results show that the introduction of morphological distinctions not present in the original tag set, whilst compromising tagger accuracy, actually leads to significantly improved parsing accuracy. This optimization allows us to bypass the additional step of morphological analysis, framing the whole preprocessing problem as a simple tagging task.

The article is structured as follows. We start out by reviewing previous work on tag set evaluation in Section 2, while Section 3 details the treebank that provides the initial gold annotations used for our experiments. Section 4 describes the experimental setup used in our work and Section 5 goes on to provide the results from our tag set optimization. Finally, Section 6 summarizes our main findings and discusses some avenues for future work.

2 Previous Work

There has been little previous work dedicated specifically to extrinsic evaluation of the effects of PoS tag sets on downstream applications. There has, however, been some work that evaluates the effects of PoS tag set granularity on PoS tagging. This includes investigation of the effects of PoS tag sets on tagging of Swedish (Megyesi, 2001; Megyesi, 2002) and English (MacKinlay, 2005).

Megyesi (2002) trained and evaluated a range of PoS taggers on the Stockholm-Umeå Corpus (SUC) (Gustafson-Capková and Hartmann, 2006), annotated with a tag set based on a Swedish version of PAROLE tags totaling 139 tags. Furthermore, they investigated the effects of tag set size on tagging by mapping the original tag set into smaller subsets designed for parsing. They argue that a tag set with complete morphological tags may not be necessary for all NLP applications, for instance syntactic parsing. They found that the smallest tag set comprising 26 tags yields the lowest tagger error rate. However, for some of the taggers, augmenting the tag set with more linguistically informative tags may actually lead to a drop in error rate. They argue that this shows that the size of the tag set as well as the type of information in the tags are crucial factors for tagger performance. However/unfortunately, they do not report results of parsing with the various PoS tag sets.

Similarly, MacKinlay (2005) investigated the effects of PoS tag sets on tagger performance in English, specifically the Wall Street Journal portion of Penn Treebank (Marcus et al., 1993). They mapped the original tag set of Penn Treebank comprising 45 tags to more fine-grained tag sets using linguistic insight to investigate whether additional linguistic information included in finer-grained tags could assist the tagger. Experimenting with both lexically and syntactically conditioned modifications such as distinguishing between count nouns and noncount nouns and between transitive and intransitive verbs, they found that more fine-grained tag sets rarely led to improvements in tagger accuracy; the most successful modification yielded an improvement in tagger accuracy of 0.05 percentage points. They did not find any statistically significant improvements using linguistically informed distinctions, arguing that their results do not support the hypothesis that it is possible to achieve significant performance improvements in PoS tagging by utilizing a finer-grained tag set.

Transition/overgang her ...

3 The Norwegian Dependency Treebank

We used the newly developed Norwegian Dependency Treebank (NDT) (Solberg et al., 2014), the first publicly available treebank for Norwegian. It was developed at the National Library of Norway in collaboration with the University of Oslo, and contains manual syntactic and morphological annotation. The treebank contains data from both varieties of Norwegian; 311 000 tokens of Bokmål and 303 000 tokens of Nynorsk. We will in the following only be using the Bokmål portion of the treebank. The annotated texts are mostly newspaper text, but also include government reports, parliament transcripts and excerpts from blogs. The annotation process of the treebank was supported by the rule-based Oslo-Bergen Tagger (Hagen et al., 2000) and then manually corrected by annotators, adding syntactic dependency analysis to the morphosyntactic annotation.

Tag	Description
`adj`	Adjective
`adv`	Adverb
`det`	Determiner
`inf-merke`	Infinitive marker
`interj`	Interjection
`konj`	Conjunction
`prep`	Preposition
`pron`	Pronoun
`sbu`	Subordinate conjunction
`subst`	Noun
`ukjent`	Unknown (foreign word)
`verb`	Verb

Table 1: Overview of the PoS tag set of NDT.

Morphological Annotation The morphological annotation and PoS tag set of NDT follows the Oslo-Bergen Tagger (Hagen et al., 2000; Solberg, 2013), which in turn is largely based on the work of Faarlund et al. (1997). The tag set consists of 12 morphosyntactic PoS tags outlined in Table 1, with 7 additional tags for punctuation and symbols. The tag set is thus rather coarse-grained, with broad categories such as `subst` (noun) and `verb` (verb). The PoS tags are complemented by a large set of morphological features, providing information about morphological properties such as definiteness, number and tense. These features are used in our tag set modifications, where the coarse PoS tag of relevant tokens is concatenated with one or more of these features to include more linguistic information in the tags.

Syntactic Annotation The syntactic annotation choices in NDT are largely based on the Norwegian Reference Grammar (Faarlund et al., 1997). The annotation choices are outlined in Table 2, taken from/courtesy of Solberg et al. (2014), providing overview of the analyses of syntactic constructions that often distinguish dependency treebanks, such as coordination and the treatment of

Head	Dependent
Preposition	Prepositional complement
Finite verb	Complementizer
First conjunct	Subsequent conjuncts
Finite auxiliary	Lexical/main verb
Noun	Determiner

Table 2: Central head-dependent annotation choices in NDT.

auxiliary and main verbs. The set of dependency relations comprises 29 dependency relations, including ADV (adverbial), SUBJ (subject) and KOORD (coordination).

4 Experimental Setup

In preparation to conducting our experiments with linguistically motivated tag set modifications, a concrete setup for the experiments needed to be established, which is presented in the following.

Our approach is somewhat semi-automatic in that we employ a "hybrid" combination of manual annotation and automatic labeling in our experiments. Our initial tags are gold standard, and using linguistic insight coupled with computational considerations, we introduce new features or feature combination to the tag set. These tags are run through a pipeline with automatically assigned tags in both training and testing, where the most promising modifications are used "further". Hence, the tag set modifications are not deterministic, nor do we try all possible combination; we only experiment with linguistically motivated distinctions. We are in reality altering the gold standard tags by the addition of more linguistic information.

Data Set Split As there was no standardized data set split of NDT due to its very recent development, we needed to establish a data set split (training/development/test). Our data set split of the treebank follows the standard 80-10-10 (training/development/test) split and will be distributed with the treebank and proposed as the new standard(?). In creating the data set split, care has been taken to preserve contiguous texts in the various data sets while keeping the split balanced in terms of genre. Our proposed data set split was used in the Norwegian contribution to the Universal Dependencies project (Øvrelid and Hohle, 2016). The split will be made available at a companion website.

Tagger For our experiments with tag set modifications, we sought a PoS tagger both reasonably fast and accurate. There is a often a considerable trade-off between the two factors, as the most accurate taggers tend to suffer in terms of speed due to their complexity. However, a tagger that achieves both close to state-of-the-art accuracy as well as very high speed is TnT (Brants, 2000). TnT was furthermore recently used to evaluate the recently proposed universal tag set (Petrov et al., 2012). The sum of these factors led to TnT being the tagger of choice for our experiments.

Parser In choosing a syntactic parser for our experiments, we considered previous work on dependency parsing of Norwegian, specifically that of Solberg et al. (2014), who found the Mate parser (Bohnet, 2010) to be the most successful parser for parsing of NDT. Furthermore, recent dependency parser comparisons (Choi et al., 2015) showed that Mate performed very well on parsing of English, beating a range of contemporary state-of-the-art parsers. Mate was consequently selected to evaluate the tag set modifications in our experiments.

Tag Set Mapping The tag set modifications are realized by mapping the existing tags to new, more fine-grained tags "based" on associated morphological features to be included in specified tags. The procedure maps the relevant existing tags to new, more fine-grained tags including more relevant morphological features for the applicable tokens, i.e., tokens that are assigned said tag and feature(s) in the gold standard data. Hence, we are in actuality altering the gold standard tags by the addition of more relevant linguistic information, producing new gold standard annotations of higher quality.

Baseline It is common practice to compare the performance of PoS taggers to a pre-computed baseline for an initial point of comparison. For PoS tagging, a commonly used baseline is the Most Frequent Tag (MFT) baseline, which we used in our experiments. This involves labeling each word with the tag it was assigned most frequently in the training. All unknown words, i.e., words not seen in the training data, are assigned the tag most frequently assigned to words seen only once in the training. Unknown and infrequent

Training	Testing	LAS	UAS
Gold	Gold	90.15%	92.51%
Gold	Auto	85.68%	88.98%
Auto	Auto	87.01%	90.19%

Table 3: Results of parsing with Mate using the various tag configurations, either using gold standard tags or automatically assigned from TnT. *Gold* denotes gold standard tags, *Auto* denotes automatically assigned tags from TnT.

words have in common that they rarely occur, and we might therefore expect them to have similar properties.

Tags & Features As we seek to quantify the effects of PoS tagging in a realistic setting, i.e., in application to raw text, we want to evaluate the parser on automatically assigned PoS tags. For the training of the parser, however, we have two options: using either gold standard or automatically assigned tags. In order to settle on a configuration, we conducted experiments with gold standard and automatically assigned tags to see how they differ with respect to performance. The results of these experiments, shown in Table 3, reveal that the combination of training and testing on automatic tags is clearly superior to training on gold standard tags and testing on automatic tags. Consequently, the parser was both trained and tested on automatically assigned tags in our experiments.

We removed any morphological features in order to simulate a realistic setting. Moreover, it is crucial that we remove these features when working with automatically assigned tags, as the features are still gold standard. For instance, if a verb token is erroneously tagged as a noun, we could potentially have a noun token with verbal features such as tense, which markedly obfuscates the training and parsing. Another important aspect is that we want to isolate the effect of PoS tags, necessitating the exclusion of morphological features.

In the following, we report on a set of experiments which evaluate increasingly fine-grained PoS tag sets. The PoS tag sets are motivated by morphosyntactic properties of Norwegian and are evaluated in terms of both tagger and parser accuracy.

5 Tag Set Optimization

By introducing more fine-grained linguistically motivated distinctions in a tag set, we increase the linguistic information represented in the tags, which may assist the parser in recognizing and generalizing syntactic patterns. However, the addition of more linguistic information to the tags and thus a more fine-grained tag set will most likely lead to a drop in tagger accuracy due to the increase in complexity. The best tagging does not necessarily lead to the best parse, hence it is interesting to investigate how linguistically informed tag set modifications may affect the interplay between tagging and parsing.

5.1 Baseline Experiments

In an initial round of experiments, we concatenated the tag of each token with its set of morphological features in order to map the original tag set to a new, more fine-grained tag set (hereafter referred to as the *full* tag set). This resulted in a total of 368 tags, which is clearly very fine-grained. The two initial tag sets, i.e., the original tag set comprising 19 tags and the full tag set comprising 368 tags, thus represent two extremes in terms of granularity. To evaluate the tag sets and investigate how the tag set granularity would affect the performance of tagging and parsing, we trained and evaluated TnT and Mate on the training and development data, respectively, on the two tag sets. In Table 4, we report the results of these experiments. We see that the tagger accuracy drastically drops when going from the original to the full tag set. TnT reports an accuracy of 97.47% on the original tag set, which is reduced to 93.48% for the full tag set. These results confirm our hypothesis that the very high linguistic quality in the full, fine-grained tag set comes at the expense of drops in tagger performance. However, the additional linguistic information provided by the full tag set improves the parser performance. With the original tag set, Mate reports an LAS of 87.01% and a UAS of 90.19%, which increases to 87.15% and 90.39%, respectively, when using the full tag set. As we are looking for linguistically informed distinctions that improve the syntactic parsing, these results are promising and serve to indicate that additional morphological information assists the syntactic parsers, motivating the optimization of the existing PoS tag set.

Tag set	MFT	Accuracy	LAS	UAS
Original	**94.14%**	**97.47%**	87.01%	90.19%
Full	85.15%	93.48%	**87.15%**	**90.39%**

Table 4: Results of tagging and parsing with the two initial tag sets.

5.2 Tag Set Experiments

We modified the tags for nouns, verbs, adjectives, determiners and pronouns in NDT by appending selected sets of morphological features to each tag in order to increase the linguistic information expressed by the tags. For each tag, we first experimented with each of the features in isolation before employing various combinations of them. We based our choices of combinations on how promising the features are and what we deem worth investigating in terms of linguistic utility, in order to see how the features might interact.

The morphological properties of the various parts-of-speech is reflected in the morphological features associated with the respective PoS tags. For instance, as nouns can take on gender, definiteness and number, additionally "being of" genitive case, the treebank operates with features for gender, definiteness, number and case with "accompanying" values. In addition to morphological properties such as definiteness, tense and number, all classes except for verbs can be distinguished "on" type, e.g., common nouns and proper nouns, while pronouns can be reflexive, reciprocal, personal or interrogative.

5.3 Experiments

"Nearly" all tag set modifications leads to a drop in tagger accuracy, as expected. Few exceptions... We observed increases in parser accuracy scores with linguistically motivated tag set modifications for all PoS tags we experimented with, the most improved being noun with an increase in LAS from the original tag set by 1.80 percentage points.

Nouns In Norwegian, nouns are assigned gender (feminine, masculine, neuter), definiteness (indefinite or definite) and number (singular or plural). There is agreement in gender, definiteness and number between nouns and their modifiers (adjectives and determiners). Additionally, NDT has a separate case feature to distinguishing nouns in genitive case.

Experimenting with nouns, we found that all tag set modifications yielded large increases in LAS and UAS. The most informative features are definiteness, which leads to an increase in LAS by 1.26 percentage points, to 88.27%, and type, yielding an LAS of 88.07%. Turning to combinations of features, we found that the combination of type and case, as well as type and definiteness, were the most promising, which led us to combine type, case and definiteness in a final experiment, resulting in LAS of 88.81% and UAS of 91.73%, constituting large increases from parsing with the original tag set, 1.80 percentage points and 1.54 percentage points, respectively.

The results from tagging and parsing with modifications to nouns are reported in Table **??**. As nouns constitute the largest class in the treebank by far (as shown in Table **??**), the effects are greatest for noun tokens in terms of overall change in performance. Apart from case, none of the tag set modifications improves the tagging. However, they all give rise to increases in parser accuracy scores. Genitive case marks possession, hence nouns marked with genitive case are quite different from other nouns, taking a noun phrase as complement. Distinguishing on type is useful and informative, as evident by the presence of separate tags for proper and common nouns in many tag sets, such as those of PTB, SUC and UD (described in Section **??**). When introducing the distinction of type, we see a large increase in parser accuracy scores, with an LAS of 88.07% and UAS of 91.11%, both exceeding the baseline by more than a percentage point. Definiteness is the most informative feature for parsing, achieving LAS of 88.27% and UAS of 91.42%.

Verbs Verbs are inflected for tense (infinitive, present, preterite, past perfect) in Norwegian and can additionally take on mood (imperative, indicative) and voice (active or passive). Note that voice and mood have only a single value, `pass` (passive) and `imp` (imperative), respectively. Verbs which are not passive are implicitly active, and verbs which are not imperative are indicative.

Introducing features in isolation, mood is the only feature leading to increases in LAS, with a reported LAS of 87.04%. Imperative clauses are fundamentally different from indicative clauses, as they lack an overt subject. Tense yields an LAS of 86.97% while voice yields an LAS of 86.96%. Combining mood and tense resulted in an LAS of 87.12% and UAS of 90.31%.

In an additional experiment, we mapped the verb tenses (mood, in the case of imperative) to finiteness. All verbs have finiteness, hence this distinction has broad coverage. This mapping is syntactically grounded as finite verbs and nonfinite verbs appear in completely different syntactic construction, and proved to greatly improve the parsing, as we saw the overall largest parser accuracy scores, with 87.30% for LAS and 90.43% for UAS, 0.29 and 0.24 percentage points higher than the baseline, respectively. This coincides with the observations seen for Swedish in Øvrelid (2008), where finiteness was found to be a very beneficial linguistic feature for parsing.

Adjectives Adjectives agree with the noun they modify in terms of gender, number and definiteness in Norwegian. Adjectives are also inflected for degree, either positive, comparative or superlative.

All features except for number (LAS of 86.99%) led to increases in parser accuracy scores. Degree being the most promising with a reported LAS of 87.29%, while distinguishing adjectives on definiteness yield an LAS of 87.14% and gender leads to LAS of 87.10%.

Combining the most promising features, we found that none of the combinations surpass the LAS with the distinction of degree alone. Definiteness and degree with an LAS of 87.23% and definiteness and number with an LAS of 87.27% and UAS identical to that of degree alone. The most fine-grained distinction, definiteness, degree and number, yields an LAS of 87.14%.

Determiners Like adjectives, determiners in Norwegian agree with the noun they modify in terms of gender, number and definiteness. Possessive pronouns are treated/analyzed as determiners in NDT.

For a number of reasons, we do not report results from combinations of features for determiners, as the features could not be combined in any meaningful way. For instance, determiners in plural never have marked definiteness, hence this combination would only apply to determiners in singular and roughly correspond to the distinction on definiteness alone. Moreover, determiners in plural are not distinguished in terms of gender (thus ruling out the combination of number and gender), neither are definite determiners. Another aspect taken into consideration is that the most promising distinction, definiteness, applies to such a small number of tokens that more fine-grained distinctions would be overly sparse.

The results from the experiments with determiners are shown in Table **??**. Introducing the type led to an increase in tagger accuracy by 0.14 percentage points to 97.61%, while marginally impacting the parsing, with LAS of 87.00%, 0.01 percentage points below the baseline, and UAS of 90.11%, 0.08 percentage points below the baseline. The increase in tagger accuracy when introducing the distinction of type is noteworthy, as we expected the finer granularity to lead to a decrease in accuracy. This serves to indicate that more fine-grained distinctions for determiners, which is a quite disparate category in the treebank, may be quite useful for tagging. However, as it has negative impact on the syntactic parsing, we can conclude that the type of a determiner does not assist in generalizing syntactic patterns, as most determiners, regardless of type, appear in the same syntactic constructions (i.e., before a noun or noun phrase).

Gender, on the other hand, improved the parsing, but complicated the tagging, as the various genders are often difficult to differentiate, especially so in the case of masculine and feminine, which share many of the same determiners. The number of a determiner, i.e., singular or plural, led to a small increase in tagger accuracy and LAS, while marginally lower UAS, 90.18%, 0.01 percentage points lower than that of the original tag set. Almost all determiners have number, and the introduction of this distinction led to small increases in tagger accuracy and LAS, but marginally lower UAS. The introduction of definiteness to the determiners led to the best parsing results, LAS of 87.30% and UAS of 90.42%, while also increasing the tagger accuracy slightly. The increase in LAS and UAS is rather interesting, as there are only 121 "determiner tokens" with marked definiteness in the development data. As this accounts for a very small number of tokens, coupled with the previously noted considerations, we did not consider further fine-grained modifications with definiteness. This goes to show that tokens with overt definiteness have noticeable impact on the syntactic parsing, and that distinguishing on definiteness is very beneficial.

Pronouns Pronouns in Norwegian include personal, reciprocal, reflexive and interrogative. They can exhibit gender, number and person. Personal pronouns have case (accusative or nominative).

The results in Table **??** show that number, person and type are the most informative features for parsing, with LAS of 87.21%, 87.22% and 87.19%, respectively. However, when combining number and person, we observe a drop by more than 0.2 percentage points, indicating that these features do not interact in any syntactically distinctive way. The most interesting observation is that all experiments exceed the baseline tagger accuracy, the most improved being the most fine-grained distinction, namely type, case and number combined, obtaining a tagger accuracy of 97.52%. This shows that the introduction of more fine-grained distinctions for pronouns is beneficial and aids the PoS tagger in disambiguating ambiguous words. While case alone yields an LAS of 87.08%, we found that the combination of type and case, which is the most successful experiment in terms of parser performance, yields the second highest tagging accuracy of 97.51%. The reason for this is that pronouns of different type and personal pronouns of different case exhibit quite different properties and appear in different constructions. Pronouns in nominative case (i.e., subjects) primarily occur before the main verb, while pronouns in accusative case (i.e., objects) occur after the main verb, as Norwegian exhibits so-called V2 word order, requiring that the finite verb of a declarative clause appears in the second position, hence its name. The combination of type and number comes in close to the performance of type and case, with LAS of 87.27% and UAS identical to that of type and case.

5.4 Optimized Tag Set

The most successful tag set modifications and their results are seen in Table 5.

An overview of the tags comprising the final, optimized tag set can be seen in Table 6.

Results from parsing with the optimized tag are shown in Table 7, compared to the results obtained with the initial tag sets.

6 Summary and Future Work

The improvements in parser performance with our optimized tag set indicate that a more fine-grained PoS tag set may assist syntactic parsers in recognizing and generalizing syntactic patterns, while potentially compromising the performance of PoS taggers.

There are several aspects of this thesis that can be further explored in future work, including extrinsic evaluation of the effects of PoS tag sets on other downstream NLP applications besides parsing, such as sentiment analysis and named entity recognition. These applications often require tagged data, but are markedly different from syntactic parsing, hence the evaluation would involve investigating an entirely different aspect of the effects of tag set granularity.

References

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, Seattle, WA, USA.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. 2015. It Depends: Dependency Parser Comparison Using A Web-Based Evaluation Tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 387–396, Beijing, China.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo, Norway.

Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0.

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. In *Proceedings of the 17th Scandinavian Conference of Linguistics*, pages 31–48, Odense, Denmark.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpanek, Pavel Straăàk, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 6th Conference on Natural Language Learning*.

Andrew MacKinlay. 2005. The Effects of Part-of-Speech Tagsets on Tagger Performance. Bachelor's thesis, University of Melbourne, Melbourne, Australia.

Christopher Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Category	Feature(s)	MFT	Accuracy	LAS	UAS
Baseline	—	94.14%	97.47%	87.01%	90.19%
Noun	Type, case & definiteness	89.61%	97.05%	88.81%	91.73%
Verb	Finiteness	93.72%	97.35%	87.30%	90.43%
Adjective	Degree	94.13%	97.41%	87.29%	90.44%
Determiner	Definiteness	94.13%	97.49%	87.30%	90.42%
Pronoun	Type & case	94.12%	97.51%	87.30%	90.41%

Table 5: Results of tagging and parsing with the most successful tag set modification for each category.

Tag	Description
`adj|komp`	Comparative adjective
`adj|pos`	Positive adjective
`adj|sup`	Superlative adjective
`det|be`	Definite determiner
`det|ub`	Indefinite determiner
`pron|pers`	Personal pronoun
`pron|pers|akk`	Personal pronoun, accusative
`pron|pers|nom`	Personal pronoun, nominative
`pron|refl`	Reflexive pronoun
`pron|res`	Reciprocal pronoun
`pron|sp`	Interrogative pronoun
`subst|appell`	Common noun
`subst|appell|be`	Common noun, definite
`subst|appell|be|gen`	Common noun, definite, genitive
`subst|appell|ub`	Common noun, indefinite
`subst|appell|ub|gen`	Common noun, indefinite, genitive
`subst|prop`	Proper noun
`subst|prop|gen`	Proper noun, genitive
`verb|fin`	Finite verb
`verb|infin`	Nonfinite verb

Table 6: Overview of the optimized tag set.

Data Set	Tag Set	MFT	Accuracy	LAS	UAS
Dev	Original	**94.14%**	**97.47%**	87.01%	90.19%
	Optimized	89.20%	96.85%	**88.87%**	**91.78%**
Test	Original	**94.22%**	**97.38%**	86.64%	90.07%
	Optimized	88.08%	96.35%	**88.55%**	**91.41%**

Table 7: Results of tagging and parsing with the optimized tag set, compared to the initial tag sets. Trained and tested using automatically assigned tags from TnT.

Mitchell Marcus, Beatrice Santorino, and Mary Ann Marcinkiewicz. 1993. Building A Large Annotated Corpus of English: The Penn Treebank. Technical report, University of Philadelphia, Philadelphia, PA, USA.

Beáta Megyesi. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 151–158, Pittsburgh, PA, USA.

Beáta Megyesi. 2002. *Data-Driven Syntactic Analysis: Methods and Applications for Swedish*. Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 915 – 932.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2089–2096, Istanbul, Turkey.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 789–795, Reykjavik, Iceland.

Per Erik Solberg. 2013. Building Gold-Standard Treebanks for Norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 459–464, Oslo, Norway.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 142–147, Stroudsburg, PA, USA.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399 – 433.

Yue Zhang and Joakim Nivre. 2011. Transition-Based Dependency Parsing with Rich Non-Local Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Lingustics: Human Language Technologies*, pages 188–193, Portland, OR, USA.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Lilja Øvrelid. 2008. Finite Matters: Verbal Features in Data-Driven Parsing of Swedish. In *Proceedings of the Sixth International Conference on Natural Language Processing*, Gothenburg, Sweden.