Optimizing a PoS Tag Set for Dependency Parsing

Anonymous ACL submission

Abstract

This paper reports on a suite of experiments that evaluates how the linguistic granularity of part-of-speech tag sets impacts the performance of tagging and syntactic dependency parsing. Our results show that parsing accuracy can be significantly improved by introducing more fine-grained morphological information in the tag set, even if tagger accuracy is compromised. Our taggers and parsers are trained and tested using the annotations of the Norwegian Dependency Treebank.

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 (Sang and 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 tag set granularity on PoS tagging itself. This includes investigation of the effects of tag sets on PoS tagging for 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, the effects of tag set size on tagging was investigated by mapping the original tag set into smaller subsets designed for parsing. Megyesi (2002) argues that a tag set with complete morphological tags may not be necessary for all NLP applications, for instance syntactic parsing. The study 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. Megyesi (2002) argues 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. Unfortunately, results for parsing with the various PoS tag sets are not reported.

Similarly, MacKinlay (2005) investigated the effects of PoS tag sets on tagger performance in English, specifically the Wall Street Journal portion of the Penn Treebank (PTB) (Marcus et al., 1993). Based on linguistic considerations, MacKinlay (2005) mapped the 45 tags of the original PTB tag set to more fine-grained tag sets to investigate whether additional linguistic information 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.

We argue, however, that it is possible to significantly improve the performance of the downstream application of syntactic parsing by introducing linguistically informed, syntactically informative distinctions to a tag set. We will now turn to the data that provided the basis for our experiments and outline its morphological and syntactic annotation choices.

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 original PoS tag set of NDT (excluding punctuation tags).

3 The Norwegian Dependency Treebank

Our experiments are based on 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 manually coded syntactic and morphological annotation for both varieties of Norwegian; 311 000 tokens of Bokmål and 303 000 tokens of Nynorsk. This paper only reports results for Bokmål, the main variety. The treebanked material mostly comprise newspaper text, but also include government reports, parliament transcripts and blog excerpts. The annotation process 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.

Morphological Annotation The morphological annotation and PoS tag set of NDT is based on the same inventory as used by 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

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.

are complemented by a large set of morphological features, providing information about morphological properties such as definiteness, number and tense. Selected subsets of these features are used in our tag set modifications, where the coarse PoS tag of relevant tokens is concatenated with one or more 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). Some central annotation choices are outlined in Table 2, taken from Solberg et al. (2014), providing overview of the analyses of syntactic constructions that often distinguish dependency treebanks, such as coordination and the treatment of auxiliary and main verbs. The annotations comprise 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.

Data Set Split As there was no existing standardized data set split of NDT due to its recent development, we first needed to define separate sections for training, development and testing. Our proposed sectioning of the treebank follows a standard 80-10-10 split. In establishing the split, care has been taken to preserve contiguous texts in the various sections while also keeping them balanced in terms of genre.

Our defined split into sections for training, development testing, and held-out testing will be distributed with the future releases of the treebank, but details for replicating the split will also be made available at a companion website for the final version of the paper.

Tagger As our experiments during development required many repeated cycles of training and testing for the different modified tag sets, we sought a PoS tagger that is 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 have the best performance for parsing of NDT. Furthermore, recent dependency parser comparisons (Choi et al., 2015) showed that Mate performed very well on parsing of English, outperforming a range of contemporary state-of-the-art parsers. We will be using Mate for gauging the effects of the tag set modifications in our experiments.

Tag Set Mapping The tag set modifications are realized through a mapping procedure that concatenates the existing tag with one or more specified features for all applicable tokens, i.e., tokens annotated with said tag and feature(s) in the gold standard data.

Evaluation To evaluate tagging and parsing with the various tag set modifications in our experiments, we employed commonly used evaluation metrics. Tagging is evaluated in terms of accuracy (reported by the TnT-included `tnt-diff` script; denoted *Acc* in the following tables), while labeled attachment score (LAS) and unlabeled attachment score (UAS) are computed by the `eval.pl`[1] script (used in the CoNLL shared tasks) to evaluate the parsing.

Baseline As a point of reference for the PoS tagging, we will be reporting the commonly used most-frequent-tag baseline, henceforth MFT. This involves labeling each word with the tag it was assigned most frequently in the training data. All unknown words, i.e., words not seen in the training

[1] `http://ilk.uvt.nl/conll/software.html`

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.

data, are assigned the tag most frequently assigned to words seen only once. The rationale is that we might expect unknown and infrequent words to have similar properties given that they both occur rarely.

Tags & Features As we seek to quantify the effects of PoS tagging in a realistic setting, i.e., in application to raw text, we choose to evaluate the parser on automatically assigned PoS tags. For training, however, we have two options: using either gold standard or automatically assigned tags. In order to settle on a configuration, we conducted experiments with training on both gold standard tags and automatically assigned tags to see how the performance differs. 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 is both trained and tested on automatically assigned tags in the remainder of our experiments.

Though some parsers make use of morphological features, we removed all morphological features beyond the PoS tags in order to simulate a realistic setting. Moreover, it is crucial that we remove these features when working with automatically assigned tags, as such features would be provided by the gold standard.

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. We would then also increase the complexity of the tagging task though, which can be expected to lead to a drop in tagger accuracy. However, the best tagging, evaluated in isolation, does not necessarily lead to the best parse, hence it is interesting to investigate how tag set modifications may affect this interplay.

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.

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 gold standard distinctions to the tag set. The data sets with 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.

As we want to investigate how we can improve the linguistic quality of a tag set, we will only consider tag set modifications we deem linguistically sensible.

5.1 Baseline Experiments

In an initial round of experiments, we concatenated the tag of each token with its full set of morphological features, thereby mapping the original tag set to a new maximally fine-grained tag set given the annotations available in the treebank. This resulted in a total of 368 tags, hereafter referred to as the *full* tag set. 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 establish some initial points of reference for how tag set granularity affects the performance of tagging and parsing on NDT, we trained and tested a full pipeline with both of these intital tags sets. The results are reported in Table 4. Unsurprisingly, we see that the tagger accuracy plummets when we move from the original to the full tag set. The MFT baseline for the original tag set is 94.14%, while it drops by almost 9 percentage points to 85.15% for the full tag set. TnT re-

Tag Set	MFT	Acc	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 our development section of NDT with the two initial tag sets. From left to right we report the tagger accuracy of the most-frequent-tag baseline, the tagger accuracy of TnT, and labeled / unlabeled attachment scores for the Mate parser.

ports 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 level linguistic information in the full, fine-grained tag set comes at the expense of reduced tagger performance. In spite of this, however, we see that the additional information in the full tag set still improves the parser performance. With the original tag set, Mate yields 87.01% LAS and 90.19% UAS, which increases to 87.15% and 90.39%, respectively, using the full tag set. As we are looking for linguistically informed distinctions that improve the syntactic parsing, these results are promising and indicate that additional linguistic information assists syntactic parsing, motivating the optimization of the existing PoS tag set.

5.2 Tag Set Experiments

We modify the tags for nouns (`subst`), verbs (`verb`), adjectives (`adj`), determiners (`det`) and pronouns (`pron`) 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 in turn first experiment with each of the available features in isolation before testing various combinations. We base our choices of combinations on how promising the features are and what we deem worth For each tag, we first experiment with each of the available features in isolation before employing various combinations of them. We base 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 are reflected in the morphological features associated with the respective PoS tags. For instance, as nouns in Norwegian can take on gender, definiteness and number, the treebank operates with features for gender, definiteness and number complementing the `subst` tag. In addition to morphological properties such as definiteness, tense and number, all classes except for verbs have a *type* feature that provides information about the subtype of the PoS, e.g., whether a noun is common or proper.

Nouns In Norwegian, nouns are assigned gender (feminine, masculine or neuter), definiteness (definite or indefinite) 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 for distinguishing nouns in genitive case. 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 Penn Treebank and Stockholm-Umeå corpus.

The results of experimenting with tag set modifications for nouns are shown in Table 5 and reveal that, apart from case, none of the tag set modifications improve the tagging. However, they all give rise to increases in parser accuracy scores. 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 improvements from parsing with the original tag set, of 1.80 percentage points and 1.54 percentage points, respectively.

Verbs Verbs are inflected for tense (infinitive, present, preterite or past perfect) in Norwegian and additionally exhibit mood (imperative or indicative; also conjunctive, however, it is very uncommon and does not occur in the treebank) and voice (active or passive). Note that both voice and mood have only a single value in the treebank; `pass` (passive) and `imp` (imperative), respectively. Verbs which are not passive are implicitly active, and verbs which are not imperative are in indicative mood.

Feature(s)	Acc	LAS	UAS
—	97.47%	87.01%	90.19%
Case	**97.48%**	87.63%	90.72%
Definiteness	97.00%	88.27%	91.42%
Gender	96.09%	87.21%	90.36%
Number	96.37%	87.97%	91.00%
Type	96.92%	88.07%	91.11%
Case & definiteness	97.03%	88.39%	91.44%
Type & case	96.92%	88.46%	91.51%
Type & definiteness	96.99%	88.44%	91.48%
Type, case & definiteness	97.05%	**88.81%**	**91.73%**

Table 5: Results of experiments with modified PoS tags for nouns.

Table 6 presents the results from tagging and parsing with modified verb tags. Imperative clauses are fundamentally different from indicative clauses as they lack an overt subject, which is reflected in the fact that mood is the only feature leading to an increase in LAS, with a reported LAS of 87.04%. Although voice is a very distinguishing property for verbs, and passive clauses are very different from active clauses, introducing this distinction in the tag set leads to a drop in LAS of 0.05 percentage points, while distinguishing between the various tenses yields an LAS of 86.97%. Combining the two most promising features of mood and tense resulted in an LAS of 87.12% and UAS of 90.31%.

In an additional experiment, we mapped the verb tenses (and 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 constructions, and proved to greatly improve the parsing, as we saw the highest parser accuracy scores of 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. Furthermore, adjectives are inflected for degree (positive, comparative or superlative).

In Table 7, we report the results of modifying the `pron` tag in NDT. All features except for number lead to increases in parser accuracy scores, the most successful of which is degree with a reported LAS of 87.29%, while distinguishing adjectives on definiteness yields an LAS of 87.14% and introducing the distinction of gender leads to LAS of 87.10%.

Feature(s)	Acc	LAS	UAS
—	**97.47%**	87.01%	90.19%
Mood	97.43%	87.04%	90.19%
Tense	97.30%	86.97%	90.18%
Voice	97.45%	86.96%	90.09%
Mood & tense	97.31%	87.12%	90.31%
Voice & tense	97.28%	86.99%	90.15%
Mood, tense & voice	97.27%	86.83%	90.05%
Finiteness	97.35%	**87.30%**	**90.43%**

Table 6: Results of experiments with modified PoS tags for verbs.

Feature(s)	Acc	LAS	UAS
—	**97.47%**	87.01%	90.19%
Definiteness	96.84%	87.14%	90.29%
Degree	97.41%	**87.29%**	**90.44%**
Gender	96.89%	87.10%	90.25%
Number	96.71%	86.99%	90.10%
Type	97.40%	87.11%	90.25%
Definiteness & degree	96.81%	87.23%	90.39%
Definiteness & gender	96.31%	87.18%	90.39%
Definiteness & number	96.78%	87.27%	**90.44%**

Table 7: Results of experiments with modified PoS tags for adjectives.

Turning to combinations of features, definiteness and number achieve the best parser accuracy scores, very close to those of degree, with 0.02 percentage points lower LAS and identical UAS. Adjectives agree with their head noun and determiner in definiteness and number, making this an expected improvement. The combination of definiteness and degree is also quite promising, obtaining LAS of 87.23% and UAS of 90.39%. It is interesting that none of the combinations surpass the experiment with degree alone, which indicates that degree does not interact with the other features in any syntactically significant way.

Determiners Like adjectives, determiners in Norwegian agree with the noun they modify in terms of gender, number and definiteness. We do not report results from combinations of features for determiners, as the features could not be combined in any meaningful way.

The results from the experiments with determiners are shown in Table 8. Introducing the distinction of type (demonstrative, amplifier, quantifier, possessive or interrogative) led to an increase

Feature	Acc	LAS	UAS
—	97.47%	87.01%	90.19%
Definiteness	97.49%	**87.30%**	**90.42%**
Gender	97.28%	87.09%	90.31%
Number	97.49%	87.04%	90.18%
Type	**97.61%**	87.00%	90.11%

Table 8: Results of experiments with modified PoS tags for determiners.

in tagger accuracy of 0.14 percentage points to 97.61%, while marginally impacting the parsing, with LAS of 87.00%, 0.01 percentage points below that of the original tag set. 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.

Gender, on the other hand, improved the parsing (87.09% LAS), 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 of 90.18%, 0.01 percentage points lower than that of the original tag set. 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, we did not consider further fine-grained modifications with definiteness. This serves to indicate that tokens with overt definiteness have noticeable impact on the syntactic parsing and that distinguishing on definiteness is very useful.

Pronouns Pronouns in Norwegian are personal, reciprocal, reflexive and interrogative. They can furthermore exhibit gender, number and person, while personal pronouns can be distinguished on case (either accusative or nominative).

The results in Table 9 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

Feature(s)	Acc	LAS	UAS
—	97.47%	87.01%	90.19%
Case	97.50%	87.08%	90.21%
Gender	97.48%	87.06%	90.23%
Number	97.49%	87.21%	90.33%
Person	97.49%	87.22%	90.32%
Type	97.48%	87.19%	90.40%
Number & person	97.49%	96.98%	90.16%
Type & case	97.51%	**87.30%**	**90.41%**
Type & number	97.49%	87.27%	**90.41%**
Type & person	97.49%	87.00%	90.14%
Type, case & number	**97.52%**	87.11%	90.36%

Table 9: Results of experiments with modified PoS tags for pronouns.

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 tagging accuracy of the original tag set, 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 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%. 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.

5.3 Optimized Tag Set

The most successful tag set modification for each PoS and the results from tagging and parsing with the respective modifications are seen in Table 10. Nouns benefit by far the most from the introduction of more fine-grained linguistically motivated distinctions, with an LAS of 88.81% and UAS of 91.73% when distinguishing on type, case and definiteness. We observe that the most promising tag set modifications for verbs, adjectives, determin-

Category	Feature(s)	MFT	Acc	LAS	UAS
Original	—	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 10: 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, def.
`subst|appell|be|gen`	Common noun, def., genitive
`subst|appell|ub`	Common noun, indef.
`subst|appell|ub|gen`	Common noun, indef., genitive
`subst|prop`	Proper noun
`subst|prop|gen`	Proper noun, genitive
`verb|fin`	Finite verb
`verb|infin`	Nonfinite verb

Table 11: Overview of the optimized tag set.

ers and pronouns all reach LAS of ~87.30% and UAS of ~90.40%. To investigate the overall effect of these tag set modifications, we tested each of the improvements in parser accuracy scores from that of the original tag set for statistical significance using Dan Bikel's randomized parsing evaluation comparator script[2], as used in the CoNLL shared tasks. For the most successful tag set modification for each of the categories seen in Table 10, the difference in LAS from the original tag set is statistically significant at significance level 0.05 (p-value < 0.05), as are all differences in UAS, except for verbs with finiteness (p-value 0.15) and pronouns with type and case (p-value 0.06).

An overview of the tags in the optimized tag set can be seen in Table 11, comprising three new tags for adjectives, two for determiners, six for pronouns, seven for nouns and two for verbs, totaling 20 tags. Appending these to the original tag set comprising 19 tags, we reach a total of 39 tags for NDT.

[2]Available as `compare.pl` at `http://ilk.uvt.nl/conll/software.html`

Data	Tag Set	MFT	Acc	LAS	UAS
Dev	Original	**94.14%**	**97.47%**	87.01%	90.19%
	Optimized	85.15%	96.85%	**88.87%**	**91.78%**
Test	Original	**94.22%**	**97.30%**	86.64%	90.07%
	Optimized	88.08%	96.35%	**88.55%**	**91.41%**

Table 12: 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.

Final Evaluation In Table 12, we show the results of parsing with the optimized tag set on the held-out test data and the development data, compared to the results obtained with the original tag set. We see significant improvements from the original tag set on both the development data and the held-out test data set. The improvement in LAS on the development data is 1.86 percentage points, while 1.91 percentage points on the held-out test data. These results indicate that the additional linguistic information in the tags of our optimized tag set greatly aids syntactic parsing and that optimizing an existing PoS tag set for a downstream application can be useful and beneficial.

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 ef-

fects 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 Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 1–18, Boulder, CO, USA.

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.

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. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, the Czech Republic.

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.

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*, pages 142–147, Stroudsburg, PA, USA.

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.

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