Optimizing a PoS Tagset for Norwegian Dependency Parsing

Petter Hohle and **Lilja Øvrelid** and **Erik Velldal**
University of Oslo
Department of Informatics
{pettehoh,liljao,erikve}@ifi.uio.no

Abstract

This paper reports on a suite of experiments that evaluates how the linguistic granularity of part-of-speech tagsets 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 tagset, 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 pre-processing 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 tagsets 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 tagsets may be an avenue worth investigating further (Manning, 2011). Clearly, the utility of a PoS tagset is tightly coupled with the downstream task for which it is performed. Even so, PoS tagsets are usually employed in a "one size fits all" fashion, regardless of the requirements posed by the task making 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; Seddah et al., 2013). Morphology interacts with syntax through phenomena such as agreement and case marking, and incorporating information on morphological properties of words can therefore often improve 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 tagset for the task of dependency parsing of Norwegian Bokmål. We report on a set of experiments where PoS tags are extended with various morphological properties and evaluated in terms of both tagging accuracy and syntactic parsing accuracy. Our results show that the introduction of morphological distinctions not present in the original tagset, whilst compromising tagger accuracy, actually leads to significantly improved parsing accuracy. This optimization also allows us to bypass the additional step of morphological analysis, framing the whole pre-processing problem as a simple tagging task. The impact on parser performance is also more pronounced in this study than in similar previous work, as surveyed in Section 2 next.

For the remainder of the paper, Section 3 details the treebank that provides the basis for our experiments, while Section 4 describes the experimental setup. Section 5 goes on to provide the results from our tagset optimization, before we finally summarize our main findings and discuss some directions for future work in Section 6.

2 Previous Work

This section reviews some of the previous work documenting the impact that PoS tagsets has on the performance of taggers and parsers.

Megyesi (2002) trained and evaluated a range of PoS taggers on the Stockholm-Umeå Corpus (SUC) (Gustafson-Capková and Hartmann, 2006),

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 142–151,
Gothenburg, Sweden, 23-24 May 2017.

annotated with a tagset based on a Swedish version of PAROLE tags totaling 139 tags. Furthermore, the effects of tagset size on tagging was investigated by mapping the original tagset 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 tagset comprising 26 tags yields the lowest tagger error rate. However, for some of the taggers, augmenting the tagset with more linguistically informative tags may actually lead to a drop in error rate (Megyesi, 2002). Unfortunately, results for parsing with the various PoS tagsets are not reported.

In a similar study, MacKinlay (2005) investigated the effects of PoS tagsets 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 tagset to more fine-grained tagsets to investigate whether additional linguistic information could assist the tagger. Experimenting with both lexically and syntactically conditioned modifications, they did not find any statistically significant improvements, 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 tagset.

Moving beyond tagging, Seddah et al. (2009) focus on syntactic constituent parsing for French and show that extending the PoS tagset with information about mood and finiteness for verbs is indeed beneficial. Similarly, the recent shared tasks on parsing morphologically rich languages has seen quite a bit of work focused on evaluating the effect of various types of morphological information on syntactic parsing (both constituent-based and dependency-based) (Tsarfaty et al., 2010; Seddah et al., 2013). They find that the type of morphological information which is beneficial for parsing varies across languages and the quality of this information (i.e. whether it is gold standard or predicted) will also influence the results.

Rehbein and Hirschmann (2013) report on experiments for parsing German, demonstrating small but significant improvements when introducing more fine-grained and syntactically motivated distinctions in the tagset, based on the Stuttgard-Tübingen Tagset (STTS). However, the scope of the changes are limited to modifier distinctions and the new tagset only includes four new PoS tags, changing two of the original STTS categories. The setup also introduces some additional complexity in the parsing pipeline in that two taggers are used; a first pass of tagging with the original STTS tagset provides context features used in a second pass of tagging using the modified tagset. When training and testing on predicted tags using the Mate dependency parser (Bohnet, 2010), Rehbein and Hirschmann (2013) report a modest increase in LAS from 86.94 using STTS to 87.13 for the modified double-tagger setup. Using the Berkeley constituent parser, there is a corresponding increase in F-score from 75.45 to 75.55 (Rehbein and Hirschmann, 2013).

Maier et al. (2014) also experiment with applying the Berkeley constituency parser to German using tagsets of varying granularity; the 12 tags of the Universal Tagset (UTS) (Petrov et al., 2012), the 54 tags of STTS and an extended version of STTS including all the morphological information from the treebanks used for training, resulting in up to 783 tags. Maier et al. (2014) experimented with six different PoS taggers, but found TnT to have the most consistent performance across different tagsets and settings. Predictably, tagger accuracy drops as granularity increases, but the best parsing performance was observed for the medium-sized tagset, i.e., the original STTS.

Müller et al. (2014) attempt to improve dependency parsing with Mate by automatically defining a more fine-grained tagset using so-called split-merge training to create Hidden Markov models with latent annotations (HMM-LA). This entails iteratively splitting every tag into two subtags, but reverting to the original tag unless a certain improvement in the likelihood function is observed. Müller et al. (2014) argue that the resulting annotations "are to a considerable extent linguistically interpretable". Similarly to the setup of Rehbein and Hirschmann (2013), two layers of taggers are used. While the modifications of the tagset are optimized towards tagger performance rather than parsing performance, the parser's LAS on the test set improves from 90.34 to 90.57 for English (using a HMM-LA tagset of 115 tags) and from 87.92 to 88.24 for German (using 107 tags).

In the current paper, we introduce linguistically motivated modifications across a range of PoS tags in the Norwegian Dependency Treebank

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 tagset of NDT (excluding punctuation tags).

and demonstrate significant – and substantial – improvements in parser performance.

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 *Bokmål* and *Nynorsk*, the two official written standards of the Norwegian language. This paper only reports results for Bokmål, the main variety. The treebanked material mostly comprises newspaper text, but also include government reports, parliament transcripts and blog excerpts, totaling 311 000 tokens for Norwegian Bokmål. The annotation process was supported by the rule-based Oslo-Bergen Tagger (Hagen et al., 2000) and then manually corrected by human annotators, also adding syntactic dependency analyses to the morphosyntactic annotation.

Morphological Annotation The morphological annotation and PoS tagset 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 tagset consists of 12 morphosyntactic PoS tags outlined in Table 1, with 7 additional tags for punctuation and symbols. The tagset 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. Selected subsets of these features are used in our tagset modifications, where the coarse PoS tag of relevant tokens is concatenated with one or more features to include more linguistic information in the 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.

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

This section briefly outlines some key components of our experimental setup.

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.[1] 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.

Tagger As our experiments during development required many repeated cycles of training and testing for the various modified tagsets, we sought a PoS tagger that is both reasonably fast and accurate. There is 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 widely used tagger that achieves both close to state-of-the-art accuracy as

[1] Our defined train/dev./test split is available for download at `http://github.com/petterhh/ndt-tools` and will be distributed with future releases of the treebank.

well as very high speed is TnT (Brants, 2000), and hence we adopt this for the current study.

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 graph-based Mate parser (Bohnet, 2010) to have the best performance for NDT. Recent dependency parser comparisons (Choi et al., 2015) show very strong results for Mate also for English, outperforming a range of contemporary state-of-the-art parsers. We will be using Mate for gauging the effects of the tagset modifications in our experiments.

Evaluation To evaluate tagging and parsing with the various tagset modifications in our experiments, we employ a standard set of measures. Tagging is evaluated in terms of accuracy (computed by the TnT-included `tnt-diff` script; denoted *Acc* in the following tables), while parsing is evaluated in terms of labeled and unlabeled attachment score (LAS and UAS; computed by the `eval.pl`[2] script from the CoNLL shared tasks).

Tagging Baseline In addition to the tagging accuracy for the original unmodified tagset, we include the most-frequent-tag baseline (MFT), as another point of reference. 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 data, are assigned the tag most frequently observed for words seen only once. The MFT baseline will also serve to indicate the ambiguity imposed by the additional tags in a given tagset modification.

Predicted vs. Gold Tags Seeking to quantify the effects of PoS tagging on parsing, we choose to both evaluate *and* train the parser on automatically predicted PoS tags.[3] Training on predicted tags makes the training set-up correspond more closely to a realistic test setting and makes it possible for the parser to adapt to errors made by the tagger. While this is often achieved using jackknifing (*n*-fold training and tagging of the labeled training data), we here simply apply the taggers to the very same data they have been trained on, reflecting the 'training error' of the taggers. In an initial experiment we also found that training on such 'silver-standard' tags actually improves parsing scores substantially compared to training on gold tags, as shown in Table 3. In fact, Straka et al. (2016) also found that this set-up actually yields higher parsing scores compared to 10-fold tagging of the training data. Of course, the test set for which we evaluate the performance is still unseen data for the tagger.

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 various configurations of PoS tag sources in training and testing. *Gold* denotes gold standard tags while *Auto* denotes tags automatically predicted by TnT.

5 Tagset Optimization

The modified tagsets used in the experiments reported in this section are defined as combinations of PoS tags and morphological features already available in the gold annotations of the treebank. In effect, we redefine the tags comprising the gold standard as provided by NDT. The best performing configuration can be seen as a tagset specifically optimized for dependency parsing. Note that the tag selection process itself can be seen as semi-automatic; while we for each introduced distinction empirically evaluate its impact on parsing performance – as to see whether it is worth including in the final configuration – the process is manually guided by linguistic considerations regarding which combinations to evaluate in the first place. Although our expressed goal is to identify a tagset optimized for the downstream task of dependency parsing, we will only consider tagset modifications we deem linguistically sensible.

The hope is that the introduction of more fine-grained distinctions in the tagset may assist the parser in recognizing and generalizing syntactic patterns. This will also increase the complexity of the tagging task, though, which can be expected to lead to a drop in tagger accuracy. However, the most accurate tagging, evaluated in isolation, does not necessarily lead to the best parse, and the aim of this section is to investigate how tagset modifications affect this interplay. In Section 5.1 we first report on a set of initial baseline experiments using the information in the treebank 'as is'. Sec-

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

[3] Though some parsers also make use of morphological features, we removed all morphological features beyond the PoS tags in order to simulate a realistic setting.

Tagset	MFT	Acc	LAS	UAS
Original	**94.14**	**97.47**	87.01	90.19
Full	85.15	93.48	**87.15**	**90.39**

Table 4: Tagging and parsing our development section of NDT with the two initial tagsets. From left to right, we report the tagger accuracy of the most-frequent-tag baseline, the tagger accuracy of TnT, and the labeled and unlabeled attachment score for the Mate parser.

tion 5.2 then details the results of tuning the selection of tags for each word class in isolation, before finally discussing overall results for the complete optimized tagset in Section 5.3.

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 tagset to a new maximally fine-grained tagset, given the annotations available in the treebank. This resulted in a total of 368 tags, hereafter referred to as the *full* tagset. The two initial tagsets, i.e., the original tagset comprising 19 tags and the full tagset comprising 368 tags, thus represent two extremes in terms of granularity. To establish some initial points of reference for how tagset granularity affects the performance of tagging and parsing on NDT, we trained and tested a full pipeline with both of these initial tagsets. The results are reported in Table 4. Unsurprisingly, we see that the tagger accuracy plummets when we move from the original to the full tagset. While the MFT baseline for the original tagset is 94.14%, it drops by almost 9 percentage points to 85.15% for the full tagset. Correspondingly, TnT achieves 97.47% and 93.48% accuracy for the original and full tagset, respectively. These results confirm our hypothesis that the high level of linguistic information in the full, fine-grained tagset comes at the expense of reduced tagger performance. In spite of this drop, however, we see that the additional information in the full tagset still improves the parser performance. With the original tagset, Mate achieves 87.01% LAS and 90.19% UAS, which for the full tagset increases to 87.15% and 90.39%, respectively. These preliminary results are encouraging in indicating that the additional linguistic information assists the syntactic parser, motivating further optimization of the tagset.

5.2 Tagset 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 in isolation and what we deem worth investigating in terms of linguistic utility.

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 inflect for gender, definiteness and number, the treebank operates with additional features for these properties. 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 (i.e., 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 can be useful, as evident by the presence of separate tags for proper and common nouns in many tagsets, such as those of Penn Treebank and Stockholm-Umeå corpus.

The results for tagset modifications for nouns are shown in Table 5 and reveal that, apart from case, none of the tagset modifications improve tagging. However, they all result in increases in parser accuracy. The most informative features are definiteness, with an increase in LAS of 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%. This

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: Tagging and parsing with modified PoS tags for nouns. The first row corresponds to using the original tagset unmodified.

constitutes an improvement 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) 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.

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 tagset 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 very different syntactic constructions, and proved to improve parsing with a 0.29 and 0.24 percentage points improvement over the baseline, for LAS and UAS, respectively. This coincides with the previous observations for

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: Tagging and parsing 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: Tagging and parsing with modified PoS tags for adjectives.

Swedish in Øvrelid (2008), where finiteness was found to be a very beneficial feature for parsing.

Adjectives Adjectives agree with the noun they modify in terms of gender, number and definiteness. Furthermore, adjectives are inflected for degree (positive, comparative or superlative).

Table 7 shows 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%.

Turning to combinations of features, definiteness and number achieve the best parser accuracy scores, very close to those of degree. 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 in isolation, 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

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: Tagging and parsing with modified PoS tags for determiners.

terms of gender, number and definiteness.

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 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 tagset. 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 parsing (87.09% LAS), but complicated tagging, as the various genders are often difficult to differentiate, in particular masculine and feminine, which share many of the same forms. The number of a determiner, i.e., singular or plural, led to a small increase in tagger accuracy and LAS, while marginally lower UAS. The introduction of definiteness gave the best parsing results, LAS of 87.30% and UAS of 90.42%, and additionally increased 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 result further underlines the importance of definiteness for parsing of Norwegian.

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

The results in Table 9 show that number, person and type are the most useful features for parsing, with LAS of 87.21%, 87.22% and 87.19%, respectively. However, when combining number and person, we observe a drop of 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 results exceed the tagging accuracy of the original tagset, with the most fine-grained distinction (type, case and number combined) provides the largest improvement (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 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.

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: Tagging and parsing with modified PoS tags for pronouns.

5.3 Optimized Tagset

Category	Feature(s)	MFT	Acc	LAS
Original	—	94.14	97.47	87.01
Noun	Type, case, def.	89.61	97.05	88.81
Verb	Finiteness	93.72	97.35	87.30
Adjective	Degree	94.13	97.41	87.29
Determiner	Definiteness	94.13	97.49	87.30
Pronoun	Type, case	94.12	97.51	87.30

Table 10: Results of tagging and parsing with the best tagset modification for each category.

The most successful tagset 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

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 pron., accusative
pron|pers|nom	Personal pron., 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., gen.
subst|prop	Proper noun
subst|prop|gen	Proper noun, genitive
verb|fin	Finite verb
verb|infin	Nonfinite verb

Table 11: Overview of the optimized tagset.

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 tagset modifications for verbs, adjectives, determiners and pronouns all reach LAS of ~87.30% and UAS of ~90.40%. To investigate the overall effect of these tagset modifications, we tested each of the improvements in parser accuracy scores from those of the original tagset for statistical significance using Dan Bikel's randomized parsing evaluation comparator script[4], as used in the CoNLL shared tasks. For the most successful tagset modification for each of the categories seen in Table 10, the difference in LAS from the original tagset is statistically significant at significance level 0.05 (p-value < 0.05), as are all differences in UAS, except for verbs with finiteness (p=0.15) and pronouns with type and case (p=0.06).

An overview of the tags in the optimized tagset 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 tagset comprising 19 tags, we reach a total of 39 tags.

Final Evaluation In Table 12, we show the results of parsing with the optimized tagset on the held-out test data and the development data, compared to the results obtained with the original tagset. We see significant improvements from the original tagset 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 tagset benefits the task of syntactic parsing.

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

Data	Tagset	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 tagset, compared to the original NDT coarse tagset. The parser is both trained and tested using automatically predicted tags from TnT.

6 Summary

This paper has reported on a range of experiments with injecting more fine-grained morphological distinctions into an existing PoS tagset, and then empirically evaluating the effects both (intrinsically) in terms of tagging accuracy and (extrinsically) in terms of parsing accuracy. Our experimental results – based on the annotations of the Norwegian Dependency Treebank and using the TnT PoS tagger and the Mate dependency parser – show that the enriched tag set leads to significantly improved parsing accuracy, even though tagging accuracy in isolation is reduced. We also observe that the improvements are more pronounced than in related previous studies for other languages. The modified tagsets in our experiments are defined as combinations of PoS tags and morphological features, using only information that is already available in the gold annotations of the treebank. The best performing tag configuration is in effect a PoS tagset optimized for dependency parsing of Norwegian. While we expect that tags that prove informative for parsing will be useful for also other downstream applications, one can of course follow the same methodology to optimize a tagset specifically for other applications instead, by using another task for the extrinsic evaluation, such as sentiment analysis, named entity recognition or any other task making use of tagged data.

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*. Stockholm, Sweden.

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.

Wolfgang Maier, Sandra Kübler, Daniel Dakota, and Daniel Whyatt. 2014. Parsing German: How much morphology do we need? In *Proeceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 1–14, Dublin, Ireland.

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

Thomas Müller, Richard Farkas, Alex Judea, Helmut Schmid, and Hinrich Schütze. 2014. Dependency parsing with latent refinements of part-of-speech tags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 963–967, Doha, Qatar.

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.

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.

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.

Ines Rehbein and Hagen Hirschmann. 2013. POS tagset refinement for linguistic analysis and the impact on statistical parsing. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories*, pages 172–183, Tübingen, Germany.

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.

Djamé Seddah, Marie Candito, and Benoît Crabbé. 2009. Cross parser evaluation and tagset variation: A french treebank study. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, pages 150–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richard Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Yuval Marton, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolinski, and Alina Wroblewska. 2013. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth*

Workshop on Statistical Parsing of Morphologically Rich Languages, pages 146–182, Seattle, 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.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL): what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*.

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.