An open source part-of-speech tagger for Norwegian: Building on existing language resources

Cristina S. Marco

Gjøvik University College
Teknologivegen 22, 2815 Gjøvik, Norway
cristina.sanchez@hig.no

Abstract

This paper presents an open source part-of-speech tagger for the Norwegian language. It describes how an existing language processing library was used to build a new part-of-speech tagger for this language. This part-of-speech tagger has been built on already available resources, in particular a Norwegian dictionary and gold standard corpus, which were partly customized for the purposes of this paper. The results of a careful evaluation show that this tagger yields an accuracy close to state-of-the-art taggers for other languages.

Keywords: open source, part-of-speech, tagger, Norwegian

1. Introduction

The web is growing multilingual, yet the availability and reliability of language resources is not the same for all languages. In the case of Scandinavian languages, the availability of basic online text processing resources such as part-of-speech taggers is quite limited (De Smedt et al., 2012; Pedersen et al., 2012; Borin et al., 2012). Thus, researchers interested in developing applications using basic resources for these languages (e.g. in machine translation systems) cannot do it to the same extent as for the English language. To overcome these limitations, this paper presents the results of an open source part-of-speech tagger for the Norwegian language, which yields results close to state-of-the-art taggers (Collins, 2002; Toutanova et al., 2003; Spoustová et al., 2009; Søgaard, 2009). This part-of-speech tagger was mainly built using available language resources, which have been partly adapted for the purposes of this paper.[1]

The contents of this paper are as follows. After the literature review of existing taggers for Norwegian in Section 2., Section 3. presents the language analyzer used for this work. The resources used to create this tool are described in Section 4. and the method followed is described in Section 5. Finally, the evaluation of this tool is presented in Section 6. The paper concludes in Section 7., with a discussion and suggestions for future work.

2. Related work

State-of-the-art taggers work rapidly and reliably with accuracies slightly over 97 percent (Collins, 2002; Toutanova et al., 2003; Spoustová et al., 2009; Søgaard, 2009). For Scandinavian languages, and specifically for Norwegian, in contrast, the availability of language analyzers is rather limited. The only available tool for Norwegian is the Oslo Bergen tagger (OBT), which is a rule-based tagger based on the Constraint Grammar formalism (Karlsson et al., 1995). This tagger yields a precision and recall of 95.4 and 99 respectively, but leaves unsolved ambiguities in the output (Johannessen et al., 2000). The last version of this tagger, available only for Norwegian Bokmål, includes a statistical module used to disambiguate the ambiguous output left by the previous tool. OBT-Stat yields an accuracy of 96 percent on the morphological tagging and 98.3 percent on the lemmas on an unseen evaluation corpus. This tagger has a rather large and complex tagset, containing 358 morphological tags and 2,000 more for full morphological analysis (Johannessen et al., 2011; Johannessen et al., 2012).

The part-of-speech tagger for Norwegian presented here yields a higher accuracy in the morphosyntactic tagging (over 97 percent) although not in the tagging of lemmas (95.2) (see Section 6.2.). Besides, this resource differs fundamentally from the OBT-Stat tagger. In particular, the part-of-speech tagger presented here is mostly based on statistics, and makes use of a simple and standard tagset (see Section 5.2.). In addition, in this work the whole resource has been adapted, including for example modules to deal with derived and compound words (see below in Section 5.).

3. The analyzer

The tool used to create this part-of-speech tagger for the Norwegian language is FreeLing.[2] FreeLing is an open source text processing tool offering a number of language analysis services, such as morphosyntactic tagging, named entity recognition, dependency parsing or sense annotation (Padró et al., 2010). On its current version, this resource provides services (to different extents) for Asturian, Catalan, English, Galician, Italian, Old Spanish, Portuguese, Russian and Spanish (Padró and Stanilovsky, 2012). This library is actively developed and maintained, highly modular, extensible and largely customizable, and thus it was particularly well suited for the purposes of this work, and also for the researchers and developers community more broadly. This work focuses on presenting the tool for part-of-speech or morphosyntactic tagging, but the other ser-

[1] Norwegian has two official written standards: Bokmål and Nynorsk. The tool presented here focuses on Norwegian Bokmål.

[2] http://nlp.lsi.upc.edu/freeling/. The tool for Norwegian is available in the development version 3.1- devel, accessible via SVN.

vices could also be customized to deal with Norwegian. In addition, FreeLing provides an application programming interface (API) that can be used to integrate language analyses into a more complex processing. The FreeLing processing pipeline for morphosyntactic tagging is illustrated in Figure 1. As shown in Figure 1, a text is submitted to the analyzer, which processes and enriches the texts with linguistic information using different modules: tokenization, dictionary, affixation, compound analysis, probability assignment and unknown-word guesser.[3]

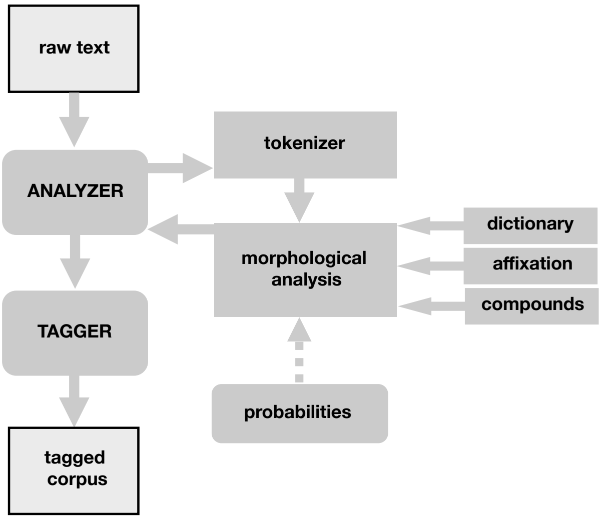

Figure 1: Text processing in FreeLing.

4. The data

In order to create this tool an existing Norwegian dictionary (4.1.) and a gold standard corpus (4.2.) were used.

4.1. Norwegian dictionary

The dictionary used is *Norsk ordbank*.[4] *Norsk ordbank* is a large database of lexical units with both morphosyntactic and argument structure information. This dictionary, totalling more than 635,712 types (1,179,629 tokens), includes the following resources: (i) word lists and patterns of inflections produced by IBM Norway,[5] (ii) entries and inflection information from *Bokmålsordboka / Nynorskordboka*, that are standard online dictionaries for Bokmål and Nynorsk produced by the Department of Linguistics and Scandinavian Studies at the University of Oslo[6], (iii) and codes for argument structure produced by project NorKompLeks at the Norwegian University of Science and Technology (Nordgård, 1996). Different versions of *Norsk ordbank* have been used by the OBT and the LOGON machine translation project (Lønning et al., 2004).

4.2. Gold standard corpus

The corpus used to train the tagger and evaluate the performance of the analyzer is the *Gullkorpus* (version 0.5). This corpus was developed by the National Library of Norway in collaboration with the Text Laboratory at the University of Oslo, under the umbrella of the project Språkbanken (Solberg, 2013).[7] *Gullkorpus* consists of texts from different newspapers from the late nineties to the present day (*Aftenposten*[8], *Dagbladet*[9] and *Klassekampen*[10]), and parliamentary records from the same period.[11] This corpus contains both morphosyntactic and syntactic annotations, and it is manually corrected by professionals.

5. Method

5.1. Dictionary adaptation

The first basic step towards the new part-of-speech tagger for Norwegian was the adaptation of the dictionary. In this adaptation the tagset was largely simplified and standardized (as described in 5.2.). Besides, some entries, such as abbreviations, affixes and multiword expressions have been incorporated into other modules (see 5.5.). After this, the Norwegian dictionary contains 622,999 words and 872,597 lemma-tag pairs. For example, the word *huset* has five lemma-tag pairs in the dictionary: two of them describe the adjective *huset* 'accommodated' in singular masculine, femenine or neutral form; the third one describes to the definite singular form of the noun *hus* 'house' and the last two the past and participial form of the verb *huse* 'to house'.

5.2. Tagset

One of the main changes in both the Norwegian gold standard corpus and dictionary is that the tagset was largely simplified and standardized. Table 1 shows the *huset* example in the dictionary of the new tool and in *Norsk ordbank*, previous modification. As can be seen from the third column in Table 1, the same information represented with a single label in the tool is represented using Norwegian terms in *Norsk ordbank*.[12] For example, in the third row from Table 1, the label *subst nøyt appell ent be normert* indicates that *huset* is a *substantiv* 'noun', *nøytrum* 'neuter', *appellativ* 'common (noun)', *entall* 'singular' and *bestemt* 'definite'. The dictionary also includes information about the syntactic type of verb (transitive in this case) and whether the word shows a normalized spelling (*normert* 'normalized'). This representation makes it difficult to comply with standardization in order to make language resources largely reusable and accessible (Ide and Romary, 2007).

The simplification of the tagset has been made in two different ways: first, the original tags have been combined into

[3] This module assigns a probability to each word analysis and, if a word has no analysis, a statistical guesser is used to find the most likely part-of-speech tags based on the word ending.

[4] The version of *Norsk orbank* used in this paper is released under the GNU General Public License and can be downloaded here: http://www.hf.uio.no/iln/om/organisasjon/edd/forsking/norsk-ordbank/.

[5] http://www.ibm.com/no/

[6] http://nob-ordbok.uio.no/

[7] The corpus can be downloaded here: http://www.nb.no/Tilbud/Forske/Spraakbanken/Tilgjengelege-ressursar/Tekstressursar

[8] http://www.aftenposten.no/

[9] http://www.dagbladet.no/

[10] http://www.klassekampen.no/

[11] More information about the project can be found here: http://www.nb.no/Tilbud/Forske/Spraakbanken/Tilgjengelege-ressursar/Tekstressursar.

[12] A similar tagset is also used in the OBT-Stat.

Word	Lemma	Adapted PoS	*Norsk ordbank* PoS
huset	*huse* ‘accommodated.A.SG.F/M’	AQ0CSP0U	adj <perf-part> m/f ub ent <trans1> normert
	huse ‘accommodated.SG.N’	AQ0NSP0U	adj <perf-part> nøyt ub ent <trans1> normert
	hus ‘house.N.SG.N.DEF’	NCNS000D	subst nøyt appell ent be normert
	huse ‘housed.PST’	VMIS	verb pret <trans1> normert
	huse ‘housed.PTCP’	VMP0	verb perf-part <trans1> normert

Table 1: Lemma and part-of-speech (PoS) of *huset* in the adapted dictionary (third column) and in *Norsk ordbank* (fourth column).

a single tag and, second, some grammatical information, such as the argument structure of verbs, has been excluded. As a result of this simplification the total number of tags included is 203, while the original *Norsk ordbank* tagset contained 358 tags.

The tagset used by this tool is mostly based on the EAGLES standard.[13] The first letter of each tag indicates the morphological class of the word, where ‘A’ stands for adjective (e.g. *raskt* ‘quick’), ‘R’ for adverb (e.g. *også* ‘also’), ‘D’ for determiner (e.g. *dette* ‘this’), ‘N’ for noun (e.g. *kultur* ‘culture’), ‘V’ for verb (e.g. *synge* ‘to sing’), ‘P’ for pronoun (e.g. *jeg* ‘I’), ‘C’ for conjunction (e.g. *eller* ‘or’), ‘I’ for interjection (e.g. *åh* ‘oh’), ‘S’ for preposition (e.g. *mot* ‘against’), ‘F’ for punctuation symbol (e.g. *;*), ‘Z’ for numbers (e.g. *5*) and ‘TO’ for the infinitive marker *å* ‘to’. The remaining letters (up to 8) specify more fine-grained morphosyntactic and semantic information, such as the gender, number and definite or indefinite character of nouns and adjectives, the tense or type (main or auxiliary) of verbs, or case information of nouns and pronouns (acussative, genitive, etc.). For example, the label ‘NCNS000D’ in Table 1 represents a singular neuter common noun in definite form describing the word *huset* ‘the house’, ‘NCMS000D’ describes a masculine singular common noun in definite form, such as in *kulturen* ‘the culture’ or *bilen* ‘the car’; ‘VAIS’ represents an auxiliary verb in the simple past, e.g. *kunne* ‘could’ or *hadde* ‘had’; and ‘PP2CSA0H’ depicts a second person personal pronoun in the singular form and accusative case used to refer to persons, such as *deg* ‘you’ or *meg* ‘me’.

5.3. Compound words

FreeLing analyzes forms not found in the dictionary through a compound and an affixation module that check whether the words are compounds or derived forms. The compound module detects whether a word is a compound formed by the concatenation of two or more dictionary words. On its current version, the compounds are detected when they are formed by words present in the dictionary and they are simply glued together (e.g. *skrivebok* ‘exercise book’, *spisestue* ‘dining room’, *søktsaker* ‘sweet things’), separated with dashes (e.g. *tv-skjermen* ‘television screen’, *EA-Sports-sjef* ‘EA sports chief’) or with epenthetic *-s-* (*aluminiumsfabrikk* ‘aliminium factory’) or *-e-* (*barnetrygd* ‘child benefit’). Compounds formed by suppletive or irregular stems (e.g. *kleskap* ‘clothes cupboard’, instead for **kledeskap*) are still not analyzed by the tool, as these forms are not found in the dictionary. The compound module, which has been recently incorporated to FreeLing and it is available on its development version (Padró and Stanilovsky, 2012), is based on the Foma format (Hulden, 2009).

5.4. Affixed words

The affixes module has been created from scratch to deal with Norwegian, as no suitable list of affixes was provided by available tools. Specifically, 10 different suffixes were added to this module; for example the Norwegian genitive *-s* for nouns (e.g. *tjueminutters treningsøkt* ‘twenty-minute workout’, *Ophras personlige trener* ‘Ophra’s personal trainer’), superlatives and comparatives *-(e)st* and *-ere* (e.g. *varmest* ‘warmest’, *penere* ‘prettier’, *dårligst* ‘worse’), nominalizing suffixes such as *-ing* (e.g. *optimisering* ‘optimization’), and adjectivizers such as *-som* (e.g. *vaktsom* ‘watchful’). In addition, 18 prefixes were included as well, such as *u-* ‘un-’ and *kjempe-* ‘great’ and *mis-* ‘miss-’ (e.g. *uviktig* ‘unimportant’, *kjempefint* ‘terrific’, *misformål* ‘wrong purpose’). Table 2 illustrates this.

Affix	Example word	Base word
-(e)st	*varmest*	*varm* ‘warm’
-ere	*penere*	*pen* ‘pretty’
-som	*vaktsom*	*vakt* ‘guard’
u-	*uviktig*	*viktig* ‘important’
kjempe-	*kjempefint*	*fint* ‘fine’
mis-	*misformål*	*formål* ‘purpose’

Table 2: Examples of affixed words.

5.5. Other modules

Tokenizer. The tokenizer module in FreeLing has also been customized, for instance to deal with compound words (e.g. *EA-Sports-sjef* ‘EA sports boss’, *tv-skjermen* ‘television screen’) and ordinal numbers (e.g. *8.* ‘8th’). Additionally, 530 abbreviations obtained from the dictionary have been included, so as not to split sentences using the dot (*osv.* ‘et cetera’, *eks.* ‘example’).

Multiword expressions. 360 multiword expressions (e.g. *i fjor* ‘last year’, *til orde* ‘in favour’, *i glemme* ‘into oblivion’, *i hele dag* ‘all day’, *for tiden* ‘currently’) with their corresponding morphological tag, also obtained from *Norsk ordbank*, have been added to the multiword expressions module included in FreeLing. This module analyzes these ex-

[13] Expert Advisory Group on Language Engineering Standards (http://www.ilc.cnr.it/EAGLES96/home.html).

pressions as single tokens, thus assigning a part-of-speech to each of them (adverbs in the above mentioned examples).

5.6. Retraining the tagger

FreeLing includes a hybrid tagger (*relax*) integrating statistical and hand-coded grammatical rules, and a Hidden Markov Model tagger (*hmm*), which is a classical trigram markovian tagger based on TnT (Brants, 2000). This paper focuses on presenting the results of the performance of this new resource using the *hmm* tagger. In order to train the tagger, an adapted version of the gold standard corpus for Norwegian was used. Similarly to the dictionary, in this corpus the tagset was largely simplified and standardized (as described in Section 5.2.) and only the morphosyntactic annotations included in the original corpus were used. This corpus contains 71,182 tokens. This is a small training corpus, compared with the corpus used to train the OBT-Stat, which consisted of 120,000 words (Johannessen et al., 2011; Johannessen et al., 2012).

5.7. An example

To illustrate the results of the new tool, Table 3 shows an example of the tokenized, morphologically analyzed, and part-of-speech-tagged output obtained from the Norwegian text in (1).[14]

(1) *Jeg hadde overhodet ikke planlagt å lage en slik*
I had at all not planned to make a such
film, men så fikk jeg tilfeldigvis høre fra noen
film but then got I accidentally hear from some
venner at det fantes en astronomiklubb
friends that there was found a astronomy club
hvor unge jenter og gutter møttes om
where young girls and boys were met on
kvelden og natten for å titte på stjernene i en
evening and night for to look at stars in a
liten landsby 800 kilometer syd for Teheran.
small town 800 kilometer south for Teheran

The first column in Table 3 shows the word input form, the second and the third columns show the lemma and part-of-speech automatically assigned by the tool and the last column shows the probability with which the part-of-speech was given. In this analysis some words have been assigned the correct tag despite the fact that they are not in the dictionary. In particular, *Teheran* is the proper name for the city of Teheran, and *astronomiklubb* 'astronomy club' is a compound noun formed by the concatenation of the nouns *astronomi* 'astronomy' and *klubb* 'club'. This output includes an error due to a word for which the morphological information has not been assigned by the tagger, despite the fact that it is included in the dictionary. In particular *slik* 'such' has been labeled as a femenine determiner ('DQ0FS00'), despite the fact that it is actually masculine in this context ('DQ0MS00'). This is probably due to the fact that this word is ambiguous, with three possible tags (masculine and feminine determiner and adverb). Probably the *relax* tagger could be used to deal with such cases, given the agreement with the following noun. In the following section the evaluation and error analysis are presented.

Word	Lemma	PoS	Probability
Jeg	jeg	PP1CSN0H	0.992932
hadde	ha	VAIS	1
overhodet	overhodet	RG	0.941667
ikke	ikke	RN	0.999703
planlagt	planlegge	VMP0	0.69863
å	å	TO	0.998538
lage	lage	VMN0	1
en	en	DQ0MS00	0.95867
slik	slik	DQ0FS00	0.250271
film	film	NCMS000U	0.990196
,	,	Fc	1
men	men	CC	0.99919
så	så	RG	0.838315
fikk	få	VAIS	1
jeg	jeg	PP1CSN0H	0.993096
tilfeldigvis	tilfeldigvis	RG	1
høre	høre	VMN0	1
fra	fra	SPS00	1
noen	noen	DQ00P00	0.396605
venner	venn	NCMP000U	0.427916
at	at	CS	0.9999
det	det	PD0NS000	0.841689
fantes	finnes	VVIS	1
en	en	DQ0MS00	0.95867
astronomiklubb	astronomiklubb	NCMS000U	1
hvor	hvor	RG	1
unge	ung	AQP0P000	0.483483
jenter	jente	NCFP000U	0.649959
og	og	CC	0.999944
gutter	gutt	NCMP000U	1
møttes	møtes	VVIS	1
om	om	SPS00	0.811223
kvelden	kveld	NCMS000D	1
og	og	CC	0.999944
natten	natt	NCMS000D	1
for	for	SPS00	0.93765
å	å	TO	0.998538
titte	titte	VMN0	1
på	på	SPS00	1
stjernene	stjerne	NCMP000D	0.55
i	i	SPS00	0.991689
en	en	DQ0MS00	0.95867
liten	liten	AQPCS00U	0.992424
landsby	landsby	NCMS000U	1
800	800	Z	1
kilometer	kilometer	NCMP000U	0.550006
syd	syd	SPS00	0.688992
for	for	SPS00	0.93765
Teheran	teheran	NP	1
.	.	Fp	1

Table 3: Tokenized, morphologically analyzed, and PoS-tagged text.

6. Evaluation

In this section the evaluation of the dictionary (6.1.) and overall tagging results (6.2.) are presented.

[14] Online news excerpt published on 19/03/2014 at aftenposten.no.

6.1. Dictionary

In order to evaluate the dictionary, two measures were used: ambiguity and coverage. *Ambiguity* measures the average number of lemma-tag pairs corresponding to each word. To compute ambiguity, each word form is assigned a score corresponding to the number of lemma-tag labels. *Coverage* measures the percentage of types and tokens in the corpus which are analysed by the dictionary. Ambiguity and coverage were measured in (i) in the dictionary and (ii) in the corpus. The results of this evaluation are presented in Table 4. As can be seen from this table, the corpus is significantly more ambiguous than the dictionary. This higher ambiguity is probably due to the fact that most function words are highly ambiguous. As for coverage, only 5.7 percent of the tokens and 16.5 percent of the types in the corpus are not covered by the dictionary. More than half of the uncovered words (64 percent) are proper names (e.g. *Øst-Europa, Alfred, Andersen, BBC*, ...). There are also many compound words (e.g. *poengstatistikken* ‘point statistics’, *høyreekstreme* ‘extreme right’, *postit-lapper* ‘post-it notes’), multiword expressions (*i går kveld* ‘yesterday evening’, *i så fall* ‘in which case’) and derived words (*kjempekompliment* ‘huge compliment’), which could be analyzed if the compound, multiword and affixes modules are used.

	Dictionary	Corpus
Ambiguity	1.4	2.8
Coverage	83.5%	94.3%

Table 4: Ambiguity and coverage of the dictionary.

6.2. Tagging

The *accuracy* in the tagging of lemmas and morphological tags was measured to evaluate the performance of the tagger. Different types of morphological information were evaluated: (i) word class or main part-of-speech (PoS-1), (ii) main part-of-speech and information about the subtype of the word class (e.g. auxiliary versus main verb) (PoS-2) and (iii) detailed morphosyntactic information given by the tag, including grammatical information such as gender, number or case (PoS-3). Table 5 illustrates this.

	Tag	Translation	Example
PoS-1	V	verb	*kunne* ‘could’
PoS-2	VA	aux. verb	*kunne* ‘could’
PoS-3	VAIS	aux. verb, past	*kunne* ‘could’

Table 5: Morphological information used to measure accuracy. *aux.* stands for ‘auxiliary verb’.

In all cases, accuracy has been obtained as a result of a 5-fold cross-validation over the gold standard corpus. The accuracy scores obtained on this corpus are summarised in Table 6. The results indicate that the performance of the tagger is quite close to those obtained by state-of-the-art taggers (between 96 and 98 percent) when the lemma, main part-of-speech and class are considered (PoS-1 and PoS-2). The tagger yields slightly lower accuracy (over 92 percent) if detailed morphological information is considered (PoS-3).

Lemma	PoS-1	PoS-2	PoS-3
95.2%	97.3%	96%	92.4%

Table 6: Accuracy obtained for lemma, PoS-1, PoS-2, and PoS-3 in the 5-fold cross-validation for the gold standard corpus.

6.3. Error analysis

The analysis of errors has been conducted over the errors obtained in the tagging during crossvalidation. This analysis shows that most of the errors in the tagging are due to the ambiguity in the dictionary. This is not surprising given the high ambiguity scores presented in Section 6.1. (recall also the *slik* example from Table 3). Specifically, more than 60 percent of the errors correspond to words for which the correct tag is available in the dictionary but the tagger has not selected it. The most frequent errors involve ambiguities in morphological features such as (i) nominative versus accusative case ambiguity in pronouns (*dere* ‘you.NOM/ACC’), (ii) number ambiguity in nouns (*år* ‘year/years’) and determiners (*ingen* ‘no.SG/PL’), (iii) gender ambiguity in adjectives (*mulig* ‘possible.M/F/N’) and determiners (*slik* ‘such.M/F’). Other frequent errors involve categorial ambiguities, such as (iv) determiner versus pronoun ambiguity of *det* ‘the/it’ or *en* ‘a/one’, (v) determiner versus adjective of *andre* ‘other/another’, and (vi) preposition versus conjunction ambiguity (e.g. *for* ‘for/because’).

26 percent of the erros correspond to (vii) words which are not included in the dictionary such as proper names (e.g. *Breivik*, *Randi*, *Sally*) or acronyms (*SFT -Statens Forurensningstilsyn-* ‘Norwegian Pollution Control Authority’) and also (viii) compounds (*multemousse* ‘cloudberry mousse’, *spillerorganisasjonen* ‘players organization’) and affixed words (*journalisters* ‘of journalists’, *kjempekompliment* ‘great compliment’) that are not correctly analyzed by the compound and affixed words modules.

The remaining 12 percent of the errors correspond to words that are present in the dictionary but in the gold standard corpus are nevertheless labeled with a different category. These are mostly cases of proper names (e.g. *Regjeringen* ‘the Government’, *Koranen* ‘Koran’).

A large part of the errors caused by ambiguous words could probably be dealt with if some rules were used (by adapting the *relax* tagger). Also proper names could correctly be analyzed using the Named Entity Recognition module. Further improvements on the compound and affixed words modules would also probably help to improve the performance of the tool in the tagging of compounds and derived or affixed words.

7. Discussion and future work

This paper presents an open source part-of-speech tagger for the Norwegian language. This tagger yields an accuracy close to state-of-the-art taggers, that is over 97 percent for the main category and 95 percent for lemmas. These results improve the results obtained by existing morphological taggers for Norwegian. In order to develop this tool, free online resources available through different projects were used, which were adapted for the purposes of this work. Apart from the higher accuracy, a significant advantage with respect to previous taggers for Norwegian is that the tagset included in the new tool has been largely simplified and standardized. Besides, the fact that this new part-of-speech tagger has been developed using FreeLing text processing library makes it easier to extend the tool or to use it for more complex processing.

There is also still room for improvement, which is left for future work. In this paper the results obtained with the statistical tagger (*hmm*) are presented, and it would be interesting to see how the *relax* tagger performs by adding rules to deal with ambiguous words. Additionally, the performance of the tagger would also probably even better if the compound module is further improved, in order to analyze compound words -very frequent in Norwegian- not formed by simply concatening words from the dictionary. Besides, the NER module could also be adapted to deal with words that are not present in the dictionary, such as proper names. Finally, it would also be interesting to compare the results with those obtained if the Norwegian Nynorsk version of *Norsk ordbank* and *Gullkorpus* are used in order to create an open source part-of-speech tagger for this standard language as well.

8. Acknowledgements

This research has been funded by the Eurostars project INFUSE.[15] I also thank Carla Parra, Gyri Smørdal Losnegaard, Lluís Padró and three anonymous reviewers for their comments and suggestions.

9. References

L. Borin, M. Brandt, J. Edlund, J. Lindh, and M. Parkvall. 2012. *The Swedish language in the digital age - Svenska språket i den digitala tidsåldern*. META-NET White Paper Series, Springer, Berlin/Heidelberg.

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

M. Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 10, pages 1–8, Philadelphia. July 2002. Association for Computational Linguistics.

K. De Smedt, G. Inger Lyse, A. Müller Gjesdal, and G. Smørdal Losnegaard. 2012. *The Norwegian language in the digital age - Norsk i den digitale tidsalderen (bokmålsversjon)*. META-NET White Paper Series, Springer, Berlin/Heidelberg.

M. Hulden. 2009. Foma: a finite-state compiler and library. In *EACL (Demos)*, pages 29–32.

N. Ide and L. Romary. 2007. Towards international standards for language resources. In *Evaluation of Text and Speech Systems Kluwer Academic Publishers*, pages 263–284.

J. B. Johannessen, K. Hagen, and A. Nøklestad. 2000. A constraint-based tagger for Norwegian. In Carl-Erik Lindberg and Steffen Nordahl Lund, editors, *17th Scandinavian Conference on Linguistics [Odense Working Papers in Language and Communication 19]*, pages 31–48. University of Southern Denmark, Odense.

J. B. Johannessen, K. Hagen, A. Nøklestad, and A. Lynum. 2011. OBT+Stat: Evaluation of a combined CG and statistical tagger. In *NEALT Proceedings Series*, pages 26–34, Oslo, Norway.

J. B. Johannessen, K. Hagen, A. Lynum, and A. Nøklestad. 2012. OBT+stat. a combined rule-based and statistical tagger. In Gisle Andersen, editor, *Exploring Newspaper Language. Using the web to create and investigate a large corpus of Modern Norwegian*, pages 51–65. John Benjamins Publishing Company, Amsterdam/Philadelphia.

F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin/New York.

J. T. Lønning, S. Oepen, D. Beermann, L. Hellan, J. Carroll, H. Dyvik, D. Flickinger, J. B. Johannsen, P. Meurer, T. Nordgård, V. Rosén, and E. Velldal. 2004. LOGON - A Norwegian MT Effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, Uppsala, Sweden.

T. Nordgård. 1996. NorKompLeks: Some Linguistic Specifications and Applications. In *ALLC-ACH'96. Abstracts*, Bergen, Norway. Universitetet i Bergen, Humanistisk Datasenter.

L. Padró and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, Istanbul, Turkey. May 2012.

L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón. 2010. Freeling 2.1: Five Years of Open-Source Language Processing Tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA, La Valletta*, Malta, May 2010.

B.S Pedersen, J. Wedekind, S. Kirchmeier-Andersen, S. Nimb, J.E. Rasmussen, L.B. Larsen, S. Bøhm-Andersen, H.Erdman Thomsen, P. J. Henrichsen, J. O. Kjærum, P. Revsbech, S.Hoffensetz-Andresen, and B. Maegaard. 2012. *The Danish Language in the Digital Age - Det danske sprog i den digitale tidsalder*. META-NET White Paper Series, Springer, Berlin/Heidelberg.

A. Søgaard. 2009. Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 205–208, Uppsala, Sweden, 11-16 July 2010. Association for Computational Linguistics.

P. E. Solberg. 2013. Building gold-standard treebanks for Norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics. NODALIDA*, Oslo,

[15] http://infuse.scan4news.com/

Norway. May 22-24.

D. j. Spoustová, J. Hajič, J. Raab, and M. Spousta. 2009. Semi-supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece. 30 March – 3 April 2009. Association for Computational Linguistics.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Confereence of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, pages 173–180, Edmonton, Canada. Association for Computational Linguistics.