Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål

Kevin Unhammer
Department of Linguistics
University of Bergen
Bergen, Norway
kun041@student.uib.no

Trond Trosterud
Department of Linguistics
University of Tromsø
Tromsø, Norway
trond.trosterud@uit.no

Abstract

We describe the development of a two-way shallow-transfer machine translation system between Norwegian Nynorsk and Norwegian Bokmål built on the Apertium platform, using the Free and Open Source resources Norsk Ordbank and the Oslo–Bergen Constraint Grammar tagger. We detail the integration of these and other resources in the system along with the construction of the lexical and structural transfer, and evaluate the translation quality in comparison with another system. Finally, some future work is suggested.

1 Introduction

The term *Norwegian* covers a variety of related spoken dialects. Up until the 1800's, Danish was the only written standard used in Norway. Bokmål emerged through various reforms which brought the written language closer to the spoken; Nynorsk however, was created from the ground up with the purpose of representing all the spoken dialects of Norway. As it is, certain dialects (especially around the Oslo area) correspond more with Bokmål, while others are closer to Nynorsk. Nynorsk is "in a minority position in Norway, with approximately 12% of the users" (Everson and Trosterud, 2000), or around 450,000 people.

Although Nynorsk is in a minority position, there are quite good linguistic resources available under Free licences, compared to many languages with the same amount of speakers. We describe the creation of `apertium-nn-nb`, a machine translation (MT) system between Nynorsk and Bokmål[1] built using these resources with the Free and Open Source Apertium platform (Armentano-Oller et al., 2006). In the following section we give an overview of the Apertium platform and Constraint Grammar. Section 3 describes how the available resources were integrated into Apertium, and how we dealt with lexical and syntactic transfer (for which we did not have Free resources available). As Bokmål and Nynorsk are mutually intelligible, a 'gisting' system would not find much use, our aim is to make the translations acceptable for *post-editing*; in the last two sections we give an evaluation of the translation quality in light of this, and a discussion of the lessons learnt and how the system may be further improved.

2 Design

2.1 The Apertium Pipeline

The Nynorsk–Bokmål language pair follows the design of the Apertium system, a highly modular, shallow-transfer pipeline MT system. Dictionaries written in XML are compiled into finite state transducers, so that word-for-word translations are possible in both directions using only two monolingual dictionaries (morphological analysis/generation) and one translational (transfer) dictionary. Both dictionary types make use of *paradigms* to e.g. generalise over common suffix sets (and their analyses), and *directional con-*

[1] Available from http://apertium.org

J.A. Pérez-Ortiz, F. Sánchez-Martínez, F.M. Tyers (eds.)
Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 35–42
Alacant, Spain, November 2009

straints, which state that a certain entry may be analysed, but not generated, or vice versa.

Hidden Markov models (HMM's) are used after analysis for part-of-speech disambiguation. The transfer module is finite state based and handles three-stage chunking transfer, although we so far only use one-stage transfer, applying operations directly on patterns of morphological categories (described in further detail in section 3.5). Output from the transfer module is fed to morphological generation. De-/reformatters applied to the beginning and end of the pipeline let us preserve formatting of various document types.

2.2 Constraint Grammar

This language pair differs from most of the other Apertium pairs in using a Constraint Grammar (CG) module[2] as a pre-disambiguator (before the HMM). CG's (Karlsson, 1990) are hand-written rules which, given ambiguously tagged input (e.g. the English word 'read' tagged both as a past and present tense verb), may SELECT one reading/analysis over all the others, or REMOVE a certain reading from the set of analyses. The last reading is never removed. We may end up with several readings if the input in fact was ambiguous or the grammar didn't manage to remove what it should. CG's may also MAP (add) new tags to readings, typically syntactic function labels. Rules may check in either direction for the existence of tags or even specific words, over absolute or undefined distances.

CG's have been shown to be robust in handling unseen text, as well as reaching high accuracy levels. CG is also the only grammar-based method to give results comparable to statistical taggers. Where statistical taggers have been shown to have a ceiling under 97%[3], Bick (2000, p. 187–188) cites 99% precision and recall when fully disambiguating with a CG tagger for Portuguese.

In an MT context the important point is that the good CG results have made it possible to present robust rule-based MT. Good examples are Bick and Hansen (2007).

[2]Using `VISL CG-3`, http://beta.visl.sdu.dk/cg3.html

[3]Leech et al. (1994); Brants (2000); Brill and Pop (1997) all cite accuracy results between 96% and 97%. Both Chanod and Tapanainen (1995) and Samuelsson and Voutilainen (1997) compare statistical and CG taggers.

In the next section we describe the development and use of the Apertium modules, including CG, in `apertium-nn-nb`.

3 Development

3.1 Resources

As our basis for the morphological analysis and generation, we used Norsk Ordbank[4], a GPL full form dictionary with over $100,000$ lemmas. We also used the morphological disambiguator of the Oslo–Bergen Tagger (OBT), a high quality GPL Constraint Grammar (Hagen et al., 2000). Both of these use the same tag scheme. They were converted into Apertium formats and tag schemes, as described below. The bilingual dictionary and the transfer rules (for syntactic differences and agreement) were built from scratch. The following sections detail the process.

3.2 Analysis and generation

Like most Apertium language pairs, we use *lttoolbox* for morphological analysis and generation, which compiles XML-formatted entries into fast finite state transducers and allows generalisations to be made across e.g. common suffix paradigms. The full form dictionary entries (with morphological information like lemma, POS, inflection, etc.) in Norsk Ordbank were semi-automatically transformed into the lttoolbox format. First, one paradigm was created per lemma (always creating the longest possible suffix), then any duplicate paradigms were merged. Closed classes (e.g. pronouns, determiners) were added manually.

3.3 Disambiguation

The OBT and Norsk Ordbank use a different tagset from Apertium. We want the data from `apertium-nn-nb` to be useful in creating new Apertium language pairs, so we converted the tags to ones which conform as much as possible to other Apertium dictionaries. Most tags could be replaced one-to-one, although some were replaced with CG sets. To exemplify the latter: the OBT uses the tags `subst.appell` and `subst.prop` for common and proper nouns, where Apertium uses `n` and `np` respectively, so rules working on the single tag `subst` were

[4]http://www.edd.uio.no/prosjekt/ordbanken/

changed to work on the *set* consisting of the tags `n` and `np`. Most of this conversion was done using simple shell scripts.

The Constraint Grammar runs as a pre-disambiguator, and does not always manage to remove all spurious analyses. We run Apertium's statistical disambiguator module after this step to make a final choice. An unsupervised bigram model (Baum-Welch algorithm, 8 iterations) was trained on Wikipedia text using the `apertium-tagger`. Although the Apertium toolset allows for more advanced statistical models (Sheikh, 2009) and methods for parameter estimation (Sánchez-Martínez et al., 2008), so far we have instead worked on improving the CG where we spotted errors. Certain errors in the disambiguation might be easier to spot when working with MT, and improvements to the CG could be of benefit to others using the OBT. When disambiguating for MT, it is important to keep in mind that we always have to end up with only one analysis, thus our version of the OBT is slightly more aggressive in removing readings. E.g. we use the following "heuristic" rule:

```
REMOVE (n) IF (0 adj)(-1 det)(1 subst);
```

to remove a noun reading if the word might also be an adjective, and is between a determiner and a noun / proper noun. When tagging a corpus for e.g. lexicographic work, this rule may be too strict, but for our purposes it is better to make a choice which will be correct most of the time than to end up with an unsolved ambiguity.

3.4 Lexical transfer

We found no freely available bilingual dictionaries between Nynorsk and Bokmål, so this we had to build from the ground up. Closed classes and some open class entries were added manually, but the bulk of the translational dictionary was created more or less automatically using the three methods described below.

First, exact matches were added where the morphology was the same for both languages. E.g., if a noun lemma was the same in both languages, and in both languages the noun could occur in the same forms (singular/plural, definite/indefinite), it was considered a translation. This quickly got us around 36,000 entries. There are two problem with this method though. One is that it may introduce a lot of false friends. For closely related languages with such high overlap in the lexicon, the benefit outweighs the risk (and lists of common false friends are not hard to come by in grammars). The other problem is that we add many "radical forms", e.g. Bokmål words which exist in the Nynorsk dictionary but are far from being the most natural sounding Nynorsk translation. We can easily put restrictions on these forms (or on all forms with a certain substring) so that they are only analysed, but not generated; but finding all such pairs involves some work.

We also added entries where there were predictable changes, e.g. the Bokmål adjective suffix *-lig* will typically be *-leg* in Nynorsk, etc. This process, also used by Tyers et al. (2009, p. 4), consists of

1. finding Bokmål entries without translations
2. running string replacements on these for typical differences in substrings
3. checking whether the altered entries actually exist in the Nynorsk analyser

Running this method gave about 2500 nouns and verbs[5].

Finally, we added some entries using automatic word alignments. We used the KDE4 corpus of software translations (about 400,000 words), and about 50,000 words of parallel text from Norwegian government web pages, automatically crawled and sentence aligned with the `bitextor` web crawler tool (Esplà-Gomis, 2009)[6].

The KDE4 translations are in the *gettext* (.po-file) format, for which there are a lot of available tools. We first used the Translate Toolkit[7] tool `poswap` to turn the English–Nynorsk and English–Bokmål .po-files into Nynorsk–Bokmål files, then we ran `poterminology`, a terminology extraction tool which gathers phrase pairs (all sub-phrases which appear together over a certain threshold), taking advantage of the short sen-

[5] A technique used in other Apertium language pairs, which we haven't tried yet, is running a target language spell checker (which gives suggestions) on the missing source language words (replacing step 2 above), and then analysing the suggestion to find the lemma.

[6] Available from http://websvn.kde.org/trunk/l10n-kde4/ and http://bitextor.sourceforge.net/ respectively.

[7] Available from http://translate.sourceforge.net/wiki/toolkit/index

tences[8] and repetition in software translations. This gave us some hundreds of new terms with very little work; however, many of the software terms were not in the monolingual dictionaries and adding them requires some manual labour.

We next ran `Giza++` (Och and Ney, 2003) on a cleaned version of the KDE4 corpus to create word alignments. We also appended our translational dictionary to the corpus 6 times[9], to improve the probability of correct alignments. We then ran these alignments through our morphological analysers, and fed them into the tool `ReTraTos` (de Medeiros Caseli et al., 2006), which extracts both phrases and single-word translations from alignments[10], and converts them into Apertium translational entries (perhaps adding directional constraints). We used the same process on the web corpus.

The `ReTraTos` method gave translations for about 3500 entries which were missing from the translation lexicon; however, these still needed a manual check, and many had to be altered slightly (e.g. changing directional constraints or morphological tags) or simply discarded. Due to the amount of noise in this data, this method required a lot of post-editing (and in our case functioned more as a source of *suggestions* for translations). Of course, a different corpus might have achieved higher quality word alignments; either a bigger corpus[11] or a corpus of a different text type (de Medeiros Caseli et al. (2006) used corpora of similar sizes, but with magazine text).

In addition to these three main methods, we also had some user-contributed entries. We created a page on the Nynorsk Wikipedia where users could suggest translations for those words which were missing from the translational lexicon (sectioned by part-of-speech), and simply added these as they trickled in.

[8]On average, the aligned translations are about 4.5 words long (including multi-sentence translations).

[9]This number was simply based on trial and error.

[10]We did not add transfer rules found by ReTraTos, although this is a possibility.

[11]Wu and Xia (1994, in de Medeiros Caseli et al., 2006, p. 230) used a 3 million word corpus to induce an English–Chinese dictionary.

3.5 Structural transfer

Nynorsk and Bokmål do not have many syntactic differences. There are some minor differences in verb phrases, and some slightly more complex differences in noun phrases[12]. One difference is that *finite bokmål passive verbs are expressed with an auxiliary in Nynorsk (infinitive passives remain unchanged):*

(1) a. Bevilgning gis oftest ikke
grant.IND give.PRES.PASS usually not

b. Løyve blir oftast ikkje gjeve
grant.IND AUX usually not give.PART
'Grants are usually not given'

c. Om høsten fylles fjorden
In fall.DEF fill.PRES.PASS fjord.DEF
med sild
with herring

d. Om hausten blir fjorden
In fall.DEF AUX fjord.DEF
fylt med sild
fill.PRES.PASS with herring
'In fall, the fjord is filled with herring"

As (1-a-b) show, we may have a string of adverbs between the Nynorsk finite and main verb. (1-c-d) demonstrates that when the subject is not the first phrase, the Nynorsk subject has to occur between the auxiliary and the finite verb. We apply a transfer rule to the finite passive verbs[13] in (1-a-b). (1-c-d) however, does not get transferred correctly yet. We need to know if the phrase after the verb is the subject; if it is not, as in *om det selges fisk* lit. 'if there is sold fish', we should transfer by the same rule as (1-a-b), to *om det blir selt fisk*. The OBT has a CG module which adds syntactic function labels, but this is still not incorporated into Apertium.

Phrase-initial Bokmål genitives are expressed periphrastically, phrase-finally in Nynorsk:

(2) a. forfatterens siste utgivelse
author.DEF.GEN last publication.IND

[12]In all the examples, Bokmål appears above Nynorsk.

[13]Norsk Ordbank uses one and the same entry for present and infinitive passive verbs in Bokmål, which we split into two entries; fortunately the OBT was already pretty good at disambiguating infinitives from finite verbs, regardless of voice. We currently do not analyse the rather infrequent past passive form.

b. den siste utgjevinga til forfattaren
the last publication.DEF of author.DEF
'the author's last publication'

c. mitt nye luftputefartøy
my new hovercraft.IND

d. det nye luftputefartøyet mitt
the new hovercraft.DEF mine
'my new hovercraft'

(2) has certain exceptions; proper names are commonly used with the genitive, as are certain frequent nouns (*dagens siste* 'today's latest', *verdas største* 'the world's greatest'). Of course, we can have strings of adjectives appearing before each noun in (2-a-d), agreeing in gender, number and definiteness.

The transfer module matches fixed-length patterns of *categories*—sets of possible part-of-speech tags and/or lemmas—on a left-to-right, longest-first basis. There are currently 33 rules for translating Bokmål to Nynorsk, and 8 for the opposite direction. The passive rules are quite simple as there is no agreement to handle; the bulk of the transfer work was on the noun phrase. We generalise over possessive determiners and genitive nouns with a single transfer category such that both (2-a-b) and (2-c-d) are handled by one rule. Although transfer rules work on fixed-length patterns, transfer macros (used by several rules) allow us to generalise over e.g. agreement or definiteness[14] operations.

Since any change to transfer (or CG) rules may introduce errors in previously error-free translations, we made extensive use of regression tests during development, essentially adding a test for each problem discovered or fixed.

In section 4, we evaluate `apertium-nn-nb` and compare its output with that of the commercial Bokmål→Nynorsk MT system Nyno.

4 Evaluation

Unknown words easily lead to errors in disambiguation or transfer, and lexical coverage is thus essential for any MT system meant for unrestricted text. We define naïve coverage as the proportion of words in a corpus which are given at least one analysis by our monolingual dictionaries[15]. Naïve coverage is 89.6% on the Nynorsk Wikipedia (5,116,174 words) and 88.2% for Bokmål (27,529,115 words)[16]. On a 7,019,526 word corpus of Bokmål newspaper texts, the naïve coverage was 91.8%.

4.1 Word Error Rate on Post-Edited text

We also tested the Word Error Rate (WER)[17] on a 3,750 word post-edited article on linguistics from the Bokmål Wikipedia. The WER was 22.06% when including unknown words, although since 64.93% of these were free-rides[18] anyway, the final WER was 10.71%. We consider this quite acceptable for post-editing.

4.2 BLEU score in comparison with Nyno

Apertium is not the only MT system between Nynorsk and Bokmål. Another system is *Nyno*[19], a commercial MT system translating from Bokmål to Nynorsk. Nyno does not translate in the other direction, probably because there is larger demand for translations to Nynorsk. Both in order to have another system to compare with, and because we deemed the task of translating into Nynorsk as harder, we confined the `apertium-nn-nb` evaluation to this direction only.

In order to compare the systems, we translated an unseen text (7,283 words) from Bokmål to Nynorsk in both Apertium and Nyno[20]. The translation was compared to two gold standards, and the BLEU score (Papineni et al., 2001) of the respective systems was calculated. One gold

[14] Although Bokmål, like Danish, allows definite determiners with indefinite nouns, Nynorsk does not.

[15] Some analyses will be missing from a naïve coverage test due to homography.
Since the dictionaries are *consistent*, we also have *translations* for all analysed words.

[16] nnwiki-20090119-pages-articles and nowiki-20090108-pages-articles from http://download.wikimedia.org.

[17] Using the tool `apertium-eval-translator`, available from http://www.dlsi.ua.es/~fsanchez/software.html#eval. WER is defined by how many editing operations are needed to change the MT output into the reference text.

[18] Ie. the same in Bokmål and Nynorsk. Typical free-rides include names, loan-words and special terminology.

[19] http://www.nynodata.no

[20] The Norwegian written languages allow many different variant forms (regarding lexical choice, inflections, etc.); Nyno provides the user with a wide range of dictionary customisation options, e.g. "radical" versus "conservative" Nynorsk, whereas `apertium-nn-nb` as of yet only outputs one standard variant. We chose the configuration which seemed to correspond most closely to the guidelines we used in designing `apertium-nn-nb`.

standard came from the bilingual websites themselves, the other was translated as a joint effort on Nynorsk Wikipedia.

Table 1: BLEU score (two reference translations) and WER (for the *O*riginal and *W*ikipedia references) for Apertium and Nyno. Numbers in parenthesis give percentage of unknown words which were free-rides.

	BLEU	WER_O	WER_W
Apertium	0.74	32.5 (36.1)	17.7 (50.5)
Nyno	0.85	29.1 (34.6)	13.3 (47.3)

With only two reference translations, both BLEU scores are quite good. The Nyno result is clearly better than Apertium, though. The main difference between the two translation programs is found in the lexicon. An unknown word rate (UWR) of 9.5% for Apertium stands in sharp contrasts to Nyno, who misses 0.8% of the words of the testbed. The singlemost important candidate for `apertium-nn-nb` improvement thus seems to be the lexicon.

4.3 Error analysis

The majority of the translation errors in `apertium-nn-nb` seem to be simply due to missing vocabulary. Of the remaining errors, most can be attributed to either *disambiguation*, *transfer* (e.g. agreement, or the problem of (1-c) above) or *lexical selection* (choosing the most natural collocations on the target language side).

Of the 20 first non-vocabulary errors in the above WER test, 13 were disambiguation problems, two were transfer problems and five related to lexical selection. In other text domains, e.g. newspaper texts, we also find some problems related to *coreference*, eg. where a pronoun refers to, and should agree with, a noun elsewhere in the text. We have not attempted to solve this.

Comparing the CG analyses (before HMM disambiguation) and translations of the first 100 sentences in the BLEU testbed, we discovered one trivially fixable transfer bug, and three pronoun gender errors which were plainly due to our lack of any coreference analysis.

A manual check of the analyses showed that 223 superfluous readings had not been removed (remained undisambiguated) without causing errors in the final translation[21], in most cases since the translations of both readings were the same. 19 readings which should have been removed, but weren't, *did* in fact cause errors. 13 readings were mistakenly removed without causing errors, while 3 where mistakenly removed and *did* cause errors.

Of the undisambiguated readings which did not cause errors, we find a lot of singular/plural ambiguity. This is common in this text type, although some times the disambiguation problems are plainly due to lack of vocabulary (e.g. an ambiguous adjective could have been disambiguated by the following unknown, but unambiguous, noun). The 22 CG errors which caused translation errors are fortunately quite easy to correct; however, 128 word tokens in these sentences were not even in the dictionary, which again shows that low coverage leads to more errors. Even if more than half were free-rides, we would still have more dictionary than disambiguation errors. More importantly, dictionary errors leak into the disambiguation and transfer components.

Below we discuss one way to improve our dictionary coverage, as well as the challenges posed by multi-word expressions such as phrasal verbs.

5 Discussion and outlook

In Norwegian, like in other Germanic languages, words combine very productively into compounds[22]. At the moment, we do not have any sort of compound analysis, but we are currently looking into ways of incorporating this in the Apertium pipeline. Some preliminary tests using left-to-right longest-match and a noun-only dictionary showed that it is quite easy to restrict the analysis of compounds to *those for which we can expect good translations*: looking at only those translations of noun-noun compounds [23] where the first part was analysed as singular indefinite, 99 out of 100 randomly selected compounds were cor-

[21] If a word has one correct (unremoved) and five mistaken (unremoved) readings, we count this as five.

[22] Munthe (1972, in Johannessen and Hauglin, 1996, p. 1) puts the number at "10.4% of all words in running text", although quite a lot of these will already be in our dictionary.

[23] Noun-noun compounds are by far the most common, from a list of 26,344 Bokmål tagged compounds (http://www.dokpro.uio.no/bokmaal/nyord/nyord_fside.html) we found only 1281 containing an adjective or verb as a member. (Such annotated compound lists may of course also easily be converted into translational lexicon entries.)

rectly translated, the remaining one—shown in (3-b)—had an ambiguous second member on the Bokmål side:

(3) a. bilkirkegård → bilkyrkjegard
car.cemetery → car.cemetery
b. postordrelager → #postordrelagar
mail.order.storage → mail.order.creator

In addition to compounding, we are also considering how to handle multi-word expressions (MWE's) such as particle verbs, consider (4) below where the particle *til* is expressed after the object[24]:

(4) a. Han anbefalte meg å gå hjem
he recommended me INF go home
b. Han rådte meg til å gå heim
he counseled me to INF go home
'He recommended that I go home'

Nynorsk text typically contains many such MWE's; but without knowing where the relevant Bokmål phrase ends, we have to take the safe route and try to find non-discontinuous Nynorsk translations. However, as mentioned earlier, the OBT has a CG which adds syntactic information; this could be used to handle MWE's, in addition to the challenge posed by (1-c) above.

We have described the development of a MT language pair using Free and Open Source tools and linguistic data. Although there are many possible improvements to be made, the system is in a usable state for post-editing MT. This would not have been possible without the reuse of existing linguistic data. Hopefully the changes we made to the CG will also benefit other users of the OBT.

Nynorsk, Bokmål, Swedish and Danish are "[morphologically] equally distant from each other" (Everson and Trosterud, 2000, p. 1) (although Nynorsk and Bokmål share more vocabulary), and we are planning on making translators for all the possible pairs. Although Nynorsk–Bokmål was the first released translator among these languages made with the Apertium toolset, there is also a group working on Swedish–Danish; the monolingual data of both these language pairs may be used directly in creating the four remaining possible pairs, as may the translational data (through the Apertium `crossdics` tool) once there is at least one connecting language pair. As translation between closely related languages has the potential to easily reach post-editable quality, the Apertium MT pairs could increase the amount of text available in these languages, and in fact, `apertium-nn-nb` has already been put to this use in drafting translations of Wikipedia articles from Bokmål to Nynorsk[25].

Except for Google's Swedish/Danish/Bokmål MT (all via English), operative general-purpose MT systems between Scandinavian languages are rule-based, and most of them are powered by CG. Being open source, Apertium broadens the Scandinavian MT horizon.

Acknowledgements

Development was funded as part of the Google Summer of Code[26] programme. Thanks to co-mentor Francis Tyers (Universitat d'Alacant, Prompsit Language Engineering) for the initial work on `apertium-nn-nb`, also to Jimmy O'Regan and the other Apertium contributors for constant help throughout. Thanks to Roar Johansen, Frode Håvik Korneliussen and Pål Julius Skogholt for reference translations, and Wikipedia contributors for suggestions, and to Paul Meurer (Aksis, Universitetet i Bergen) and Kristin Hagen (Universitetet i Oslo) for help with the OBT.

References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese-Spanish machine translation. In Vieira, R., Quaresma, P., Nunes, M., Mamede, N., Oliveira, C., and Dias, M., editors, *Computational Processing of the Portuguese Language*, pages 50–59. Springer-Verlag, Itatiaia, Brazil.

Bick, E. (2000). *The Parsing System "Palavras"*.

[24] And sometimes after adverbs, although here the source may have an inherent attachment ambiguity.

[25] http://nn.wikipedia.org/wiki/Kategori:Omsett_med_Apertium

[26] http://code.google.com/soc/

Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Århus.

Bick, E. and Hansen, J. A. (2007). The Fyntour Multilingual Weather and Sea Dialogue System–Description and Assessment. In *Proceedings of DECALOG (SEMDIAL07)*, pages 157–158, Trento, Italy.

Brants, T. (2000). TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231, Seattle, Washington. ACL.

Brill, E. and Pop, M. (1997). *Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging*, volume 11 of *Text, Speech and Language Technology*, pages 27–43. Kluwer Academic Press, Dordrecht.

Chanod, J.-P. and Tapanainen, P. (1995). Tagging French - comparing a statistical and a constraint-based method. In *EACL 1995*, pages 149–156, Belfield, Dublin, Ireland. University College Dublin.

de Medeiros Caseli, H., das Graças Volpe Nunes, M., and Forcada, M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245.

Esplà-Gomis, M. (2009). Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites. In *Proceedings of MT Summit XII*, Ottawa, Canada. Association for Machine Translation in the Americas.

Everson, M. and Trosterud, T. (2000). Software localization into nynorsk norwegian. Technical report, The Norwegian Language Council.

Hagen, K., Johannessen, J., and Nøklestad, A. (2000). A Constraint-Based Tagger for Norwegian. In Lindberg, C.-E. and Lund, S. N., editors, *17th Scandinavian Conference of Linguistics*, volume 19 of *Odense Working Papers in Language and Communication*, pages 31–48. Syddansk Universitet, Odense.

Johannessen, J. and Hauglin, H. (1996). An automatic analysis of Norwegian compounds. In Haukioja, T., editor, *Papers from the 16th Scandinavian Conference of Linguistics*, Turku/Åbo, Finland.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In Karlgren, H., editor, *Proceedings of the 13th Conference on Computational Linguistics*, volume 3, pages 168–173, Helsinki, Finland.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th Intl. Conference on Computational Linguistics*, pages 622–628, Kyoto, Japan. ACL.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL '02*, pages 311–318, Morristown, NJ, USA. ACL.

Samuelsson, C. and Voutilainen, A. (1997). Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of the 8th conference of the EACL*, pages 246–253, Morristown, NJ, USA. ACL.

Sánchez-Martínez, F., Pérez-Ortiz, J. A., and Forcada, M. L. (2008). Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66. DOI: 10.1007/s10590-008-9044-3.

Sheikh, Z. M. A. W. (2009). A Trigram Part-of-Speech Tagger for the Apertium Machine Translation Platform. Technical report.

Tyers, F., Wiechetek, L., and Trosterud, T. (2009). Developing Prototypes for Machine Translation between Two Sámi Languages. In Màrquez, L. and Somers, H., editors, *EAMT-2009*, pages 120–127, Barcelona, Spain. Universitat Politècnica de Catalunya.