# Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change

**Eiríkur Rögnvaldsson**[1], **Sigrún Helgadóttir**[2]
Department of Icelandic, University of Iceland[1], Árni Magnússon Institute for Icelandic Studies[2]
{Árnagarði við Suðurgötu, IS-101[1], Neshaga 16, IS-107[2]}, Reykjavík, Iceland
E-mail: {eirikur,sigruhel}@hi.is

**Abstract**

We describe experiments with morphosyntactic tagging of Old Norse narrative texts using different tagging models for the TnT tagger (Brants, 2000) and a tagset of almost 700 tags. It is shown that by using a model that has been trained on both Modern Icelandic texts and Old Norse texts, we can get 92.65% tagging accuracy which is considerably better than the 90.36% that have been reported for Modern Icelandic. In the second half of the paper, we show that the richness of our tagset enables us to use the morphosyntactic tags in searching for certain syntactic constructions and features in a large corpus of Old Norse narrative texts. We demonstrate this by searching for – and finding – previously undiscovered examples of two syntactic constructions in the corpus. We conclude that in an inflectional language like Old Norse, a morphologically tagged corpus like this can be an important tool in studying syntactic variation and change.

## 1. Introduction

In a previous project (Helgadóttir 2004; 2007), we have trained the TnT tagger written by Brants (cf. Brants, 2000) on a corpus of Modern Icelandic. The corpus used in that project was created in the making of the Icelandic Frequency Dictionary (*Íslensk orðtíðnibók*, henceforth IFD; Pind et al., 1991). The IFD corpus is considered to be a carefully balanced corpus consisting of 590,297 tokens with 59,358 types – both figures including punctuation.

The corpus contains 100 fragments of texts, approximately 5,000 tokens each. All the texts were published for the first time in 1980–1989. Five categories of texts were considered, i.e. Icelandic fiction, translated fiction, biographies and memoirs, non-fiction (evenly divided between science and humanities) and books for children and youngsters (original Icelandic and translations). No two texts could be attributed to the same person (as author or translator).

The texts were pre-tagged using a specially designed computer program and the tagging was then carefully checked and corrected manually. Thus, this corpus is ideal as training material for data-driven statistical taggers, such as the TnT tagger.

In the present project, we applied the TnT tagger trained on the Modern Icelandic corpus to Old Norse (Old Icelandic) texts.[1] This paper describes the results of this experiment, and also describes our experiments with using the morphologically tagged Old Norse corpus to search for syntactic constructions.

## 2. Tagging Modern Icelandic

In this section, we describe the tagset used in our research, and give a brief overview of our experience with the training of the TnT tagger on Modern Icelandic texts.

### 2.1 The tagset

The tagset developed for the IFD is very large, compared to tagsets designed for English at least, such as the Penn Treebank tagset (Marcus et al., 1993). The size of the tagset of course reflects the inflectional character of Icelandic, since it is for the most part based on the traditional Icelandic analysis of the parts of speech and grammatical categories, with some exceptions where that classification has been rationalized.

In the tag strings, each character corresponds to a single morphosyntactic category. The first character always marks the part of speech. Thus, the sentence *Hún hefur mætt gamla manninum* 'She has met the old man' will be tagged like this:

| | | |
|---|---|---|
| (1) | Hún | *fpven* |
| | hefur | *sfg3eþ* |
| | mætt | *ssg* |
| | gamla | *lkeþvf* |
| | manninum | *nkeþg* |

The meaning of the tags is as follows:

(2) *fpven:* pronoun (f) – personal (p) – feminine (v) – singular (e) – nominative (n)
*sfg3eþ:* verb (s) – indicative (f) – active (g) – 3rd person (3) – singular (e) – past (þ)
*ssg:* verb (s) – supine (s) – active (g)
*lkeþvf:* adjective (l) – masculine (k) – singular (e) – dative (þ) – definite (v) – positive (f)
*nkeþg:* noun (n) – masculine (k) – singular (e) – dative (þ) – suffixed article (g)

[1] It is customary to use the term 'Old Norse' for the language spoken in Norway, Iceland and the Faroe Islands up to the middle of the 14th century. The overwhelming majority of existing texts written in this language is either of Icelandic origin or only preserved in Icelandic manuscripts. For the purposes of this paper, 'Old Norse' is thus synonymous with 'Old Icelandic'.

Of the word forms in the IFD corpus, 15.9% are ambiguous as to the tagset within the IFD. This figure is quite high, at least compared to English, which reflects the fact that the inflectional morphology of Icelandic is considerably more complex than English. Icelandic nouns can have up to 16 grammatical forms or tags, verbs up to 106 different tags, and adjectives up to 120 tags. Altogether, 639 different tags occur in the IFD corpus, but the total sum of possible tags is around 700.

Some of the ambiguity is due to the fact that inflectional endings in Icelandic have many roles, the same ending often appearing in many places (e.g. *-a* in *penn**a*** for all oblique cases in the singular (acc., dat., gen.), and accusative and genitive in the plural of the masculine noun *penni* 'pen', producing 5 different tags for one form of the same word). The most ambiguous of word forms in the IFD, *minni*, has 24 tags in the corpus, and has not exhausted its possibilities (Bjarnadóttir, 2002).[2]

## 2.2 Training the tagger

The computer files for the IFD corpus each contain one text excerpt. Each file was divided into ten approximately equal parts. From these, ten different disjoint pairs of files were created. In each pair there is a training set containing about 90% of the tokens from the corpus and a test set containing about 10% of the tokens from the corpus. Each set should therefore contain a representative sample from all genres in the corpus. The test sets are independent of each other whereas the training sets overlap and share about 80% of the examples. All words in the texts except proper nouns start with a lower case letter.

Results for ten-fold cross-validation testing for the TnT tagger are shown in table 1 (cf. Helgadóttir, 2005; 2007). It is worth noticing that these results show lower performance rates when the tagger is applied to the Icelandic corpus than is achieved for example for Swedish as reported in Megyesi (2002). In that study, TnT was applied to and tested on the SUC corpus with 139 tags compared to the Icelandic tagset of over 600 tags. Performance rates are also considerably lower than have been reported for the systems trained on the Penn treebank.

| *Type* | *Accuracy %* |
|---|---|
| All words | 90.36 |
| Known words | 91.74 |
| Unknown words | 71.60 |

Table 1: Mean tagging accuracy for all words, known words and unknown words for TnT.

Table 1 shows results for known words, unknown words and all words. Mean percentage of unknown words in the ten test sets was 6.84. This is similar to what was seen in the experiment on Swedish text (Megyesi, 2002) and indicates that the major difficulty in annotating Icelandic words stems from the difficulty in finding the correct tag for unknown words. Words belonging to the open word classes (nouns, adjectives and verbs) account for about 96% of unknown words in the test sets whereas words in these word classes account for just over 51% of all words in the test sets.

# 3. Tagging Old Norse texts

Having trained the TnT tagger on Modern Icelandic texts, we wanted to find out whether the tagger could be of help in tagging Old Norse narrative texts, with the purpose of facilitating the use of these texts in research on syntactic variation and change. To create a manually annotated training corpus for Old Norse from scratch would have been a very time-consuming task. Thus, the possibility of using the bootstrapping method that we describe in this section was a key factor in realizing this project.

Bootstrapping is of course a common approach in training taggers and parsers. To our knowledge, however, this approach has not been used in historical linguistics to develop tagging models for a different stage of language than the tagger was originally trained on. Our method somewhat resembles the experiments of Hwa et al. (2005), who used parallel texts to build training corpora by projecting syntactic relations from English to languages for which no parsed corpora were available. The training corpora created using this method were then in turn used to develop stochastic parsers for the languages in question. The whole process took only a small fragment of the time it would have taken to create a manually corrected corpus to train the parsers.

The common factor in our project and the work reported by Hwa et al. (2005) is the use of another language, or (in our case) another stage of the same language, as a starting point in the bootstrapping process. Our experiments with bootstrapping the tagging of Old Norse texts are described in this section.

## 3.1 Old Norse vs. Modern Icelandic

At a first glance, it may seem unlikely that a tagger trained on 20th century language could be applied to 600-700 years old texts. However, Icelandic is often claimed to have undergone relatively small changes from the oldest written sources up to the present. The sound system, especially the vowel system, has changed dramatically, but these changes have not led to radical reduction or simplification of the system and hence they have not affected the inflectional system, which has not changed in any relevant respects. Thus, the tag set developed for Modern Icelandic can be applied to Old Norse without any modifications.

The vocabulary has also been rather stable. Of course, a great number of new words (loanwords, derived words and compounds) have entered the language, but the majority of the Old Norse vocabulary is still in use in Modern Icelandic, even though many words are confined to more formal styles and may have an archaic flavor.

[2] *minni* can be a noun meaning 'memory', present tense of the verb *minna* 'remind', comparative of the (irregular) adjective *lítill* 'small'. In all of these words we find extensive syncretism, resulting in many different tag strings for this word form in each part of speech.

On the other hand, many features of the syntax have changed (cf. Faarlund, 2004; Rögnvaldsson, 2005). These changes involve for instance word order, especially within the verb phrase, the use of phonologically "empty" NPs in subject (and object) position, the introduction of the expletive *það* 'it, there', the development of new modal constructions such as *vera að* 'be in the process of' and *vera búinn að* 'have done/ finished', etc.

In spite of these changes, we found it worthwhile to try to adapt the tagging model that we had trained for Modern Icelandic to our Old Norse electronic corpus. Our motive was not to get a 100% correct tagging of the Old Norse texts, but rather to facilitate the use of the texts in syntactic research, cf. Section 4 below.

## 3.2 The Old Norse corpus

Our Old Norse corpus consists of a number of narrative prose texts (sagas), which are assumed to have been written in the 13th and 14th centuries – a few of them probably later. Among these are many of the most famous Old Norse sagas. The division of the corpus is shown in Table 2:

| *Text* | *Tokens* |
|---|---|
| Family Sagas (around 40 sagas)<br>(*Íslendingasögur*) | 1,074,731 |
| Sturlunga Saga<br>("Contemporary Sagas") | 283,002 |
| Heimskringla<br>(Sagas of the Kings of Norway) | 250,920 |
| The Book of Settlement<br>(*Landnámabók*) | 42,745 |
| Total | 1,651,398 |

Table 2: Division of the Old Norse corpus.

The texts we use are (with the exception of The Book of Settlement) taken from editions, which were published between 1985 and 1991 (Halldórsson et al., 1985-86; Kristjánsdóttir et al., 1988; Kristjánsdóttir et al., 1991). In these editions, the text has been normalized to Modern Icelandic spelling. This involves, for instance, reducing the number of vowel symbols ('æ' is used for both 'ae ligature' (æ) and 'oe ligature' (œ), 'ö' is used for both 'o with a slash' (ø) and 'o with a hook'), inserting *u* between a consonant and a word-final *r* (*maðr* 'man' > *maður*), shortening word-final *ss* and *rr* (*íss* 'ice' > *ís*, *herr* 'army' > *her*), changing word-final *t* and *k* in unstressed syllables to *ð* and *g*, respectively (*þat* 'it' > *það*, *ok* 'and' > *og*), etc. Furthermore, a few inflectional endings are changed to Modern Icelandic form.

It must be emphasized, however, that these changes do not in any way simplify the inflectional system or lead to the loss of morphological distinctions in the texts. Thus, the texts are just as good as sources of syntactic evidence as texts that are published in the normalized Old Norse spelling.

On the other hand, we must point out that the original versions of these texts do not exist; the texts are mostly preserved in vellum manuscripts from the 13th through the 15th centuries, but some of them only exist in paper manuscripts from the 16th and 17th centuries. This makes it extremely difficult to assess the validity of these texts as linguistic evidence, since it is often impossible to know whether a certain feature of the preserved text stems from the original or from the scribe of the preserved copy, or perhaps from the scribe of an intermediate link between the original and the preserved manuscript. It is well known that scribes often did not retain the spelling of the original when they made copies; instead, they used the spelling that they were used to. In many cases, two or more manuscripts of the same text are preserved, and usually they differ to a greater or lesser extent. Furthermore, it is known that not all of the editions that our electronic texts are based on are sufficiently accurate (cf., for instance, Degnbol, 1985).

Even though this may to some extent undermine the validity of the texts as sources of syntactic evidence, it does not directly concern the main subject of this paper, which is to show that we can use a tagging model developed for Modern Icelandic to assist us in making the Old Norse corpus a usable tool in studies of syntactic variation and change. There is no reason to believe that possible inaccuracies and errors in the texts – cases where they fail to mirror correctly the syntax of the manuscripts – have any effects on the tagging accuracy. That is, the use of more accurate editions would not lead to less accurate tagging.

## 3.3 Training the tagger on the Old Norse corpus

We started by running TnT on the whole Old Norse corpus using the tagging model developed for Modern Icelandic (cf. Helgadóttir, 2005; 2007). We then measured the accuracy by taking four samples of 1,000 words each from different texts in the corpus – one from the *Family Sagas*, one from *Heimskringla*, and two from *Sturlunga Saga* – and checking them manually. Counting the correct tags in these samples gave 88.05% correct tags, compared to 90.36% for Modern Icelandic.

Even though these results were worse than those we got for Modern Icelandic, we considered them surprisingly good. The syntax of Old Norse differs from Modern Icelandic syntax in many ways, as mentioned above, and one would especially expect the differences in word order to greatly affect the performance of a trigram based tagger like TnT. However, sentences in the Old Norse corpus are often rather short, which may make them easier to analyze than the longer sentences of Modern Icelandic.

We then selected seven whole texts (sagas) and two fragments from the *Sturlunga* collection for manual correction – around 95,000 words in all. This amounts to one third of the *Sturlunga* collection. The manual correction was a time-consuming task, but the time and effort spent on checking and correcting the output of TnT was only a small fragment of the time and effort it would have taken to tag the raw text.

We trained TnT on the corrected text (95,000 words), tagged the whole corpus again with the resulting model, and measured the accuracy on the same four samples of 1,000 words each as in the first experiment. Now the results were much better – 91.73% correct tags, which is better than the 90.36% accuracy that we got for Modern Icelandic. It may seem surprising how much the accuracy improved when we used this model, especially when we consider that the training corpus was much smaller than the training corpus for Modern Icelandic (95,000 words compared to more than 500,000). On a closer look, however, this is understandable.

First, many of the errors occurring in the first experiment could be predicted and were easy to correct. For instance, the word *er* was always classified as a verb in the third (or first) person singular present indicative ('is, am'), as it usually is in Modern Icelandic. In Old Norse, however, this word is very often a subordinate conjunction ('when') or a relative particle ('that, which'). When the tagger was trained on a corrected Old Norse text, it could quickly and easily learn the correct tagging of these words, due to their frequency.

Second, it is well known that tagging accuracy is usually very much lower for unknown words than for known words, and the number of unknown words was much lower in the second experiment. In the first experiment, using the model for Modern Icelandic, the unknown word rate was 14.64%, reflecting the fact that a number of Old Norse words are rare or do not occur in Modern Icelandic. In the second experiment, using the model for Old Norse, the unknown word rate dropped to 9.63%, even though the training corpus was much smaller as pointed out above. This reflects the relatively small vocabulary of the Old Norse texts, which in turn reflects the narrow universe that the texts describe (cf. also Rögnvaldsson, 1990).

Finally, we trained TnT on a union of the corrected Old Norse texts and the Modern Icelandic texts. Thus, the training set for the final experiment consists of around 500,000 words from Modern Icelandic texts plus 95,000 words from Old Norse texts. When we tagged the Old Norse corpus using this model, we got 92.65% accuracy for the same four samples as in the first two experiments. The results of the three experiments are shown in Table 3:

| *Tagging model* | *Accuracy %* |
|---|---|
| Modern Icelandic model | 88.05 |
| Old Norse model | 91.73 |
| MI + ON model | 92.65 |

Table 3: Tagging accuracy for Old Norse texts using three different tagging models.

It is possible to improve the results by tagging the texts using all three models and combining the results of different models in various ways. All three models agree on the tags for 84.55% of the words. In 80.88% of the cases, they agree on the correct tag, but for 3.68% of the words, all three models agree on a wrong tag.

For 15.45% of the words, the models disagree. In most cases, two of them assign the same tag and the third model assigns a different tag. In a few cases, each model assigns a separate tag. Thus, if we assume that the tag is correct when all three models agree, we only need to look at 15.45% of the whole corpus. This means that the highest possible accuracy to be obtained using this method is 96.32%, since all models agree on a wrong tag in the remaining cases as pointed out above.

We could also choose to disregard the model that is trained only on Modern Icelandic texts, since it gives much lower accuracy than the other two models. The remaining models agree on the tagging of 93.52% of the words – incorrectly for 4.28% of the words. If we only look at 6.48% where the models disagree, we are down to around 107,000 words that we have to correct manually. This is a manageable task, which we intend to finish in the near future. We think that performance may exceed 95% after manual revision of the training set, assuming that about half of the disagreements can be correctly resolved. This is an acceptable result in our view, and should be sufficient for most uses of the corpus.

In this connection, it must be pointed out that a majority of the tagging errors only involve one morphosyntactic feature. Thus, nouns are often tagged as accusative instead of dative, or vice versa, whereas gender and number are correctly tagged; verbs are often tagged as 3rd person instead of 1st person, whereas mood, voice, number, and tense are correctly tagged; etc. This means that by using fuzzy search, we should in many cases be able to find what we are looking for, even if the words are not quite correctly tagged.

## 4. Tagged texts in syntactic research

Over the past two decades, interest in historical syntax has grown substantially among linguists. Accompanied by the growing amount of electronically available texts, this has led to the desire for – and possibility of creating – syntactically parsed corpora of historical texts, which could be used to facilitate search for examples of certain syntactic features and constructions. A few such corpora have been developed, the most notable being the Penn Parsed Corpora of Historical English, developed by Anthony Kroch and his associates (Kroch and Taylor, 2000; Kroch et al., 2004). These corpora have already proven their usefulness in a number of studies of older stages of English (cf., for instance, Kroch et al., 2000; Kroch and Taylor, 2001).

We wanted to know whether our tagged Old Norse corpus could be used in syntactic research in a similar manner as syntactically parsed corpora. We had been using the raw unannotated texts for this purpose (cf., for instance, Rögnvaldsson, 1995; 1996) but the search for certain syntactic constructions and features had proven to be cumbersome and give insufficient results. Although our tagging is morphological in nature, the tags carry a substantial amount of syntactic information and the tagging is detailed enough for the syntactic function of words to be more or less deduced from their morphology and the adjacent words. Thus, for instance, a noun in the nominative case can reasonably safely be assumed to be a subject, unless it is preceded by the copula *vera* 'to be' which is in turn preceded by another noun in the nominative, in which case the second noun is a predicative complement. A noun in the accusative or dative case can in most instances be assumed to be a (direct or indirect) object, unless it is immediately preceded by a preposition (cf. also Rögnvaldsson, 2006). As is well known, Modern Icelandic also has accusative and dative subjects, and even some nominative objects (Thráinsson, 2007), but these can easily be identified from their accompanying verbs.

To test the usefulness of the tagging of Old Norse texts in syntactic research, we have made a small study of two controversial and disputed features of Old Norse syntax; Object Shift and Passive. These studies are described in this section.

## 4.1 Object Shift

As originally described by Holmberg (1986), Object Shift is the process of moving a (direct or indirect) object to the left across a negation. In Modern Icelandic, this process applies both to pronouns and full NPs (or DPs), as shown in (3), whereas in the "Mainland" Scandinavian languages (Danish, Norwegian, and Swedish), it only applies to pronouns, as (4) shows (examples from Thráinsson, 2007). The "shifted" object is underlined whereas the negation is in boldface and the "place of origin" of the shifted object is shown by an underscore:

(3) Nemandinn las bókina **ekki** ___
the student read book not
'The student didn't read the book'
Nemandinn las hana **ekki** ___
the student read she not
'The student didn't read it'

(4) *Studenten læste bogen **ikke** ___
the student read book not
'The student didn't read the book'
Studenten læste den **ikke** ___
the student read she not
'The student didn't read it'

It has been suggested that this difference between Icelandic and the Mainland Scandinavian languages is somehow related to the fact that Icelandic has a much richer case morphology than the Mainland Scandinavian languages (cf. Holmberg and Platzack, 1995). If this were so, one would expect to find both types of Object Shift in Old Norse, since the case system of Icelandic is in all relevant respects the same as in Old Norse. The Mainland Scandinavian languages would then be assumed to have lost Object Shift of full DPs due to the loss of case inflections.

However, it has been claimed that Object Shift of full DPs does not occur in Old Norse. Mason (1999) claims to have found two examples of shifted full DP objects in his study of nine Old Norse sagas. Sundquist (2002), on the other hand, concludes "that these two examples do not provide evidence for a full DP Object Shift like in modern Icelandic". Haugan (2001) did not find any examples of full DP Object Shift in his study of Old Norse, and neither did Sundquist (2002) in a study of Middle Norwegian. Thus, Sundquist concludes that "full DP Object Shift is not an option in earlier stages of Mainland Scandinavian".

It is therefore of considerable theoretical interest to search for examples of full DP Object Shift in Old Norse texts. However, this is a tedious and time-consuming task. Even though this is a perfectly grammatical construction in Modern Icelandic, it appears to be very rare in texts. Thus, one can read dozens or even hundreds of pages without finding a single example. When the constructions that we are looking for are that rare, it is easy to overlook the few examples that actually occur in the texts that we read. Given the rarity of full DP Object Shift in Modern Icelandic, one may wonder whether those who have studied Object Shift in Old Norse have looked at a large enough corpus.

We have searched for examples of full DP Object Shift in our morphologically tagged Old Norse corpus. In this search, we use a simple program that searches for a verb in the indicative or the subjunctive, followed by a noun, an adjective, or a demonstrative pronoun in an oblique case, followed by a negation (one of the words *eigi*, *ei*, *ekki* 'not', *aldrei*, *aldregi* 'never'). We allow for up to two words between the noun/adjective/demonstrative pronoun and the negation. Thus, in addition to simple sentences with a noun immediately following the verb and preceding the negation, we will find sentences where both a demonstrative pronoun and an adjective precedes the noun, and sentences where a prepositional phrase consisting of a preposition and a noun follows the head noun.

Of course, we will neither get 100% precision nor 100% recall by using this pattern. It will miss some potential examples of Object Shift; for instance, sentences with an adverb modifying a prenominal adjective when a demonstrative pronoun is also present, or sentences with an adjective modifying an object of a preposition, which follows the head noun. Furthermore, this search pattern will return a number of sentences that are not instances of Object Shift.

When we run this search pattern on the Old Norse corpus, it returns 245 examples. The majority of these examples do not show Object Shift. These are for instance sentences like (5):

(5) hann skal þetta fé **aldregi** fá ___ síðan
'he shall this money never get since'
'he shall never have this money again'

In this sentence, the fronted NP *þetta fé* is not an object of the verb *skal*, but rather an object of the verb *fá*. Thus, this is not an instance of Object Shift but rather shows OV order in the VP, which is quite a different matter (see, for instance, Rögnvaldsson, 1996; Hróarsdóttir, 2000).

However, it doesn't take long to clean the search results and throw away the sentences that do not show Object

Shift. When we have finished this cleaning, it appears that we really are left with some genuine examples of full DP Object Shift:

(6) a. Nú leita þeir um skóginn og finna Gísla **eigi** ___
now search they about the forest and find Gisli not
'Now they search through the forest and don't find Gisli'
b. er hann dræpi Þórð **eigi** ___og förunauta hans
when he killed Thord not and companions his
'if he didn't kill Thord and his companions'
c. og fundu Þórð **eigi** ___ sem von var að
and found Thord not as expectance was at
'and not surprisingly, they didn't find Thord'

Using this method, we found at least 9 indisputable examples of full DP Object Shift. This may not be the exact number of such sentences in our corpus. First, in addition to these examples, there are some borderline cases, which may or may not be interpreted as instances of Object Shift. Second, our searching method does not guarantee 100% recall, as explained above. However, this doesn't really matter for our purposes. We have shown conclusively that full DP Object Shift existed in Old Norse, contrary to what has previously been claimed in the literature; and we have demonstrated the efficiency of our searching method.

### 4.2 Passive

Another controversial feature of Old Norse syntax is the nature of the passive. It has sometimes been claimed (Dyvik, 1980; Faarlund, 1990) that all passive sentences in Old Norse are lexical but not derived by NP-movement (or chain-formation). This claim has been disputed, for instance by Benediktsson (1980), and it has been claimed that the existence of agentive prepositional phrases (*by*-phrases) would be an argument against this analysis, since such phrases presuppose a derivational analysis of passive sentences (Rögnvaldsson, 1995).

Be that as it may, it is quite clear that agentive prepositional phrases in passives are rather rare in Modern Icelandic, and hence, one would not expect to find many of them in Old Norse. Faarlund (2004), for instance, quotes two such examples but concludes: "This is very rarely found, however."

It is not easy to search for such examples in an unannotated electronic text. One would have to search for the preposition *af* 'by', but this preposition is one of the most frequent words in Old Norse so this search would return thousands of sentences. However, once we have a morphologically tagged text, it is relatively easy to search for agentive prepositional phrases. We can search for a past participle, followed by *af*, followed by a nominal (noun, pronoun, adjective) in the dative. Since the distinction between past participle forms and adjectives in the neuter singular is not always clear, and the tagger makes a number of errors in this classification, we also search for the adjectives in addition to the past participles. This search returns some 130 sentences. Most of them are not instances of agentive phrases, since the preposition *af* can also have other functions. Nevertheless, we have found at least 15 sentences with agentive prepositional phrases, only a few of which have previously been quoted in the literature on this subject. Three of these sentences are shown below – the agentive phrases in boldface:

(7) a. að Þorvarður Spak-Böðvarsson hafi skírður verið **af Friðreki biskupi**
that Thorvard Spak-Bodvarsson has baptized been by Fridrek bishop
'that Thorvard Spak-Bodvarsson has been baptized by bishop Fridrek'
b. Og er þetta mál var rannsakað **af lögmönnum**
and when this case was investigated by lawyers
'and when lawyers investigated this case'
c. Óttar gerði sem honum var boðið **af Sighvati**
Ottar did as him was ordered by Sighvat
'Ottar did what Sighvat ordered him'

Thus, our searching method has enabled us to strengthen the evidence for the existence of derivational passive in Old Norse.

## 5. Conclusion

In this paper, we have demonstrated that it is possible to use a tagging model trained on Modern Icelandic texts to facilitate tagging of Old Norse narrative texts. By using this method, we are able to tag a large corpus of Old Norse with acceptable accuracy in a relatively short time – only a fragment of the time it would have taken to build a tagging model for Old Norse from scratch.

Furthermore, we have shown that a corpus tagged using a rich tagset based on morphosyntactic features can fruitfully be used in the search for a number of syntactic constructions, and hence is a valuable tool in studying syntactic variation and change. Of course, a morphologically tagged corpus like the one we have built doesn't amount to a fully parsed corpus. Several syntactic features cannot be searched for using our method. However, given the tremendous effort it would take to build a parsed corpus of this size, we think our method is an alternative that must be taken seriously.

Later this year, we intend to make the tagged Old Norse texts available on the web using the Xaira program (www.oucs.ox.ac.uk/rts/xaira/) from the British National Corpus. This will enable users to search the corpus for complex patterns using both words and tags in the search text. Thus, the corpus will hopefully be of use to anyone studying Old Norse language, literature, and culture.

## 6. Acknowledgements

This project was partly supported by a grant from the University of Iceland Research Fund to the project "The syntactic use of Old Icelandic POS tagged texts".

We would like to thank three anonymous reviewers for their valuable comments, which helped us improve the paper considerably, although we have not been able to follow all their suggestions.

## 7. References

Benediktsson, H. (1980). The Old Norse Passive: Some Observations. In Hovdhaugen, E. (Ed.), *The Nordic Languages and Modern Linguistics* 4. Universitetsforlaget, Oslo, Norway, pp. 108-119.

Bjarnadóttir, K. (2002). The Icelandic µ-TBL Experiment: Preparing the Corpus. Paper presented at NLP1 final session, January 9. GSLT, Växjö, Sweden.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000,* Seattle, WA, pp. 224-231.

Degnbol, H. (1985). Hvad en ordbog behøver – og andre ønsker [What a Dictionary Needs – and Others Wish for]. In *The Sixth International Saga Conference. Workshop Papers* I. Det arnamagnæanske institut, University of Copenhagen, Copenhagen, Denmark, pp. 235-254.

Dyvik, H. (1980). Har gammelnorsk passiv? [Does Old Norse have the Passive?] In Hovdhaugen, E. (Ed.), *The Nordic Languages and Modern Linguistics* 4. Universitetsforlaget, Oslo, Norway, pp. 82-107.

Faarlund, J.T. (1990). Syntactic Change. *Toward a Theory of Historical Syntax.* Mouton, Berlin, Germany.

Faarlund, J.T. (2004). *The Syntax of Old Norse.* Oxford University Press, Oxford, UK.

Halldórsson, B., Torfason, J., Tómasson, S., Thorsson, Ö. (Eds.). (1985-86). *Íslendinga sögur* [The Icelandic Family Sagas]. Svart á Hvítu, Reykjavik, Iceland.

Haugan, J. (2001). Old Norse Word Order and Information Structure. Doctoral dissertation, NTNU, Trondheim, Norway.

Helgadóttir, S. (2005). Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 257-265.

Helgadóttir, S. (2007). Mörkun íslensks texta [Tagging Icelandic Text]. *Orð og tunga*, 9, pp. 75-107.

Holmberg, A. (1986). Word Order and Syntactic Features in the Scandinavian Languages and English. Doctoral dissertation, University of Stockholm, Stockholm, Sweden.

Holmberg, A., Platzack, C. (1995). *The Role of Inflection in the Syntax of Scandinavian Languages.* Oxford University Press, Oxford, UK.

Hróarsdóttir, Þ. (2000). *Word Order Change in Icelandic: from OV to VO.* John Benjamins, Amsterdam, The Netherlands.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3), pp. 311-325.

Kristjánsdóttir, B., Halldórsson, B., Sigurðsson, G., Grímsdóttir, G.Á., Ingólfsdóttir, G., Torfason, J., Tómasson, S., Thorsson, Ö. (Eds.). (1988). *Sturlunga saga* [The Sturlunga Collection]. Svart á Hvítu, Reykjavik, Iceland.

Kristjánsdóttir, B., Halldórsson, B., Torfason, J., Thorsson , Ö. (Eds.). (1991). *Heimskringla* [The Sagas of the Kings of Norway]. Mál og Menning, Reykjavik, Iceland.

Kroch, A., Santorini, B., Delfs, L. (2004). Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME- RELEASE-1/

Kroch, A., Taylor, A. (2000). Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/PPCME2- RELEASE-2/

Kroch, A., Taylor, A. (2001). Verb-Object Order in Early Middle English. In Pintzuk, S., Tsoulas, G., Warner, A. (Eds.), *Diachronic Syntax: Models and Mechanisms.* Oxford University Press, Oxford, UK, pp. 132-163.

Kroch, A., Taylor, A., Ringe, D. (2000). The Middle English Verb-Second Constraint: a Case Study in Language Contact and Language Cange. In Herring, S., Schoesler, L., van Reenen, P. (Eds.), *Textual Parameters in Older Language*. John Benjamins, Philadelphia, pp. 353-391.

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), pp. 313-330.

Mason, L. (1999). Object Shift in Old Norse. MA thesis, University of York, York, UK.

Megyesi, B. (2002). Data-Driven Syntactic Analysis – Methods and Applications for Swedish. Doctoral dissertation, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Pind, J. (Ed.), Magnússon, F., Briem, S. (1991). *Íslensk orðtíðnibók* [Icelandic Frequency Dictionary, IFD] Orðabók Háskólans, Reykjavík, Iceland.

Rögnvaldsson, E. (1990). Orðstöðulykill Íslendinga sagna [The Concordance to the Icelandic Family Sagas]. *Skáldskaparmál*, 1, pp. 54-61.

Rögnvaldsson, E. (1995). Old Icelandic: A Non-Configurational Language? *NOWELE*, 26, pp. 3-29.

Rögnvaldsson, E. (1996). Word Order Variation in the VP in Old Icelandic. *Working Papers in Scandinavian Syntax*, 58, pp. 55-86.

Rögnvaldsson, E. (2005). Setningafræðilegar breytingar í íslensku. [Syntactic Changes in Icelandic.] In Thráinsson, H. (Ed.) *Setningar. Handbók um setningafræði* [Sentences: A Handbook on Syntax]. (Íslensk tunga III.) Almenna bókafélagið, Reykjavík, Iceland, pp. 602-635.

Rögnvaldsson, E. (2006). The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation. In Henrichsen, P.J., Skadhauge, P.R. (Eds.): *Treebanking for Discourse and Speech, Proceedings of the NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse*. Copenhagen Studies in Language 32. Samfundslitteratur, Copenhagen, Denmark, pp. 133-145.

Sundquist, J.D. (2002). Object Shift and Holmberg's Generalization. In Lightfoot, D. (Ed.), *Syntactic Effects of Morphological Change*. Oxford University Press, Oxford, UK, pp. 326-347.

Thráinsson, H. (2007). *The Syntax of Icelandic.* Cambridge University Press, Cambridge, UK.