The Icelandic m-TBL Experiment:

m-TBL Rules for Icelandic Compared to English Rules

Eiríkur Rögnvaldsson

University of Iceland
and GSLT

January 2002

This paper reports on a work in progress. Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Sigrún Helgadóttir are experimenting on the possibilities of using Torbjörn Lager's µ-TBL program to build a PoS tagger for Icelandic. Kristín writes about the preparation of the corpus and the tagsets; Auður writes about the templates for Icelandic; Sigrún writes about the testing of the tagsets and the templates; and Eiríkur writes about the rules learned by the program. Although each paper presents work that has primarily been carried out by the author, the four authors have worked closely together on this experiment. Each paper should be more or less self-contained, but they all refer heavily to work described in one or more of the other papers.

m-TBL Rules for Icelandic Compared to English Rules

1. Introduction

The purpose of this paper is to look at the **rules** that the μ-TBL tagger (Lager 1999, 2000) extracts from an Icelandic training corpus, and to compare these rules to the **rules** that the same program generates for English. μ-TBL is based on Brill's (1995) methodology, which is "mixed" in the sense that it "draws inspiration from both rule-based and stochastic taggers" (Jurafsky & Martin 2000:307).

By investigating the rules, I hope (i) to gain a better understanding of the mechanism underlying the learning process; (ii) to be in a better position to suggest modifications to the set of templates used; and (iii) to figure out the possibilities of manually revising the rules, or adding to the set of automatically learned rules.

In this experiment, I used a rich tagset for Icelandic (a slightly modified version of "Version 2", cf. Kristín Bjarnadóttir 2001). In addition to the part of speech, all the standard grammatical categories are tagged, such as gender, number, case, and definiteness in nouns, gender, case, number, definiteness and grade in adjectives, class, person, gender, number, and case in pronouns, and voice, mood, tense, person, and number in verbs. Prepositions and adverbs are treated as one class. For English, the Penn Treebank Tagset was used (Marcus, Santorini & Marcinkiewicz 1993).

My Icelandic training corpus contained 47,673 words, and the test corpus had 11,923 words. Both are taken from the source files to the Icelandic Frequency Dictionary (*Íslensk orðtíðnibók*; Pind, Magnússon & Briem 1991). For English, I used a training corpus with 60,000 words and a test corpus with 10,000 words. Both are taken from the *Wall Street Journal Corpus* (Marcus, Santorini & Marcinkiewicz 1993).

Having experimented a bit with different sets of templates (see also Auður Þórunn Rögnvaldsdóttir 2001), I decided to use the following seventeen templates:

(1)

```
      tag:A>B <- tag:C@[-1].
      tag:A>B <- tag:C@[1].
      tag:A>B <- tag:C@[-1,-2].
      tag:A>B <- tag:C@[-1,-2,-3].
      tag:A>B <- tag:C@[-1] & tag:D@[1].
      tag:A>B <- tag:C@[-1] & tag:D@[-2].
      tag:A>B <- tag:C@[-1] & tag:D@[-2] & tag:E@[-3].
      tag:A>B <- tag:C@[1,2].
      tag:A>B <- tag:C@[-1] & tag:D@[1,2].
      tag:A>B <- wd:C@[0].
      tag:A>B <- wd:C@[1].
      tag:A>B <- wd:C@[-1].
      tag:A>B <- wd:C@[0] & wd:D@[-1].
      tag:A>B <- wd:C@[0] & tag:D@[-1].
      tag:A>B <- wd:C@[0] & tag:D@[1].
      tag:A>B <- wd:C@[-1,-2].
      tag:A>B <- wd:C@[0] & wd:D@[-1] & wd:E@[-2].
```

Nine of these templates only refer to tags; six of them are lexical, that is, they only refer to words; whereas two templates refer to both words and tags.

The results of the tests were quite satisfactory. I ran the program twice in succession with the score threshold set to 4. Then I composed the rules into one stack, lowered the score threshold to 2, and ran the program once again on the training corpus. After composing the rules again, I ran the resulting set of rules on the test corpus. The results were as follows (see also Sigrún Helgadóttir 2001):

(2)

	Baseline	# of rules	# of errors	% correct
English	95,9	127	189	98,0
Icelandic	89,0	339	613	95,0

Even though the end results for Icelandic are not as good as those for English, it must be kept in mind that the initial baseline for Icelandic was much lower. Thus, if we look at the percentages, the improvement is actually much greater in the Icelandic test corpus than in the English one. However, we must also keep in mind that when the score gets higher, it becomes more and more difficult to improve on it. The results, then, may quite well be similar.

2. The Rules

Let us now look at the rules that the system learns. I have counted the number of rules that conform to each individual template. The results are shown in the following table:

(3)

Total	# Ic	% Ic	# En	% En	
15,9	50	14,7	24	18,9	tag: A>B <- tag: C @ [-1]
12,9	56	16,5	4	3,1	tag: A>B <- tag: C @ [1]
9,9	19	5,6	27	21,3	tag: A>B <- wd: C @ [0] & tag: D @ [1]
9,2	40	11,8	3	2,4	tag: A>B <- tag: C @ [1,2]
8,8	37	10,9	4	3,1	tag: A>B <- tag: C @ [-1,-2]
7,3	27	8,0	7	5,5	tag: A>B <- tag: C @ [-1,-2,-3]
6,2	21	6,2	8	6,3	tag: A>B <- wd: C @ [0]
5,6	17	5,0	9	7,1	tag: A>B <- tag: C @ [-1] & tag: D @ [1,2]
5,4	23	6,8	2	1,6	tag: A>B <- wd: C @ [-1,-2]
4,9	13	3,8	10	7,9	tag: A>B <- wd: C @ [0] & tag: D @ [-1]
4,7	15	4,4	7	5,5	tag: A>B <- wd: C @ [-1]
3,0	4	1,2	10	7,9	tag: A>B <- tag: C @ [-1] & tag: D @ [1]
2,4	6	1,8	5	3,9	tag: A>B <- wd: C @ [1]
2,4	7	2,1	4	3,1	tag: A>B <- tag: C @ [-1] & tag: D @ [-2]
1,1	3	0,9	2	1,6	tag: A>B <- tag: C @ [-1] & tag: D @ [-2] & tag: E @ [-3]
0,2		0,0	1	0,8	tag: A>B <- wd: C @ [0] & wd: D @ [-1] & wd: E @ [-2]
0,2	1	0,3		0,0	tag: A>B <- wd: C @ [0] & wd: D @ [-1]
100,0	339	100,0	127	100,0	

The first column shows the percentage of the rules for both Icelandic and English (339+127) that conform to each template. The next two columns show the number and the percentage, respectively, of the Icelandic rules conforming to each template; and the fourth and fifth columns show the corresponding figures for English.

By and large, there is a reasonably good correspondence between Icelandic and English, but there are some interesting differences, however. We see that the tag of the preceding word (`tag:C@[-1]`) is an important factor in both languages, accounting for 14,7% of the rules in Icelandic and 18,9% in English. On the other hand, the tag of the following word (`tag:C@[1]`) seems to be a much stronger factor in predicting correct tags in Icelandic than in English; in Icelandic, it accounts for 16,5% of the rules, whereas it only accounts for 3,1% of the rules in English.

In comparing the predictive power of the different parts of the context, it may give a better picture to group related templates together. First, we look at six templates that refer to the preceding tag(s):

(4)

Total	# Ic	% Ic	# En	% En	
15,9	50	14,7	24	18,9	`tag: A>B <- tag: C @ [-1]`
8,8	37	10,9	4	3,1	`tag: A>B <- tag: C @ [-1,-2]`
7,3	27	8,0	7	5,5	`tag: A>B <- tag: C @ [-1,-2,-3]`
4,9	13	3,8	10	7,9	`tag: A>B <- wd:  C @ [0]  & tag: D @ [-1]`
2,4	7	2,1	4	3,1	`tag: A>B <- tag: C @ [-1] & tag: D @ [-2]`
1,1	3	0,9	2	1,6	`tag: A>B <- tag: C @ [-1] & tag: D @ [-2]& tag: E @ [-3]`
40,3	137	40,4	51	40,2	

These templates account for almost exactly the same percentage of the rules in Icelandic and English. However, it is clear that the immediately preceding tag (`tag:C@[-1]` and `wd:C@[0] & tag:D@[-1]`) is most important in English, accounting for 26,8% of the rules vs. 18,5% in Icelandic. On the other hand, the tags that either can be immediately preceding the word in question, or with one or two words intervening (`tag:C@[-1,-2]` and `tag:C@[-1,-2,-3]`), are also important in Icelandic, accounting for 18,9% of the rules vs. only 8,6% in English.

Next we look at three templates that refer to the following tag(s):

(5)

Total	# Ic	% Ic	# En	% En	
12,9	56	16,5	4	3,1	`tag: A>B <- tag: C @ [1]`
9,2	40	11,8	3	2,4	`tag: A>B <- tag: C @ [1,2]`
9,9	19	5,6	27	21,3	`tag: A>B <- wd:  C @ [0]   & tag: D @ [1]`
32,0	115	33,9	34	26,8	

As pointed out above, the predictive power of the following tag(s) appears at first sight to be much greater in Icelandic than in English. Taken together, the templates `tag:A>B <- tag:C@[1]` and `tag:A>B <- tag:C@[1,2]` account for 28,3% of the rules in Icelandic, but only 5,5% in English. However, the picture changes drastically when we add the third similar template; `tag:A>B <- wd:C@[0] & tag:D@[1]`. This template turns out to be very useful in English, but of much less importance in Icelandic. This is reminiscent of the situation in the first group of rules, those referring to the preceding context; the template `tag:A>B <- wd:C@[0] & tag:D@[-1]` is also much more important in English than in Icelandic, although the difference is not as great as here. As with the preceding context, we can see here that the immediately following tag is most important for English, whereas the template `tag:A>B <- tag:C@[1,2]`, where a word can intervene between the word in question and the affecting tag, is also important for Icelandic.

Next we have two templates where both the preceding and the following tag(s) seem to matter:

(6)

Total	# Ic	% Ic	# En	% En	
3,0	4	1,2	10	7,9	`tag: A>B <- tag: C @ [-1] & tag: D @ [1]`
5,6	17	5,0	9	7,1	`tag: A>B <- tag: C @ [-1] & tag: D @ [1,2]`
8,6	21	6,2	19	15,0	

It is clear that these templates are more important in English than in Icelandic, but I cannot point to a particular reason for that. Note, however, that we see the same pattern here as in the other sets of templates; the template where a word is allowed to intervene is relatively important in Icelandic.

The following two templates refer to the preceding words:

(7)

Total	# Ic	% Ic	# En	% En	
4,7	15	4,4	7	5,5	`tag: A>B <- wd:  C @ [-1]`
5,4	23	6,8	2	1,6	`tag: A>B <- wd:  C @ [-1,-2]`
10,1	38	11,2	9	7,1	

The importance of the preceding word is relatively similar in Icelandic and English, but as usual, the template where a word is allowed to intervene is much more important in Icelandic.

The following template does not refer to any feature in the context, be it word or tag; it merely states that the tag of a certain word should be changed from A to B in all cases.

(8)

Total	# Ic	% Ic	# En	% En	
6,2	21	6,2	8	6,3	`tag: A>B <- wd:  C @ [0]`

This is an instance of a purely statistical template. It is easy to see that rules based on it are bound to produce some wrong results, given the premise that our initial annotator works properly. If the annotator has assigned two different tags to a certain word, we must assume that the word can in fact have two different analyses. Thus, a rule that always replaces analysis A with analysis B cannot always be right. However, it might very well be right in such an overwhelming number of cases that this type of rule is in fact justified, despite its shortcomings.

Finally, we have the three remaining templates:

(9)

Total	# Ic	% Ic	# En	% En	
2,4	6	1,8	5	3,9	tag: A>B <- wd: C @ [1]
0,2		0,0	1	0,8	tag: A>B <- wd: C @ [0] & wd: D @ [-1]& wd: E @ [-2]
0,2	1	0,3		0,0	tag: A>B <- wd: C @ [0] & wd: D @ [-1]
2,8	7	2,1	6	4,7	

The following word has some value in English, but very small in Icelandic. The other two templates are only used in one rule each, one in English and the other in Icelandic. Thus, there is very little to say about these templates.

3. The Differences between Icelandic and English

There appear to be three main differences between the sets of rules for Icelandic and those for English. First, the templates that only refer to the immediately preceding or following word or tag play a bigger role in English than in Icelandic. This is shown in the following table:

(10)

Total	# Ic	% Ic	# En	% En	
9,9	19	5,6	27	21,3	tag: A>B <- wd: C @ [0] & tag: D @ [1]
15,9	50	14,7	24	18,9	tag: A>B <- tag: C @ [-1]
4,9	13	3,8	10	7,9	tag: A>B <- wd: C @ [0] & tag: D @ [-1]
3,0	4	1,2	10	7,9	tag: A>B <- tag: C @ [-1] & tag: D @ [1]
6,2	21	6,2	8	6,3	tag: A>B <- wd: C @ [0]
4,7	15	4,4	7	5,5	tag: A>B <- wd: C @ [-1]
2,4	6	1,8	5	3,9	tag: A>B <- wd: C @ [1]
12,9	56	16,5	4	3,1	tag: A>B <- tag: C @ [1]
0,2	1	0,3		0,0	tag: A>B <- wd: C @ [0] & wd: D @ [-1]
60,1	185	54,6	95	74,8	

We see here that these templates are responsible for ¾of the rules in English, but only 54,6% in Icelandic. Conversely, the templates that take a larger context into account are relatively more important in Icelandic:

(11)

Total	# Ic	% Ic	# En	% En	
9,2	40	11,8	3	2,4	tag: A>B <- tag: C @ [1,2]
8,8	37	10,9	4	3,1	tag: A>B <- tag: C @ [-1,-2]
7,3	27	8,0	7	5,5	tag: A>B <- tag: C @ [-1,-2,-3]
5,4	23	6,8	2	1,6	tag: A>B <- wd: C @ [-1,-2]
5,6	17	5,0	9	7,1	tag: A>B <- tag: C @ [-1] & tag: D @ [1,2]
2,4	7	2,1	4	3,1	tag: A>B <- tag: C @ [-1] & tag: D @ [-2]
1,1	3	0,9	2	1,6	tag: A>B <- tag: C @ [-1] & tag: D @ [-2]& tag: E @ [-3]
0,2		0,0	1	0,8	tag: A>B <- wd: C @ [0] & wd: D @ [-1]& wd: E @ [-2]
39,9	154	45,4	32	25,2	

The reason for this difference can probably for the most part be traced to the agreement properties of Icelandic. There is internal agreement in noun phrases; we find, for instance, the following pattern:

(12)

Hér	eru	þessi	litlu	börn
here	are	these	small	children

Ég	sá	þessi	litlu	börn
I	saw	these	small	children

In the first example, all the words in the string *þessi litlu börn* are in the nominative, whereas they are in the accusative in the second example. Unfortunately, the nominative and accusative forms of all these words are identical; but we can still predict the correct tag by referring to the verb, since the verb *vera* ‘be’ always takes a predicate phrase in the nominative, whereas the verb *sjá* ‘see’ takes an object in the accusative. Furthermore, Icelandic has agreement between subjects and predicate adjectives. Many of the following examples seem to involve agreement of these types.

(13)

```
tag: lhensf > lheosf  <- tag: fp1en    @ [-1,-2,-3]
tag: fphen  > fpheo   <- tag: fp2en    @ [-1,-2,-3]
tag: nheo   > nhen    <- tag: fphen    @ [-1,-2,-3]
tag: sþghen > ssg     <- tag: fpken    @ [-1,-2,-3]
tag: nheng  > nheog   <- tag: fpkeo    @ [-1,-2,-3]
tag: nveþ-m > nvee-m  <- tag: fpvee    @ [-1,-2,-3]
tag: fpkeþ  > fpveþ   <- tag: fpven    @ [-1,-2,-3]
tag: fpkeþ  > fpveþ   <- tag: nven-m   @ [-1,-2,-3]
tag: fovfn  > fovfo   <- tag: sfg3eþ   @ [-1,-2,-3]
tag: lhfnsf > lvensf  <- tag: sfg3eþ   @ [-1,-2,-3]
tag: lvensf > lhfnsf  <- tag: sfg3fn   @ [-1,-2,-3]
tag: sþgven > sþghfn  <- tag: sfg3fn   @ [-1,-2,-3]
tag: foheo  > fohen   <- tag: sfm3eþ   @ [-1,-2,-3]
tag: nhfng  > nhfog   <- tag: sng      @ [-1,-2,-3]
tag: tfhen  > tfheo   <- tag: sng      @ [-1,-2,-3]
tag: tfken  > tfkeo   <- tag: sng      @ [-1,-2,-3]
```

In the rule `tag:nhfng>nhfog <- tag:sng @[-1,-2,-3]`, for instance, the tag of a noun is changed from nominative (`n`) to accusative (`o`) if a verb in the infinitive precedes it; and one or two words may intervene. (In fact, this rule is of course much more general, since other forms of the verb also affect the case of a following noun; but we may assume that such combinations have not occurred often enough to be learned as rules).

Another difference between the Icelandic and English rules lies in the importance of lexical templates, referring to individual words. Admittedly, the templates that exclusively refer to words account for a similar percentage of the rules in both languages:

(14)

Total	# Ic	% Ic	# En	% En	
6,2	21	6,2	8	6,3	tag: A>B <- wd: C @ [0]
4,7	15	4,4	7	5,5	tag: A>B <- wd: C @ [-1]
2,4	6	1,8	5	3,9	tag: A>B <- wd: C @ [1]
5,4	23	6,8	2	1,6	tag: A>B <- wd: C @ [-1,-2]
0,2	1	0,3		0,0	tag: A>B <- wd: C @ [0] & wd: D @ [-1]
0,2		0,0	1	0,8	tag: A>B <- wd: C @ [0] & wd: D @ [-1]& wd: E @ [-2]
19,1	66	19,5	23	18,1	

The two "mixed" templates, those referring to both words and tags, play a much bigger role in English than in Icelandic:

(15)

Total	# Ic	% Ic	# En	% En	
4,9	13	3,8	10	7,9	tag: A>B <- wd: C @ [0] & tag: D @ [-1]
9,9	19	5,6	27	21,3	tag: A>B <- wd: C @ [0] & tag: D @ [1]
14,8	32	9,4	37	29,1	

Conversely, the templates that only refer to tags account for more than 70% of the rules in Icelandic, whereas in English they only account for a little more than half of the set of rules:

(16)

Total	# Ic	% Ic	# En	% En	
15,9	50	14,7	24	18,9	tag: A>B <- tag: C @ [-1]
3,0	4	1,2	10	7,9	tag: A>B <- tag: C @ [-1] & tag: D @ [1]
5,6	17	5,0	9	7,1	tag: A>B <- tag: C @ [-1] & tag: D @ [1,2]
7,3	27	8,0	7	5,5	tag: A>B <- tag: C @ [-1,-2,-3]
12,9	56	16,5	4	3,1	tag: A>B <- tag: C @ [1]
8,8	37	10,9	4	3,1	tag: A>B <- tag: C @ [-1,-2]
2,4	7	2,1	4	3,1	tag: A>B <- tag: C @ [-1] & tag: D @ [-2]
9,2	40	11,8	3	2,4	tag: A>B <- tag: C @ [1,2]
1,1	3	0,9	2	1,6	tag: A>B <- tag: C @ [-1] & tag: D @ [-2]& tag: E @ [-3]
66,1	241	71,1	67	52,8	

The question is, then: Why is the role of the current word (`wd:C@[0]`) much bigger in English than in Icelandic? The answer is probably also related to the inflectional character of Icelandic. In testing his tagger, Brill (1995:23) was surprised that "the addition of lexicalized transformations did not result in a much greater improvement in performance". He points out that "[w]hen transformations are allowed to make reference to words and word pairs, some relevant information is probably missed due to sparse data". If this is the case in English, one might expect the problem to be much bigger in Icelandic. Since the same lexeme can appear in many different forms there, chances are that each form does not appear often enough in the corpus for a rule to be established. Thus, most of the individual words that appear in the Icelandic rules are those that have invariable form, such as adverbs, prepositions, and conjunctions. This is evident from the following exhaustive list of rules that apply the "mixed" templates `tag:A>B <- wd:C@[0] & tag:D@[-1]` and `tag:A>B <- wd:C@[0] & tag:D@[1]`:

(17)

```
tag: af     > cn     <- wd: að     @ [0]  & tag: sng    @ [1]
tag: c      > cn     <- wd: að     @ [0]  & tag: sng    @ [1]
tag: c      > cn     <- wd: að     @ [0]  & tag: sng    @ [1]
tag: af     > c      <- wd: að     @ [0]  & wd:  þó     @ [-1]
tag: nkee   > nkeþ   <- wd: afa    @ [0]  & tag: c      @ [-1]
tag: fohen  > foheo  <- wd: allt   @ [0]  & tag: af     @ [-1]
tag: af     > sfg1en <- wd: á      @ [0]  & tag: fp1en  @ [1]
tag: af     > sfg1en <- wd: á      @ [0]  & tag: fp1en  @ [-1]
tag: ssg    > sfg2en <- wd: átt    @ [0]  & tag: fp2en  @ [-1]
tag: sfg3eþ > sfg1eþ <- wd: brosti @ [0]  & wd:  til    @ [1]
tag: lhensf > ssg    <- wd: búið   @ [0]  & tag: af     @ [1]
tag: foheþ  > lheþsf <- wd: einu   @ [0]  & wd:  í      @ [-1]
tag: fpkeo  > fpken  <- wd: hann   @ [0]  & tag: sfg3eþ @ [1]
tag: aam    > sfg3en <- wd: heldur @ [0]  & wd:  að     @ [1]
tag: af     > lhensf <- wd: mikið  @ [0]  & tag: nhen   @ [1]
tag: nhen   > nheo   <- wd: orð    @ [0]  & tag: af     @ [1]
tag: nkeo   > nkee   <- wd: pabba  @ [0]  & tag: c      @ [1]
tag: af     > lhensf <- wd: rétt   @ [0]  & tag: .      @ [1]
tag: sfg3eþ > faken  <- wd: sá     @ [0]  & tag: lkenvf @ [1]
tag: sfg3eþ > faken  <- wd: sá     @ [0]  & tag: nken   @ [1]
tag: fpkeþ  > fpveþ  <- wd: sér    @ [0]  & wd:  á      @ [-1]
tag: fpkeo  > fpveo  <- wd: sig    @ [0]  & wd:  í      @ [1]
tag: nkeþ   > nkeo   <- wd: stað   @ [0]  & wd:  í      @ [-1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: .      @ [-1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: c      @ [-1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: sfg1fn @ [1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: sfg1fn @ [-1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: sfg1fþ @ [1]
tag: af     > fp1fn  <- wd: við    @ [0]  & tag: sfg1fþ @ [-1]
tag: lkensf > lvensf <- wd: viss   @ [0]  & tag: af     @ [1]
tag: af     > fpkfo  <- wd: þá     @ [0]  & tag: af     @ [-1]
tag: af     > fpkfo  <- wd: þá     @ [0]  & tag: sng    @ [-1]
```

Of these, the words *að*, *af*, *á*, *í*, *við*, *þá* are typical prepositions/adverbs/conjunctions, even though they also can belong to other lexemes. The 98 Icelandic rules that refer to individual words only mention 65 different word forms, whereas the 60 English rules that refer to individual words mention 51 different word forms.

The third difference lies in the type of errors corrected by the rules. In English, almost 2/3 of the rules change one major part of speech to another, whereas in Icelandic, only 18% of the rules do that. Thus, 72% of the Icelandic rules only serve to identify the correct grammatical categorization of the word in question, such as number, gender, case, and so on.

By far the greatest number of rules in Icelandic have the role of correcting the case of nouns, adjectives and pronouns. These rules are 151, or 44,5% of the total number of rules. 102 of those rules change nominative to accusative or vice versa. 44 rules, or 13%, change the gender of adjectives and pronouns. Other rule types, such as those that change the person of verbs, or the number of adjectives and pronouns, are much more rare.

4. The Remaining Errors

As far as I can see, a considerable proportion (170-180 instances, including most of the common types) of the **remaining 613 errors** in the Icelandic test corpus involves case. This means that a lot would be gained if we could modify the rules such that they would do a better job in correctly identifying the cases. This is, however, often rather difficult. Let us for instance look at one of the three most common types of errors; nouns in the dative which are however tagged as accusative:

10 occurrences tagged as nveo that should be nveþ:

327:	að þeir séu sleipir í	**sögu**	. það er eiginlega ekki
489:	að fara í andaglas í	**tölvu**	. hvað þessir strákar þykjast
1372:	ofan í sig , þessari	**túpu**	sem átti að liggja slétt
3130:	sem hún fylgdist með hverri	**hreyfingu**	hetju sinnar . Alli gaf
5883:	einhverjum ástæðum var ömmu í	**nöp**	við þessa iðju . Jóra
6586:	týndist Stóri-Jón reyndi eftir bestu	**getu**	að segja frá ferðalaginu niður
10351:	súkkulaði og kaffi inni í	**stofu**	. meira að segja Guðmundur
4101:	ætlaði að koma við í	**sjoppu**	og . kaupa eitthvert sælgæti
5469:	. einar dyr voru á	**framhlið**	hans en þær voru sjaldan
5049:	var farið að svima af	**eftirvæntingu**	. hún rétti út höndina

In all but one of these examples, the noun in question is preceded by a preposition (sometimes another word intervenes between the preposition and the noun). The problem is, however, that these prepositions sometimes govern accusative and sometimes dative. It is of course usually possible to determine from the context which case is the correct one, but it is not clear to me whether it would be possible to write rules for that. These rules would have to be rather complicated, and it is quite possible that they would lead to more errors than correct analyses.

Another class of the most common errors is furnished by the neuter personal pronoun *það*. It has identical forms in the nominative and accusative, and is often incorrectly tagged; in the following examples, the accusative form is tagged as nominative:

10 occurrences tagged as fphen that should be fpheo:

6662:	átti ég að vita ,	**það**	muldraði . lögregluþjónninn má ég
7981:	þeirra í skyndilega . þögn	**það**	gerði hungrið og . þorstinn
8193:	undir fótum sér , sá	**það**	, heyrði það , fékk
8199:	, heyrði það , fékk	**það**	í fangið þegar hann datt
9864:	þrjóskur . flugmaðurinn sagði mér	**það**	. þá lýgur hann því
10563:	skotinn í henni , sérðu	**það**	ekki ? hvíslaði Kata spekingslega
10600:	í bók að maður sjái	**það**	á augunum í fólki .
10873:	jólafríið kemur . kennarinn segir	**það**	. ætlar hann að kenna
11426:	já . amma mín kallaði	**það**	að krossa sig , en
11870:	skapað líf . Jesús getur	**það**	. af því að hann

In all but one of these examples, *það* is preceded by a verb. This makes it very likely that *það* is an object here, and hence not in the nominative. However, the rules only refer to whole tags, but not to major parts of speech. Although *það* is preceded by verbs in most of these cases, these verbs are tagged differently, since they have different moods, persons, tenses, etc. Hence, the instances involving each individual tag are not many enough for the program to establish a rule.

There are two ways to remedy this situation. One is to enlarge the training corpus. If the training corpus is large enough, we can expect that at least some of the different verb tags occur often enough preceding *það* for the correct rule to be established. The other solution would be to enable the program to refer either to a part of the tag (in this case, only the first letter, the s that shows that we are dealing with a verb), or to the lexeme. This would lead to much more general rules, as pointed out by Brill (1995:23-25). If we could combine these two ways, i.e., a larger training corpus and more general rules, we ought to be able to get reasonably good results in PoS tagging of Icelandic texts.

References

Auður Þórunn Rögnvaldsdóttir. 2002. The Icelandic µ-TBL Experiment: Templates for Icelandic. Term paper in NLP 1, GSLT.

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* 21: 543-566.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing.* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey.

Jörgen Pind (ed.), Friðrik Magnússon and Stefán Briem. 1991. *Íslensk orðtíðnibók.* Orðabók Háskólans, Reykjavík.

Kristín Bjarnadóttir. 2002. The Icelandic µ-TBL Experiment: Preparing the Corpus.Term paper in NLP 1, GSLT.

Lager, Torbjörn. 1999. The µ-TBL System: Logic Programming Tools for Transformation-Based Learning. In *Proceedings of the Third International Workshop on Computational Natural Language Learning* (CoNLL'99), Bergen.

Lager, Torbjörn. 2000. The µ-TBL system. User's manual. Version 0.9.

Marcus, Mitch P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19:313-330.

Sigrún Helgadóttir. 2002. The Icelandic µ-TBL Experiment: Learning rules from four different corpora by using the µ-TBL System – Further developments. Term paper in NLP 1, GSLT.