The Icelandic μ-TBL Experiment: Preparing the Corpus

Kristín Bjarnadóttir
Institute of Lexicography
Reykjavík

December 31, 2001

1 Introduction

My part in the Icelandic μ-TBL project was preparing an existant tagged corpus of modern Icelandic for the trial runs with the μ-TBL PoS tagger, i.e., adapting the format of the corpus and introducing the appropriate errors for the training and testing corpora by the use of a mock annotator.[1] Four sets of corpora were prepared, by gradually simplifying the tagsets.

2 The Icelandic Frequency Dictionary

The corpus used in our μ-TBL experiment is a part of the corpus created in the making of the *Icelandic Frequency Dictionary* (IFD, *Íslensk orðtíðnibók*), published in 1991 by the Institute of Lexicography in Reykjavík. The editors were Jörgen Pind, Friðrik Magnússon, & Stefán Briem, and work on the dictionary began in 1985. The IFD is modelled on the Brown Corpus and related projects, the editors quoting Garside, Leech, & Sampson 1987 as an important source (among others). For reasons of space and time I will not go into details of the IFD here, except when it touches on our work on the μ-TBL experiment, referring those interested to the English version of the preface to the IFD (cf. information on the Institute of Lexicography's website: http://www.lexis.hi.is.)

The IFD corpus is a carefully selected and balanced one, consisting of half a million running words. For our experiment we chose to use only a tenth of it, aiming for 50,000 words from texts from books for children and youngsters. To give you an idea of how this part of the corpus fits in with the larger scheme, the

[1] This paper is the 1st part of a description of the project, the following papers describing other aspects: Auður Rögnvaldsdóttir: *Templates for Icelandic*, Sigrún Helgadóttir: *Learning rules from four different training corpora by using the μ-TBL System — Further developments*, and Eiríkur Rögnvaldsson: *μ-tbl Rules for Icelandic compared to English Rules.*

criteria for the selection of texts for the corpus used in the *Icelandic Frequency Dictionary* are these:

1. 100 fragments of texts, approximately 5,000 running words each.
2. All the texts were published for the first time in 1980–1989.
3. Five categories of texts:
   (a) Icelandic fiction.
   (b) Translated fiction.
   (c) Biographies and memoirs.
   (d) Non-fiction (evenly divided between science & humanities).
   (e) Books for children and youngsters (original Icelandic and translations).
4. No two texts can be attributed to the same person, i.e., as author or translator.

In our experiment with μ-TBL we used half of the section of books for children and youngsters, 10 fragments of original Icelandic texts (i.e., not translations), approximately 5,000 words each. These texts were chosen as they are fairly straight-forward, and in ordinary, everyday language for the most part.[2] We deemed these texts to be complicated enough for our first experiment with a Brill type tagger for Icelandic, perhaps bearing in mind the good results obtained in assignment 2 in the NLP1 course by the use of a homogenous text from the Wall Street Journal. The cohesive style and subject matter of that publication is probably one of the reasons why people like to use it for testing.

I will now describe the characteristics of the IFD texts and the modifications of the analyzed texts needed for our work.

2.1 The Characteristics and Format of the IFD Corpus

The IFD was published over ten years ago, and in the intervening period all the people working on that project have left the Institute of Lexicography, where I am a member of staff. It appears that the texts in the IFD corpus were to a slight degree modified in the process of the work on the dictionary. The main characteristics of the texts and the modifications are these:

1. All texts start and finish with a complete sentence.
2. The texts were read carefully and obvious spelling errors, etc., corrected:
   (a) All punctuation marks were deleted. (These were re-inserted into the analysed text before the μ-TBL trials.)

[2] My experience in lexicography favoured this material; it is on the whole a very good source of examples of everyday language as it is usually not stilted, artificial or pretentious.

(b) The spelling was to a limited extent normalized (e.g. *z* was replaced by *s*, etc.).

(c) Word boundaries were sometimes changed in the text, making for a more uniform spelling.

(d) The texts were originally obtained from printers, and it seems that not all files were of the actual published material. In such cases corrected texts were used for the printed dictionary, leaving a discrepancy between the Institute of Lexicography's text collection and the files containing the actual analysis. In other words, the analysis was corrected according to the printed material, but the unannotated text files were not.

The tagging for the IFD was done in three stages, starting with hand-tagging, followed by two stages of automatic tagging. The material was then painstakingly corrected by hand.

1. I stage: 54,000 running words were handtagged in a pilot project (cf. Friðrik Magnússon. 1988. „Hvað er títt?“ *Orð & tunga* 1:1–49.)

2. II stage: Automatic tagging, based on the handtagged material in stage I; 50 texts, 250,000 running words.

3. III stage: Automatic tagging, with improvements to the tagger based on stage II; 50 texts, 250,000 running words.

The program for the automatic tagging was especially created by Stefán Briem for the project. In the preface of the IFD, the editors estimate the results to be a correct analysis for just over 80% of the word forms in the corpus, giving both the correct grammatical tags and the appropriate lemmatization. (Subsequent work on the program improved its performance on ordinary texts to just below 90%.)

2.2 The IFD Tagset

The tagset used in the printed IFD is more or less based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized, mainly where the traditional division into word classes seems not to be substantiated by syntactic facts. Thorough explanations of the tagset are found in the introduction to the IFD, and the full set of grammatical features included in the tagset are shown in the English version of the preface to the IFD as follows (p. 40):

Column	Category	Analytical symbol - information
1	Word class	**N-noun**
2	Gender	**K**-masculine, **V**-feminine, **H**-neuter, **X**-gender unspecified
3	Number	**E**-singular, **F**-plural
4	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
5	Article	**G**-with suffixed definite article
6	Proper noun	**M**-name of person, **Ö**-place name, **S**-other proper noun
1	Word class	**L-adjective**
2	Degree	**F**-positive, **M**-comparative, **E**-superlative
3	Declension	**S**-strong declension, **V**-weak declension, **O**-indeclinable
4	Gender	**K**-masculine, **V**-feminine, **H**-neuter
5	Number	**E**-singular, **F**-plural
6	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
1	Word class	**F-pronoun**
2	Subcategory	**A**-demonstrative pronoun, **B**-indefinite demonstrative pronoun, **E**-possessive pronoun, **O**-indefinite pronoun, **P**-personal pronoun, **S**-interrogative pronoun, **T**-relative pronoun
3	Gender/Person	**K**-masculine, **V**-feminine, **H**-neuter / **1**-1st pers., **2**-2nd pers.
4	Number	**E**-singular, **F**-plural
5	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
1	Word class	**G-article**
2	Gender	**K**-masculine, **V**-feminine, **H**-neuter
3	Number	**E**-singular, **F**-plural
4	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
1	Word class	**T-numeral**
2	Category	**F**-cardinal number
3	Gender	**K**-masculine, **V**-feminine, **H**-neuter
4	Number	**E**-singular, **F**-plural
5	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
1	Word class	**S-verb** (except for past participle)
2	Voice	**G**-active, **M**-middle
3	Mood	**N**-infinitive, **B**-imperative, **F**-indicative, **V**-subjunctive, **S**-supine, **L**-present participle
4	Tense	**N**-present, **Þ**-past
5	Number	**E**-singular, **F**-plural
6	Person	**1**-1st person, **2**-2nd person, **3**-3rd person
1	Word class	**S-verb** (past participle)
2	Voice	**G**-active, **M**-middle
3	Mood	**Þ**-past participle
4	Gender	**K**-masculine, **V**-feminine, **H**-neuter
5	Number	**E**-singular, **F**-plural
6	Case	**N**-nominative, **O**-accusative, **Þ**-dative, **E**-genitive
1	Word class	**A-adverb**
2	Degree	**M**-comparative, **E**-superlative
3	Category/ case governor	**A**-does not govern case, **U**-exclamation / **O**-governs accusative, **Þ**-governs dative, **E**-governs genitive
1	Word class	**C-conjunction**
2	Category	**N**-sign of infinitive, **T**-relative conjunction
1	Word class	**E**-foreign word
1	Word class	**X**-unanalysed word

The features in the tagset appear in columns in the printed book and in the computer files, where they are separated by a blank space. This did cause some problems, as the tags are not of a fixed lenght, and blank spaces are also used as field separators in the remainder of the line where the word form and the lemma appear. Further complications were caused by the fact that the tagsets were rearranged before the dictionary was printed, leaving a different set of tags in the corpus from that of the printed book. As the tagset we were using for our annotator was based on the material from the printed book this caused some problems.

The rearrangement of the tagset simply consisted in switching the order of columns within the tagset around (as in adjectives: *leshen* > *lhense, lfohfo* > *loffoh*, etc.) However, the tagset was also simplified before printing by removing information on the case-assignment of verbs from the tags for the verbs. This simplifies the strings — and leaves us with a discrepancy in the data in the two sources.[3] No comments on the changes have been found yet at the Institute, and no lists of the tags in the corpus, but we did manage to work them out (as far as we can tell). The lack of documentation caused quite a bit of confusion and a great deal of extra work.

There are 621 tags in the tagset for the printed IFD, even though there are gaps in the set which would probably be filled with a larger corpus. This is admittedly a very large tagset, and thus a candidate for reductions, as I will come back to, but for our first experiments with the μ-TBL tagger this was the tagset we used. We then proceeded to simplify the tagset, as I will get back to, but first a look at the ambiguity of word forms in Icelandic is necessary.

2.3 Ambiguous Word Forms in the IFD

There are 59,343 word forms in the entire IFD. 15.9% of the word forms are ambiguous as to tagsets within the IFD. This figure is quite high, at least compared to English, but then the inflectional morphology of Icelandic is considerably more complex than in English. An Icelandic noun can have up to 16 grammatical forms or tags, an adjective up to 120 tags, and a verb over a hundred tags. Some of the ambiguity is due to the fact that inflectional endings in Icelandic do a very heavy duty, the same ending often appearing in many places (e.g. *-a* in *penna* 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 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; there are gaps in the paradigm shown in the occurrences in the IFD:

[3] Very many verbs in Icelandic can assign more than one case to their objects. The difference in structure can be meaning specific, as in *ræna manninn* 'rob the man', and *ræna manninum* 'kidnap the man'. For a PoS tagger this difference may not always be of crucial importance; for a lexicographer it is, as it must be for semantic disambiguation.

Word form: minni

Lexeme:	*Tags & grammatical features:*
lítill adjective	*minni* 'smaller', comparative, weak declension:
	4 tags: LMVKEN LMVKEO LMVKEÞ LMVKEE
	(masc.sg. nom., acc., dat., gen.)
	3 tags: LMVKFO LMVKFÞ LMVKFE
	(masc.pl. acc., dat., gen.)
	4 tags: LMVVEN LMVVEO LMVVEÞ LMVVEE
	(fem.sg. nom., acc., dat., gen.)
	3 tags: LMVHFN LMVHFO LMVHFE
	(neut.pl. nom., acc., gen.)
	(gaps: masc.pl.nom. and neut.pl.dat.)
minni noun, neut.	'memory':
	5 tags: NHEN NHEO NHEÞ NHFN NHFO
	(sg.nom., acc., dat.; pl.nom., acc.)
minn possess. pron.	'mine'
	1 tag: FEVEÞ (fem.sg.dat.)
minna verb	*minni* 'remind' (sby of sth.), 'remember'
	1 tag: SGVNE3
	(active voice, subjunct., pres., sing. 3rd pers.)
	(a number of gaps in the paradigm)

The ambiguous word forms are thus both within one lexeme and between lexemes, sometimes involving quite a few lexemes in fact, as seen in the following examples of ambiguous word forms, where the numbers refer to different tags within each lexeme:

Word form	**Lexeme and number of tags in the IFD**
eins	*eins* adj. (12), *einn* pron. (2), *einn* adj. (2), *einn* num. (2), *eins* adv. (1)
líka	*líki* n.masc. (3), *líka* adv. (1), *líka* v. (1), *líkur* adj. (1), *líkur* n.masc. (1)
rétt	*réttur* adj. (5), *rétt* n.fem. (2), *rétta* v. (2), *rétt* adv. (1), *réttur* n.masc. (1)
vara	*vara* v. (2), *vari* n.masc. (2), *vara* n.fem. (1), *vara* v. (1), *vör* n.fem. (1)
Þá	*sá* pron. (2), *hann* pron. (1), *þá* adv. (1), *þá* conj. (1), *þiggja* v. (1)

This proliferation of ambiguous tags from the IFD is the raw material on which to train the μ-TBL tagger.[4]

[4]The large number of tags for each word form does make one wonder whether it would be possible to include the lemma itself as a part the information for the μ-TBL tagger, especially as it is included in the analysis for the IFD which always includes the base form of a word as

3 The μ-TBL Corpus

The section used for our test runs is approximately a tenth of the complete IFD corpus, 52,403 running words. Roughly 20% of the material was used as a test corpus (10,503 running words, 11,923 tagged lines, including the reintroduced punctuation marks), and the remainder made up the training corpus (41,900 running words, 47,673 tagged lines). Each of the ten texts in the trial corpus was split into two parts in the right proportions, breaking off on complete sentences.

3.1 Adapting the data for the μ-TBL tagger

I will not discuss the problems of adapting the files in any detail. Here is an example from the IFD files:

```
n k e n - m   Emil      Emil    |à
s f g 3 e þ   horfði    horfa   |à
l k e n s f   heillaður heillaður       |à
a o           inní      inní    |à
n h f o       augu      auga    |à
n k e e g     hvolpsins hvolpur |à
s f g 3 e þ þ gleymdi   gleyma  |à
n v e þ       stund     stund   |à
c             og        og      |à
n k e þ       stað      staður  |à
f o h e n     ekkert    enginn  |à
s f g 3 e þ   var       vera    |à
a a           til       til     |à
c             nema      nema    |à
f a h f n     þessi     þessi   |à
l h f n v f   brúnu     brúnn   |à
l h f n v f   saklausu  saklaus |à
n h f n       augu      auga    |à
```

The third column in the IFD files contains the lemma for the word form being analyzed, and the fourth column is a marker for the part of the corpus the text belongs in. The corresponding part of the text in the μ-TBL format looks like this:

```
wd(9709,'Emil').
tag(9709,'nken-m').
tag('nken-m','nken-m',9709).
wd(9710,'horfði').
tag(9710,'sfg3eþ').
tag('sfg3eþ','sfg3eþ',9710).
wd(9711,'heillaður').
tag(9711,'lkensf').
tag('lkensf','lkensf',9711).
```

```
wd(9712,'inní').
tag(9712,'ao').
tag('ao','ao',9712).
wd(9713,'augu').
tag(9713,'nhfo').
tag('nhfo','nhfo',9713).
wd(9714,'hvolpsins').
tag(9714,'nkeeg').
```

part of the analysis.

```
tag('nkeeg','nkeeg',9714).
wd(9715,',').
tag(9715,',').
tag(',',',',9715).
wd(9716,'gleymdi').
tag(9716,'sfg3eþ').
tag('sfg3eþ','sfg3eþþ',9716).
wd(9717,'stund').
tag(9717,'nveo').
tag('nveo','nveþ',9717).
wd(9718,'og').
tag(9718,'c').
tag('c','c',9718).
wd(9719,'stað').
tag(9719,'nkeþ').
tag('nkeþ','nkeþ',9719).
wd(9720,',').
tag(9720,',').
tag(',',',',9720).
wd(9721,'ekkert').
tag(9721,'fohen').
tag('fohen','fohen',9721).
wd(9722,'var').
tag(9722,'sfg3eþ').
tag('sfg3eþ','sfg3eþ',9722).
wd(9723,'til').
tag(9723,'ae').
tag('ae','aa',9723).
wd(9724,'nema').
tag(9724,'c').
tag('c','c',9724).
wd(9725,'þessi').
tag(9725,'faven').
tag('faven','fahfn',9725).
wd(9726,'brúnu').
tag(9726,'lhfþvf').
tag('lhfþvf','lhfnvf',9726).
wd(9727,',').
tag(9727,',').
tag(',',',',9727).
wd(9728,'saklausu').
tag(9728,'lhfnvf').
tag('lhfnvf','lhfnvf',9728).
wd(9729,'augu').
tag(9729,'nhfo').
tag('nhfo','nhfn',9729).
```

By comparing these two samples, you can see that the blank spaces have been removed from the IFD tags, and the punctuation marks have been reintroduced into the files, as they are necessary for the tagger.[5]

3.2 The Annotator

For the purpose of this exercise a mock annotator was necessary, in order to introduce the necessary errors for training the μ-TBL tagger. In order not to make life too difficult, this was done by treating the IFD corpus as a closed one, simply changing all ambiguous tags to the most frequent tag in the whole IFD corpus for the appropriate word form. I am sorry to say that I do not have the figures of how many ambiguous forms there were in our corpus. Information on the error rate in our material I will leave for my colleagues to tell you about.

We are perfectly aware that we will need a proper annotator in the future, and the vocabulary of the IFD is much to small to serve as more than an indication of the direction to take. There are only 31,876 lexemes in the IFD, appearing in 59,343 word forms. This served us well in the present project, as we only had to cope with a subset of the IFD anyway. A much better source for an Icelandic annotator is a project on a full form morphological database at the Institute of Lexicography, with inflections of 100,000 lexemes, combined

[5] The big shock was finding discrepancies between the two versions of texts in the corpus and in the Institute's text files, which had to be used to reintroduce the punctuation. This is probably due to the diligence of the editors of the IFD who did check the printed (and corrected) versions of the texts.

to access to the Institute's archives of well over 700,000 lexemes, and lists of approx. 500,000 word forms from over 25 million running words in the Institute's text collections. This material will be available to us, if and when we need it. A word guesser is of course also on our wish list, but we have not started on that project yet.[6]

4 Modifying the Tagsets

As the tagset in the IFD is very large, it is tempting to experiment with simplified tagsets to see the effects on the tagger. I made four versions of the tagsets, and these are briefly described below. As the trial runs with the μ-TBL tagger were being conducted by a group of people in various locations, and perhaps in less than optimally controlled fashion due to the pressure of time, these versions are not necessarily exactly the same as the ones my colleagues used in their experiments, as they made some further modifications by themselves. However, these versions do show the basic lines we were contemplating in regard to the complexity of the tagsets.

- **Version 1.2:** 660 tags appear in our corpus (excluding punctuation tags); the complete number of tags for this version of the IFD is not available.
  This version contains the data used in the prototype of the IFD, including case assignment of verbs. The format of the data was changed, and minimal corrections were made, such as filling in a few 'empty' tags.
  As the tagset used for the annotator in our experiment was the one in version 2 of the corpus (no case features on the verbs), running the μ-TBL on this data is perhaps not optional. Using a tagset as complicated as this might be worthwhile with a very large corpus, as it would give important information on case assignment, which might be of use in disambiguation. It might therefore be worth while to try a version as complex as this, as Icelandic word order is fairly free, if one could thereby take one step towards a correct analysis of the syntax. In a small corpus, the scarcity of data would probably make a tagset of this complexity more of a hindrance than help. The annotator would also make proportionally more mistakes, as the ambiguity in the tags for individual words would increase, as the case assignment of verbs is lexeme based.

- **Version 2:** 488 tags appear in our corpus, out of the 621 tags in the full tagset for the IFD (excluding punctuation tags).
  The case assignment features were removed from the tags for verbs. This is the version used in the printed form of the IFD, and therefore also the one used for our mock annotator.

[6] This is where my own research project touches on this experiment, i.e. in the analysis of compounds. The rules of compounding in Icelandic are both extremely productive and very complex; using the μ-TBL tagger for that analysis would be of great interest to me.

The analysis obtained by this version is the one most in line with the traditional syntactic analysis of Icelandic.

- **Version 3:** 198 tags appear in our corpus (excluding punctuation tags); the corresponding number of tags for the whole IFD is not available.
  Gender was removed from the tagset, as applicable (in nouns, adjectives, pronouns, the definite article, numerals, and past participles). The classification of pronouns was also removed.
  The idea behind this simplification was that some generalizations might be captured in this way, such as the fact that a verb will of course assign case irrespective of gender (and number). Thus the verb *sjá* 'to see' assigns the accusative case, irrespective of gender in these sentences:

  Ég sá manninn 'I saw the man' (acc.masc.)
  Ég sá konuna 'I saw the woman' (acc.fem.)
  Ég sá barnið 'I saw the child' (acc.neut.)

  Being able to trace the case assignment irrespective of gender (and number) might be useful when dealing with verbs for which the data is scarce, which is of course not the case with the verb *sjá*. The danger is, though, that any simplification of the tagset will cause a loss of important data. A tagset leaving out gender, such as the one in version 3, would for example not capture facts on gender agreement between adjective and noun, e.g. $góði_{masc.}$ $maðurinn_{masc.}$ 'the good man', $góða_{fem.}$ $konan_{fem.}$ 'the good woman'.

- **Version 9:** 10 tags: Word class only.
  This version was included in order to see the difference in the running of the tagger; it is hard to see that it would be very useful in real life.

More work is needed on the tagsets, especially when we come to dealing with a larger part of the IFD data. A balance has to be found between the complexity of the tagsets, the time it takes to run the system, and our end result. In the end, a great deal hinges on the purpose of the exercise, i.e., the use of the tagged corpus we are aiming for. I, as a lexicographer, am not greatly worried at having to wait overnight for results, as long as these results are fairly accurate. My colleagues may have other needs.

Bibliography

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

Frequency Dictionary of Icelandic. English translation of preface and chapter introductions. 1993. Rannsóknar- og fræðslurit 4. Orðabók Háskólans, Reykjavík.

Friðrik Magnússon. 1988. Hvað er títt? *Orð & tunga*1:1–49.

Garside, R., G. Leech, and G. Sampson. 1987. *The computational analysis of English: A corpus-based approach.* Longman, London.

Jurafsky, Daniel, & James H. Martin. 2000. *Speech and Language Processing.* Prentice-Hall, New Jersey.

Jörgen Pind, Friðrik Magnússon & Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík. [Referred to as the Icelandic Frequency Dictionary (IFD), or Frequency Dictionary of Icelandic.]

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