The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation

Eiríkur Rögnvaldsson

University of Iceland
eirikur@hi.is

Abstract

We describe the Corpus of Spoken Icelandic (ÍS-TAL) which is made up of 15 hours of spontaneous naturally occurring conversations, 31 conversations in all. The corpus comprises 184,080 tokens, 14,297 types and 9,221 lemmas. It has been transcribed using standard orthography. We present a list of the 30 most common lemmas in the corpus and compare it to a list of the most frequent lemmas in the written language, concluding that the differences between the two lists are smaller than expected. We have tagged the corpus morphologically with a statistical tagger that had been trained on written texts. The results are much better than we expected, and the tagging accuracy is as least as high as for the written texts. The final part of the paper is a report on a work in progress. We have been experimenting with converting the morphological tagging into a shallow syntactic markup by applying a few simple hand-written rules. Even though the analysis we get by using this procedure is bound to be incomplete and contain several errors, we conclude that the results are promising and we can use this method to build a simple yet useful treebank with minimal effort.

1. Introduction

Studies of regional differences in pronunciation have a long tradition in Iceland, but apart from that, research on the spoken language has been very little, and no corpus of spoken Icelandic has been available to researchers. This led seven researchers from three academic institutions to embark on the task of building a corpus of spoken Icelandic. These researchers have different background and different interests, comprising fields such as phonetics, phonology, morphology, syntax, lexicography, sociolinguistics, discourse studies, conversational analysis, language acquisition, corpus linguistics, and psychology.

In this paper, I will first briefly describe the spoken language corpus in section 2. In section 3, I report on our experiments with tagging the corpus morphologically by using a statistical tagger that had been trained on written Icelandic texts, and present some frequency figures from the corpus and compare them to frequency figures for written Icelandic texts. Finally, section 4 describes preliminary attempts at converting this morphological tagging into shallow syntactic parsing by using only minimal effort, thus making a simple low-level treebank.

2. The corpus

The Corpus of Spoken Icelandic (ÍS-TAL) is the first and only of its kind in Iceland (see http://www.hi.is/~eirikur/istal). It is a collaborative project between participants from the Iceland University of Education (Kennaraháskóli Íslands), the University of Iceland (Háskóli Íslands), and the Institute of Lexicography (Orðabók Háskólans). The project leader is Þórunn Blöndal, Assistant Professor at the Iceland University of Education. The project has received generous grants from the Icelandic Research Council.

The project started in 1999, and the recordings were made in 2000. The corpus contains approximately 15 hours of spontaneous naturally occurring conversations, 31 different conversations in all. This material has been transcribed using standard Icelandic orthography. Overlapping, interruption and latching is shown, and several types of comments have been inserted. In designing the corpus and the transcription system, we looked at descriptions of several spoken language corpora, especially the Swedish Spoken Language Corpus in Göteborg (Allwood 1999, see also http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3) and the British National Corpus (http://www.natcorp.ox.ac.uk/), but neither of those has been closely followed.

A short sample from the transcribed corpus is shown in (1) below.

(1) **Transcription sample**

B: svo þarftu líka að skrifa undir að þú sért samþykk <A>þessu ((hlær))</A>
A: að ég sé samþykk að þú notir samtalið mitt</B> jájá <B>(x)</B>
B: ég á</A> ekkert von á því að við tölum um eitthvað sem að
A: neinei ekki svona neitt sérstakt=
B: =ekki háalvarlegt alla vega=
A: =ekki háalvarleg mál
B: nei
A: en hvað <TS>segirðu</TS>
B: bara allt=
A: =það er svo hryllilega langt síðan að <B>ég</B> hef séð þig <TS>hefurðu verið eitthvað í leikfiminni</TS>
B: já
B: voðalega lítið=
A: =já <B>ég hef</B> svo lítið séð þig ég hef nefnilega verið á öllum mögulegum tímum
B: ég fór
B: já ég hef mjög lítið verið núna í einar tvær þrjár vikur ég hef bara ekki mátt vera að því
A: nei ((raddir í fjarska))
B: svo átti ég nú ekki <Á>kort</Á> í eina viku eða tíu daga
A: já
B: en ég á nú orðið kort ég keypti mér sko gatakort <HM>tuttuguogfjögra</HM> tíma

(( ))	comment
(x)	incomprehensible
=	latching
<A>, <B>	overlapping
<TS>	interrogative intonation
<Á>	emphasis
<HM>	high speech rate

The transcriptions have revealed that the corpus is made up of 184,080 running words, and 14,297 different word-forms. Thus, the corpus is obviously rather small, so this must be regarded as a pilot project. However, the corpus has already proved its usefulness in teaching, especially in courses on discourse analysis. Pilot studies of the conversational structure of the material have been presented publicly (Blöndal 2003, 2005).

3. Morphological tagging and lemmatization of the corpus

From the beginning, we planned to make a detailed morphological and syntactic description of the corpus. However, the fundings we had only allowed us to make the recordings and transcriptions. It was only after we had trained three publicly available taggers on Icelandic written texts that we found a way to advance our work. In that project (Helgadóttir 2005), the taggers were trained on a 500,000-word corpus of written Icelandic from five different genres (the source files for *Íslensk orðtíðnibók* (Icelandic Frequency Dictionary), Pind et al. 1991).

Due to the inflectional character of Icelandic, the tagset we used is rather large (for a Western European language at least), containing 639 different tags in all. Each noun belongs to one of three genders, inflects for two numbers and four cases, and can appear with or without a suffixed article. This means that nouns can have 3×2×4×2=48 different tags; verbs can have 106 different tags; adjectives can have 120 different tags; etc.

On the first pass, we got 90.36% accuracy for the written texts using the TnT tagger (Brants 2000), but we were able to reach 93.65% by simplifying the tags a bit, using a backup lexicon, and comparing the outcome of TnT with the outcome of two other taggers and developing methods for selecting the most likely analysis (cf. Helgadóttir 2005).

When we had finished training the TnT tagger on Icelandic written texts, we got the idea to try the tagger on the spoken language corpus. We expected to get considerably worse results from that experiment. First, because the conversations that make up the spoken language corpus are radically different from the texts that the tagger had been trained on, and our previous experiments had shown much worse results when the tagger was applied to different genres than it was trained on (see Helgadóttir 2005). Second, the spoken language corpus contains a lot of incomplete sentences, repetitions, speech errors, and all kinds of inconsistencies, which we expected to hamper the performance of the tagger.

We first had to tokenize the text and strip of all comments and indications of speaker, overlapping, interruption etc. Thus, the input to the tagger was only raw continuous text. When the sample in (1) above was fed into the tagger, it looked like (2).

(2) **Spoken language input to the TnT tagger**

svo
þarftu
líka
að
skrifa
undir
að
þú
sért
samþykk
þessu
.
að
ég
sé
samþykk
að
þú
notir
samtalið
mitt
jájá
.
ég
á
ekkert
von
á
því
að
við
tölum
um
eitthvað
sem
að
.
neinei
ekki
svona
neitt
sérstakt
.
ekki
háalvarlegt
alla
vega
.
ekki
háalvarleg
mál
.
nei
.
en
hvað
segirðu
.
bara
allt
.
það
er
svo
hryllilega
langt
síðan
að
ég
hef
séð
þig
hefurðu
verið
eitthvað
í
leikfiminni
.
já
.
voðalega
lítið
.
já
ég
hef
svo
lítið
séð
þig
ég
hef
nefnilega
verið
á
öllum
mögulegum
tímum
.
ég
fór
.
já
ég
hef
mjög
lítið
verið
núna
í
einar
tvær
þrjár
vikur
ég
hef
bara
ekki
mátt
vera
að
því
.
nei
.
svo
átti
ég
nú
ekki
kort
í
eina
viku
eða
tíu
daga
.
já
.
en
ég
á
nú
orðið
kort
ég
keypti
mér
sko
gatakort
tuttuguog-
fjögra
tíma

The results of the tagging were a pleasant surprise. On the first pass, prior to any simplification of tags etc., the accuracy for the spoken language corpus was around 92.5%, which means that it was considerably higher than for the written texts that the tagger was trained on.

This was quite the opposite of our initial expectations. However, on a closer look, this should not be very surprising. Even though spoken language differs radically from written language in many respects, many of the differences are really favorable for the tagger. We can mention two obvious features.

First, the conversations that make up the spoken language corpus are for the most part about daily life and not about any technical or idiosyncratic subjects. This means that they do not contain a lot of words that are unknown to the tagger. The unknown words rate is considerably lower in the spoken language corpus than in the written language corpus (4.89% compared to 6.84, cf. Helgadóttir 2005).

Second, even though the sentences in the spoken language are often incomplete, they are usually short and relatively simple. They do not contain many complex phrases and they do not exhibit many cases of long distance dependencies etc. Written Icelandic exhibits considerable freedom in word order and this freedom often makes it difficult for statistical PoS taggers to analyze sentences since the analysis of a certain word is often dependent on another word, which is far away in the sentence. This is usually not the case in the spoken language.

Nivre and Grönquist (2001) describe an experiment where they trained a statistical tagger on data from written Swedish and then adapted it to the characteristics of spoken language data. They found that with relatively small modifications, they could reach tagging accuracy for the spoken language on a similar level with the accuracy reported for written texts. We have not retrained the TnT tagger on corrected spoken language data, but obviously, that would be a logical step to take.

After tagging the corpus with TnT, we applied the CST lemmatizer (Jongejan and Haltrup 2005) to the tagged output. The result was quite good, and only minimal manual corrections were needed to get the correct lemmatization. It turns out that the corpus is composed of some 9.221 different lemmas. We have previously published some preliminary frequency studies of the spoken language corpus (Svavarsdóttir 2003). However, the morphological tagging and the subsequent lemmatization has enabled us to carry out much more extensive studies of the vocabulary of the corpus and compare it with the vocabulary of Icelandic written texts according to Pind et al. (1991).

A comparison of the 30 most frequent lemmas in spoken and written Icelandic is shown in (3). We had anticipated that the difference between spoken and written language would be substantial, but in fact, it is perhaps not as great as one would have thought. Most of the words that are frequent in the spoken language are also among the most frequent words in the written language. The exceptions are a few words that we know to be typical of spoken language – words like *já* ‘yes’ and *nei* ‘no’, and a few words that are not easy to translate – *sko* (‘you see’, ‘well’, ‘uh’, etc.), *bara* (‘just’), *hérna* (‘here’, ‘well’, ‘uh’, etc.).

(3) **The 30 most frequent lemmas in spoken and written Icelandic**

Spoken language				Written language			
#	**Lemma**	**Gloss**	**wr.l.**	**#**	**Lemma**	**Gloss**	**sp.l.**
1	vera	‘be’	2	1	og	‘and’	6
2	að	‘that’	3	2	vera	‘be’	1
3	það	‘it, there’	6	3	að	‘that’	2
4	já	‘yes’	179	4	í	‘in’	7
5	ég	‘I’	8	5	á	‘on’	12
6	og	‘and’	1	6	það	‘it, there’	3
7	í	‘in’	4	7	hann	‘he’	9
8	þessi	‘this’	15	8	ég	‘I’	5
9	hann	‘he’	7	9	sem	‘who, that’	18
10	sko	‘you see’ etc.	-	10	hafa	‘have’	23
11	bara	‘just’	-	11	hún	‘she’	16
12	á	‘on’	5	12	en	‘but’	21
13	ekki	‘not’	13	13	ekki	‘not’	13
14	þú	‘you’	39	14	til	‘to’	39
15	svona	‘such’	176	15	þessi	‘this’	8
16	hún	‘she’	11	16	við	‘at’	41
17	einhver	‘someone’	59	17	um	‘about’	40
18	sem	‘who, that’	9	18	með	‘with’	25
19	nei	‘no’	-	19	af	‘of’	36
20	svo	‘so’	28	20	að	‘to’	31
21	en	‘but’	12	21	sig	‘x-self’	56
22	þá	‘then’	44	22	koma	‘come’	34
23	hafa	‘have’	10	23	verða	‘become’	42
24	fara	‘go’	29	24	fyrir	‘for’	45
25	með	‘with’	18	25	segja	‘say’	28
26	vita	‘know’	67	26	allur	‘all’	30
27	hérna	‘here’	-	27	svo	‘so’	65
28	segja	‘say’	25	28	sá	‘that’	20
29	nú	‘now’	49	29	fara	‘go’	24
30	allur	‘all’	26	30	þegar	‘when’	47

("wr.l." refers to the frequency rank of the word in the Icelandic Frequency Dictionary. "sp.l." refers to the frequency rank of the word in the Icelandic spoken language corpus.)

If we look at the top thirty list for the written language, it turns out that all the words on that list, except two, are among the fifty most frequent words in the spoken language. Thus, even though there are a few words that are typical of spoken language, there seem to be no high frequency words that are mainly used in the written language but are rare in the spoken language.

4. Syntactic analysis

When we were making the original plans for the spoken language corpus, we intended to annotate it syntactically. However, it soon became clear to us that this was a far too ambitious task. We did not have the necessary resources. Thus, the plans of syntactically annotating the corpus were abandoned. However, having received the positive results of the PoS tagging, we realized that we had in fact taken the first step towards a syntactic annotation of the spoken language corpus. The fact is that even though the tagging made with our tagset is of course morphological in nature, it carries a substantial amount of syntactic information also. 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 verb *vera* ‘to be’ which is in turn preceded by another noun in the nominative, in which case the second noun is a predicate. A noun in the accusative or dative case is usually an object, unless it is immediately preceded by a preposition. True, Icelandic also has oblique subjects, but they can easily be identified from their accompanying verbs. Within noun phrases, for instance, nouns, pronouns, adjectives, and numerals all agree in number, gender, and case – and crucially, they show overt signs that make it possible for the tagger to assign the correct tags to them. To illustrate this, we show below one sentence from the tagged corpus, where each individual letter in the tag represents a single morphological category or the value of a single category.

(4) A tagged sentence from the spoken language corpus

Word	**Tag**	**Gloss**	**Content of the tag**
Helgi	nken-m	‘Helgi’	noun (n) – masc (k) – sing (e) – nom (n) – proper noun (m)
minn	feken	‘mine’	pronoun (f) – possessive (e) – masc (k) – sing (e) – nom (n)
farðu	sbg2en	‘go-you’	verb (s) – imp (b) – active (g) – 2nd pers (2) – sing (e) – pres (n)
niður	a	‘down’	adverbial (a)
og	c	‘and’	conjunction (c)
skoðaðu	sbg2en	‘check-you’	verb (s) – imp (b) – active (g) – 2nd pers (2) – sing (e) – pres (n)
nýja	lkeovf	‘new’	adjective (l) – masc (k) – sing (e) – acc (o) – weak (v) – positive (f)
tölvuleikinn	nkeog	‘computer game’	noun (n) – masc (k) – sing (e) – acc (o) – article (g)
þinn	fekeo	‘your’	pronoun (f) – possessive (e) – masc (k) – sing (e) – acc (o)

‘Dear Helgi, please go downstairs and try your new computer game.’

As can be seen, all the necessary information for analyzing the syntactic structure of this sentence is present here. The nouns, pronouns and adjectives are marked for case, which makes it possible to deduce their syntactic functions. Two or more adjacent words, which receive the same tags for gender, number, and case, like *nýja tölvuleikinn þinn* ‘your new computer game’, can safely be assumed to belong to the same phrase. Since this phrase is in the accusative case, it can also safely be assumed to be an object of the preceding verb *skoðaðu* ‘check-you’. And so on.

One might think that tagging written texts using the same tagset would give us comparable syntactic information. To a certain extent, that is of course the case – but only to a certain extent. Sentences from written texts are usually longer and more complicated, not least with respect to word order. Many types of syntactic movement and other types of long distance dependencies are much more common in the written language than they are in the spoken language, so that words that go together syntactically are not necessarily adjacent in written texts. Therefore, it is often much more difficult to read the syntactic structure of sentences off the morphology and the word order in written texts.

We have written a number of simple scripts to convert some of the morphological information into syntactic information. The challenge is twofold: First, to assign the correct syntactic category to each word, and then to group the words into larger units, syntactic phrases. These tasks are of course related. We started by automatically assigning syntactic features (represented by uppercase letters) to the words in the corpus according to their morphological tags. Thus, we added the feature ‘S’ (= subject) to every inflected word (nouns, pronouns, adjectives, and numerals) bearing the ‘n’ tag (for nominative). We also added the feature ‘O’ (= object) to every inflected word bearing the tags ‘o’ (for accusative) or ‘þ’ (for dative); and so on.

The next step was to write rules for correcting these tags according to the syntactic environment, and to write rules that try to assign some hierarchical structure to the sentences by grouping words together. We are currently working on these rules and still think we can make them more efficient. However, I want to stress that we are not going to spend much time on this, and the rules will not be many – 10-15 at most. Some of the most important rules are the following:

(5) **A few contextual rules for changing syntactic tags**

- Change the tag of a word in the nominative from ‘S’ to ‘P’ (for predicate) if the word is immediately preceded by the verb *vera* ‘be’, which is in turn immediately preceded by a word in the nominative.
- Change the tag of a word from ‘O’ to ‘P’ if it is immediately preceded by a word tagged as a preposition.
- Change the tag of a word from ‘O’ to ‘S’ if it immediately precedes or follows a verb from the list of verbs taking oblique subjects.
- Group all adjacent words bearing the same syntactic feature (‘S’, ‘O’, etc.) into a single unit.
- Group words separated by the conjunctions *og* ‘and’ and *eða* ‘or’ into a single unit if they bear the same syntactic feature.
- Change the tag of verbs ending in –*u* and having ‘b’ as the second character in the tag into ‘VS’ (these are imperative verb forms that have a cliticized subject).

When applied to the sentence in (4) above, the rules give us the following output:

(6) **Output of the rules**

S	Helgi	nken-m
+	minn	feken
VS	farðu	sbg2en
A	niður	a
C	og	c
VS	skoðaðu	sbg2en
O	nýja	lkeovf
+	tölvuleikinn	nkeog
+	þinn	fekeo

We have tried the rules that we have written so far on the corpus and they seem to work pretty well. It would be meaningless to calculate exact percentages at this point, since we still haven’t finished writing the rules, but it appears that we will get a reasonably good analysis out of this.

When we have finished writing the rules, we will apply them to the corpus and correct, say, half of the output manually. After that, we plan to feed the corrected results into TnT to train the tagger on the syntactic tags.

We will then test the tagger on the remaining part of the corpus and see how it succeeds in the syntactic annotation.

One may of course object to our approach by pointing out that by using this procedure, we are treating the spoken language just as if it were written texts. We are disregarding all features that characterize spoken language as opposed to written language. We are not taking any notice of discourse functions etc. To that I can only say that our purpose by making this experiment was only to try a certain method of syntactic annotation. I am the syntactician in the research group, and this reflects my interests. The group also comprises specialists in discourse studies who are responsible for that kind of studies, and who have in fact already used the corpus in various research projects (Blöndal 2003, 2005, etc.). However, it must of course be admitted that it is not at all clear that it is correct or has any relevance to assign syntactic structure to all sorts of sentence fragments that are typical for spoken language.

5. Conclusion

In this paper, I have given an overview of the Corpus of Spoken Icelandic and shown that it is possible to use a statistical tagger that has been trained on written Icelandic texts (Helgadóttir 2005) to successfully tag this corpus. This is in accordance with the results reported for Swedish by Nivre and Grönqvist (2001). Furthermore, I have argued that it is possible to convert the morpological information contained in the tags into syntactic annotation which can be considerably enhanced by adding a few simple hand-written rules.

Finally, we may ask: Can we use the term ***treebank*** for the syntactically annotated corpus resulting from applying this procedure? That is a matter of definition, I guess. It is clear that we will not get a consistent and elaborated treebank out of this. The result is bound to be a rather flat structure with lots of errors and inconsistencies. But it is far better than having no syntactic annotation at all. This might be called a low budget treebank. Those who have been to Iceland will realize that the trees in this treebank are very much in line with the trees you find in nature there – low, crooked and stunted, but nevertheless nice to have.

References

Allwood, J. (1999) *The Swedish Spoken Language Corpus at Göteborg University*, in *Fonetik 99: Proceedings from the Twelfth Swedish Phonetics Conference*, pp. 5–9. (Gothenburg Papers in Theoretical Linguistics 81)

Blöndal, Þ. (2003) *Repeats in Icelandic Conversations – some preliminary observations. NordicResearch on Relation Between Utterances*; in Henrichsen, P. J. (ed.) *Proceedings of the NordTalk Symposium at CMOL (CBS) December 2002*, pp. 98–109. (Copenhagen Working Papers in LSP)

Blöndal, Þ. (2005) *Feedback in Conversational Storytelling. Feedback in Spoken Interaction*, in *Nordtalk 2003.* Gothenburg Papers in Theoretical Linguistics, pp. 1–17

Brants, T. (2000) *TnT - A Statistical Part-of-Speech Tagger. Version 2.2.* `http://www.coli.uni-sb.de/~thorsten/tnt/`

Helgadóttir, S. (2005) *Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic*; in Holmboe, H. (ed.) *Nordisk Sprogteknologi 2004*, pp. 257–265, Museum Tusculanums Forlag

Jongejan, B.; D. Haltrup (2005) *The CST Lemmatiser;* `http://www.cst.dk/download/cstlemma/current/doc/cstlemma.pdf`

Nivre, J.; L. Grönqvist (2001) *Tagging a Corpus of Spoken Swedish*; International Journal of Corpus Linguistics 6, 47–78

Pind, J. (ed.); F. Magnússon; S. Briem (1991) *Íslensk orðtíðnibók*; Orðabók Háskólans

Svavarsdóttir, Á. (2003) *Ordbogen og den daglige tale. Om den islandske talesprogsbank (ISTAL) og dens betydning i ordbogsredaktion*; in Hansen, Z. S.; A. Johansen, (eds.). *Nordiske studier i leksikografi 6*, pp. 43–48