The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation

Special session on treebanks for spoken language and discourse

NoDaLiDa, Joensuu, May 19, 2005

Eiríkur Rögnvaldsson

Aim of the project

- This is a report of a work in progress
- The aim of this work is to find out to what extent it is possible to use a statistical PoS tagger, supplemented by a few hand-written rules, to assign a shallow syntactic structure to a corpus of spoken Icelandic

The Corpus of Spoken Icelandic

- A collaborative project between researchers from three academic institutions
  - Iceland University of Education
  - University of Iceland
  - Institute of Lexicography
- Project leader:
  - Þórunn Blöndal, MSc
    - Assistant Professor, Iceland University of Education

Transcription

- Project period: 1999-2002
  - Recordings finished: 2000
  - Transcriptions finished: 2001
- The transcription uses standard orthography
  - but several nonlinguistic features are shown
    - overlapping
    - interruption
    - latching
    - etc.

The size of the corpus

- 15 hours of spontaneous conversations
- 31 different conversations
- 185,000 running words
- 14,000 wordforms
- 9,000 lexemes

Frequency in spoken language

- 1 vera ‘be’ 2
- 2 að ‘that’ 3
- 3 það ‘it, there’ 6
- 4 já ‘yes’ 179
- 5 ég ‘I’ 8
- 6 og ‘and’ 1
- 7 í ‘in’ 4
- 8 þessi ‘this’ 15
- 9 hann ‘he’ 7
- 10 sko -
- 11 bara -
- 12 á ‘on’ 5
- 13 ekki ‘not’ 13
- 14 þú ‘you’ 39
- 15 svona ‘such’ 176
- 16 hún ‘she’ 11
- 17 einhver ‘someone’ 59
- 18 sem ‘that’ 9
- 19 nei ‘no’ -
- 20 svo ‘so’ 28
- 21 en ‘but’ 12
- 22 þá ‘then’ 44
- 23 hafa ‘have’ 10
- 24 fara ‘go’ 29
- 25 með ‘with’ 18
- 26 vita ‘know’ 67
- 27 hérna -
- 28 segja ‘say’ 25
- 29 nú ‘now’ 49
- 30 allur ‘all’ 26

Frequency in written language

- 1 og ‘and’ 6
- 2 vera ‘be’ 1
- 3 að ‘that’ 2
- 4 í ‘in’ 7
- 5 á ‘on’ 12
- 6 það ‘it, there’ 3
- 7 hann ‘he’ 9
- 8 ég ‘I’ 5
- 9 sem ‘that’ 18
- 10 hafa ‘have’ 23
- 11 hún ‘she’ 16
- 12 en ‘but’ 21
- 13 ekki ‘not’ 13
- 14 til ‘to’ 39
- 15 þessi ‘this’ 8
- 16 við ‘at’ 41
- 17 um ‘about’ 40
- 18 með ‘with’ 25
- 19 af ‘of’ 36
- 20 að ‘to’ 31
- 21 sig ‘x-self’ 56
- 22 koma ‘come’ 34
- 23 verða ‘become’ 42
- 24 fyrir ‘for’ 45
- 25 segja ‘say’ 28
- 26 allur ‘all’ 30
- 27 svo ‘so’ 65
- 28 sá ‘that’ 20
- 29 fara ‘go’ 24
- 30 þegar ‘when’ 47

Tagging written texts

- Three different taggers have been trained
  - on written texts from five genres
- The tagset contains more than 600 tags
  - nouns can have 48 different tags
  - verbs can have 106 different tags
  - adjectives can have 120 different tags
- TnT gave the best results
  - 90.36% correct tags on the first pass

Tagging spoken language

- We expected to get worse results from the tagging of the spoken language because:
  - the spoken texts differ radically from the written texts
    - the taggers had proven to be sensitive to types of text
  - the spoken language corpus contains lot of “irregularities”
    - incomplete sentences, repetitions, speech errors, inconsistencies

A 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

Input to the tagger

svo	að	á	.	ekki	það	verið
þarftu	ég	ekkert	neinei	háalvarleg	er	eitthvað
líka	sé	von	ekki	mál	svo	í
að	samþykk	á	svona	.	hryllilega	leikfiminni
skrifa	að	því	neitt	nei	langt	.
undir	þú	að	sérstakt	en	síðan	já
að	notir	við	.	hvað	að	.
þú	samtalið	tölum	ekki	segirðu	ég	
sért	mitt	um	háalvarlegt	.	hef	
samþykk	jájá	eitthvað	alla	bara	séð	
þessu	.	sem	vega	allt	þig	
.	ég	að	.	.	hefurðu	

The results

- The results were in fact quite good
  - around 92.5% correct tags on the first pass
- The reasons appear to be at least two:
  - the conversations concern daily life
    - unknown words were only 4.89%
  - the sentences are usually short and simple
    - do not contain many complex phrases or long distance dependencies

A tagging sample

Helgi	nken-m	‘Helgi’	n=noun k=masc e=sing n=nom m=proper noun
minn	feken	‘mine’	f=pronoun e=possessive k=masc e=sing n=nom
farðu	sbg2en	‘go-you’	s=verb b=imp 2=2nd pers e=sing n=pres
niður	a	‘down’	a=adv
og	c	‘and’	c=conj
skoðaðu	sbg2en	‘check-you’	s=verb b=imp 2=2nd pers e=sing n=pres
nýja	lkeovf	‘new’	l=adj k=masc e=sing o=acc v=weak f=positive
tölvuleikinn	nkeog	‘computer game’	n=noun k=masc e=sing o=acc g=article
þinn	fekeo	‘your’	f=pronoun e=possessive k=masc e=sing o=acc

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

Examples of rules

- 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 is immediately preceded by 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.

Output of the rules

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

Next steps

- After finishing the rules, we will
  – apply them to the whole corpus
  – manually correct a part of the corpus
  – train TnT on the corrected results
  – apply TnT to the the rest of the corpus
  – see how it succeeds in syntactic annotation
- Can we call the resulting corpus a treebank?