Coping with variation in the Icelandic Diachronic Treebank

Eiríkur Rögnvaldsson
Anton Karl Ingason   Einar Freyr Sigurðsson
eirikur,antoni,einasig@hi.is

University of Iceland

RILiVS Workshop, September 18th 2009
University of Oslo
Outline

1 Introduction
   - The project
   - Contents of the treebank

2 Building trees
   - Open source policy
   - IceNLP: Tagging, Shallow parsing, Lemmatizing
   - CorpusSearch: Rule-based parsing

3 Diachronic issues of Icelandic syntax
   - Case study I: The New Passive
   - Case study II: Quirky subjects

4 Conclusion
The project

- **Viable Language Technology beyond English** – Icelandic as a test case
- A three year project funded by a grant of excellence from the Icelandic Research Fund (RANNÍS)
- **Objective:** Make it realistic to develop three particular types of LT modules with limited resources without sacrificing the quality of the work
- A parsed corpus is one of those three types of resources
- http://iceblark.wordpress.com/
Contents of the treebank

- Modern Icelandic written texts
  - of different genres
- Modern Icelandic spoken language
  - Spontaneous conversations
- Old Icelandic narrative texts
  - Icelandic Sagas, Heimskringla, Sturlunga saga, etc.
- Selected texts from the 16th - 20th centuries
Homework

- Are we ready to share our tools and data with others even if they might do brilliant things that we never thought of (Krauwer yesterday)?
- Absolutely
- (And we will try to use those brilliant results of others to do something even more brilliant)
Open source policy

- IceNLP (pos-tagger, shallow parser, lemmatizer, segmentizer, tokenizer, data format management etc.) was recently made open source (LGPL)
  - http://sourceforge.net/projects/icenlp/
  - http://nlp.ru.is/

- We use the output of IceNLP as an input to rule-based CorpusSearch (MPL) parsing
  - http://corpussearch.sourceforge.net/

- We run everything on Linux
  - still, Java, platform independent

- The data we create will be mostly free and open too
  - although this may not be possible for all the modern texts
The sentence in (1) is from Sturlunga saga.

(1) Rannveig og Hergerður voru dætur þeirra
    Rannveig and Hergerður were daughters their
‘Rannveig and Hergerður were their daughters’
**Step 1 - Part-of-Speech tagging (IceTagger)**

**Input:**

Rannveig og Hergerður voru dætur þeirra.

**Output:**

Rannveig nven-m
og c
Hergerður nven-m
voru sfg3fp
dætur nvfn
þeirra fphfe
. .
Step II - Shallow parsing (IceParser)

Input:

Rannveig nven-m
go c
Hergerður nven-m
voru sfg3fp
doetur nvfn
þeirra fphfe
.
.

Output:

{*SUBJ> [NPs [NP Rannveig nven-m NP] [CP og c CP]
[NP Hergerður nven-m NP] NPs] *SUBJ>}
[VPb voru sfg3fp VPb] {*COMP< [NP doetur nvfn NP] *COMP<}
{*QUAL [NP þeirra fphfe NP] *QUAL} . .
Step III - Lemmatize (Lemmald)

... and translate tagset and convert to labeled bracketing (Formald)

Input:

{ *SUBJ> [NPs [NP Rannveig nven-m NP] [CP og c CP] 
[NP Hergerður nven-m NP] NPs] *SUBJ>} 
[VPb voru sfg3fp VPb] { *COMP< [NP dætur nvfn NP] *COMP<} 
{ *QUAL [NP þeirra fphfe NP] *QUAL} . .

Output:

( (IP-MAT (NP-SBJ (NP (N-FSNIP Rannveig-rannveig) ) ) 
(CP (C og-og) ) (NP (N-FSNIP Hergerður-hergerður) ) ) ) 
(VPb (V-IA3PD voru-vera) ) 
(NP-COMP (N-FPNIC dætur-dóttir) ) 
(NP-QUAL (PRO-PNPG þeirra-það) ) (; .-.) ) )
Structure now looks like this

(lemmas and the final period omitted from picture)

IP-MAT

NP-SBJ

NP

CP

N-FSNIP

Rannveig

og

Hergerður

VPb

V-IA3PD

voru

N-FPNIC

dætur

NP-COMP

NP-QUAL

PRO-PNPG

þeirra
Step IV - CorpusSearch revision queries

- Minor revisions of labeling conventions
- Build more structure (by referring to structure)
  - CorpusSearch is designed for linguists
  - precedes, iPrecedes, dominates, iDominates, hasSister, cCommands, ...
- Correct mistakes based on structure
  - IP should dominate only one subject
- Some of this functionality may (and should) end up in other modules
- Example revisions on following slides
Finite verb should be the head of IP-MAT

IP-MAT

NP-SBJ
  NP
    NP-FSNIP
    Rannveig
  CP
    C
    og
  NP-FSNIP
    Hergerður

VPb
  V-IA3PD
    voru

NP-PRD
  N-FPNIC
    dætur

NP-QUAL
  PRO-PNPG
    þeirra
Finite verb should be the head of IP-MAT
query: (IP-MAT iDoms {1}[1]VP*)
    AND ([1]VP* iDoms finiteVerb)

delete_node{1}:

**finiteVerb** is defined as any tag that matches: V-I*|V-S*|V-M* (I=indicative, S=subjunctive, M=imperative)
Move NP-QUAL under immediately preceding NP
Move NP-QUAL under immediately preceding NP
The actual revision query

query: ({1}[1]NP* hasSister {2}[2]NP-QUAL) 

extend_span{1, 2}:
Step V - Manual correction using CorpusDraw

(this tree doesn’t actually need manual corrections)
Variation as a problem for Generative Syntax

- Real world data is not as clear cut as one might expect if one believes in Principles and Parameters
- We aim to test recent theories on language acquisition, variation and productivity against our diachronic data (e.g. [Yang2009])
  - Is the successful acquisition of a UG parameter value based on the ratio of unambiguous evidence of the relevant pattern? (token frequency)
  - Does the acquisition of other productive patterns rest on a rule having a relatively low rate of exceptions? (type frequency)
- Treebank statistics! (Quirky Subjects, New Passive, etc.)
The New Passive

Canonical passive:

(2) Það var barinn lítill
    it was beaten.M.SG.NOM little.M.SG.NOM
    strákur
    boy.M.SG.NOM
    ‘A little boy was beaten’

The New Passive:

(3) Það var barið lítinn strák
    it was beaten.N.SG little.ACC boy.ACC
The New Passive with accusative objects:

- Contains *vera* ‘be’ or *verða* ‘will, become’
- The finite verb is 3sg
- Contains a past participle
- Contains an object
- The object is in accusative case
- The past participle c-commands the object
The New Passive

node: IP*

query: (IP* iDoms [1]V-IA3SD )
AND ([1]V-IA3SD iDoms [2]*-vera)
AND (IP* doms VPP)
AND (VPP iDoms [4]V-DANSN)
AND (IP* doms [3]NP-OBJ)
AND ([2]*-vera precedes [3]NP-OBJ)
AND ([3]NP-OBJ iDoms N-..A..)
The New Passive

- [Eythórsson2008] suggests a parametric variation: case feature [± accusative] assignment
- Increased frequency of the expletive það ‘it, there’ in the first half of the 19th century ([Hróarsdóttir1998], [Rögnvaldsson2002])
- Why does a child reanalyse passive data in the 20th century (but not the 19th ...)?
- With other words: what are the origins of the New Passive?
The New Passive

How did it emerge?

Some proposals:

- Reanalysis of the passive of intransitive verbs; the first step after that being among inherently reflexive verbs ([Maling and Sigurjónsdóttir 2002])
- “The New Passive is [...] closely related to the highly frequent and productive impersonal P[repositional]-passive” ([Sigurðsson 2009]; cf. also Kjartansson 1991)
- Lack of Definiteness Effect ([Guðmundsdóttir 2000])
- “Weakening” (or non-agreement, cf. DAT-NOM verbs) of the past participle ([Árnadóttir and Sigurðsson 2008])

We need (more) empirical evidence!
Quirky subjects

- Found in Modern Icelandic but not in Old Icelandic?
- Word order: an indication of the subject
- Statistics should show different results for the 12th than the 20th century
Quirky subjects

[Rögnvaldsson1996]; Gísla saga Súrssonar:

(4) Hún síndist honum ríða grám hesti
   she.NOM seemed him.DAT ride grey horse
   ‘It looked like to him she was riding a grey horse’

(5) Honum síndist hún ríða grám hesti
    him.DAT seemed she.NOM ride grey horse
The Icelandic treebank will contain a lot of variation, both synchronic and diachronic.

In order to study this variation thoroughly, we need a properly annotated phrase structure.

We build the treebank by combining and re-using existing open source tools.

A sophisticated query language and search software enables us to deal with the variation.

The treebank will open up new possibilities in the study of Icelandic syntax.
References I


References II


Eiríkur Rögnvaldsson. 2002. ÞaÐ í fornu máli — og síðar. [ÞaÐ (’it, there’) in old icelandic — and later.]. Íslenskt mál, (24).

References IV

Charles Yang.
2009.
Three factors in language variation.
To appear in Lingua.