Extending the Comparative Dimension of Diachronic Syntax: A Parsed Corpus of Icelandic from the 12th Century to Modern Times

Joel Wallenberg
Einar Freyr Sigurðsson       Anton Karl Ingason
jwall,einasig,antoni@hi.is

University of Iceland

May 11th 2010
UMass Amherst
Outline

1. Introduction
   - Why build a treebank?

2. The project
   - How does one build a treebank?
   - Current state of the project

3. Preliminary results
   - Passives and by-phrases
   - Postnominal possessive pronouns
   - From Object-Verb syntax to Verb-Object syntax
   - Relationship between topicalization and V2

4. Conclusion
What is a treebank?

- A corpus of selected texts, annotated for Part of Speech (PoS) and syntactic structure.
  - Samples of comparable size are taken from texts written in a variety of genres, over a number of centuries.

- In our case, we annotate full phrase structure (as opposed to dependency parsing) with labeled bracketing.
  e.g. (IP-MAT ... )
Why build a treebank?

- An important resource for various NLP applications:
  - To generate conference papers?
  - Write, train, and test PoS-taggers for Icelandic and other morphologically rich languages.
  - Write, train, and test statistical phrase structure parsers.
  - Write, train, and test other technologies (e.g. lemmatizers, machine translation systems).
- ... but that is not our main motivation
Why build a treebank?

- To study (morpho-)syntax in a rigorous and quantitative manner:
  - Hypotheses about statistical tendencies in morphosyntax can be developed and tested.
  - Results can be easily replicated by different researchers.
  - Absolutely necessary for studying syntactic variation and syntactic change.
- To study whatever else you want. For example, we would love for philologists, lexicographers, and students of literature to use it as well. In general, more users improves the corpus over time.
Research questions

- How has the syntax of Icelandic changed (12th century - ModIce)?
- Can we investigate changes like this in detail?
- We would like to work towards developing a field of comparative quantitative diachronic syntax.
- For example: Are English and Icelandic extremely different or nearly identical?
Old English and Old Icelandic were extremely similar in both syntax and morphology. Modern English syntax and modern Icelandic syntax appear to be very similar at first glance. On the other hand, they differ strikingly in subdomains of the syntax, and nearly all of the morphology. So what happened?

English and Icelandic went through very similar syntactic changes (e.g., OV to VO), and ended up with very similar results. But the changes and results were not identical, so we have a chance to learn something quite precise about the set of morphosyntactic parameters that can differ from language to language.
What we want to do

- Parse at least 1 million words of texts that span the history of Icelandic.
  (ca. 100 000 - 150 000 words per century, from 12th through early 20th.)
- We are currently working on:
  - The First Grammatical Treatise (12th century)
  - Íslensk hómilíubók (12th century homilies)
  - Piltur og stúlka (19th century novel)
- The following slides illustrate the annotation process.
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 I - Part-of-Speech tagging (IceTagger) (Loftsson 2008)

**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) (Loftsson & Rögnvaldsson 2007)

Input:

Rannveig nven-m
og c
Hergerður nven-m
voru sfg3fþ
dærur nvfn
þeirra fphfe
...

Output:

{*SUBJ> [NPs [NP Rannveig nven-m NP] [CP og c CP]
[NP Hergerður nven-m NP] NPs] *SUBJ>}
[VPb voru sfg3fþ VPb] {*COMP< [NP dætur nvfn NP] *COMP<}
{*QUAL [NP þeirra fphfe NP] *QUAL} . .
... and translate tagset and convert to labeled bracketing (Formald)

Input:

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

Output:

( (IP-MAT (NP-SBJ (NPR-N Rannveig-rannveig))
(CONJ og-og) (NPR-N Hergerður-hergerður) )
(VPB (BEPI voru-vera) )
(NP-PRD (NS-N dætur-dóttir) )
(NP-POS (PRO-G þeirra-það) ) (; .-. ) ) )
CorpusSearch (Randall 2005) revision queries

- CorpusSearch, basic functions: recursively searches syntactic trees with labeled bracketing
- CorpusSearch, automated corpus revision: searches and manipulates trees
- CorpusSearch, CorpusDraw: manual correction

query: (IP-MAT iDoms {1}[1]VP*)
AND ([1]VP* iDoms finiteVerb)

delete_node{1}:
CorpusSearch revision queries
CS revision queries

```
IP-MAT
  NP-SBJ
    NPR-N  CONJ  NPR-N
    Rannveig  og  Hergerður
  BEPI
    voru
    dætur
  NP-PRD
    NS-N
    dætur
  NP-POS
    PRO-G
    þeirra
```
CS revision queries

IP-MAT

NP-SBJ

NPR-N

Rannveig

CONJ

og

NPR-N

Hergerður

BEPI

voru

NP-PRD

NS-N

dætur

NP-POS

PRO-G

þeirra
Manual correction using CorpusDraw
Current state of the project

- We have parsed 13184 words (goal is 1 million)
  - 5319 come from Íslensk hómilíubók (12th century)
  - 7865 come from Piltur og stúlka (19th century)

- Why the relatively slow start?
  - Time-consuming training process.
  - The challenge to get us to agree with each other on a parse for each sentence.
  - A lot of documentation work:
    - [www.linguist.is/wiki](http://www.linguist.is/wiki)
  - Extensive reviews: 3 rounds of correction per chunk of text.

- We aim to reach full production speed in July, leaving 1-1.5 years to complete the project.

- ... But we have some preliminary results
Passives and by-phrases

- *By-phrases* show interesting differences in usage in modern English and Icelandic.
  - The house was destroyed by fire. (Sigurjónsdóttir & Maling 2001)
- The current version of the Icelandic corpus has 60 passives, of which 4 have *by*-phrases
  - Íslensk hómilíubók: 3 in 25 IP-MAT/IP-SUB (12%)
  - Piltur og stúlka: 1 in 35 (2.86%)
- The difference in frequency of *by*-phrases between OldIce and ModIce is not statistically significant (Fisher; p-value = 0.2984) – but the corpus will get larger
Passives and by-phrases

(2) Þeir menn eru og mikils virtir af Guði, er
Those men are also much respected by God, who
þatningar taka af vondum mönnum fyr hans sakar.
torture take from bad men for his sake.
‘Those men who undergo torture from bad men for his sake
are also much respected by God ...’
(Ísl. hómilíubók)

(3) og var honum ritað af hreppstjórunum
and was him-DAT written by the sheriffs
‘and he was written to by the sheriffs’ (Piltur og stúlka)
Passives and by-phrases

- Comparison with Penn parsed corpora of historical English (PPCME2, PPCEME):
  - Middle English: 794 by-phrases in 13830 passives (5.74%)
  - Early Modern English: 2498 by-phrases in 20103 passives (12.43%)
- The difference is statistically significant ($\chi$-squared; p-value < 2.2e-16)
- Conclusion: the usage of by-phrases increased in frequency in English, and may have stayed the same in the history of Icelandic.
Postnominal possessive pronouns in English

- Almost nonexistent throughout the history of English (Old English = 9/19552, 0.046%)
- Sporadic examples of insignificant frequency from all periods, but by Middle English they only appear in vocative contexts

(4) Ond his healle & þa cynelican geseto his sceawedon.
And his hall & the royal seat his showed
‘And showed his hall and his throne’ (OldEng)

(5) ‘Broþir myn Edmund’, he seith
Brother my Edmund, he said
‘My brother Edmund, he said’ (MidEng)

(6) O scholler myne, happy art thou ...
Oh scholar my, happy are you ...
‘Oh my scholar, you are happy …’ (Early ModEng)
In Icelandic, however, we get pre- and postnominal possessive pronouns at all stages of the language.

(7) En þó er hennar hreinlífi dýrlegra en
But still is her clean-living more wonderful than annarra
others-GEN
‘But still, her clean living is more wonderful than that of others’
(Hómilíubók)

(8) ... en ég vona, að guð minn góður heyri mína bæn
... but I hope, that God my good hears my prayer
‘... but I hope, that my good God hears my prayer’
(Piltur)
Postnominal possessive pronouns in Icelandic

(9) ... þótt lítið sé í mein gjört barni þeirra
... although little is in harm done child theirs
‘... although not much is done to harm their child’
(Hómilíubók)

(10) Fagur ertu, dalur fósturjarðar minnar
Beautiful are you, valley motherland my
‘You are beautiful, valley of my motherland’
(Piltur)
What is the diachronic development of this pattern?
The corpus has 135 possessive pronouns.
- Íslensk hómilíubók: 62 postnominal out of 81 (76.54%)
- Piltur og stúlka: 48 postnominal out of 54 (88.89%)

The difference is not statistically significant (\(\chi^2\)-squared; p-value = 0.1134)

A real change? A sociolinguistic contrast?

A lot more data from different genres is needed.
Very well-studied fact: the syntax of English changed from OV to VO, resulting in nearly uniform VO word-order by 1500.

This change has been studied in quantitative detail in studies such as Kroch & Taylor (2000), showing that the change gradually spread across geographical dialects.

(11) hie wolden hit admodeliche þolien.
    They would it humbly suffer.
    ‘They would humbly suffer it.’

(Trinity Homilies; CMTRINIT,185.2576; Southeast Midlands, date: 1225)
We know that Old Icelandic showed some OV and some VO clauses, but we do not have detailed and reliable quantitative information about the different possible OV and VO patterns. Modern Icelandic is VO, but we do not have very much information at all about the time periods between the oldest and most recent stages of Icelandic syntax.
OV example with NP object

(12) að vérr skulum hans erindi reka við yður
    that we shall his errand run with you
    ‘that we shall run his errand with you’

( (IP-SUB-2 (NP-SBJ (PRO-N vér-ég))
    (MDPI skulum-skulu)
    (NP-OB1 (NP-POS (PRO-G hans-hann))
       (N-A erindi-erindi))
    (VB reka-reka)
    (PP (P við-við)
       (NP (PRO-A yður-þú)))))
)
(13) Jóan kunni og vildi velja sér inn besta hlut.
    John could and wanted choose himself the best thing.
    ‘John could and wanted to choose the best thing for himself.’
**OV example with pronoun object**

(14) Kristur sjálfur hefir oss til boðið  
    Christ himself has us to offered  
    ‘which Christ himself has offered us’

\[
\begin{align*}
\text{(IP-SUB (NP-OB1 *T*-4)} & \\
\text{(NP-SBJ (NPR-N Kristur-kristur)} & \\
\text{(NP-PRN (PRO-N sjálfur-sjálfur)))} & \\
\text{(HVPI hefir-hafa)} & \\
\text{(NP-OB2 (PRO-D oss-ég))} & \\
\text{(RP til-til)} & \\
\text{(VBN boðið-bjóða))} & \\
\end{align*}
\]
From OV to VO

- We know that Old Scandinavian influenced English profoundly, especially in phonology and vocabulary, but we do not know exactly how much English syntax was influenced (though we have lots of guesses).
- A major hurdle is not being able to compare earlier stages of English and Scandinavian in a precise, quantitative way...until now!
  - Sort of...keep in mind that the numbers are very small right now.
  - Hómilíubók:
    - Clauses with auxiliary verb, main verb, and NP object = 12
    - Clauses with auxiliary verb, main verb, and pronoun object = 7
NP Objects in two Early Middle English dialects
NP Objects, adding Íslensk hómilíubók

The diagram shows the distribution of sentence types in three regions: Homiliubok, Southeast Midlands, and West Midlands. The chart represents the proportion of clauses that are either matrix (light blue) or subordinate (dark blue) sentences.

- **Homiliubok**: A significant number of clauses are subordinate, with a smaller portion being matrix sentences.
- **Southeast Midlands**: The distribution is more balanced, with a roughly equal number of matrix and subordinate clauses.
- **West Midlands**: Here, matrix clauses dominate over subordinate clauses.

The type of sentence is indicated by the color: light blue for VO (matrix) and dark blue for OV (subordinate).
We know that the estimate for OV/VO is not entirely accurate for English, because of NP extraposition (Heavy NP Shift)

(15) Þa æfter þam þe hi gewyld hæfdon eall
Then after that-DAT that they controlled have all
heora feonda land
their enemies’ land
‘After that time when they conquered all of their enemies’ land...’

(Saint Eustace and his Companions, date: c. 11th century, in the YCOE, Taylor, Warner, Pintzuk & Beths 2003)
Note that this still occurs in modern English, though there is no OV anymore:

(16) "Nothing changes tragedy into comedy like gayness. It’s what we call in the entertainment world the GAY EX MACHINA."

(from the “That’s Gay” feature on the TV program infoMania)

We do not have a good way of estimating the frequency of NP extraposition in Middle English. We could use pronoun objects, except that they have a different, very complicated syntax. But not in Icelandic...
NP objects and pronoun objects
Direct object topicalization in English (Speyer 2008)
Object topicalization in modern V2 languages

- Dutch, 17.5% (3283/18795) (Bouma 2008)
- German, 14% (rough estimate from the TIGER corpus, Anthony Kroch p.c.)
- Old (12th c.) French, 15% (Kroch & Santorini 2009)
- Icelandic, 13.6% (9/65) (Us right now)
Conclusion

- With a very small amount of data, we have been able to find highly suggestive results.
- However, none of the results are conclusive, and very few approach statistical significance. There is simply not enough data.
- We need more data, and given a little more time, our team will produce it. Don’t worry.
- Eventually, we also need more users for the corpus. This is not only how the results develop, but how the corpus is maintained.