12th century homilies – the cutting edge in parsing

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

University of Iceland

Íslensk máltækni, April 15th 2010
Reykjavík University
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

4. Conclusion
What is a treebank?

- A corpus of selected texts (by time period, genre, etc.), annotated for Part of Speech (PoS) and syntactic structure.
- In our case, we annotate full phrase structure (as opposed to dependency parsing) with labelled bracketing.
  e.g. (IP-MAT ... )
Why build a treebank?

- To develop resources for Natural Language Processing:
  - Write, train, and test PoS-taggers.
  - Write, train, and test phrase structure parsers.
  - Write, train, and test other technologies (e.g. lemmatizers, machine translation systems).

- To study (morpho-)syntax:
  - Quantitatively.
  - Results can be easily replicated.
  - 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.
Research questions

- How has the syntax of Icelandic changed (12th century - ModIce)?
- Can we investigate the change in detail?
- Are English and Icelandic extremely different or nearly identical?
  - ModEng and ModIce don’t appear to be very similar but OldEng and OldIce appear extremely similar, does that mean that English changed and Icelandic didn’t?
  - Well, no. Clearly Icelandic has changed. In fact, 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 some very fundamental things.
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:
  - Íslensk hómilíubók (12th century)
  - Piltur og stúlka (19th century)
- 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 sfg3fp
dætur 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 dætur nvfn NP] *COMP<}
{*QUAL [NP þeirra fphfe NP] *QUAL} . .
Step III - Lemmatize (Lemmald) (Ingason et al. 2008)

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

**Input:**

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

**Output:**

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

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

delete_node{1}:
CorpusSearch revision queries

IP-MAT

NP-SBJ
  NPR-N Rannveig
  CONJ og
  NPR-N Hergerður

VPB BEPI
  NPR-N voru
  NS-N dætur

NP-PRD NP-POS
  PRO-G þeirra
CS revision queries

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

IP-MAT

NP-SBJ  BEPI  NP-PRD

NPR-N  CONJ  NPR-N  voru  NS-N  NP-POS

Rannveig  og  Hergerður  dætur  PRO-G  þeirra
Manual correction using CorpusDraw
Current state of the project

- We have parsed 8329 words (goal is 1 million)
  - 3220 come from Íslensk hómilíubók (12th century)
  - 5109 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
  - 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 have an important role in studying passives.
- The 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

(2) Þeir menn eru og mikils virtir af Guði, er píningar taka af vondum mönnum fyr hans sakar.  
(Ísl. hómilíubók)

(3) og var honum ritað af hreppstjórunum  
(Piltur og stúlka)
Passives and by-phrases

- Comparison with Penn parsed corpora of historical English
- 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 (Fisher; p-value < 2.2e-16)
Postnominal possessive pronouns in English

- Almost nonexistent throughout the history of English
- Sporadic examples of insignificant frequency from all periods

(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)
Postnominal possessive pronouns in Icelandic

- In Icelandic, however, we get pre- and postnominal possessive pronouns at all stages of the language

(7) En þó er hennar hreinlífi dýrlegra en annarra ... (Hómilíubók)
(8) ... en ég vona, að guð minn góður heyri mína bæn ... (Piltur)
(9) ... þótt lítið sé í mein gjört barni þeirra þeirra ... (Hómilíubók)
(10) Fagur ertu, dalur fósturjarðar minnar ... (Piltur)
Rise of postnominal possessive pronouns in Icelandic?

- What is the diachronic development of this pattern?
- The corpus has 135 possessive pronouns.
  - Íslensk hómilíubók: 62 postnominal out of 81 (76.54%)
  - Píltur 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)
From OV to VO

- We know that Old Icelandic showed some OV and some VO clauses, but we do not have quantitative information.

- 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 vör skulum hans erindi reka við yður

( (IP-SUB-2 (NP-SBJ (PRO-N vör-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.

( (IP-MAT (NP-SBJ (N-N Jóan-jóan)))
  (MDDI (MDDI kunni-kunna))
  (CONJ og-og)
  (MDDI vildi-vilja))
 (VB velja-velja)
 (NP-OB2 (PRO-D sér-sig))
 (NP-OB1 (D-A inn-hinn))
  (ADJS-A besta-góður)
   (N-A hlut-hlutur))
  (. .-.))
)

VO example with NP and pronoun object
(14) Kristur sjálfur hefir oss til boðið

( (IP-SUB (NP-OB1 *T*-4)
    (NP-SBJ (NPR-N Kristur-kristur)
      (NP-PRN (PRO-N sjálfur-sjálfur)))
    (HVPI hefir-hafa)
    (NP-OB2 (PRO-D oss-ég))
    (RP til-til)
    (VBN boðið-bjóða))
)
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

<table>
<thead>
<tr>
<th>Clause</th>
<th>Southeast Midlands</th>
<th>West Midlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>matrix</td>
<td>VO: 0.2</td>
<td>VO: 0.2</td>
</tr>
<tr>
<td></td>
<td>OV: 0.8</td>
<td>OV: 0.8</td>
</tr>
<tr>
<td>subordinate</td>
<td>VO: 0.2</td>
<td>VO: 0.2</td>
</tr>
<tr>
<td></td>
<td>OV: 0.8</td>
<td>OV: 0.8</td>
</tr>
</tbody>
</table>
NP Objects, adding Íslensk Hómilíubók
From OV to VO, NP extraposition

- We know that the estimate for OV/VO is not entirely accurate for English, because of NP extraposition:

  (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

<table>
<thead>
<tr>
<th>Text and Type of Object</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Midlands - All NPs</td>
<td>1.0</td>
</tr>
<tr>
<td>Southeast Midlands - All NPs</td>
<td>0.6</td>
</tr>
<tr>
<td>Homiliubok - All NPs</td>
<td>0.6</td>
</tr>
<tr>
<td>Homiliubok Pronoun Objects</td>
<td>1.0</td>
</tr>
</tbody>
</table>
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.