The Icelandic Diachronic Treebank

Einar Freyr Sigurðsson

Department of Icelandic, University of Iceland, Reykjavík

UNIVERSITY OF ICELAND

About the treebank

- A diachronic phrase structure treebank – from Old to Modern Icelandic
  - ≈ 200.000 words per century
- In co-operation with the University of Pennsylvania – compatible with the Penn Parsed Corpora of Historical English
  - Anthony Kroch and Beatrice Santorini
- A part of a bigger project
  - IceBLARK – http://iceblark.wordpress.com
- The treebank team:
  - Eiríkur Rögnvaldsson (eirikur@hi.is), project leader,
  - Anton Karl Ingason (anton.karl.ingason@gmail.com),
  - Einar Freyr Sigurðsson (einasig@hi.is),
  - Joel Wallenberg (joelcw@babel.ling.upenn.edu)
- An ongoing work
- Website and documentation: http://linguist.is/wiki/

Why do we need a diachronic treebank?

- With a diachronic corpus it is possible to find how languages change over time
- We are able to ...
  - ... compare (relative) frequencies of different constructions
  - ... do quantitative historical linguistics that can be replicated
- A diachronic treebank is needed to understand ...
  - ... the **spread** of a change over time, such as OV to VO
  - ... the **actuation** of a syntactic change, e.g. of the New Passive in Icelandic

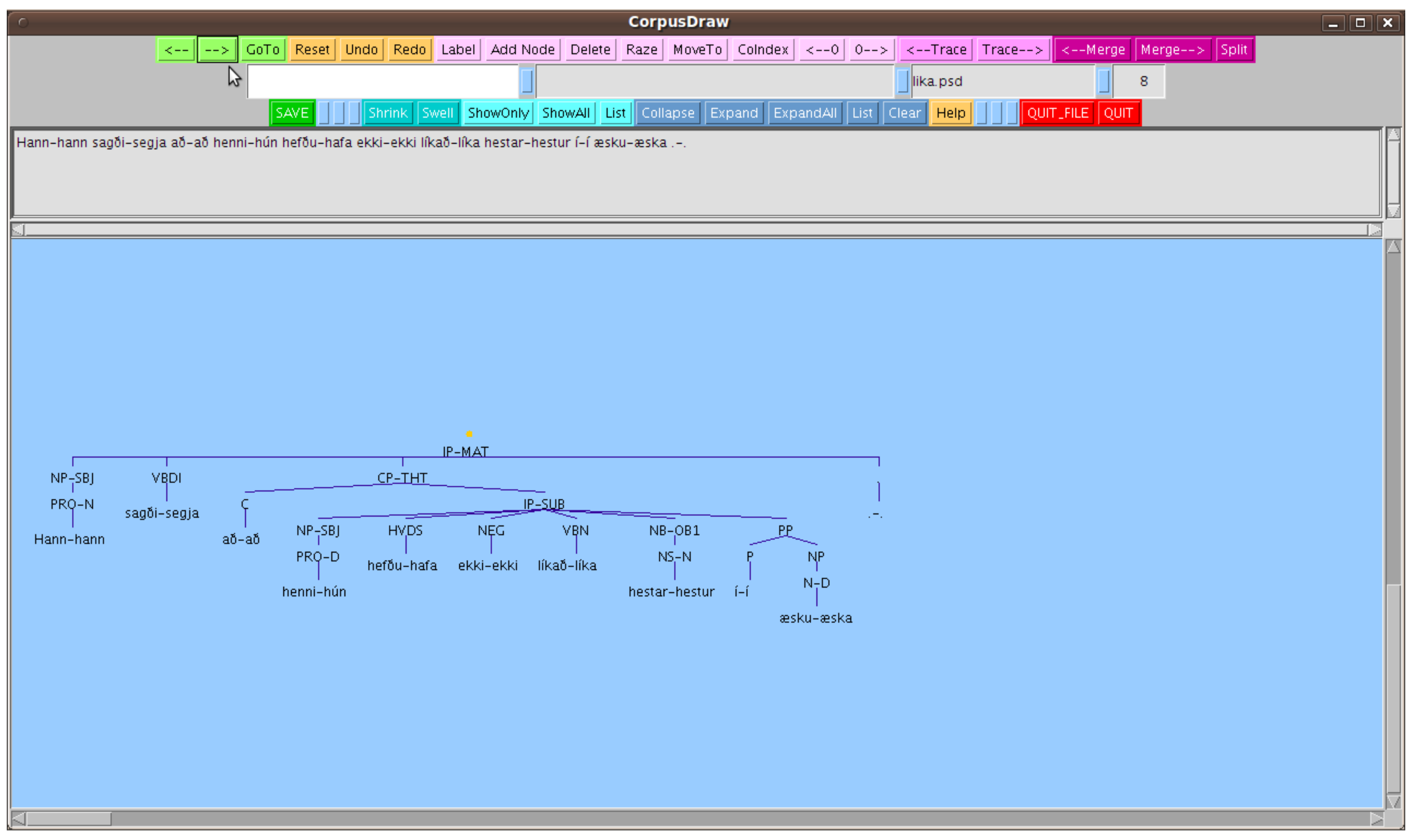
Tools

- The main tools we use are:
  - IceNLP (pos-tagger, shallow parser, lemmatizer, etc.)
    - http://sourceforge.net/projects/icenlp/
  - CorpusDraw and CorpusSearch
    - Developed by Beth Randall
- IceNLP transforms sentences from plain text and gives:
  - Lemmas
  - Part of speech tags
  - Basic phrase structure

CorpusDraw

- Then CorpusDraw is used to manually correct the output from IceNLP (e.g. a sentence like (1)):

(1) Hann sagði að **henni** **hefðu** **ekki líkað hestar** í æsku
he said that she.DAT had.PL not liked horse.PL.NOM in youth
‘He said that she didn’t like horses when she was young’



Hann-hann sagði-segja að-að henni-hún hefðu-hafa ekki-ekki líkað-líka hestar-hestur í-í æsku-æska .-.

Compatibility with the Penn Parsed Corpora

- The raw data we get from CorpusDraw looks like this:

```
( (IP-MAT (NP-SBJ (PRO-N Hann-hann))
          (VBDI sagði-segja)
          (CP-THT (C að-að)
                  (IP-SUB (NP-SBJ (PRO-D henni-hún))
                          (HVDS hefðu-hafa)
                          (NEG ekki-ekki)
                          (VBN líkað-líka)
                          (NB-OB1 (NS-N hestar-hestur))
                          (PP (P í-í)
                              (NP (N-D æsku-æska)))))
          (. .-.)))
```

- This is compatible with the Penn Parsed Corpora (Kroch, Santorini and Delfs 2004):

```
( (IP-MAT (NP-SBJ (PRO I))
          (VBP believe)
          (CP-THT (C 0)
                  (IP-SUB (NP-SBJ (PRO I))
                          (MD shall)
                          (VB like)
                          (NP-OB1 (PRO$ your) (N cook))
                          (ADVP (ADV very) (ADV well))))
          (. .))   (ID FHATTON-E3-H,I,148.34))
```

CorpusSearch

- Has the agreement of DAT-NOM verbs with plural nominative objects, cf. (1), changed over the ages?
  - Both agreement and non-agreement found in Old and Modern Icelandic
  - DAT-NOM > DAT-ACC marginal in MIce, not found in OIce
    - What is the relative frequency of agr. vs. non-agr.?
- We use CorpusSearch to search for certain patterns or phrase structure, such as DAT-NOM verbs.
  - A set of DAT-NOM verbs defined: *áskotnast* ‘acquire’, *líka* ‘like’, etc.
  - Here the lemmas come into play: líkað-**líka**
- CorpusSearch uses syntactic terms as *(immediately) dominates, C-Commands, has sister* ...

CorpusSearch query

- A query that finds agr. as well as non-agr. with DAT-NOM verbs:

```
node:  IP*
query: (HV[PD][IS]|MD[PD][IS]|VB[PD][IS]
       hasSister NP-OB1)
       AND (VB* iDominates *-áskotnast|*-líka)
       AND (NP-OB1 iDominates NS-N)
```

i) Searches within every IP (IP-MAT, IP-SUB ...)

ii) HV|MD|VB (the have-verb or a modal or a main verb) which is either in present (P) or past (D) tense, and either in indicative (I) or subjunctive (S) mood, is sister to NP-OB1 (an object)

iii) The main verb (whether it’s finite or not) immediately dominates the lemma *áskotnast* or *líka*

iv) NB-OB1 immediately dominates NS-N (a plural nominative noun)

Current status

- Annotation of the first 200.000 words (in 19th century texts) is underway
- Documentation and guidelines are written as the project evolves (http://linguist.is/wiki/)
- The goal is to finish the annotation process in the next 1–1$^1/_2$ years

This work was supported by the Icelandic Research Fund and the University of Iceland Research Fund

References

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/