Creating a dual-purpose treebank

Eiríkur Rögnvaldsson, Anton Karl Ingason
Einar Freyr Sigurðsson & Joel Wallenberg
www.linguist.is/icelandic_treebank

University of Iceland, University of Pennsylvania, Newcastle University

ACRH
January 5th, 2012
Heidelberg University

Overview

1. Introduction
2. Data
   - The diachronic dimension
   - Text selection
   - Text quality
3. Methods
   - Text conversion
   - The annotation process
4. License policy
5. Conclusion

The Icelandic Parsed Historical Corpus

Dual-purpose treebank

- Modern Icelandic Language Technology
- Diachronic comparative quantitative syntax

The Icelandic Parsed Historical Corpus

- Diachronic treebank spanning 12th through 21st centuries
- 1 003 532 words, with samples from 61 different texts
- All texts part-of-speech tagged, fully parsed, and lemmatized
- The entire annotation (pos-tags, parse, and lemmas) has been hand-corrected
- (We are now on the second round of correction)
- Texts samples for each century are balanced for genre; primarily narrative and religious texts

Funding

- **RANNÍS, Icelandic Research Fund, grant of excellence:**
  Viable language technology beyond English – Icelandic as a test case
- **U.S. National Science Foundation (NSF):**
  Evolution of Language Systems: a comparative study of grammatical change in Icelandic and English
- **University of Iceland research fund:**
  Historical Icelandic Treebank

Project members and collaborators

PIs:

- Eiríkur Rögnvaldsson (RANNÍS)
- Joel C. Wallenberg (NSF)

Annotators:

- Anton Karl Ingason
- Brynhildur Stefánsdóttir
- Einar Freyr Sigurðsson
- Hulda Óladóttir
- Joel C. Wallenberg

IceNLP:

- Hrafn Loftsson

International collaborators:

- Tony Kroch (UPenn)
- Beatrice Santorini (UPenn)

Typing army:

- Andri Gunnar Hauksson
- Eyrún Lóa Eiríksdóttir
- Guðrún Ingólfsdóttir
- Hulda María Frostadóttir
- Vignir Árnason

The diachronic dimension

Why is Icelandic a good candidate for a project of this kind?

- Continuous supply of texts from at least two distinct genres (narratives, religious texts) from a long period
- Icelandic morphosyntax remains very similar from the 12th century to the present
  - Morphology: Basically identical
  - Syntax: Limited word order changes (some only in a quantitative sense)
- 12th century Icelandic is readable by Modern Icelanders

Text selection

	nar	rel	bio	sci	law	Total
12th	0	40871	0	4439	0	**45310**
13th	93463	21196	0	0	6183	**120842**
14th	77370	21315	0	0	0	**98685**
15th	111560	0	0	0	0	**111560**
16th	35733	60464	0	0	0	**96197**
17th	46281	28134	52997	0	0	**127412**
18th	63322	22963	22099	0	0	**108384**
19th	100362	20370	0	3268	0	**124000**
20th	103921	21234	0	0	0	**125155**
21st	43102	0	0	0	0	**45310**
Total	**675114**	**236547**	**75096**	**7707**	**6183**	**1000647**

Text quality

Common challenges in historical linguistics

- Not all texts are accurately dated
- Spelling and perhaps even word order may in some cases reflect the period of the manuscript rather than the date of composition
- We used accurately dated texts when possible
- For more comprehensive sampling we relaxed this requirement and relied on philological estimates
- Ultimately, each user has to decide how she wants to approach dating issues

Text conversion

- Spelling has been modernized for practical reasons
- Our language processing tools assume Modern Icelandic input
- Matching highly variable spelling in search is complicated
- The corpus comes with information about printed editions and it contains page numbers that can be used to track down examples
- Aligning the corpus with a more detailed representation of the original manuscripts is left for future work

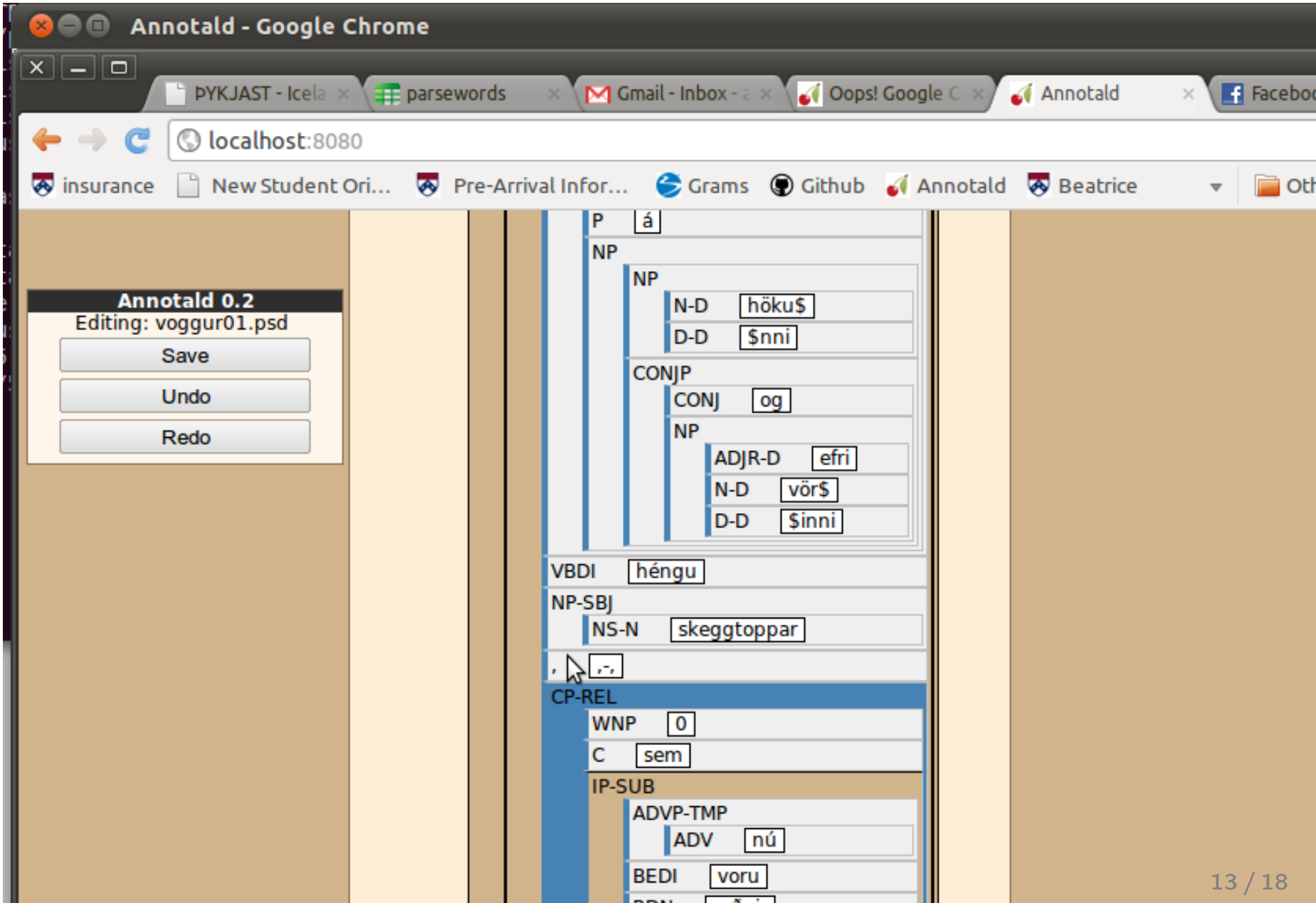
Annotation scheme

```
( (IP-MAT (NP-SBJ (PRO-N Hann-hann))
          (VBDI spurði-spyrja)
          (CP-QUE (WADVP-1 (WADV hvernig-hvernig))
                  (C 0)
                  (IP-SUB (ADVP *T*-1)
                          (NP-SBJ (NPR-D Grími-grímur))
                          (VBDS liði-líða))))
  (ID 1888.GRIMUR.NAR-FIC,.301))
```

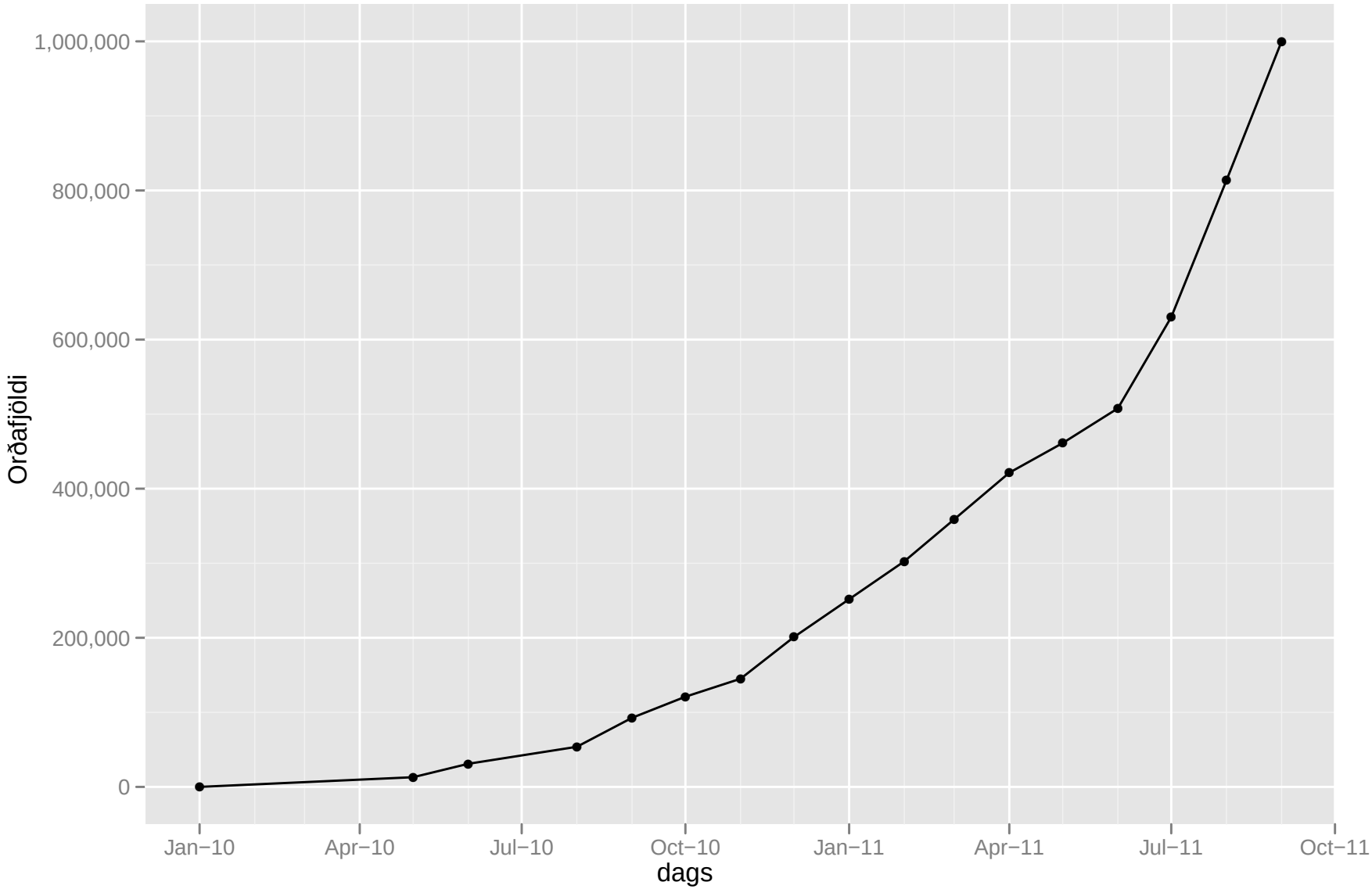
The annotation process

- Conversion to modern spelling
- Manual sentence (tree) boundary annotation
- PoS-tagging, shallow parsing and lemmatization using the IceNLP toolkit (Loftsson 2008; Loftsson and Rögnvaldsson 2007; Ingason et al. 2008)
- Conversion to Penn Treebank format (Python scripts)
- Automatic adjustments to phrase structure (CorpusSearch revision queries)
- Manual phrase structure annotation
- Automatic error checking (CorpusSearch "sanity checks")

Annotald – manual correction (Beck et al. 2011)



Size of IcePaHC over the course of the project



10 basic types of user freedom

- Raw data available can be downloaded for local use (corpus not hidden behind a search interface)
- Comprehensive documentation freely available online
- Available without registration, user identification of some sort, or signing of contracts
- Development process of corpus relies only on free/open source software tools (for transparent replication of annotation process)
- Open development (annotation is carried out in an open online version control repository for transparency regarding the actual steps taken in the development and immediate access to work-in-progress)

10 basic types of user freedom

- Regular scheduled releases of numbered versions during development as well as for more permanent milestone versions so that researchers can always produce replicable results on a recent version of the corpus
- Users can improve the corpus and release modified versions without special permission
- Free of cost to academia
- Free of cost to commercial users
- Corpus released under a standard free license of some sort for straightforward compatibility with other projects (GPL, LGPL, CC, etc.)

The value of use freedom

"as one of the GSoC tasks is to make a CG converter into LT formalisms, Michael was looking for the CG rules for English, but there aren't really rich sets of rules that are not proprietary. This is why it seems to make much more sense to test the conversion on Icelandic."

Marcin Milkowski, Polish Academy of Sciences, 13th July 2011

Conclusion

- As of August 2011, all main goals of the project have been reached
- IcePaHC is currently available for download in labeled bracketing format for anyone who wants to run experiments in statistical parsing, etc.
  - http://linguist.is/icelandic_treebank/Download
- A number of papers on historical syntax have already taken advantage of IcePaHC (including 4 papers at the last DiGS conference)
- Our user freedom policy has encouraged hundreds of downloads of the corpus and we look forward to seeing more researchers apply IcePaHC to diverse problems