The Balanced Tagged Corpus of Icelandic and Other Icelandic Language Technology Resources

Eiríkur Rögnvaldsson, University of Iceland
Sigrún Helgadóttir, Árni Magnússon Institute for Icelandic Studies

Nordic Seminar on CALL and Corpora
Reykjavík, September 24-25, 2010

Icelandic LT in 2000

- Ten years ago, Icelandic language technology was virtually non-existant
- There was a relatively good spell checker, a not-so-good speech synthesizer, and that was all
- There were no programs or even individual courses on language technology or computational linguistics at any Icelandic university
- There was no ongoing research in these areas
- No software companies were working on LT

The LT Program

- A special LT committee was appointed in 1998
  - to investigate the status of language technology in Iceland
  - to come up with proposals for strengthening Icelandic LT
- The committee published its report in 1999
  - containing several proposals for actions
- In 2000, the Icelandic Government launched a special Language Technology Program
  - with the aim of supporting instutions and companies to create basic resources for Icelandic LT work

Products of the LT Program

- A full-form morphological database of Modern Icelandic inflections (Kristín's talk today)
- A balanced morphosyntactically tagged corpus of 25 million words (Sigrún's part of this talk)
- A training model for data-driven POS taggers (*TnT*)
- A new text-to-speech system (*Ragga*)
- An isolated word speech recognizer (*Hjal*)
- An improved spell checker (*Púki*)

IceNLP

- IceTagger, a linguistic rule-based tagger
- IceParser, a shallow parser
  - both written by Hrafn Loftsson
- Lemmald, a lemmatizer
  - written by Anton Karl Ingason
- These programs make up the IceNLP package
  - which is open source, http://icenlp.sourceforge.net/
  - and online, http://nlp.cs.ru.is/IceNLPWeb/icenlp.html