Icelandic Language Resources and Technology: Status and Prospects

Eiríkur Rögnvaldsson, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Matthew Whelpton, Anna Björk Nikulásdóttir, Anton Karl Ingason

Outline of the talk

- We give an overview of Icelandic language technology since its inception ten years ago and describe briefly its main achievements.
- Then we outline the research program of the Icelandic Language Technology community for the next few years, which is being implemented thanks to a large grant which has just been allotted to the program by the Icelandic Research Fund.
- Finally, we discuss the need for Nordic cooperation within Language Technology and put forward some concrete proposals for enhanced cooperation.

Icelandic LT in 1999

- 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

Starfshópur um tungutækni

- In 1998, a special LT committee was appointed
  - by the Minister of Education, Science, and Culture
- Tasks:
  - 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 2001, 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

Proposed actions

- The development of common linguistic resources that can be used by companies as sources of raw material for their products
- Investment in applied research in the field of language technology
- Financial support for companies for the development of language technology products
- Development and upgrading of education and training in language technology and linguistics

Products of the LT Program

- A full-form morphological database of Modern Icelandic inflections
- A balanced morphosyntactically tagged corpus of 25 million
- A training model for data-driven POS taggers
- A text-to-speech system
- A speech recognizer
- An improved spell checker

Tasks of the ICLT

- maintaining an information center for Icelandic language technology by running a website (www.tungutaekni.is
- encouraging cooperation on LT projects between universities, institutions and private companies;
- organizing and coordinating university education in language technology
- taking part in Nordic, European and international cooperation in the field of language technology
- initiating and participating in research projects in language technology
- initiating and participating in commercial projects in language technology
- keeping track on resources and products in the field of language technology
- holding an annual LT conference with the participation of LT researchers, companies and the public
- supporting the growth of Icelandic language technology in all possible ways

- Recent LT projects
  - supported by the Icelandic Research Fund
  - and the Icelandic Technical Development Fund
- IceTagger, a linguistic rule-based tagger
- IceParser, a shallow parser
- Lemmald, a lemmatizer
- A context-sensitive spell checker

A new project

- We have initiated a new LT Project
  - in order to build basic resources for Icelandic LT
  - and develop methods for less-resourced languages
- *Viable Language Technology beyond English*
  - *Icelandic as a test case*
- A three year interdisciplinary project
  - which has received a Grant of Excellence
  - from the Icelandic Research Fund

- Languages other than English face two main problems in LT:
  - They have less resources to develop LT modules
  - They may differ from English in important linguistic ways and therefore the established methods from English LT need adaptation
- It is essential to develop new methods for constructing LT modules in more efficient ways

- Three types of LT modules will be developed within the project:
  - A database of semantic relations
  - A shallow transfer machine translation system
  - A pilot treebank
- These modules are central to current LT work and prerequisites for further research and development in Icelandic LT

Points of emphasis, 1

- Developing methodologies for creating resources for new languages more efficiently, with focus on semi-automatic/ machine assisted resource generation
- An inquiry into linguistic issues that are of little relevance for English LT but crucial for many other languages, with a special focus on general methods to deal with morphological richness and morphological ambiguity
- A case study of Icelandic where we use the tools and methods developed to build a treebank, a database of semantic relations and a machine translation system

Points of emphasis, 2

- Evaluation of the tools and methods developed – focusing on quality of output as well as the output/manpower ratio
- Writing and publishing guidelines for creating similar LT modules for less-resourced and/or morphologically rich languages
- Enhancing research training in the field by giving graduate students the opportunity to work on research projects, as it is vital for the future of Icelandic LT to educate and train young researchers in the field

- The LT Program was very successful
  - LT education has started
  - various resources have been created
  - several R&D projects have been initiated
  - Nordic cooperation has been firmly established
- Icelandic LT is not yet self-sustained
  - now that the LT Program has ended
  - and more funding is needed for R&D

The cost of Icelandic LT

- Estimated cost
  - of making Icelandic LT self-sustained:
- ISK 1,000,000,000 (€ 10,000,000)
  - distributed over 4-5 years
- Total budget of the LT program
  - from 2001-2004:
- ISK 133,000,000
  - 1/8 of the estimated cost

Norræn samvinna

- Nordic Language Technology Research Programme (2001-2004) – ýmis net
- Nordic Graduate School of Language Technology (NGSLT, 2004-2009)
- Northern European Association for Language Technology (NEALT, stofnað 2006)
- Þátttaka í margvíslegum umsóknum
  - sem fæstar hafa hlotið brautargengi

Proposals for cooperation

- A common website
  - containing accessible and standardized information on available language resources and tools for the Nordic languages
- A Nordic Summer School in LT
  - where graduate students and researchers could meet, exchange ideas, attend practical training sessions and pass on technical skills

Summary

- We have demonstrated how joined efforts of the government, research communities, and commercial companies, enhanced by Nordic cooperation, have succeeded in establishing the basis for Icelandic language technology in a relatively short time
- We have outlined the research plan of the Icelandic LT community, which aims at developing low-cost methods for building language resources for less-resourced languages, in addition to contributing to the building of an Icelandic BLARK
- We have discussed some ideas for Nordic cooperation on LT, especially as regards compilation and dissemination of information and on LT teaching