# Icelandic Language Technology Ten Years Later

Eiríkur Rögnvaldsson
University of Iceland

SALTMIL Workshop
LREC, Marrakech,
May 27, 2008


HÁSKÓLI ÍSLANDS

# Icelandic Language Technology in 1998

- Ten years ago, Icelandic LT did not exist
  - and nobody was working on it
- We had
  - a good spell checker
  - a relatively primitive speech synthesizer
- We didn't have
  - university programs or courses in LT
  - academic research on Icelandic LT
  - software companies developing LT products

# The LT committee

- In 1998, a special LT committee was appointed
  - by the Minister of Education, Science, and Culture, Mr. Björn Bjarnason
    - Rögnvaldur Ólafsson (chairman), Eiríkur Rögnvaldsson, Þorgeir Sigurðsson
- Tasks:
  - to investigate the status of language technology in Iceland
  - to come up with proposals for strengthening Icelandic LT

# The LT Program

- 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
- This initiative has led to several projects
  - either directly or indirectly

# 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

# LT education

- An interdisciplinary Master's program
  - started in 2002 at the University of Iceland
  - admitting students with background in either language and linguistics or computer science
- The program was relaunched last fall
  - as a joint program between the University of Iceland (Dept. of Icelandic) and Reykjavik University (Dept. of Computer Science)
  - in cooperation with the NGSLT (Nordic Graduate School of Language Technology)

# Priority tasks in Icelandic LT

- For Icelanders, the main aim must be that it should be possible to use Icelandic, written with the proper characters, in as many contexts as possible in the sphere of computer and communication technology. Naturally, however, they will have to adjust their expectations to practical considerations.
- To make it possible to use Icelandic in all areas, under all circumstances, would be an immense task. Therefore, the main emphasis must be put on those areas that touch on the daily life and work of the general public, or are likely to do so in the near future.

# 1. Software translation

- *The main computer programs on the general market (Windows, Word, Excel, Netscape, Internet Explorer, Eudora,...) should be available in Icelandic*
- Icelandic versions of Windows XP and Vista and Microsoft Office now exist
  - and several other application programs have been translated
- Open-source software is also being translated

# 2. Icelandic characters

- *It should be possible to use the Icelandic non-ASCII characters (áéíóúýðþæöÁÉÍÓÚÝÐÞÆÖ) in all circumstances: in computers, mobile telephones, teletext and other applications used by the public*
- Here the situation has improved considerably
  - due to reduced use of 7 bit character tables
  - however, there still seem to be problems with some mobile phones

# 3. Natural language parsing

- *Work should proceed on the parsing of Ice-landic, with the aim that it should be possible to use computer technology to analyse Icelandic texts grammatically and syntactically*
- The LT program funded three major projects related to this objective:
  - a full-form morphological database
  - a grammatical tagger
  - an HPSG-based syntactic parser

# 3.1 Corpus – 3.2 Lexicon

- *A large computerized text corpus including Icelandic texts of a wide variety of types should be established*
- A 25 million word corpus is now being built
  - shall be finished later this year
- *A grammatically and semantically annotated lexicon should be established*
- No such lexicon exists, nor is being prepared
  - however, several items of raw material can be found

# 4. Authoring tools

- *Good auxiliary programs should be developed for textual work in Icelandic, i.e. for hyphen-ation, spell-checking, grammar correction, etc.*
- Polderland has developed a new spell checker for use with Microsoft Office
  - the *Púki* spell checker has been improved
  - and there exists an open source spell checker
- No grammar or style checker programs exist
  - but preparatory work is being done

# 5. Speech synthesis

- *A good Icelandic speech synthesizer should be developed. It should be capable of reading Icelandic texts with clear and comprehensible pronunciation and natural intonation that is understandable without special training*
- A new synteziser has been developed
  - in cooperation between the University of Iceland, Iceland Telecom, and Hex Software
  - its quality appears to be very good

# 6. Speech recognition

- *Work should be done on speech recognition for Icelandic, the aim being to develop programs that can understand normal Icelandic speech*
- An isolated word recognition system has been developed
  - by the University of Iceland and four private companies
  - but continuous speech recognition is far away

# 7. Machine translation

- *Work should be done on the development of translation programs between Icelandic and other languages, one of the aims being to simplify searches in databases*
- Almost nothing has happened in this area
  - some isolated experiments have been made
  - several people use translation memory
  - no usable translation programs are underway

# Participation in Nordic cooperation

- Nordic Language Technology Research Program (2001-2004)
  - and several networks funded by the program
- Nordic Graduate School of Language Technology (NGSLT, 2004-2008)
  - a number of students have attended its courses
- Northern European Association for Language Technology (NEALT, funded 2006)
- Several applications to funding bodies

# Recent developments

- Icelandic Centre for Language Technology (ICLT)
  - a consortium of researchers from three institutes
- Recent projects connected to the ICLT
  - seen as a contribution to an Icelandic BLARK
    - IceTagger
    - IceParser
    - Context-sensitive spell checker
  - partly supported by the Icelandic Research Fund

# 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 (€ 1,350,000)
  - 1/8 of the estimated cost

# After the LT Program

- 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

# Who is going to pay?

- We will have to pay for Icelandic LT
  - but how?
- Will Icelanders pay a higher price
  - for having LT products in their native language?
- Are we willing to pay higher taxes
  - to support Icelandic LT companies?
- Will the politicians realize the importance
  - of having language technology in Icelandic?

# IT and the future of Icelandic

- Information technology has become
  - an important and integrated feature
  - of the daily life
  - of almost every single Icelander
- What if our native language is not usable
  - within new technologies
  - in fields of innovation and creativity
  - in areas where new job opportunities are offered?

# Why Icelandic LT?

- Why do we need Icelandic LT?
  - it is not only a question of language protection
  - it is also a question of human rights
- We should be able to use our native language
  - anytime and anywhere
  - within the Icelandic language community
- Everything else is an unacceptable retreat
  - which will pave the way to the death of Icelandic