Eiríkur Rögnvaldsson: Evrópskt máltæknisamstarf

META-NET/META-NORD, 2011-2013

META NET

White Paper Series | Hvítbókaröð

THE ICELANDIC ÍSLENSK
LANGUAGE IN TUNGA Á
THE DIGITAL STAFRÆNNI
AGE ÖLD

Eiríkur Rögnvaldsson
Kristín M. Jóhannsdóttir
Sigrún Helgadóttir
Steinþór Steingrímsson

Springer
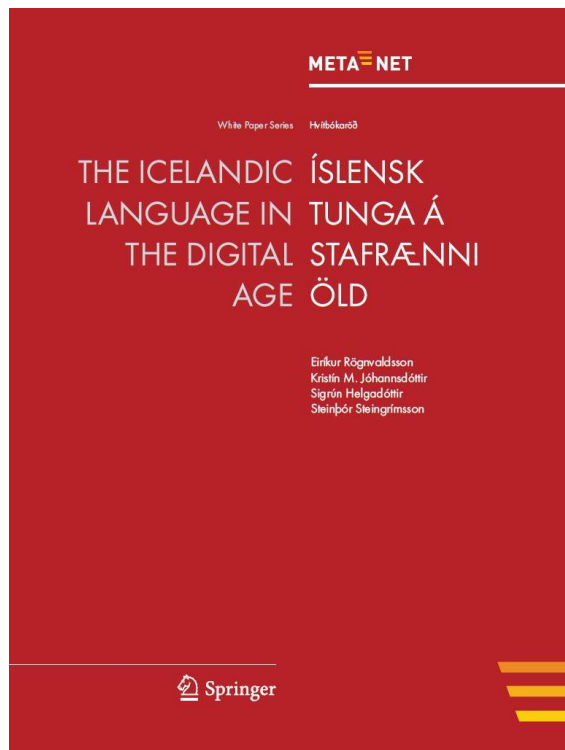
Weak/no support

Croatian
Icelandic
Latvian
Lithuanian
Maltese
Romanian

Speech processing

Weak/no support

Basque
Bulgarian
Croatian
Czech
Danish
Estonian
Finnish
Galician
Greek
Icelandic
Irish
Latvian
Lithuanian
Maltese
Norwegian
Portuguese
Serbian
Slovak
Slovene
Swedish

Machine translation

Weak/no support

Croatian
Estonian
Icelandic
Irish
Latvian
Lithuanian
Maltese
Serbian

Text analysis

Weak/no support

Icelandic
Irish
Latvian
Lithuanian
Maltese

Speech and text resources

ELE Deliverables 2022

- **34 reports on European Languages**
  - **Report on the Icelandic Language**
  - **Reports on the Nordic Minority Languages**
- **State of the art in Language Technology and Language-centric AI**
- **Digital Language Equality (preliminary definition)**

Chart axis labels (left): Good support, Moderate support, Fragmentary support, Weak or no support

Chart x-axis labels: English, German, Spanish, French, Finnish, Italian, Portuguese, Dutch, Swedish, Polish, Catalan, Greek, Hungarian, Romanian, Estonian, Danish, Norwegian, Lithuanian, Czech, Latvian, Bulgarian, Slovak, Slovenian, Croatian, Basque, Welsh, Galician, Icelandic, Serbian, Maltese, Irish, Luxembourgish, Bosnian

Achievements of the past few years

- **Language Technology Programme for Icelandic (LTPI), 2019-2022**
  - Budget ≈20 M€ – Mostly government funded
- **Iceland became a member of CLARIN and an Icelandic CLARIN centre was established**
  - All deliverables of the LTPI are available for free from CLARIN-IS under standard open licenses
- **The Government published an AI Strategy document in 2021**
  - The importance of developing LT resources and tools for Icelandic is explicitly mentioned
- **An LT consortium comprising academic institutions, NGOs and private companies was established**
  - The members of the consortium have cooperated closely on developing LT resources and tools
- **LT education and research has been strengthened considerably**
  - The number of researchers, students and developers working on LT has multiplied

Strengths and highlights

- **A number of new resources have been created within the LTPI, such as:**
  - *Icelandic Gigaword Corpus*, a general national corpus, 1.9 billion tokens
  - *ParIce*, aligned parallel corpus between Icelandic and English, 3.5 million translation units
  - *Samrómur*, audio corpus, 2.85 million sentences from 28,000 speakers, 247,800 minutes
- **A number of new tools have been created within the LTPI, such as:**
  - *Greynir*, a package that parses text and generates constituency trees, finds lemmas, inflects NPs, assigns PoS tags, and more
  - *GreynirTranslate*, IS-EN and EN-IS translation models based on a multilingual BART model
- **A number of tools are under development and will be finished by the end of the LTPI, such as:**
  - A new text-to-speech system
  - A new speech recognition system
  - A new machine translation system between Icelandic and English

Threats and shortcomings

- **The LTPI will end later this year and its future is uncertain**
  - The Government has indicated that the LTPI will be prolonged
  - However, no firm promises have been made and future funding is not secured
- **A great number of important resources and tools is still lacking, such as:**
  - spoken language corpora; parallel corpora between Icelandic and other languages than English; corpora for different purposes (sentiment analysis, question answering, summarisation), annotated multimodal corpora; term lists
  - sentiment analysis; summarization; question answering; natural language understanding; natural language generation; dialogue management; disambiguation; translation between Icelandic and other languages than English; speech translation; automatic subtitling; advanced speech synthesis (intonation, empathy); specialised grammar checking
- **More cooperation between academia and industry and more investment from industry is needed**

European Language Grid / European Language Equality, 2019-2022

- **ELG will become the primary platform for the Language Technology community in Europe**
- ELG is an initiative for the European LT community, including industry, innovation and research
  - We will create a digital marketplace on which European companies can showcase and offer their Language Technologies to customers
- **The ELE project is a direct response to the [resolution](#) "Language equality in the digital age", which was passed by the European Parliament in a landslide vote – 592 votes in favour and only 45 against – in September 2018**
- The report provides more than 40 recommendations, including:
  - 25. Establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands
  - 27. Europe has to secure its leadership in language-centric AI

Recommendations for the SRIA

- **Small languages need special dedicated LT programmes**

- **Special support for low-resourced languages where the LT market is small and not sustainable**

- **Products of the major international IT companies must be compatible with small languages**

A delegation headed by the President of Iceland and the Minister of Culture visited five major international IT companies in May