Eiríkur Rögnvaldsson

Icelandic language technology: an overview

Abstract

We describe the establishment and development of Icelandic language technology since its very beginning ten years ago. The ground was laid with a report from an Expert Group appointed by the Minister of Education, Science and Culture in 1998. In this report, which was delivered in the spring of 1999, the group proposed several actions to establish Icelandic language technology. This paper reviews the concrete tasks that the group listed as important and their current status. It is shown that even though we still have a long way to go to reach all the goals set in the report, good progress has been made in most of the tasks. Icelandic participation in Nordic cooperation on language technology has been vital in this respect. In the final part of the paper, we speculate on the cost of Icelandic language technology and the future prospects of a small language like Icelandic in the age of information technology.

1. Introduction

At the turn of the century, Icelandic language technology (henceforth LT) was virtually non-existent.[1] 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, and no Icelandic software companies were working on language technology.

All of this has now changed and Icelandic language technology has been firmly established. In the fall of 1998, the Minister of Education, Science and Culture, Mr. Björn Bjarnason, appointed an Expert Group to investigate the situation in language technology in Iceland. Furthermore, the group was supposed to come up with proposals for strengthening the status of Icelandic language technology. The members of the group were Rögnvaldur Ólafsson, Associate Professor of Physics, Eiríkur Rögnvaldsson, Professor of Icelandic Language, and Þorgeir Sigurðsson, electrical engineer and linguist.

The Expert Group handed its report to the Minister in April 1999 (Ólafsson et al. 1999). It took a while to get things going, but in 2000, the Icelandic Government launched a special Language Technology Program (Arnalds 2004; Ólafsson 2004), with the aim of supporting institutions and companies to create basic resources for Icelandic language technology work. In the report, four types of actions were proposed in order to establish Icelandic language technology:

- 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 computational linguistics.

[1] This paper is a revised and updated version of material in Rögnvaldsson (2008); cf. also Rögnvaldsson et al. (2009).

This has all been done, to some extent at least (Arnalds 2004; Ólafsson 2004; Rögnvaldsson 2005), and this initiative resulted in several projects which have had profound influence on Icelandic LT. In this paper, we will give an overview of this work and other activities in the field during the past ten years, and then speculate on the prospects of language technology in Iceland and the future of the language in the age of information technology.

2. Priority tasks and their implementation

In this section, we give an overview of the most important resources, research projects and language technology products that the LT program initiated. The Expert Group report stated the following (Ólafsson et al. 1999, 33):

> 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.

Following this statement, the LT Expert Group proposed a list of priority tasks for Icelandic language technology during the following five years. Those tasks are listed here in italics at the beginning of each subsection, and in the text that follows, we try to estimate to what extent each task has been fulfilled (cf. also Arnalds 2004; Ólafsson 2004; Rögnvaldsson 2005).

2.1 Software translation

The main computer programs on the general market (*Windows, Word, Excel, Netscape, Internet Explorer, Eudora,...*) *should be available in Icelandic.*

In 2004, Icelandic versions of Windows XP (including Internet Explorer) and Microsoft Office 2003 came on the market. These versions do not seem to suffer from any technical bugs, as was the case with the first translation of Windows (Windows 98) into Icelandic a few years earlier. However, the translations have not met with great success, and most people, except perhaps the older generation, seem to prefer the English version. The reason is probably that people had grown used to having these programs in English and saw no reason for adopting the Icelandic version. An Icelandic translation of Windows 7 and Microsoft Office 2010 has just been finished, and it will be interesting to see whether these versions gain more popularity than their predecessors.

In addition to this, special interest groups have been formed in order to translate open-source software for GNU/Linux. Thus, there exists an Icelandic version of the KDE (K Desktop Environment; www.is.kde.org/), and the Ubuntu operating system (www.ubuntu.com/) is currently being translated. The Firefox browser has also been translated into Icelandic, together with the interfaces of popular websites such as Facebook.

2.2 Icelandic characters

It should be possible to use the Icelandic non-ASCII characters (á é í ó ú ý ð þ æ ö Á É Í Ó Ú Ý Ð Þ Æ Ö) in all circumstances: in computers, mobile telephones, teletext and other applications used by the public.

When this was written, the ISO 8859-1 standard, which includes all the above-mentioned characters, had already been in existence for a number of years. However, many TV sets lacked special Icelandic characters in teletext pages, and mobile phones could not show any non-ASCII characters since they used a 7-bit character table. Nowadays, most TV sets and mobile phones can show all Icelandic characters although there seem to be some exceptions. Thus, the situation has improved considerably during the last decade.

2.3 Morphological and syntactic parsing

Work should proceed on the parsing of Icelandic, with the aim that it should be possible to use computer technology to analyze Icelandic texts grammatically and syntactically.

The LT Program funded three major projects in this area. The Institute of Lexicography received a grant for building a full-form morphological database of Icelandic (Bjarnadóttir 2005). This database is still growing and now contains around 260,000 lexemes and 5.6 million inflectional forms (http://bin.arnastofnun.is). In another project at the Institute of Lexicography, three data-driven taggers of different types (TnT, MXPOST and fnTBL) were trained and evaluated on a manually tagged Icelandic corpus of 500,000 words (Helgadóttir 2005).

A commercial company, Frisk Software (http://frisk.is/), also received a grant for developing an HPSG-based parser with the future aim of building grammar and style checking software for Icelandic (Albertsdóttir/Stefánsson 2004). Unfortunately, this project has not been finished.

After the LT Program ended, Hrafn Loftsson, Assistant Professor in Computer Science at Reykjavik University, developed a rule-based PoS tagger, *IceTagger* (Loftsson 2006). Loftsson is also the main author of a shallow syntactic parser, *IceParser* (Loftsson/Rögnvaldsson 2007). A mixed method lemmatizer for Icelandic, *Lemmald*, has been developed by Anton Karl Ingason, a Language Technology student (Ingason et al. 2008). These three programs make up the IceNLP package which is online at http://nlp.cs.ru.is.

Furthermore, the LT Expert Group (Ólafsson et al. 1999) mentioned two prerequisites for further progress in this field, which are listed in 2.3.1 and 2.3.2.

2.3.1 A balanced corpus

A large computerized text corpus including Icelandic texts of a wide variety of types should be established.

In 2004, the Institute of Lexicography received a grant from the LT Program for building a balanced morphosyntactically tagged corpus of Modern Icelandic (Helgadóttir

2004). This corpus will contain 25 million words of different genres, including transcribed spoken language, and shall be finished in 2011. A preliminary version is online at http://mim.hi.is.

2.3.2 A semantically annotated lexicon

A grammatically and semantically annotated lexicon should be established.

This lexicon was meant to be something similar to the PAROLE/SIMPLE lexicon (www.ub.es/gilcub/ SIMPLE/simple.html). No such lexicon has been built yet. However, many types of raw material for building a lexicon of this type do exist, especially in various collections and databases at the Institute of Lexicography, such as the ISLEX database which comprises 50,000 entries for Icelandic and their equivalents in Danish, Norwegian, and Swedish (www.arnastofnun.is/page/arnastofnun_ord_islex) and will be finished in late 2011.

2.4 Spelling and grammar checkers

Good auxiliary programs should be developed for textual work in Icelandic, i.e. for hyphenation, spell-checking, grammar correction, etc.

When this was written (Ólafsson et al. 1999), we had the spell-checking program *Púki* from Frisk Software (http://frisk.is), which has now been improved with support from the LT Program (Skúlason 2004). In 2002, the Dutch company Polderland (http://www.polderland.nl/) developed a spell-checking program for the Microsoft Office package. Furthermore, there exists an open source spell checker for Icelandic based on Aspell (http://aspell.net/) which can be used with GNU/Linux applications. These programs (as most spell checkers) are word-based, and hence cannot cope with many common spelling errors.

No grammar checking or style checking programs exist, but a prototype of a context-sensitive spell checker has been developed which could hopefully lay the ground for a basic grammar checker (Ingason et al. 2009). This prototype has been integrated into LanguageTools (www.languagetool.org) and works with OpenOffice (www.openoffice.org).

2.5 Text-to-speech system

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 formant-based Icelandic speech synthesizer was originally made around 1990 (Carlson et al. 1990) and improved around 2000. Even though this synthesizer was very useful for blind and visually impaired people, its quality was far from being satisfactory for use in commercial applications for the general public.

The last project that the LT Program supported was a new text-to-speech system, which was made in cooperation between the University of Iceland, Iceland Telecom, and Hex Software. The system was trained by Nuance and uses their technology. For several rea-

sons, the system has not been put to use in commercial applications and many users, especially among the blind, do not find the voice quality of the system satisfying.

As a result, the Icelandic Organization of Blind and Partially Sighted is now planning to develop a new text-to-speech system in cooperation with the University of Iceland, Reykjavik University, and the Ivo software company (www.ivona.com/). If everything goes as planned, this system will be finished in 2012.

2.6 Speech recognition

Work should be done on speech recognition for Icelandic, the aim being to develop programs that can understand normal Icelandic speech.

In 2003, the University of Iceland and four leading companies in the telecommunication and software industry joined efforts to build an isolated word speech recognizer for Icelandic, with support from the LT Program and in cooperation with ScanSoft (now Nuance) (Rögnvaldsson 2004). The performance of the system has turned out to be quite satisfying; the recognition rate appears to be at least 97% (Rögnvaldsson 2004). However, no attempts have been made to develop a system for recognizing continuous speech.

2.7 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.

The development in this area has been limited, although some isolated experiments have been made. In 2008, Stefán Briem, an independent researcher, launched a free web-based service, which offers translations between Icelandic and three other languages (English, Danish, and Esperanto; http://tungutorg.is/). Hrafn Loftsson and his associates have been developing a rule-based shallow transfer translation system from Icelandic to English (Brandt et al. 2011), based on the Apertium platform (http://www.apertium.org/). A preliminary version of the system is available online at http://nlp.cs.ru.is/ApertiumISENWeb/.

Since 2009, Google Translate (http://translate.google.com) has offered translation to and from Iclandic. The quality of the translation was rather poor in the beginning, but is constantly getting better.

3. The current status of Icelandic LT

After the LT Program ended six years ago, LT researchers from three institutes (University of Iceland, Reykjavik University and the Árni Magnússon Institute for Icelandic Studies), who had been involved in most of the projects funded by the LT Program, decided to join forces in a consortium called the Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the Program. The main roles of the ICLT are to:

– serve as an information centre on Icelandic LT by running a website (http://iclt.is);
– encourage cooperation on LT projects between universities, institutions and commercial companies;

– organize and coordinate university education in LT;
– participate in Nordic, European and international cooperation within LT;
– initiate and participate in R&D projects in LT;
– keep track of resources and products in the field of Icelandic LT;
– hold LT conferences with the participation of researchers, companies and the public;
– support the growth of Icelandic LT in all possible manners.

Over the past six years, the ICLT researchers have initiated several new projects which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. The most important product of these projects is the IceNPL package (IceTagger, IceParser and Lemmald) mentioned in section 2.3 above. In 2009, the ICLT received a relatively large three year Grant of Excellence from the Icelandic Research Fund for the project "Viable Language Technology beyond English – Icelandic as a test case" (http://iceblark.wordpress.com). Within that project, three types of LT resources are being developed:

– a database of semantic relations (a pilot WordNet; Nikulásdóttir/Whelpton 2010);
– a prototype of a shallow-transfer machine translation system (Brandt et al. 2011);
– a treebank with a historical dimension (Rögnvaldsson et al. 2011).

These resources were chosen because they were considered central to current LT work and prerequisites for further research and development in Icelandic LT.

For a small language community and a small research environment like the Icelandic one, it is vital to cooperate, not only on the national level but also internationally. Since 2000, Icelandic researchers and policy makers have taken an active part in Nordic co-operation on language technology. This has been of major importance in establishing the field in Iceland. The Nordic Language Technology Research Programme 2000-2004 was instrumental in this respect. Icelandic researchers also take part in the Northern European Association for Language Technology (NEALT, http://omilia.uio.no/nealt/), and the bi-annual Nordic-Baltic conferences of computational linguistics (NODALIDA). In 2003, the 14th NODALIDA conference was held at the University of Iceland in Reykjavík.

Iceland has just recently entered the CLARIN consortium (http://clarin.eu), and takes part in the EU-funded META-NORD project which starts February 1st, 2011, and aims to establish an open linguistic infrastructure in the Baltic and Nordic countries. We sincerely hope that our participation in these projects will help us to develop, standardize and make available several important LT resources and thus contribute to the growth of Icelandic language technology.

4. The price and prospects of Icelandic LT

Twelve years ago, the LT Expert Group estimated that it would cost around one billion Icelandic krónas (which then amounted to about ten million Euros) to make Icelandic language technology self-sustained. After that, the free market should be able to take

over, since it would have access to public resources that would have been created by the LT Program, and that would be made available on an equal basis to everyone who was going to use these resources in their commercial products.

However, the total budget of the government-funded LT program over its lifespan (2000-2004) was only 133 million Icelandic krónas – that is, around ⅛ of the sum that the Expert Group estimated would be needed. It should therefore come as no surprise that we still have a long way to go. There are only 320,000 people speaking Icelandic, and that is not enough to sustain costly development of new products. At present, no commercial companies are working in the LT area because they don't see it as profitable. It is thus extremely important to continue public support for Icelandic language technology for some time, but given the current financial situation, it does not seem likely that such support will come from the state budget in the near future.

When we try to estimate the importance of Icelandic language technology we must realize that ICT has become an important and integrated feature of the daily life of almost every single Icelander. If Icelandic cannot be used within ICT, speakers will be faced with a completely new situation, without parallels earlier in the history of the language. We will have an important area of the daily life of ordinary people where they cannot use their native language. How is that going to affect the speakers and the language community? What will happen when the native language is no longer usable within new technologies and in other new and exciting areas; in fields of innovation and creativity; and in areas where new job opportunities are offered? We don't have to think long about this scenario to see the signs of imminent danger.

In 2009, the Icelandic Parliament (Alþingi) unanimously approved an official language policy which had been prepared by the Icelandic Language Council (Íslenska til alls 2009). The policy document contains a section on ICT and the Icelandic language, where it is explicitly stated that Icelandic should be useable – and used – in all areas within information and communications technology that touch upon the daily life of the public. It remains to be seen what the government is going to do in order to implement this policy.

But the need for native language technology is not, and should not be, only driven by people's wish to protect and preserve their language. It is equally – or even more – important to look at this from the user's point of view. Ordinary people should not be forced to use foreign languages in their everyday lives. They have the right to be able to use their native language anytime and anywhere within their language community, in all possible contexts. Otherwise, they will be linguistically oppressed in their own language community.

5. References

Albertsdóttir, M./Stefánsson, S.E. (2004): Beygingar- og málfræðigreinikerfi [A system for morphological and syntactic parsing]. In: *Samspil tungu og tækni*. Reykjavík: Ministry of Education, Science and Culture, 16-19.

Arnalds, A. (2004): Language technology in Iceland. In: Holmboe, H. (ed.): *Nordisk Sprogteknologi. Årbog 2003*. Copenhagen: Museum Tusculanums Forlag Københavns Universitet, 41-43.

Bjarnadóttir, K. (2005): Modern Icelandic inflections. In: Holmboe, H. (ed.): *Nordisk Sprogteknologi. Årbog 2005*. Copenhagen: Museum Tusculanums Forlag Københavns Universitet, 49-50.

Brandt, M.D./Loftsson, H./Sigurþórsson, H./Tyers, F. (2011): Apertium-IceNLP: a rule-based Icelandic to English machine translation system. In: Forcada, M.L./Depraetere, H./Vandeghinste, V. (eds.): *EAMT 2011: Proceedings of the 15th Conference of the European Association for Machine Translation, 30-31 May 2011*. 217-224.

Carlson, R./Granström, B./Helgason, P./Thráinsson, H./Jensson, P. (1990): An Icelandic text-to-speech system for the disabled. In: *STL-QPSR* 31, 4, 55-56.

Helgadóttir, S. (2004): Mörkuð íslensk málheild [A tagged Icelandic corpus]. In: *Samspil tungu og tækni*. Reykjavík: Ministry of Education, Science and Culture, 67-71.

Helgadóttir, S. (2005): Testing data-driven learning algorithms for PoS tagging of Icelandic. In: Holmboe, H. (ed.): *Nordisk Sprogteknologi. Årbog 2004*. Copenhagen: Museum Tusculanums Forlag Københavns Universitet, 257-265.

Ingason, A.K./Helgadóttir, S./Loftsson, H./Rögnvaldsson, E. (2008): A mixed method lemmatization algorithm using a Hierarchy of Linguistic Identities (HOLI). In: Nordström, B./Ranta, A. (eds.): *Advances in natural language processing*. (= Lecture Notes in Computer Science 5221). Berlin: Springer, 205-216.

Ingason, A.K./Jóhannsson, S.B./Rögnvaldsson, E./Loftsson, H./Helgadóttir, S. (2009): Context-sensitive spelling correction and rich morphology. In: Jokinen, K./Bick, E. (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. (= NEALT Proceeding Series 4). Tartu: NEALT, Tartu University Library, 231-234.

Íslenska til alls [Icelandic for all purposes] (2009): Tillögur íslenskrar málnefndar að íslenskri málstefnu samþykktar á Alþingi 12. mars 2009. Reykjavík: Ministry of Education, Science and Culture.

Loftsson, H. (2006): Tagging a morphologically complex language using heuristics. In: Salakoski, T./Ginter, F./Pyysalo, S./Pahikkala, T. (eds.): *Advances in natural language processing, 5th International Conference on NLP, FinTAL 2006, Proceedings.* (= Lecture Notes in Computer Science 4139). Berlin: Springer, 640-651.

Loftsson, H. (2007): *Tagging and parsing Icelandic text*. Doctoral dissertation. Sheffield: Department of Computer Science, University of Sheffield.

Loftsson, H./Rögnvaldsson. E. (2007): IceParser: an incremental finite-state parser for Icelandic. In: Nivre, J./Kaalep, H.-J./Muischnek. K./Koit, M. (eds.): *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. Tartu: University of Tartu, 128-135.

Nikulásdóttir, A.B./Whelpton, M. (2010): Extraction of semantic relations as a basis for a future semantic database for Icelandic. In: *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*. Valetta: SALTMIL, 33-39.

Ólafsson, Rögnvaldur (2004): Tungutækniverkefni menntamálaráðuneytisins [The Language Technology Program of the Ministry of Education, Science and Culture]. In: *Samspil tungu og tækni*. Reykjavík: Ministry of Education, Science and Culture, 7-13.

Ólafsson, Rögnvaldur/Rögnvaldsson, E./Sigurðsson, Þ. (1999): *Tungutækni. Skýrsla starfshóps* [Language Technology. Report of an expert group]. Reykjavík: Ministry of Education, Science and Culture.

Rögnvaldsson, E. (2004): The Icelandic speech recognition project *Hjal*. In: Holmboe, H. (ed.): *Nordisk Sprogteknologi. Årbog 2003*. Copenhagen: Museum Tusculanums Forlag Tusculanums, 239-242.

Rögnvaldsson, E. (2005): Staða íslenskrar tungutækni við lok tungutækniátaks [The Status of Icelandic Language Technology at the End of the Language Technology Program]. In: *Tölvumál* 24-2. www.sky.is/index.php?option=com_content&task=view&id=55&Itemid=85).

Rögnvaldsson, E. (2008): Icelandic Language Technology ten years later. In: *Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages. SALTMIL workshop, LREC 2008*. Marrakech: SALTMIL, 1-5.

Rögnvaldsson, E./Ingason, A.K./Sigurðsson, E.F. (2011): Coping with Variation in the Icelandic Diachronic Treebank. In: Johannessen, J.B. (ed.): *Language variation infrastructure. Papers on selected projects.* (= Oslo Studies in Language 3.2). Oslo: University of Oslo, 97-111.

Rögnvaldsson, E./Loftsson, H./Bjarnadóttir, K./Helgadóttir, S./Nikulásdóttir, A.B./Whelpton, M./Ingason, A.K. (2009): Icelandic language resources and technology: status and prospects. In: Domeij, R./Koskenniemi, K./Krauwer, S./Maegaard, B./Rögnvaldsson, E./de Smedt, K. (eds.): *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Tartu: NEALT, Tartu University Library, 27-32.

Skúlason, F. (2004): Endurbætt tillögugerðar- og orðskiptiforrit Púka [Improved suggestions and hyphenations in the Púki Spell Checker]. In: *Samspil tungu og tækni*. Reykjavík: Ministry of Education, Science and Culture, 29-31.

6. Acknowledgements

As working at the University of Iceland as Professor of Icelandic Language, I would would like to thank the University and especially the Faculty of Icelandic and Comparative Cultural Studies for providing infrastructural assistance to my work and this text.