Icelandic Language Technology Ten Years Later

Eiríkur Rögnvaldsson
Department of Icelandic, University of Iceland
Árnagarði við Suðurgötu, IS-101, Reykjavík, Iceland
E-mail: eirikur@hi.is

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 a committee appointed by the Minister of Education, Science and Culture in 1998. In this report, which was delivered in the spring of 1999, the committee proposed several actions to establish Icelandic language technology. This paper reviews the concrete tasks that the committee 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

Ten years ago, Icelandic language technology (LT) virtually didn't exist. 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 or college, 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 a special committee to investigate the situation in language technology in Iceland. Furthermore, the committee was supposed to come up with proposals for strengthening the status of Icelandic language technology. The members of the committee were Rögnvaldur Ólafsson, Associate Professor of Physics, Eiríkur Rögnvaldsson, Professor of Icelandic Language, and Þorgeir Sigurðsson, electrical engineer and linguist.

The committee 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. This initiative resulted in several projects which have had profound influence on the field. 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.

The purpose of the paper is to show how the authorities, industry, and academia can fruitfully cooperate to build language technology resources and tools from scratch in a relatively short time for a relatively small budget. We think our experience may be useful for other small language communities where language technology is in its infancy and needs to be established.

2. Proposals of the LT Committee

In the report of the Language Technology Committee (Ólafsson et al., 1999), 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 linguistics.

This has all been done, to some extent at least (Arnalds, 2004; Ólafsson, 2004; Rögnvaldsson, 2005). An overview of the most important resources, research projects and language technology products is given in section 3 below.

In the fall of 2002, the University of Iceland launched a new Master's program in Language Technology. This is a two-year interdisciplinary program (120 ECTS credits), and the applicants can either have a B.A. degree in the humanities (languages and linguistics) or a B.Sc. degree in computer science (or electrical or software engineering). Due to lack of resources, both financial and human, students were only admitted to the program twice, in 2002 and 2003.

Last fall, the program was relaunched, now as a joint program between the Department of Icelandic at the University of Iceland and the School of Computer Science at Reykjavik University. We hope that this cooperation will enable the two universities, in cooperation with the Nordic Graduate School of Language Technology (NGSLT), to offer sound and solid education, and to recruit enthusiastic students who will engage in research and development on Icelandic language technology.

In addition to this, a few Icelandic students have studied language technology abroad in recent years, and the first Icelandic Ph.D. in the field received his degree last year from the University of Sheffield (Loftsson 2007).

3. Priority tasks and their implementation

The above-mentioned report on Icelandic language technology (Ólafsson et al., 1999) stated the following:

> 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 Language Technology Committee 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).

3.1 Software translation

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

In 2004, an Icelandic version 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 see no reason for adopting the Icelandic version. An Icelandic translation of Windows Vista and Microsoft Office 2007 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; http://www.is.kde.org/), and the new Hardy Heron version of the Ubuntu operating system (www.ubuntu.com) is currently being translated.

3.2 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.

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.

3.3 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 Language Technology Project 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 259,000 lexemes and 5.6 million inflectional forms (iceland.spurl.net/tunga/VO/). 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 (www.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 and Stefánsson, 2004). Unfortunately, this latter project has not been finished.

The Language Technology Committee (Ólafsson et al., 1999) mentioned two prerequisites for further progress in this field, which are listed in 3.3.1 and 3.3.2.

3.3.1 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 Language Technology Program for building a balanced morphologically 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 later this year.

3.3.2 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 (http://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 is being built and will comprise 50,000 entries for Icelandic and their equivalents in Danish, Norwegian, and Swedish (www.lexis.hi.is/islex-ohvefur/islex-meira.html).

3.4 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 nine years ago (Ólafsson et al., 1999), we had the spell-checking program *Púki* from Frisk Software, which has now been improved with support from the Language Technology Program (Skúlason, 2004). In 2002, the Dutch company Polderland (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 (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 current work on a context-sensitive spell checker mentioned in Section 5 below will presumably lay the ground for a basic grammar checker.

3.5 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 Language Technology 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. People seem to agree that the quality is very good. The system came on the market last year and appears to be a success, especially due to a recently launched online service which uses the system for reading web pages and text entered by users (http://www.hexia.net/upplestur).

3.6 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 Language Technology 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.

3.7 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. Just recently, Stefán Briem, an independent researcher, has launched a free web-based service, which offers translations between Icelandic and three other languages (English, Danish, and Esperanto; www.tungutorg.is). Furthermore, the Icelandic Technical Development Fund has given a grant to a private company that works on translation software for translating from Icelandic to English, but this software has not been marketed yet and the status of its development is unclear. Iceland has also taken part in a Nordic project which aims at enabling multilingual web search (Dalianis et al., 2007).

4. Nordic cooperation

Since 2000, Icelandic researchers and policy makers have taken active part in Nordic cooperation on language technology. This participation has been of major importance in establishing the field in Iceland. From 2001-2004, the Nordic Language Technology Research Programme (Holmboe, 2005) funded language technology Documentation Centers in the five Nordic countries (www.nordoknet.org). At the end of 2004, the Icelandic center merged with the website www.tungutaekni.is, which the Language Technology Program had been running since its start in 2001. This website is now run by the ICLT (see Section 5 below). Thanks to the documentation center, we now have a good and accessible overview of people, projects, products, materials, companies, organizations, etc. having to do with Icelandic language technology.

Through the documentation center, we have also made contacts with several people and institutions in the Nordic and Baltic countries (cf. Fersøe et al., 2005). As a result of those contacts, Icelandic researchers have participated in several applications to Nordic and European funding bodies during the past few years. Even though most of these applications have not been successful, we have gained invaluable experience from taking part in them and cooperating with Nordic colleagues.

Another important aspect of the Nordic cooperation in language technology is the Nordic Graduate School of Language Technology (NGSLT, www.ngslt.org), funded by NorFA – now NordForsk (Nordic Research Board, www.nordforsk.org). The activities of the school started in 2004 and will run for five years. Even though the school is primarily intended for doctoral students, master's level students from Iceland have been admitted to the courses. This is absolutely crucial for the Icelandic universities, since they do not have the capacity to give the students high-quality education in language technology at home.

Icelandic researchers also take part in other Nordic and Baltic activities in the field, such as the newly established Northern European Association for Language Technology (NEALT, omelia.uio.no/nealt), and the biannual Nordic conferences of computational linguistics (NODALIDA). In 2003, the 14th NODALIDA conference was held at the University of Iceland in Reykjavík.

5. The price and prospects of Icelandic LT

After the Language Technology Program ended by the end of 2004, researchers from three research institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies) decided to join forces in a consortium called Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the Program. During the past three years, these researchers, who had been involved in most of the projects supported by the Language Technology Program, have initiated several new projects, three of which should be especially mentioned: *IceTagger*, a linguistic rule-based tagger (Loftsson, 2006, 2007), *IceParser*, a shallow parser (Loftsson and Rögnvaldsson, 2007; Loftsson, 2007), and a context-sensitive spell checker which shall be finished later this year. These programs are seen as a contribution to the establishment of a BLARK (Basic Language Resource Kit; cf. Krauwer, 2003) for Icelandic, and the group has made plans for the next steps towards that goal.

These projects have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. However, much more money is needed in order to create a BLARK for Icelandic. The Language Technology Committee estimated that it would cost around one billion Icelandic krónur, about ten million Euros, to make Icelandic language technology self-sustained (Ólafsson et al., 1999). After that, the free market should be able to take over, since it would have access to public resources that would have been created for money from the Language Technology Program, and that would be made available on an equal basis to everyone who were going to use these resources in their commercial products.

Even though the Language Technology Program was very successful and had a great impact on the development of Icelandic language technology, the fact remains that its total budget over the lifespan of the program (2000-2004) was only 133 million Icelandic krónur (Ólafsson, 2004), or around 1.35 million Euros – that is, 1/8 of the sum that the committee estimated would be needed. It should therefore come as no surprise that we still have a long way to go. There are only 300,000 people speaking Icelandic, and that is not enough to sustain costly development of new products. It costs just as much to build language resources for Icelandic as for languages with hundreds of millions of speakers. Therefore, we feel it is extremely important to continue public support for Icelandic language technology for some time, in order to make the most out of the money that has been spent up to now, and utilize the knowledge and experience that researchers and companies have gained.

One way to do this would be to make more use of free/open source licenses, both for software and linguistic resources. It has recently been argued convincingly by several authors (cf., for instance, Forcada, 2006; Streiter et al., 2007; Alegria et al., 2008) that it is essential for minor/non-central/less-resourced languages to adopt open source policy with respect to LT resources in order to survive the Information Age.

Unfortunately, many Icelandic resources such as dictionaries and corpora are privately owned, either by commercial companies or individual authors or researchers, and it can be difficult and expensive, or even impossible, to get permission to use them even for research, not to mention for commercial applications. All grants from the Language Technology program were given with the condition that the resources developed would be accessible for anyone wanting to use them in language technology products. However, these resources are not distributed under an open source license and most of them are not free. Even though the license to use them is usually not very expensive, the license fee acts as a barrier for the use of these resources in LT research and development. It would obviously be beneficial for the future of Icelandic LT to implement open source policy, and this has recently been strongly advocated (Trosterud, 2008; Gíslason, 2008).

6. Conclusion - LT and the future of Icelandic

In this paper, 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.

When we try to estimate the importance of Icelandic language technology we must realize that information technology has become an important and integrated feature of the daily life of almost every single Icelander. If Icelandic cannot be used within information technology, 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.

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.

7. Acknowledgements

Thanks to three anonymous reviewers whose comments have influenced (and hopefully improved) the paper considerably, especially by suggesting the addition of paragraphs on BLARK and open source license and translation.

8. References

Albertsdóttir, M., and Stefánsson, S.E. (2004). Beygingar- og málfræðigreinikerfi [A System for Morphological and Syntactic Parsing]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 16-19.

Alegria, I., Arregi, X., Artola, X., Diaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., and Sarasola, K. (2008). Strategies for Sustainable MT for Basque: Incremental Design, Reusability, Standardization and Open-source. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, pp. 59-64.

Arnalds, A. (2004). Language Technology in Iceland. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 41-43.

Bjarnadóttir, K. (2005). Modern Icelandic Inflections. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2005*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 49-50.

Carlson, R., Granström, B., Helgason, P., Thráinsson, H., and Jensson, P. (1990). An Icelandic Text-to-Speech System for the Disabled. In *Proceedings of ECART (European Conference on the Advancement of Rehabilitation Technology).* Maastricht, the Netherlands.

Dalianis, H., Rimka, M., and Kann, V. (2007). Using Uplug and SiteSeeker to Construct a Cross Language Search Engine for Scandinavian. Paper presented at the workshop *The Automatic Treatment of Multilinguality in Retrieval, Search and Lexicography*, Copenhagen, Denmark, April 26. (people.dsv.su.se/~hercules/papers/scanduplug.pdf)

Fersøe, H., Rögnvaldsson, E., and de Smedt, K. (2005). NorDokNet – Network of Nordic Documentation Centres – Contacts to Future Baltic Partners. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2005*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 13-23.

Forcada, M.L. (2006). Open Source Machine Translation: an Opportunity for Minor Languages. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages*, Genoa, Italy, May 23. (www.mt-archive.info/LREC-2006-Forcada.pdf)

Gíslason, H. (2008). Gögn og gaman: jarðvegur nýþróunar í tungutækni [The Ground for Innovation in Language Technology]. Paper presented at the workshop *Á íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavík, Iceland, March 7.

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

Helgadóttir, S. (2005). Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 257-265.

Holmboe, H. (2005). *Nordisk sprogteknologisk forskningsprogram 2000-2004. Epilog*. NordForsk, Oslo, Norway.

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia, pp. 8-15.

Loftsson, H. (2006). Tagging a Morphologically Complex Language Using Heuristics. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T. (Eds.), *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Proceedings.* Turku, Finland, pp. 640-651.

Loftsson, H. (2007). Tagging and Parsing Icelandic Text. Doctoral dissertation, Department of Computer Science, University of Sheffield, UK.

Loftsson, H., and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Nivre, J., Kaalep, H-J., Muischnek, K., and Koit, M. (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. Tartu, Estonia, pp. 128-135.

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

Ólafsson, R., Rögnvaldsson, E., and Sigurðsson, Þ. (1999). *Tungutækni. Skýrsla starfshóps* [Language Technology. Report of a Committee]. Ministry of Education, Science and Culture, Reykjavík, Iceland.

Rögnvaldsson, E. (2004). The Icelandic Speech Recognition Project *Hjal*. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 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]. *Tölvumál*, February 24.

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

Streiter, O., Scannell, K.P., and Stuflesser, M. (2007). Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers. To appear in *Machine Translation*. (borel.slu.edu/pub/mt.pdf)

Trosterud, T. (2008). Grammar-based Language Technology as an Answer to the Challenges Facing Icelandic and other Circumpolar Languages. Paper presented at the workshop *Á íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavík, Iceland, March 7.