Icelandic Language Resources and Technology: Status and Prospects

Eiríkur Rögnvaldsson
University of Iceland
eirikur@hi.is

Hrafn Loftsson
Reykjavik University
hrafn@ru.is

Kristín Bjarnadóttir
Árni Magnússon Institute for Icelandic Studies
kristinb@lexis.hi.is

Sigrún Helgadóttir
Árni Magnússon Institute for Icelandic Studies
sigruhel@hi.is

Matthew Whelpton
University of Iceland
whelpton@hi.is

Anna Björk Nikulásdóttir
University of Iceland
abn@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

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.

1 Introduction

Ten years ago, Icelandic language technology (henceforth LT) 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, 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 appointed a special committee to investigate the situation in language technology in Iceland and come up with proposals for strengthening the status of Icelandic language technology. The committee handed its report to the Minister in April 1999 (Ólafsson et al., 1999) and in 2000, the Government launched a special Language Technology Program (Arnalds, 2004; Ólafsson, 2004), with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in several projects which have had profound influence on the field (cf. Rögnvaldsson, 2008).

In this paper, we will first give an overview of this work and other activities in the field during the past ten years. Then we will briefly outline the research program of the Icelandic LT research community for the next few years and point out the importance of open source policy for less-resourced languages. Finally, we will discuss the importance of Nordic cooperation within LT and put forward some concrete proposals to this effect, especially concerning education and dissemination of information.

2 Icelandic LT Work 1999-2009

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 (Rögnvaldsson, 2008). The main direct products of the LT Program are the following:

- A full-form morphological database of Modern Icelandic inflections (Bjarnadóttir, 2004, 2005).
- A balanced morphosyntactically tagged corpus of 25 million words (Helgadóttir, 2004).
- A training model for data-driven POS taggers (Helgadóttir, 2005, 2007).
- A text-to-speech system (Rögnvaldsson, Kristinsson and Þorsteinsson, 2006).
- A speech recognizer (Rögnvaldsson, 2004; Waage, 2004).
- An improved spell checker (Skúlason, 2004).

After the government-funded LT Program ended, 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 the *Icelandic Centre for Language Technology* (ICLT), in order to follow up on the tasks of the Program. The ICLT serves its role by:

- maintaining an information center for Icelandic language technology by running a website (cf. Rögnvaldsson, 2005);
- 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.

Over the past four years, researchers connected to the ICLT, who had been involved in most of the projects funded by the LT Program, have initiated several new projects, which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. The most important projects are: IceTagger, a linguistic rule-based tagger (Loftsson, 2007, 2008), IceParser, a shallow parser (Loftsson and Rögnvaldsson, 2007, 2008), Lemmald, a lemmatizer (Ingason et al., 2008) and a context-sensitive spell checker (Ingason et al., 2009). These projects are seen as a contribution to the establishment of a BLARK (Basic LAnguage Resource Kit, cf. Krauwer, 2003) for Icelandic.

The Icelandic LT research group is now in a position to make a research plan for the next few years, building on the resources created and the experience gained in the group's previous work. We know what kinds of resources, tools and methods are most urgently needed, and we believe we know what kind of research needs to be carried out in the near future. We have just received a relatively large Grant of Excellence ("Viable Language Technology beyond English - Icelandic as a test case") from the Icelandic Research Fund to carry out our research plan.

3 Research Plan for Icelandic LT

The existence of LT for any given language could be a deciding factor in whether that language survives the 21st century. The problem is that language resources like treebanks and wordnets are expensive to build and as the corresponding resources for English and other dominant languages become more advanced, the gap between the minority language and the "state of the art" grows. And as English continues to lead the field onwards, even the other dominant languages could struggle to keep up.

Languages other than English face two main problems in LT:

- They have less resources than English to develop LT modules (people and money);
- They may differ from English in important linguistic ways (morphology, syntax, etc.) and therefore the established methods from English LT need adaptation.

Solutions and innovations which address these two problems form the foundation of viable LT for all languages other than English. Although the first problem is a general one, it is particular-

ly acute for languages with small speech communities, such as Icelandic or Faroese, and languages spoken only in countries where economic conditions are unfavourable, such as various African languages. The second problem is moderate or acute depending on the typological distance from English. For instance, English has only sparse morphological inflection and established solutions therefore largely ignore this linguistic property; however, many languages (like Icelandic) have an extremely rich morphology which poses special challenges.

The second problem also relates to how linguistic knowledge is generally harnessed in LT. The rise and success of statistical methods have made the field look like just a branch of applied machine learning in recent years. However, much of the difference between proposed systems lies in the selection of features fed into the machine – but selecting a good feature set is about good linguistics, not good statistics. The tradition in the literature of opposing data-driven statistical methods to hand-crafted linguistic rule methods could therefore be both misleading and harmful (cf. also Trosterud, 2008).

To address the problems for LT viability discussed above, it is essential to develop new methods for constructing LT modules, such as treebanks and semantic databases, in more efficient ways. Our primary objective is to make it realistic to develop three particular types of LT modules with limited resources without sacrificing the quality of the work. The three types of modules are a **database of semantic relations** (Nikulásdóttir and Whelpton, 2009**), a shallow transfer machine translation system**, and a **pilot treebank**. These modules are chosen because they are central to current LT work and prerequisites for further research and development in Icelandic LT. The project will emphasize the following points:

- 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;
- 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.

In short, the project emphasizes the development of viable research methods and practical solutions that will strengthen Icelandic LT and serve as a model for other less-resourced languages.

4 The Prospects of Icelandic LT

The Language Technology Committee estimated that it would cost around one billion Icelandic krónas (then 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 was 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 krónas – that is, 1/8 of the sum that the committee estimated would be needed. Since then, the LT group has received a number of research grants which amount to approximately 15 million Icelandic krónas. 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. If Icelandic is to survive as a viable national language in the developed world, it must be able to meet IT demands. Consequently, investment in language technology must form an essential part of its language preservation policy. Furthermore, continued public support for Icelandic language tech-

nology will guarantee exploitation of the tools already developed and the knowledge and experience of researchers and companies which has already been accrued. A further 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) 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).

In our project, we adhere to the recent open source policy of the Icelandic Government. The source code of our research results will be available under different licenses dependant on their intended usage. Most of it will probably be freely available for the development of Open Source software under the GNU General Public License versions 2 and 3 (http://www.gnu.org/licenses/#GPL). In accordance with our general policy, the source code of the main programs that we have developed, IceParser, IceTagger, and Lemmald (cf. Section 2) will be made open source in the course of the next few months.

5 Proposals for Nordic Cooperation

Since 2000, Icelandic researchers and policy makers have taken an active part in Nordic cooperation on language technology. This has been of major importance in establishing the field in Iceland. For a small language community and a small research environment like the Icelandic one, cooperation on LT education, research, use of infrastructures, etc., is vital. The Nordic Language Technology Research Programme 2000-2004 (Holmboe, 2005) was very important in this respect and the continuation of that program or a similar one is absolutely essential.

Some of the smaller language communities in the Nordic/Baltic area still do not have even the most basic LT modules and resources. It is just as expensive to build these modules and resources for the small language communities as for the larger ones, and enough national funding for such development may not be available. For fruitful cooperation involving all the languages in question to be possible, it is necessary to create some minimal common ground, and that means that the smaller language communities need some external support in the beginning. This support can be in the form of direct funding from Nordic funds or programs, but it can also involve exchange of research and knowledge – which then, of course, must be easily accessible.

From 2001-2004, the Nordic Language Technology Research Programme funded language technology Documentation Centers in the five Nordic countries and their cooperation network (NorDokNet; Fersøe, 2005). One of the main goals of the Centers was to collect information on people, projects, products, materials, companies, organizations, etc. having to do with LT in the Nordic countries. Unfortunately, the Centers are no longer funded, and although their web pages still exist, they are not updated as regularly as one would wish and their common website, which has moved to www.cst.dk/nordoknet, is not updated at all.

In 2005, the Nordic Council of Ministers commissioned a ten-year plan in the form of an expert panel report for making the Nordic Countries a leading region in LT (Lindén et al., 2006). One of the main recommendations of the report was the compilation of BLARK reports for the Nordic languages and subsequent funding of LT tools and resources to fill the gaps revealed by the reports. We believe that it would be extremely beneficial to enhanced cooperation to have a common website containing accessible and standardized information on available language resources and tools for the Nordic languages. This could be in the form of a simple table (perhaps on a wiki page for anyone to fill out) with lines for the tools and resources (POS tagger, lemmatizer, monolingual corpus, dictionary, etc.) and columns for the languages. Much of this information can be found on the web pages of the

Nordic Documentation Centers but it does not have a common format, it takes time to collect it, and sometimes it is outdated.

Another aspect of cooperation is education. In 2002, the University of Iceland launched an interdisciplinary Master's program in LT. This is now a joint program between the Department of Icelandic at the University of Iceland and the School of Computer Science at Reykjavik University. The students in the program have had the opportunity to take courses in the Nordic Graduate School of Language Technology (NGSLT). Participation in NGSLT has been absolutely crucial for the Icelandic universities, since they do not have the capacity to give the students high-quality education in LT at home. Unfortunately, the funding period of the school has expired, so this opportunity will not be available after this academic year. It is unclear whether and how we will be able to continue our Master's Program without the availability of the NGSLT courses.

A Nordic Summer School in LT where graduate students and researchers could meet, exchange ideas, attend practical training sessions and pass on technical skills would be very effective in disseminating knowledge and encouraging mutual awareness of ongoing projects, especially if a small number of inspiring international experts were invited to participate in events.

We need to increase and emphasize cooperation in LT teaching and research training – both cooperation between universities and countries, and also cooperation between different fields such as linguistics, computer science, statistics, etc. There have been proposals to start a common Nordic Master's Program but due to lack of funding, it has not been possible to put them into action. It is essential for Nordic LT to find some ways to continue cooperation in this area.

Although both the ICLT and the Linguistic Institute of the University of Iceland are members of CLARIN, Iceland is unfortunately not a member of the CLARIN consortium and thus does not get any funding from the project. Due to lack of domestic resources, Icelandic members have therefore been unable to participate in CLARIN activities. Iceland would obviously have much to gain from the ongoing and planned cooperation within CLARIN, but as things stand, it does not look as if we will be able to take active part in this cooperation in the foreseeable future. It must be a priority task for us to find ways to change this.

6 Conclusion

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. We have also outlined the research plan of the Icelandic LT community for the next few years. In addition to its contribution to the building of an Icelandic BLARK, the project aims at developing low-cost methods for building language resources for less-resourced languages. In this respect, we emphasize the importance of open source policy for language resources. Finally, we discuss some ideas for Nordic cooperation on Language Technology, especially as regards compilation and dissemination of information and on LT teaching.

References

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

Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. [Morphological Description of Modern Icelandic.] In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 23-25.

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

Hanne Fersøe. 2005. Network of Nordic Language Technology Documentation Centres (NorDokNet). In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 13-16.

Mikel L. Forcada. 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.

Hjálmar Gíslason. 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.

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

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

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. [Tagging Icelandic Text.] *Orð og tunga* 9, pp. 75-107.

Henrik Holmboe. 2005. *Nordisk sprogteknologisk forskningsprogram 2000-2004. Epilog*. NordForsk, Oslo, Norway.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In Bengt Nordström and Aarne Ranta (Eds.), *Advances in Natural Language Processing*. Lecture Notes in Computer Science, Vol. 5221. Springer, Berlin, Germany, pp. 205-216.

Anton Karl Ingason, Skúli Bernhard Jóhannsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir, and Hrafn Loftsson. 2009. Context-Sensitive Spelling Correction and Rich Morphology. *Proceedings of NODALIDA 17*.

Steven Krauwer. 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.

Krister Lindén, Kimmo Koskenniemi, and Torbjørn Nordgård (Eds.). 2006. *The Nordic Countries - A Leading Region in Language Technology*. http://forums.csc.fi/kitwiki/pilot/view/Main/LTExpertPanelReport.

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

Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1), 47-72.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, Estonia, pp. 128-135.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2008. Linguistic Richness and Technical Aspects of an Incremental Finite-State Parser. *Partial Parsing 2008. Between Chunking and Deep Parsing*. LREC 2008 Workshop, Marrakech, Morocco, pp. 1-6.

Anna Björk Nikulásdóttir and Matthew Whelpton. 2009. Automatic Extraction of Semantic Relations for Less-Resourced Languages. In *Proceedings of WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. Workshop at NODALIDA 17.*

Rögnvaldur Ólafsson. 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.

Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson, and Þorgeir Sigurðsson. 1999. *Tungutækni. Skýrsla starfshóps.* [Language Technology. Report of a Committee]. Ministry of Education, Science and Culture, Reykjavík, Iceland.

Eiríkur Rögnvaldsson. 2004. The Icelandic Speech Recognition Project *Hjal*. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 239-242.

Eiríkur Rögnvaldsson. 2005. Icelandic Documentation Center for Language Technology. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 31-33.

Eiríkur Rögnvaldsson. 2008. Icelandic Language Technology Ten Years Later. *Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages*. SALTMIL Workshop, LREC 2008, Marrakech, Morocco, pp. 1-5.

Eiríkur Rögnvaldsson, Björn Kristinsson, and Sæmundur Þorsteinsson. 2006. Nýr íslenskur þulur að koma á markað. [A New Icelandic Text-to-Speech System.] *Morgunblaðið*, January 20th.

Friðrik Skúlason. 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.

Trond Trosterud. 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.

Helga Waage. 2004. Hjal – gerð íslensks stakorðagreinis. [The Making of an Icelandic Isolated Word Recognizer.] In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 47-53.