The status and prospects of Icelandic Language Technology

Eiríkur Rögnvaldsson, University of Iceland

0. Introduction

The title of my talk is "The status and prospects of Icelandic Language Technology". If I had given a talk on this subject four years ago, it would only have lasted a couple of minutes. At that time, Icelandic language technology hardly existed at all. We had a relatively good spell checker, a not so good speech synthesizer, and that was all. There were no programs or even single courses on language technology or computational linguistics at the University, there was no research in these areas, and no software companies were working on language technology. Fortunately, all of this has now changed, to some extent at least. In 2001, the Icelandic Government launched a special Language Technology Project, with the aim of supporting institutions and companies to create basic resources for Icelandic language technology work. This initiative has resulted in several projects which are either finished or well underway. In my talk, I will give an overview of the most important of these projects, and then speculate a little on the prospects of language technology in Iceland.

1. Proposals of the Language Technology Committee

In the fall of 1998, the Minister of Education, Science and Culture, Mr. Björn Bjarnason, appointed a special committee to investigate the situation in Icelandic Language Technology. 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. In the report, four types of actions were proposed in order to establish Icelandic language technology:

1. The development of common linguistic databases that can be used by companies as sources of raw material for their products.
2. Investment in applied research in the field of language technology.
3. Financial support for companies for the development of language technology products.
4. Development and upgrading of education and training in the field of language technology and linguistics.

All of this has been done to some extent. The majority of the money has been spent on the first item. The Institute of Lexicography received a grant for building a full-form morphological database of Icelandic. This database was finished in early 2004 and contains almost 180,000 lexemes and 2.4 million inflectional forms. The Institute of Lexicography has also received a grant for building a balanced tagged corpus of Modern Icelandic. This corpus will contain 25 million words and shall be finished in mid-2007. Furthermore, a consortium consisting of the University of Iceland and four private companies has collected material and created a transcribed wordlist for training speech recognizers.

Money has also been spent on applied research projects. Frisk Software has been working on an HPSG-based syntactic parser with the aim of developing grammar and style checking software for Icelandic. The Institute of Lexicography and the University have also been evaluating and adapting PoS taggers for Icelandic. Furthermore, Frisk Software has received a grant for improving their spell checker, *Púki*, which has been on the market for several years.

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 students have either a BA degree in the humanities (languages and linguistics) or a BS degree in computer science. The students have had the opportunity to participate in most of the language technology projects that have been going on for the past three years. The first student graduated with an MA in Language Technology last October. A few more students will graduate this year and the next, but due to lack of funding, the future of the program is rather uncertain, even though it has been decided that some language technology courses will be taught during the next academic year.

2. Individual priority tasks and their implementation

In the report on Icelandic language technology, the following was stated:

> 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, and in the following text, I try to estimate to what extent each task has been fulfilled.

1. *The main computer programs on the general market (Windows, Word, Excel, Netscape, Internet Explorer, Eudora,...) should be available in Icelandic.*
   Last year, an Icelandic version of Windows XP (including Internet Explorer) and Microsoft Office came on the market. I have used the Icelandic Windows XP on my computers since last fall and I really like it; it does not seem to suffer from any technical bugs, as was the case with the first attempt to translate Windows into Icelandic.

2. *It should be possible to use the Icelandic non-ASCII characters (áéíóúýðþæö ÁÉÍÓÚÝÐÞÆÖ) in all circumstances: in computers, mobile telephones, teletext and other applications used by the public.*
   Here the situation has improved somewhat, in part because of the extended use of Unicode. Many mobile phones now have Icelandic characters in the menus, but they can't always be used in SMS messages, for instance.

3. *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 has funded two major projects in this area: A grammatical tagger for Icelandic and an HPSG-based syntactic parser. However, the committee mentioned two prerequisites for progress in this field:

   3.1 *Establishment of a large computerized text corpus including Icelandic texts of a wide variety of types.*
   Work on such a corpus has recently started, as mentioned earlier in this talk.

   3.2 *Establishment of a grammatically and semantically analyzed lexicon.*
   No such lexicon exists, nor is being prepared. However, many types of raw material for such a lexicon do exist, both in the morphological database built by the Institute of Lexicography and in many collections of that institute.

4. *Good auxiliary programs should be developed for textual work in Icelandic, i.e. for hyphenation, spell-checking, grammar correction, etc.*
   When this was written six years ago, we had the spell-checking program *Púki* from Frisk Software, which has now been improved. The Dutch company Polderland has also developed a spell-checking program, which comes with Microsoft Office. No grammar checking or style checking programs exist, but the syntactic parser developed by Frisk Software is meant to lay a foundation for the development of such programs.

5. *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.*
   An Icelandic speech synthesizer that was originally made around 1990 has been improved. It now uses a more recent technology than the original version, but nevertheless, its pronunciation is far from being satisfactory for use in commercial applications. Preparatory work for the making of a new speech synthesizer has been going on for some time, and it is to be hoped that the actual work can start very soon so that the product will be ready by the end of this year.

6. *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 their efforts to build the first Icelandic speech recognizer in cooperation with ScanSoft, Inc. The performance of the system has turned out to be quite satisfying; the recognition rate appears to be at least 97%. However, no attempts have been made to develop a system for recognizing continuous speech.

7. *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 and several people have used auxiliary software such as translation memory, but no usable translation software is being developed, as far as I know.

8. *Certain parties (institutions or companies) should be given responsibility for individual projects.*
   The 1999 report led to the establishment of the Language Technology Project and its steering committee, which should initiate language technology projects and coordinate actions. It is safe to say that this was a fortunate decision. However, the Language Technology Project ended at the end of last year, and it is unclear how and by whom its work will be continued.

3. The price and prospects of Icelandic Language Technology

The Language Technology Committee estimated that it would cost around one billion, one thousand million Icelandic krónur, or about 12.5 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 for money from the Language Technology Project, and that would be made available on an equal basis to all who were going to use these resources in their commercial products. Even though the Language Technology Project has had a great impact on the development of Icelandic language technology, the fact remains that its total budget over the lifespan of the project was only 133 million Icelandic krónur, or around 1.6 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 still less than 300,000 people speaking Icelandic, and that is not enough to sustain costly development of new products.

Unfortunately, the Language Technology Project suffered from the same two fundamental flaws as a number of other government-funded time-limited projects in Iceland – and no doubt also elsewhere. The first flaw is the short time span of the projects. It takes time to build education, research, development and industry from scratch. Four or five years are simply not enough time. The second flaw lies in the distribution of funding over the project period. Usually, the majority of the money is used during the first part of the project period, before the research community and the industry is really prepared to use the money in the most sensible and profitable manner. In the latter part of the project period, the money available becomes less and less, and finally, when people have been educated and a research and development environment has been created, no money is left in the project. To be sure, the Language Technology Project has been very successful, but 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.

It costs just as much to build language resources for a language with 300,000 speakers like Icelandic as for languages with hundreds of millions of speakers. This means that if we want Icelandic language technology, we will have to pay for it some way or another. This raises a basic question: Should this extra price be paid from the state budget, or should we leave it to the individuals to decide whether they buy Icelandic

language technology products, or cheaper products that can only handle English? I think the answer is quite clear in this case. Obviously, Icelandic language technology products must be competitive in price. We can protect – or try to protect – our agriculture by customs, import restrictions, etc., but we cannot protect our language in the same manner. People may be ready to pay a little more for language technology programs and tools in their own language, but if the difference is substantial, people will resort to the foreign – almost always English – products. That is no big deal for most people. We are quite used to having English around us all day anyway, in all kinds of situations. Then we come to another equally important question: Are we prepared to pay this extra price in our taxes? I really don't know. The decision is of course not up to the general public, except indirectly through general elections, but rather up to the politicians – the parliament and the government. Do they realize the importance of language technology? That remains to be seen.

If private companies don't get more public support, I'm afraid we won't see the development of many new Icelandic language technology products in the near future. However, I believe language technology research will survive. Our participation in Nordic cooperation has been vital in this respect. The Nordic Language Technology Research Programme has funded language technology Documentation Centers in the five Nordic countries. At the end of last year, the Icelandic center merged with the website www.tungutaekni.is, which the Language Technology Project had been running since its start in 2001. 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. As a result of having made those contacts, we now participate in several applications to Nordic and European research funding bodies. If we are successful in some of these applications, it will be a great boost to Icelandic language technology research in the next few years.

Another important aspect of the Nordic cooperation in language technology is the Nordic Graduate School of Language Technology, funded by NorFA – now NordForsk. The activities of the school started last year and will run for five years. Even though the school is primarily intended for doctoral students, our master's level students have been admitted to its courses. This is absolutely crucial for us, since we do not have the capacity to give the students high-quality education in language technology at home.

4. Language technology and the future of Icelandic

When we try to estimate the importance of Icelandic language technology we have to realize that information technology has become an important and integrated feature of the daily life of almost every single Icelander. If we cannot use Icelandic within information technology we 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 happens 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 much about this scenario to see the signs of imminent danger.

But the need for Icelandic language technology is not, and should not, only be driven by our wish to protect and preserve our language. It is equally or even more important to look at this from the speakers' point of view. They 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 the Icelandic language community. Everything else is a retreat, which we should not accept. It is of course clear that we will never have everything in Icelandic. The small size of the language community means that we will always have to make some compromises. We don't object to having the letters R, N, and P on the gear shift lever in our cars, and many of us don't even realize that these letters stand for the English words *reverse*, *neutral*, and *park*. In this environment, these letters are just signs, independent of any particular language. But a language in dynamic use, language in context, cannot be disconnected from its origins like single letters can. Therefore, we must be able to use Icelandic in as many and diverse contexts as possible. Otherwise, we will be linguistically oppressed in our own language community.

Last October, I was a member of a delegation from the Nordic documentation centers in language technology, which visited the Baltic countries in order to make contacts with people involved in language technology in these countries. This was an extremely interesting trip and we learned about many ambitious projects. One thing we found particularly interesting was a white paper entitled "Development strategy of the Estonian language", which has recently been approved by the Estonian government. This paper is accompanied by "Estonian HLT Roadmap for 2004-2111", with detailed descriptions of actions to be taken each year. The Estonians are a small nation, with only about one million people having Estonian as their native language. As far as I can see, they are on about the same level with respect to language technology as we are. They are more advanced in certain areas but lag behind us in others. However, there is one important, and in fact crucial, difference. Their Language Technology program is just starting whereas our program has ended. Thus, we must ask ourselves: Are we satisfied with that?