# Language Resources for Icelandic

*Sigrún Helgadóttir*[1]*,Eiríkur Rögnvaldsson*[2]

(1)Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík Iceland
(2)University of Iceland, Reykjavík Iceland

sigruhel@hi.is, eirikur@hi.is

ABSTRACT

We describe the current status of Icelandic language technology with respect to available language resources and tools. The recent META-NET survey of the state of language technology support for 30 languages clearly demonstrated that Icelandic lags behind almost all European languages in this respect. However, it is encouraging that as a result of the META-NORD project, almost all basic language resources for Icelandic are now available through the META-SHARE repository and the local site http://www.málföng.is/, many of them in standard formats and under standard CC or GNU licenses. This is a major achievement since many of these resources have either been unavailable up to now or only available through personal contacts. In this paper, we describe briefly most of the major resources that have been made accessible through META-SHARE; their type, content, size, format, and license scheme. It is emphasized that even though these resources are extremely valuable as a basis for further R&D work, Icelandic language technology is far from having become self-sustaining and the Icelandic language technology community will need support from partners in the Nordic countries and Europe if Icelandic is to survive in the Digital Age.

KEYWORDS: Icelandic, Language Resources, Repositories, Licenses.

# 1 Introduction

According to the survey of language technology support for European languages recently conducted by META-NET (http://meta-net.eu) and published in the series "Europe's Languages in the Digital Age", Icelandic is among the European languages that have the least support (Rögnvaldsson et al., 2012). This is not surprising. The Icelandic language community with its 320,000 speakers is by far the smallest in the survey – only Maltese comes close in the number of speakers, and is also on the same level as regards language technology support. It is well known that the cost of preparing a language for the digital age is independent of the number of speakers. The same basic resources are needed, irrespective of the size and capacity of the language community. Icelandic lacks both financial and human resources to be able to follow the changes that the digital revolution has made – and will make in the near future – to the use of human language within information technology, and on human-computer interaction.

That doesn't mean, however, that the situation of Icelandic is hopeless. Even though most advanced and high-level tools and resources are lacking, a number of basic resources have been built during the last decade. This includes a PoS tagger, a lemmatizer, and a parser; a morphological database containing 270,000 paradigms; a tagged corpus of 25 million words; a multilingual dictionary with 50,000 entries; a large collection of terminologies; a treebank containing one million words; and a number of other smaller yet valuable resources. These resources form a solid ground to build on. Of course, Icelandic will always lag behind languages with millions of speakers in language technology support. It will be necessary to prioritize and select carefully which resources it is absolutely vital to develop.

For the future of Icelandic, it is extremely important that all of the resources mentioned above have now been made accessible and open to the linguistic and language technology communities at large. This should enforce R&D work on Icelandic and contribute to securing the survival of the language in the digital age. In this respect, the situation has changed dramatically during the last two years due to Iceland's participation in the META-NORD project (http://meta-nord.eu). Before the project started, some of the above-mentioned resources were indeed available, but information on their availability was not widely spread. META-NORD has made a large contribution to finalizing and standardizing some of these resources, and the META-NORD team spent a lot of effort convincing the owners of some other of these resources to make them public.

In this paper, we give an overview of the current status of language resources for Icelandic. Section 2 summarizes the main results of the META-NET survey of language technology support. Section 3 deals with language resource repositories, both META-SHARE and the Icelandic repository http://www.málföng.is/. In Section 4, we describe briefly the most important Icelandic language resources; their content, format, availability, etc. Finally, Section 5 is a conclusion.

# 2 Icelandic language resources in a European perspective

In September 2012, Springer Verlag published a series of 31 white papers entitled "Europe's Languages in the Digital Age" (http://www.meta-net.eu/whitepapers/overview). These volumes present the result of a study conducted by META-NET, a European Network of Excellence dedicated to building the technological foundations of a multilingual European information society. Each white paper describes one European language – its characteristics

| | Quantity | Availability | Quality | Coverage | Maturity | Sustainability | Adaptability |
|---|---|---|---|---|---|---|---|
| **Language Technology: Tools, Technologies and Applications** | | | | | | | |
| Speech Recognition | 1 | 1 | 1 | 1.5 | 1 | 0 | 1 |
| Speech Synthesis | 1 | 1 | 2.5 | 2.5 | 2 | 1 | 1 |
| Grammatical analysis | 2 | 5.5 | 4 | 3 | 3.5 | 3.5 | 3 |
| Semantic analysis | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Text generation | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Machine translation | 1 | 4 | 1 | 1.5 | 1.5 | 1.5 | 2 |
| **Language Resources: Resources, Data and Knowledge Bases** | | | | | | | |
| Text corpora | 1.5 | 4 | 3 | 2.5 | 2.5 | 4.5 | 3 |
| Speech corpora | 1 | 2 | 1.5 | 1.5 | 1 | 1.5 | 1.5 |
| Parallel corpora | 1 | 1 | 1 | 0.5 | 1 | 1 | 1 |
| Lexical resources | 1 | 2 | 2.5 | 2.5 | 2 | 2 | 2 |
| Grammars | 1 | 4 | 2.5 | 2 | 2.5 | 2.5 | 2 |

Figure 1: State of language technology support for Icelandic

and particularities, its status in the society and in an international context. The main purpose of the papers, however, is to describe the status of each language with respect to technological support – language resources and tools. Experts were asked to estimate the status of the language in 11 different subfields on a scale ranging from 0 (very low) to 6 (very high) using seven different criteria. The results for Icelandic are shown in Figure 1 (Rögnvaldsson et al., 2012, page 60).

In these white papers, the state of language technology support is also compared across all 31 languages. This comparison is based on four key areas: speech processing, machine translation, text analysis, and speech and text resources. The languages were placed in one of five possible categories for each area. It turns out that Icelandic is one of only four languages that are placed in the bottom category (categorized as having weak or no support) for all these four areas – the other three being Latvian, Lithuanian and Maltese. If we compare the eight META-NORD languages (Nordic and Baltic) across all four areas, Icelandic ranks lowest of them as shown in Figure 2 (Vasiljevs et al., 2012).

The low ranking of Icelandic in this comparison is hardly surprising. Serious work on Icelandic language technology only started in the beginning of the century (Rögnvaldsson

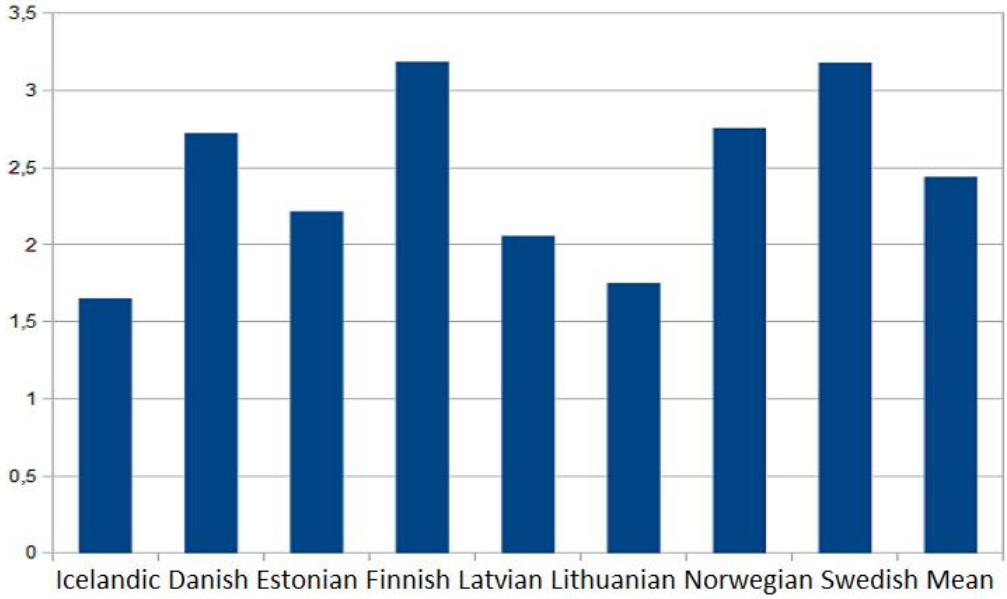


Figure 2: Average scores for each of the META-NORD languages

et al., 2009). At that time, no tools and almost no digitized language resources existed. During the last decade, the tiny Icelandic language technology community has managed to build a number of important tools and resources. Valuable language resources have also been built by Icelandic lexicographers and terminologists. Thus, even though the relative ranking of Icelandic is not encouraging, Icelanders can be proud of what they have achieved.

Two projects have been essential in providing the atmosphere and the financial basis for this work. One is the Language Technology Programme which the Minister of Education, Science and Culture initiated in 2000 (Rögnvaldsson et al., 2009). This program, which lasted for four years, funded the building of several basic resources and tools. The other important project is META-NORD. As a result of Iceland's participation in that project, most of the basic language resources for Icelandic are now easily accessible under open source licenses.

## 3 Málföng.is

During the last two years, Iceland has participated in META-NET through the META-NORD project which comprises all the Nordic and Baltic countries (Vasiljevs et al., 2012). The main goals of META-NORD, as of the sister projects CESAR in Eastern Europe and METANET4U in Southern Europe, were to strengthen the status of language technology in Europe, to increase awareness about the opportunities and challenges of language technology, and to make language resources and tools for all European languages more open and accessible.

One of the main aims of META-NET and the related projects was to build data repositories where language resources could be stored centrally. This aim was fulfilled by the launching of META-SHARE (`http://www.meta-share.eu/`). META-SHARE is a distributed repository with a number of nodes which are meant to be synchronized. Presently there is a number of managing nodes where all metadata which has been recorded in any META-SHARE node can

be accessed. The network will gradually be extended to encompass additional nodes and centers and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

With the exception of Iceland, META-SHARE nodes have been established in all the META-NORD countries. Due to lack of human, technical and financial resources, the Icelandic META-NORD partner, the University of Iceland, has not yet been able to establish its own META-SHARE node. Instead, the META-SHARE node established by Tilde serves as a mother node for Icelandic language resources.

However, the University of Iceland is working on the formation of a national consortium, consisting of UI, the Árni Magnússon Institute for Icelandic Studies (AMI), the Reykjavik University, the National and University Library, and perhaps a few others to maintain an Icelandic national repository of language resources. These plans will hopefully be realized in the course of 2013.

Instead of establishing a local META-SHARE node, it was decided to launch a new website called `http://www.málföng.is/` (málföng is a neologism for 'language resources'). The purpose and structure of this website is partly different from that of META-SHARE. First, the user interface and all the documentation is in both Icelandic and English. Second, the scope is meant to be wider than that of META-SHARE. The website `http://www.málföng.is/` will not only accept resources that are "complete" and "standard" in some sense, but also incomplete resources, work in progress, resources that do not follow any established standards, etc. Moreover, `http://www.málföng.is/`will contain resources of different types than those in META-SHARE – all kinds of research results, language technology papers, etc.

# 4 Main types of Icelandic language resources and tools

## 4.1 Overview

The Icelandic META-NORD team managed to get hold of almost all of the most important language resources for Icelandic as regards tools and textual resources. As for spoken language resources, a number of valuable resources have been recorded in META-SHARE and are available through `http://www.málföng.is/`, but others are either proprietary or not yet available.

Metadata was entered into the META-SHARE node at Tilde for 23 language resources (11 corpora, 9 lexical conceptual resources and 3 tools services). Of these 2 are tools for processing Icelandic text, one is a language independent tool and 20 are resources containing Icelandic text. None of the resources available for download were uploaded to the META-SHARE node. Instead links are provided as part of the metadata to `http://www.málföng.is/` where some of the downloadable resources are located or to websites such as `http://sourceforge.net/` and `https://github.com/`. At the time of writing 7 additional resources are available through META-SHARE that contain Icelandic text.

In the following, we will review the main types of tools and resources that were harvested, and describe the work carried out by the META-NORD team in order to prepare these resources for inclusion in META-SHARE.

## 4.2 Tools

The first PoS tagger for Icelandic was developed by Stefán Briem during the preparatory work for the *Icelandic Frequency Dictionary* (IFD, Íslensk orðtíðnibók (Pind et al., 1991)), see Section 4.3. This tagger was never publicly released and information on its structure and performance is scanty. During the years 2001–2003, several taggers were trained on Icelandic texts (the source files from the IFD). Thorsten Brant's TnT gave the best results (Helgadóttir, 2007). The training model developed in this project, together with the source texts from the IFD, has been available for researchers under certain conditions. However, it has not been publicly advertised nor uploaded to any software repository.

Hrafn Loftsson started developing an open source software package for analyzing and processing Icelandic texts during his Ph.D. studies from 2004–2007 (Loftsson, 2007, 2008; Loftsson and Rögnvaldsson, 2007). Since then, students at the University of Reykjavík and the University of Iceland have helped in developing individual components. The software, which goes by the name of IceNLP, is rule-based and uses heuristic methods which guess prepositional phrases and syntactic functions and use the acquired knowledge to force feature agreement where appropriate.

IceNLP is implemented in Java and consists of the following components: tokenizer, unknown word guesser, part-of-speech tagger, lemmatizer, parser and named-entity recognizer. Anton Karl Ingason is the main author of the lemmatizer (*Lemmald*, cf. (Ingason et al., 2008)). Individual components of IceNLP can be run independently or the JAVA clusters in question connected directly to software that is being developed.

IceNLP can be used for various tasks, such as breaking up text into individual tokens, tagging each token with its morphosyntactic tag, finding the lemma of a particular word and returning a shallow phrase structure and labels indicating syntactic functions. The package is downloadable under the LGPL (GNU Lesser General Public License), either directly from sourceforge (`http://icenlp.sourceforge.net/`) or via META-SHARE or `http://www.málföng.is/`.

Two rule-based machine translation systems between Icelandic and other languages have been developed. One is *Tungutorg* (`http://tungutorg.is/`) which translates between Icelandic and English, both ways, and also from Icelandic to Danish and from Esperanto to Icelandic. This system, which was developed by Stefán Briem, is closed and the source has not been released. Hence, it has not been registered in META-SHARE.

The other system is *Apertium-is-en* (`http://nlp.cs.ru.is/ApertiumISENWeb/`), a prototype of a shallow-transfer machine translation system that translates Icelandic text into English. The system is based on the Apertium translation system (`http://www.apertium.org/`). It was developed in the years 2009–2010 at the University of Reykjavík as the MSc project of Martha Dís Brandt as well as in independent projects of two other students, under the guidance of Hrafn Loftsson (Brandt et al., 2011). The Apertium system is downloadable under the GPL (GNU General Public License) from sourceforge (`http://sourceforge.net/projects/apertium/`) or via META-SHARE or `http://www.málföng.is/`.

## 4.3 Corpora

In the META-NORD project metadata for 6 Icelandic text corpora were entered into META-SHARE and the corpora made available through `http:www.málföng.is/`. One of those, IcePaHC, is a treebank and will be described in Section 4.4. The other five corpora will be described briefly in this section. Five corpora containing both text and sound files were also made available. These will be described in Section 4.8.

The largest of the text corpora, *Íslenskur Orðasjóður*, is a very large corpus of modern Icelandic that was compiled in two research projects: *Leipzig Corpora Collection* and *Frequency Dictionary Icelandic* (Hallsteinsdóttir et al., 2007; Quasthoff et al., 2012). The corpus consists of 5 sub-corpora. The two largest portions are texts from domains ending in .is collected in the autumn of 2005 by the National and University Library of Iceland (ca. 227 million running words) and the same collected in the autumn of 2010 (ca. 336 million running words). The three remaining portions contain text from a newspaper (Morgunblaðið) collected in 2001 (ca. 18.1 million running words), newspaper text from the Internet crawled in 2011 (ca. 22.6 million running words) and texts from the Icelandic edition of Wikipedia (ca. 2.5 million running words).

The corpus comes with an automatically generated monolingual lexicon, comprising frequency statistics, samples of usage, cooccurring words and a graphical representation of the word's semantic neighbourhood (Hallsteinsdóttir et al., 2007). Despite some limitations, this corpus is the only very large corpus of Icelandic in existence and it has proven to be useful in several projects. Of these, it is worth mentioning a project to create a Database of Semantic Relations (Nikulásdóttir and Whelpton, 2010), and projects to develop context sensitive spelling correction for Icelandic and the correction of OCR texts obtained from old print (ongoing unfinished projects).

The oldest Icelandic tagged corpus is the IFD corpus (Pind et al., 1991) which was compiled for the making of the Icelandic Frequency Dictionary, *Íslensk orðtíðnibók*, published in 1991. The IFD corpus consists of just over half a million running words, containing 100 fragments of texts, approximately 5,000 running words each. The corpus has a heavy literary bias as about 80% of the texts are fiction. The tagset of the IFD is more or less based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized. The underlying tagset contains about 700 tags, of which 639 tags actually appear in the corpus. The tags are character strings where each character has a particular function, denoting a (specific value of a) grammatical category. The tagging and lemmatization of the IFD corpus was manually corrected and hence the corpus can be used as a gold standard for training part-of-speech (PoS) taggers and lemmatizers. All data-driven taggers used now for tagging Icelandic text are trained on the IFD corpus and it was used for the development of the rule-based tagger IceNLP (Loftsson, 2008).

*The Tagged Icelandic Corpus* (MÍM) was finished during the META-NORD project period (Helgadóttir et al., 2012). The corpus was originally financed by the Language Technology Programme initiated in 2000. The MÍM corpus is a synchronic corpus that contains about 25 million running words compiled at the AMI during the years 2004–2012. The texts were taken from different genres of contemporary Icelandic, i.e. texts produced in 2000–2010. The corpus is intended for use in Language Technology projects and for linguistic research. The aim of the project was to produce a balanced collection of contemporary texts, mor-

phosyntactically tagged and lemmatized and supplied with metadata in TEI-conformant XML format (Burnard and Bauman, 2008). The texts are now available for download and search via META-SHARE or http://www.málföng.is/.

To make the corpus as useful as possible in LT projects it was considered of utmost importance to secure copyright clearance for the texts to be used. It was anticipated that most of the texts would be protected by copyright (final figure is about 88.5%). Permission was sought from all owners of copyrighted texts included in the MÍM corpus. Official texts (e.g. law, judicial texts, regulations and directives) are not copyrighted (11.5%). All copyright owners signed a special declaration and agreed that their material may be used free of licensing charges. In turn, AMI agrees that only 80% of each published text is included and that copies of the MÍM corpus are only made available under the terms of a standard license agreement. The crucial point in the license agreement is that the licensee can use his results freely, but may not publish in print or electronic form or exploit commercially any extracts from the corpus, other than those permitted under the fair dealings provision of copyright law. Data induced from the corpus, for example by a statistical PoS tagger, is considered results and may be used in commercial products. The license granted to the licensee is non-transferable.

The budget of the project did not allow for extensive collection and transcription of spoken language. Through collaboration with other projects, it was, however, possible to secure some spoken language data. It consists of about 500,000 running words of transcribed text which is about 2.2% of the corpus. The spoken data was obtained through four different projects (Thráinsson et al., 2007) and it includes transcriptions of about 54 hours of natural speech, recorded in different settings in the period 2000–2006. The collection contains monologues, interviews and spontaneous conversations between adults of both sexes and with different backgrounds. All the recordings have been carefully transcribed in a predefined format. The transcribed texts are a part of the downloadable MÍM corpus and are available for search with the rest of the corpus. All names have been substituted with pseudonyms, and other personal data has been removed. The monologues part of the spoken text is debates from unprepared sessions in the Icelandic Parliament, recorded in 2004-2005. The transcribed texts together with the sound files of the spoken text will be made available for search at a later date. Search in all texts except the parliamentary speeches will be password protected. The transcribed text files and sound files with the parliamentary speeches form a separate corpus that will be described in Section 4.8.

Annotation of the corpus was performed in three steps: sentence segmentation and tokenization, morphosyntactic tagging and lemmatization and transcription into TEI-conformant XML format together with relevant metadata. The procedure and software used for sentence segmentation, tokenization, morphosyntactic tagging and lemmatization was explained by (Loftsson et al., 2010) in their work on the *MIM-GOLD* corpus. The tagset used was developed for the *IFD* corpus. The automatic morphosyntactic tagging accuracy has been estimated as 88.1-95.1%, depending on text type (Loftsson et al., 2010). The corpus was transferred into TEI-conformant XML format as a part of the META-NORD project. The Norwegian search interface Glossa (Johannessen et al., 2008), which in turn uses the IMS Corpus Workbench (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/) as a search engine, was adapted to be used with the MÍM corpus and other tagged corpora that are available through http://www.málföng.is/.

The MIM-GOLD corpus (Loftsson et al., 2010) is a corpus of about 1 million running words which has been sampled from MÍM. This corpus is intended as a reliable standard for the development of LT tools. The intention is that tagging and lemmatization of this subcorpus will be manually corrected. At the time of writing the tagging has been corrected by one annotator. The files are made available as version 0.9 through META-SHARE and `www.málföng.is/` for download. In later phases of the MIM-GOLD project tagging will be checked and accuracy estimated. Lemmatization will also be revised.

The last text corpus to be mentioned that is available via META-SHARE and `http://www.málföng.is/` is the *Saga corpus*. The corpus consists of 44 electronic texts of some of the Old Icelandic sagas: Family Sagas (Íslendingasögur), Sturlunga Saga, Sagas of the Kings of Norway (Heimskringla) and the Book of Settlement (Landnámabók). The texts have been normalized to Modern Icelandic spelling (Rögnvaldsson and Helgadóttir, 2011). Several inflectional endings were also changed to Modern Icelandic form.

## 4.4 Treebanks

The Icelandic Parsed Historical Corpus or IcePaHC (Wallenberg et al., 2011; Rögnvaldsson et al., 2011) is a one million word treebank containing material from both Modern Icelandic and older stages of the language. It is the product of three different projects which originally had different aims. The earliest and largest of these projects was a subpart of a large language technology project which had the aim of developing three different basic language resources for Icelandic. The aim of this subproject was to build a treebank of Modern Icelandic for use in language technology and to develop efficient parsing methods and tools for less resourced languages. Since some of the participants had been involved in historical syntax research, they also wanted to include a few texts from older stages of the language. However, the main emphasis was on language technology use – the corpus was intended to be a suitable training material for a statistical parser for Modern Icelandic.

At the same time, two other projects with the aim of developing resources for studying diachronic Icelandic syntax were in preparation. After some discussion, the participants in these three projects decided to join forces and make a combined effort to build a large parsed corpus covering the history of Icelandic syntax from the earliest sources up to the present. This corpus thus serves the dual purpose of being one of the cornerstones of Icelandic language technology and being an invaluable tool in Icelandic diachronic syntax research.

IcePaHC is a phrase structure treebank which uses the same general type of labeled bracketing as the Penn Treebank (with dash-separated lemmata added). The Penn annotation scheme had already been adapted for Old English (Taylor et al., 2003), which is rather similar to Icelandic in many respects, both as regards the syntax and the morphological system. Thus, the scheme could be applied to Icelandic with only slight modifications.

IcePaHC is designed from the beginning to serve both as a language technology tool and a syntactic research tool, and developed by people with research experience in both diachronic syntax and computational linguistics. Most parsed corpora are developed either for language technology use (such as the Penn Treebank, `http://www.cis.upenn.edu/~treebank/`) or for syntactic research (such as the Penn Parsed Corpora of Historical English, PPCHE, `http://www.ling.upenn.edu/hist-corpora/`.

The usefulness of the corpus as a tool for diachronic syntax research has already been demonstrated in a number of papers. The corpus has not yet been put to use within

language technology but there is no reason to doubt that it can serve that purpose too. The corpus contains around 300,000 words which can safely be considered Modern Icelandic – texts from the $19^{th}$, $20^{th}$ and $21^{st}$ centuries. That is more than enough material to train a statistical parser.

The corpus is completely free and open without any registration or paperwork, and the same goes for all the software that has been used to build it and the software that was developed within the project. Both the software and the corpus itself are distributed under the LGPL license and can be downloaded from the IcePaHC home page (`http://www.linguist.is/icelandic_treebank/Download`) or via META-SHARE. The treebank has also been uploaded to the INESS repository at the University of Bergen (`http://iness.uib.no`) where it may be viewed and searched.

In addition to IcePaHC, two small Icelandic treebanks exist and have been uploaded to META-SHARE. One is a dependency treebank containing Icelandic translation of the first part of the Norwegian novel *Sophie's world* (*Sofies verden*) by Jostein Gaarder. This text was annoted within the Nordic Treebank Network. The other is a small fragment of the JRC-Aquis text which was annotated within META-NORD. Both texts have been aligned with texts from several other languages.

## 4.5 Dictionaries

In November 2011 *ISLEX – Icelandic Scandinavian Web Dictionary* was opened to the public. The project began in 2005 and is a collaboration of five institutes: AMI in Reykjavík, Iceland, The Danish Society for Language and Literature (DSL) in Copenhagen, Denmark, The Department of Linguistic, Literary and Aesthetic Studies at Bergen University, Norway, and the Department of Swedish at Gothenburg University, Sweden. The project has mainly been financed by the governments of these countries. The administration of ISLEX is in Reykjavík and the Icelandic part of the dictionary is compiled and processed by AMI, and the development of the database and the software is also centred there. The editing of the target languages takes place in the participating countries, each editorial team being responsible for their own target language (Sigurðardóttir et al., 2008).

The dictionary was from the start designed for the web where the possibilities offered by that medium could be used. ISLEX thus contains many images, sounds and hyperlinks. The pronunciation of all Icelandic headwords is given as sounds, and all nouns, adjectives and verbs are linked to the DMII (see Section 4.7). ISLEX is a medium sized dictionary with 50,000 headwords. It describes modern Icelandic with an emphasis on phrases, fixed expressions and examples of use, all of which are translated into the target languages. It is the first comprehensive Scandinavian online dictionary which combines so many languages.

The Icelandic META-NORD team secured consent from the five institutes to make the ISLEX database available through META-SHARE and `http://www.málföng.is/`. A link to the ISLEX search page is provided and the database is downloadable in LMF (Lexical Markup Framework) format (`http://www.lexicalmarkupframework.org/`). The database contains the Icelandic headwords, phrases and fixed expressions and their translations into Danish, Swedish and the two Norwegian language standards and grammatical information such as part-of-speech and gender and number for nouns.

The conversion of the database to the LMF format was done as a part of the META-NORD project. When converting multilingual dictionaries into LMF format, a special record would

usually be made for each sense of every word in all the languages of the dictionary. A so-called "Sense Axis" would then be used to link closely related senses in different languages. For ISLEX we had to take a different route. Special records were made for each sense of the words in the source language, Icelandic, which in turn had translations for that sense in each of the target languages.

## 4.6 Terminologies

A Term Bank was established by the Icelandic Language Institute in November 1997 (Thorbergsdóttir, 2003). The bank contained terminologies from terminological committees and individuals. Some of the terminologies had been published in printed books. In 2006, the Icelandic Language Institute became a part of AMI which now is a curator of the bank.

One of the roles of the Term Bank is to standardize the use of terms within related and unrelated subject fields. The aim is to hinder that many different terms are used for the same concept or phenomenon. The Term Bank provides an overview of Icelandic terminology and topical neologisms and thereby makes it easier to coordinate and standardize term usage. Additionally, the Term Bank provides access to Icelandic translations of foreign terms, and access to definitions of terms in Icelandic and other languages. The Term Bank thus benefits all those who write about specialized topics, such as translators, teachers, students, journalists, government agencies, businesses and any interested people, and last but not least compilers of dictionaries.

As a part of the META-NORD project an agreement was reached with the copyright owners of 41 terminologies in the Term Bank that their terminologies could be made available through META-SHARE and `http://www.málföng.is/`. The terminologies are made available in one package with 41 terminologies with the CC BY-SA license.

Most of the terminologies contain terms in English and Icelandic, just over 103 thousand Icelandic terms in total and just over 104 thousand English terms. Some terminologies also contain terms in other languages e.g. the Nordic languages and German and French. A total of 16 languages are represented in the Term Bank. Some of the terminologies contain definitions or explanations and some even examples of usage and cross-reference between concepts.

The terminologies were transferred into TBX (TermBase eXchange, `http://www.tbxconvert.gevterm.net/`) format which is the standard format used for terminologies. By using the standard the interchange of terminological data including detailed lexical information is made easier. The terminologies from the Icelandic Term Bank in TBX format were easily imported into Eurotermbank.

## 4.7 Grammars

In 2004 an online version of the *Database of Modern Icelandic Inflection* (DMII; Beygingarlýsing íslensks nútímamáls) was opened to the public on the website of the AMI (`http://bin.arnastofnun.is/`). Earlier the same year the database was made available for use in language technology and lexicography (Bjarnadóttir, 2012). The database was created as a multipurpose resource to serve both the general public, teachers and linguists and the language technology community. The database contains about 270,000 paradigms from Modern Icelandic with over 5.8 million inflectional forms. The DMII was originally

financed by the Language Technology Programme initiated in 2000.

As the necessary data for making a productive rule system was not available, the DMII was produced as a database containing the full paradigms for as large a portion of the Icelandic vocabulary as possible. The original source for the DMII was the electronic version of the *Dictionary of Icelandic* (Árnason, 2000) with about 135,000 headwords and the lexicographic archives of the AMI. The author of the DMII, Kristín Bjarnadóttir, has compared the vocabulary of DMII with the vocabulary of the MIM corpus (Section 4.3) and is in the process of adding paradigms to the DMII based on that comparison (Bjarnadóttir, 2012). The vocabulary of the DMII will also be augmented with vocabulary from the Icelandic Term Bank (Section 4.6) (personal communication).

The Icelandic inflectional system is very rich, with up to 16 inflectional forms for nouns, 120 for adjectives and 107 for verbs, not including variants. For each paradigm in the database each lemma is shown in full, including variants. In the version that is used for language technology projects each word form is shown together with the lemma and a morphosyntactic tag.

The database has proven to be extremely useful. The number of visits to the online version of the database has risen every year since its opening. It is used by native speakers of Icelandic, the general public, students and teachers, and students of Icelandic as a foreign language. The DMII has also been used extensively in LT projects such as for search engines, PoS tagging, context sensitive correction, in language teaching and lexicography.

The database has been available for download from the AMI website (`http://ordid.is/forsida/`) free of charge with a proprietary license since 2009. The Icelandic META-NORD team secured permission to include the DMII in META-SHARE and on `http://www.málföng.is/` in such a way that links are provided to the appropriate pages of the website of the AMI.

## 4.8 Spoken language

Five corpora containing transcribed speech and sound are made available through META-SHARE and `http://www.málföng.is/`. Three of these were developed by the researcher Arnar Jensson (Jensson et al., 2008). All three corpora are based on a read bi-phonetically balanced text. The *Jensson Corpus* is 3.8 hours in length with 5,612 utterances from 20 speakers. The *Thor Corpus* is 2 hours in length with 4000 utterances from 20 speakers, 10 female and 10 male. The *RÚV Corpus* is 46 minutes in length with 400 utterances from 20 speakers and contains read news items that include a large vocabulary. No two speakers read the same text. The files belonging to these three projects can be obtained from the developer under the license CC BY-NC-SA.

As mentioned in Section 4.3 transcribed text and sound files with parliamentary debates form a separate corpus, the *Parliament Speech Corpus* which is available for search and download under CC BY 3.0 license via META-SHARE and `http://www.málföng.is/`. The corpus contains twenty hours of speeches from the Icelandic Parliament during the winter of 2004–2005, in synchronized text and sound files. Information about the recordings and the speakers, such as their age and gender, are provided as well. The data is intended to reflect natural spoken Icelandic under formal conditions. The discussion periods were chosen as they primarily consist of unprepared speeches that are unlikely to have been written in

advance and read out loud. In addition, the aim was on diversity of topics and speakers (w.r.t. their origin, age and gender) (Thráinsson et al., 2007).

The *Hjal* corpus is the product of a project to build the first Icelandic speech recognizer during the years 2002–2003 (Rögnvaldsson, 2004). The project was financed by the Language Technology Programme initiated in 2000 and was performed in cooperation with ScanSoft, Inc. Their role was to train the speech recognizer on the basis of the material prepared in the project. The goal of the project was to collect sufficient material to train a speaker independent isolated word recognition system. Since the project was government funded, the data produced are open to all that want to develop a speech recognizer for Icelandic.

Caller sheets were prepared containing words, phrases and sentences for the participants to read. They were to include words and phrases that are likely to be used in ASR (automatic speech recognition) applications; a certain number of person names, place names, company names, numerals, numbers (money amounts etc.), commands, and meaningful fillers (*OK*, *please*, etc.). Furthermore, each sheet should contain five phonetically rich sentences and three strings of isolated letters. The sentences were to be composed in such a way as to get enough samples of all occurring diphones and common triphones in Icelandic.

A word frequency list for Icelandic was also made. The minimum size of the list was to be 30,000 word forms, but due to the inflectional character of Icelandic, it was decided to include about 50,000 word forms in the list. Volunteers were recruited to call in and read the sentences. When valid recordings from 2000 speakers, sufficiently well distributed with respect to gender, age groups, regional dialects, and type of telephone (mobile vs. fixed line), had been obtained, data collection was stopped. The Hjal corpus consists of recordings from 883 speakers of these 2000 speakers together with the transcribed speech. The corpus is available under the license CC BY 3.0.

The word frequency list contains 50–60 thousand word forms. These word forms were transcribed phonetically with both the SAMPA and IPA standards. The list together with the transcriptions is available as an Excel-file under the name *Pronunciation Dictionary for Icelandic* under the license CC BY 3.0.

One of the resources made available through META-SHARE is the Icelandic – Scandinavian web dictionary ISLEX (Section 4.5). The Icelandic headwords have been recorded and the recordings can be accessed through the website (`http://islex.is/`). The resource *ISLEX Recordings* is available through META-SHARE and `http://www.málföng.is/` under the license CC BY-NC-ND.

Finally we will mention two projects that were not a part of META-NORD but are nevertheless very important LT prjojects in the Icelandic context.

The first is *Almannarómur*, a project to develop an Icelandic speech recognizer that was carried out during the winter 2011–2012 (Guðnason et al., 2012). The Almannarómur project is a part of an open source speech project, hosted by Google. The aim of the project was to enable small language communities to generate an open source speech corpus that can be used for research. In the project a database of spoken sentences was created to aid development of automatic speech recognition for Icelandic. However, the database can be used for many other types of spoken language technologies. During the project time 113,547 sentences were recorded by 563 participants through Android phones made available by Google.

The database is not yet available but will be made available to the public via http://www.málföng.is/ in order to develop spoken language technologies for the Icelandic language. For example, the database will be particularly suitable for short utterances in a mobile environment. Google has already used the database to train an acoustic model for Icelandic and made a speech recognizer for Icelandic available through their Android phones.

The second is a project to develop a new speech synthesizer for Icelandic. The Society of the Blind in Iceland instigated in 2010 the development of a new speech synthesizer and made a contract with the Polish company Ivona software (http://www.ivona.com/en/) to carry out the task. The synthesizer was ready in 2012 and contained two voices, a male voice and a female voice. Two actors recorded extensive text material that had been prepared for this purpose. Unfortunately these recordings are not available for other researchers and developers of spoken language technologies for the Icelandic language.

## 5 Conclusion

Even though considerable achievements have been made in building language resources for Icelandic during the last decade, it is clear that Icelandic language technology is not self-sustaining. As pointed out in the *Strategic Research agenda* (SRA) recently published by META-NET (http://www.meta-net.eu/sra-en; (Rehm and Uzkoreit, 2012)), Iceland needs external support in order to be able to follow the rapid development in language technology.

"Not all countries have the required expertise or human resources to take care of the technology support for their languages. For example, in Iceland there is not a single position in LT at any Icelandic university or college and there is only one company that works in this area. Those colleagues who work on LT at universities and research institutes come from either language or computer science departments; their main duties are not related to LT, still they managed to produce a few basic technologies and resources but advanced types of resources do not exist at all for Icelandic, nor do they for many other under-resourced languages. This is why we need to intensify research and establish techniques, methods and instruments for research and knowledge transfer so that colleagues in countries such as Iceland can benefit as much as possible for their own language from the research carried out in other countries for other languages. Bootstrapping the set of core language technologies and resources for all languages spoken in Europe is not a matter of a few countries joining forces but a challenge on the European scale that must be addressed accordingly to avoid digital exclusion and secure future business development." (Rehm and Uzkoreit, editors, 2012, p. 66).

It remains to be seen how such bootstrapping and knowledge transfer methods can be implemented and whether they will suffice to secure the establishment of necessary language resources, tools and technologies for survival in the Digital Age.

## Acknowledgments

The META-NORD project was supported by the EU ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no 270899 (META-NORD). Th authors would like to thank Kristín Bjarnadóttir, Steinþór Steingrímsson and Halldóra Jónsdóttir for valuable assistance.

## References


Árnason, M., editor (2000). *Íslensk orðabók [Dictionary of Icelandic]. 3rd edition, electronic version*. Edda hf., Reykjavík.

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT2012*, pages 13–184, Istanbul.

Brandt, M. D., Loftsson, H., Sigurþórsson, H., and Tyers, F. M. (2011). Apertium-IceNLP: A rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-2011)*, pages 217–224. Leuven.

Burnard, L. and Bauman, S. (2008). *Guidelines for Electronic Text Encoding and Interchange P5 edition*. Text Encoding Initiative. `http://www.tei-c.org/Guidelines/P5/`.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. (2012). Almannarómur: An Open Icelandic Speech Corpus. In *Proceedings of SLTU '12, 3rd Workshop on Spoken Languages Technologies for Under-Resourced Languages*, Cape Town, South Africa.

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., and Richter, M. (2007). Íslenskur orðasjóður – Building a Large Icelandic Corpus. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M., editors, *NODALIDA 2007 Conference Proceedings*, pages 288–291, Tartu. University of Tartu.

Helgadóttir, S. (2007). Mörkun íslensks texta [Tagging Icelandic Text]. *Orð og tunga*, 9:75–107.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MIM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT2012*, pages 67–72, Istanbul.

Ingason, A. K. Loftsson, H., Helgadóttir, S., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In Raante, A. and Nordström, B., editors, *Advances in Natural Language Processing, Lecture Notes in Computer Science*, volume 5221, pages 205–216. Springer, Berlin.

Jensson, A. T., Iwano, K., and Furui, S. (2008). Language model adaptation using machine-translated text for resource-deficient languages. *Eurasip Journal on Audio, Speech, and Music Processing*, 2008. Article ID 573832.

Johannessen, J. B., Nygaard, L., Priestley, J., and Nøklestad, A. (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of LREC 2008*, pages 617–621, Marrakesh, Morocco.

Loftsson, H. (2007). *Tagging and Parsing Icelandic Text*. PhD thesis, Department of Computer Science, University of Sheffield.

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

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., editors, *NODALIDA 2007 Conference Proceedings*, pages 128–135, Tartu. University of Tartu.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, pages 53–60, Valetta.

Nikulásdóttir, A. B. and Whelpton, M. (2010). Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, pages 33–39, Valetta.

Pind, J., Magnússon, F., and Briem, S. (1991). *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik.

Quasthoff, U., Fiedler, S., and Hallsteinsdóttir, E., editors (2012). *Frequency Dictionary Icelandic / Íslensk tíðniorðabók*. Leipziger Universitätsverlag, Leipzig.

Rehm, G. and Uzkoreit, H., editors (2012). *Strategic Research Agenda for Multilingual Europe 2020*. Presented by the META Technology Council. Springer. Berlin.

Rögnvaldsson, E. (2004). The Icelandic Speech Recognition Project Hjal. In Holmboe, H., editor, *Nordisk Sprogteknologi. Nordic Language Technology. Årbog 2003*, pages 239–242. Museum Tusculanums Forlag, Copenhagen.

Rögnvaldsson, E. and Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, C., van den Bosch, A. P. J., and Zervanou, K. A., editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76. Springer, Berlin.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2011). Creating a Dual-Purpose Treebank. In Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

Rögnvaldsson, E., Jóhannsdóttir, K. M., Helgadóttir, S., and Steingrímsson, S. (2012). *The Icelandic Language in the Digital Age*. Series editors Uzkoreit, H. and Rehm, G. Springer. Berlin.

Rögnvaldsson, E., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Nikulásdóttir, A. B., Whelpton, M., and Ingason, A. K. (2009). Icelandic Language Resources and Technology: Status and Prospects. In Domeij, R., Koskenniemi, K., Krauwer, S., Maegaard, B., Rögnvaldsson, E., and de Smedt, K., editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, pages 27–32. Northern European Association for Language Technology (NEALT), Tartu University Library, Tartu.

Sigurðardóttir, A., Hannesdóttir, A. H., Jansson, H., Jónsdóttir, H., Trap-Jensen, L., and Úlfarsdóttir, Þ. (2008). ISLEX – an Icelandic-Scandinavian Multilingual Online Dictionary. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII Euralex International Congress (Barcelona, 15-19 July 2008)*, pages 779–790, Barcelona. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. http://www.users.york.ac.uk/~lang22/YcoeHome1.htm.

Thorbergsdóttir, Á. (2003). Íslenskt íðorðastarf og orðabanki íslenskrar málstöðvar [Icelandic terminological work and the word bank of the Icelandic Language Institute]. *Málfregnir*, 13:3–12.

Thráinsson, H., Angantýsson, Á., Svavarsdóttir, Á., Eythórsson, T., and Jónsson, J. G. (2007). The Icelandic (Pilot) Project in ScanDiaSyn. *Nordlyd*, 34(1):87–124.

Vasiljevs, A., Forsberg, M., Gornostay, T., Hansen, D. H., Jóhannsdóttir, K. M., Lindén, K., Lyse, G. I., Offersgaard, L., Oksanen, V., Olsen, S., Pedersen, B. S., Rögnvaldsson, E., Rozis, R., Skadina, I., and de Smedt, K. (2012). Creation of an Open Shared Language Resource Repository in the Nordic and Baltic Countries. In *Proceedings of LREC 2012*, pages 1076—-1083, Istanbul.

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)*. http://www.linguist.is/icelandic_ treebank/.