The Tagged Icelandic Corpus (MÍM)

Sigrún Helgadóttir*, Ásta Svavarsdóttir*, Eiríkur Rögnvaldsson†,
Kristín Bjarnadóttir*, Hrafn Loftsson‡

*The Árni Magnússon Institute for Icelandic Studies, †University of Iceland, ‡Reykjavík University
Reykjavík, Iceland
sigruhel@hi.is, asta@hi.is, eirikur@hi.is, kristinb@hi.is. hrafn@ru.is

Abstract

In this paper, we describe the development of a morphosyntactically tagged corpus of Icelandic, the *MÍM* corpus. The corpus consists of about 25 million tokens of contemporary Icelandic texts collected from varied sources during the years 2006–2010. The corpus is intended for use in Language Technology projects and for linguistic research. We describe briefly other Icelandic corpora and how they differ from the *MÍM* corpus. We describe the text selection and collection for *MÍM*, both for written and spoken text, and how metadata was created. Furthermore, copyright issues are discussed and how permission clearance was obtained for texts from different sources. Text cleaning and annotation phases are also described. The corpus is available for search through a web interface and for download in TEI-conformant XML format. Examples are given of the use of the corpus and some spin-offs of the corpus project are described. We believe that the care with which we secured copyright clearance for the texts will make the corpus a valuable resource for Icelandic Language Technology projects. We hope that our work will inspire those wishing to develop similar resources for less-resourced languages.

Keywords: corpus, tagging, Icelandic

1. Introduction

This paper describes the Tagged Icelandic Corpus (the *MÍM* corpus) and how it was created. The project has been developed at The Árni Magnússon Institute for Icelandic Studies (AMI)[1]. The *MÍM* corpus is a synchronic corpus that will contain about 25 million running words. The texts are taken from different genres of contemporary Icelandic, i.e. texts produced in 2000–2010. All the texts have already been collected, part of the corpus has been tagged and is available for search (about 17.7 million tokens in October 2011).[2] The texts have already been used for various Language Technology (LT) projects. The *MÍM* corpus will be available in its entirety, both for search and download, in the summer of 2012.

Work on the corpus building started in 2004. It was one of the main projects of an LT Program launched by the Minister of Education, Science and Culture in 2000 (Rögnvaldsson et al., 2009). From the beginning, the *MÍM* corpus was mainly intended for use in LT, and the product of the work should be a balanced collection of contemporary texts, morphosyntactically tagged and lemmatised and supplied with metadata in TEI-conformant XML format (Burnard and Bauman, 2008). However, it soon became apparent that it would also be necessary to supply a web-based search interface to the corpus, for the benefit of researchers, teachers, students and lexicographers.

The paper is structured as follows. In Section 2., we describe briefly other Icelandic corpora. In Section 3., we give an account of the *MÍM* corpus and how it was created. The availability and use of the corpus is described in Section 4., and related projects are mentioned in Section 5. Finally, we conclude with a summary in Section 6.

2. Icelandic Corpora

At the turn of the century Icelandic LT virtually did not exist (Rögnvaldsson et al., 2009). In a report, written for the Minister of Education, Science and Culture in 1999 (Ólafsson et al., 1999), the lack of corpora for the development of LT tools is given a particular mention. The compilation of a balanced morphosyntactically tagged corpus of 25 million words was therefore one of the projects supported by the special LT Program launched in 2000.

However, a small corpus, annotated with morphosyntactic tags and lemmata, existed at the Institute of Lexicography (now a part of the AMI). This corpus had been compiled for the making of the Icelandic Frequency Dictionary (*IFD*), *Íslensk orðtíðnibók*, published in 1991 (Pind et al., 1991). The *IFD* corpus[3] 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 lemmatisation 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.

Íslenskur orðasjóður[4] is an Icelandic corpus of more than 250 million running words collected from all domains ending in *.is* during the autumn of 2005, together with an auto-

[1] http://www.arnastofnun.is/

[2] http://mim.hi.is/

[3] http://www.malfong.is

[4] http://wortschatz.uni-leipzig.de/ws_ice/

matically 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). The web texts were cleaned substantially before inclusion in the corpus. Since the corpus is neither balanced nor morphosyntactically tagged, its usefulness for certain types of linguistic research and LT projects is limited. 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 *Icelandic Parsed Historical Corpus* (*IcePaHC*)[5] is a diachronic treebank that was released in version 0.9 in August 2011 and contains about one million running words from every century between the 12^{th} and the 21^{st} centuries inclusive (Rögnvaldsson et al., 2011). The texts are annotated for phrase structure, PoS-tagged and lemmatised. The corpus is designed to serve both as an LT tool and a syntactic research tool. The corpus is completely free and open since most of the texts are no longer under copyright.

3. Creating the *MÍM* Corpus

In this section, we describe the creation of the *MÍM* corpus. We describe text collection, procedures for securing consent from copyright holders to use their material, text sources for written and spoken texts, methods for cleaning and annotation of the texts, and, finally, the creation of the metadata.

3.1. Text collection

Since the *MÍM* corpus is the first large balanced and tagged corpus with Icelandic text, one of the main criteria for its compilation was that it should contain a "balanced" or a "representative" text collection. However, researchers do not always agree on what is meant by these concepts. "Representativeness" has been defined as either "representing the population of texts or representing the structure of readership" (Przepiórkowski et al., 2010). Either of these criteria is difficult to establish. Following the population of texts would for instance mean that, for the period in question, most of the texts should have been sampled from the web. Following the structure of readership would require a survey of readership to be undertaken which was not practical at the time. A very pragmatic approach to the text collection was, however, adopted. An attempt was made to collect texts from different genres and from different sources. Only texts that were available in digital form were acquired. The texts were to have been written in the 21^{st} century, i.e. during the years 2000–2010, and be original writings in Icelandic. The texts were also to be morphosyntactically tagged and supplied with metadata.

In planning the text collection, the British National Corpus (*BNC*)[6] project (Aston and Burnard, 1998) was used as a model. However, with the advent of the Internet and the World Wide Web, the publishing scene has changed dramatically since the early nineties when the *BNC* was created. All the texts in the *BNC* corpus came from printed sources, apart from the spoken component. Since the budget of the *MÍM* project did not allow for the typing of text, the main restriction on the text collection was that the texts should be electronically available. Great care was taken in securing permission from copyright owners to use their text. The second restriction is thus that if a permission was not obtained for a particular text it was not included in the corpus.

Source	%
Printed newspapers	27.9
Printed books	22.3
Printed periodicals	8.7
Blog	7.6
Text from www.visindavefur.is	6.8
Text from government websites	6.4
Text from websites of organizations	6.2
Legal texts and adjudications	4.1
Texts written-to-be-spoken	2.9
School essays	2.6
Spoken language	2.2
Online newspapers and periodicals	1.5
Miscellany	0.8
Total	100.0

Table 1: Texts in *MÍM* by source

Table 1 shows the contribution of texts from the various text sources (media in *BNC* terminology) to the corpus material. Over one third of the texts were harvested directly from the World Wide Web. The spoken component, which comprises about 2.2% of the corpus texts, was made available by other projects.

3.2. Permissions clearance

Since the *MÍM* corpus was originally intended mainly for use 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%). Early on in the project, cooperation was secured from the *Writer's Union of Iceland*[7], the *Association of Non-fiction and Educational Writers in Iceland*[8] and the *Icelandic Publishers' Association*[9]. All these associations recommended to their members that they should cooperate with the project. The most important of these, and the most difficult to secure, was the recommendation of the publishers' association, since publishers are normally the keepers of digital copies of published material.

Permission was sought from all owners of copyrighted texts included in the the *MÍM* corpus. Official texts (e.g. law, judicial texts, regulations and directives) are not copyrighted

[5] `http://www.linguist.is/icelandic_treebank/`

[6] `http://www.natcorp.ox.ac.uk/`

[7] `http://rsi.is/`

[8] `http://hagthenkir.is/`

[9] `http://bokautgafa.is/`

(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.

With the help of a solicitor, legal documents were drawn up: A declaration for copyright holders to sign and a user license for prospective users of the corpus. Copyright holders were contacted either by e-mail or ordinary mail and received a copy of the declaration to sign, a copy of the user license, and a leaflet describing the *MÍM* project. Copyright holders were usually contacted twice. If there was no response after the second contact their text was discarded.

3.3. Written texts

It was decided that about 20–25% of the texts should be taken from **printed books**. Again, a very pragmatic approach had to be adopted. Publishers that were willing to cooperate were contacted. Books were selected from their catalogues and the authors contacted. If a positive answer was not obtained within a reasonable time limit another book was substituted and the procedure repeated. When the copyright owner had given his or her consent the publisher was contacted to obtain a digital copy of the book. It was soon found that the publishers only had digital copies available of books that had been published during the last few years. It was therefore not possible to include the texts of all books that permission was obtained for. Texts from books comprise about 22% of the corpus material and are taken from 117 books (47 novels, 12 biographies and memoirs and 58 books containing non-fiction).

The largest portion of text, about 28%, is taken from newspapers, mostly from **printed newspapers** (less than 1% from two **online newspapers**). The printed newspapers are *Morgunblaðið* (20%) and *Fréttablaðið* (8%). It is relatively easy to obtain permission to use text from newspapers since it is sufficient to get a signature from the editor. The texts from *Morgunblaðið* were obtained directly from their database, classified by content. The text was sampled so as to reflect seasonal variation in the topics under discussion. The text files from *Morgunblaðið* contained some metadata that could be removed automatically. The text from *Fréttablaðið* was obtained in PDF files. The text was extracted from the PDF files and had to be rearranged to a certain extent. The text from the two online newspapers was harvested directly from the web as clean text.

Text from **printed periodicals** (8.7%) was obtained from various sources. Most of the texts came from two publishers who each publishes a number of periodicals. Permission was obtained from the publishers and all the texts were delivered on a CD as either Word files or PDF files. A number of specialized periodicals were also sampled. They cover subjects like farming, aviation, immigrants, linguistics, medicine, natural sciences, computing, literature, history, fishing, education, and mathematics and sciences. Each editor had to be approached, and in some instances it was necessary to approach the author of each article in these periodicals. The texts were delivered as Word files, PDF files or harvested directly from the web.

Blog texts comprise about 7.6% of the corpus and they were harvested directly from the web. Each blogger was approached by e-mail and asked to consent to his text being used in the corpus. The blog texts in the corpus will be anonymous, only classified by type of writer, i.e. as texts written by politicians, theologians and what was called "general bloggers".

The University of Iceland operates a website where the public can post questions on any subject (**www.visindavefur.is**). The answers are written by university academic staff and they cover most subjects taught at the university. The editors very kindly made answers from 38 writers available to the *MÍM* project and they also secured their permission to use the texts. The material comprises about 6.8% of the corpus texts and covers diverse subjects like meteorology, nursing, philosophy and anthropology.

About 11.5% of the texts in the corpus are official texts and therefore not covered by copyright. These are speeches from the Icelandic Parliament (Alþingi), (about 1% of the corpus texts, part of the texts **written-to-be-spoken** in Table 1), **legal texts and adjudications** (4.1%), and texts from the **websites of government ministries** (6.4%). All these texts, apart from the parliamentary speeches that were obtained from the database of Alþingi, were harvested directly from the respective websites.

Text was obtained from the **websites** of 14 **organizations** (6.2%). Permission was secured from the directors of these organizations and the text harvested directly from their websites. These websites represent diverse organizations like The Icelandic Road Administration, Save the Children in Iceland, and The Icelandic Tourist Board.

Texts classified in Table 1 as texts **written-to-be-spoken** comprise 2.9% of the corpus and are divided between the parliamentary speeches, radio and TV news scripts, and speeches harvested from various websites. These speeches are sermons delivered by ministers of the church in Reykjavík, addresses delivered at meetings, and radio scripts. The parliamentary speeches are not protected by copyright, but each of the other authors had to be contacted individually.

School essays (2.6%) are both essays from university students and papers written as a part of final examinations in Icelandic in a grammar school in Reykjavík. University students were contacted by e-mail, and they sent their essays back by e-mail, either as Word files or PDF files. The examination papers were obtained from the school office. Each student was contacted individually. Papers were not included in the corpus unless the writers had given their consent.

Only a small portion of the text was harvested from **online newspapers and periodicals** (1.5%). Permission was obtained from the editors.

In the category **miscellany** there are various small text excerpts, e.g. from teletext, leaflets, program notes from the Icelandic Symphony Orchestra, and text from electronic mailing lists.

3.4. Spoken texts

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. The monologues are speeches from unprepared sessions in the Icelandic Parliament, recorded in 2004-2005. The interviews come from a sociolinguistic project and include several sessions, each with an interviewer and three interviewees. The conversations were recorded in informal settings, such as the homes or work places of one or more of the participants. 2–5 persons took part in each conversation. All the recordings have been carefully transcribed in a predefined format.

Permission was sought from each speaker to use the recordings anonymously for the purpose of language research. In the transcriptions, all names have been substituted by pseudonyms, and other personal data has been removed, since the permission is conditional upon not revealing personal information. The transcribed text from all the recordings will be made a part of the *MÍM* corpus. Moreover, the transcriptions aligned with the sound files will form a separate corpus which will be made searchable on a special website. This corpus will be protected by username and password. One part of the spoken language corpus, which contains transcribed recorded debates from the Icelandic Parliament, can be used more freely as restrictions regarding public data are not as strict as in the case of private dialogues.

3.5. Cleaning the text

As already mentioned in Section 3.3., texts obtained for the *MÍM* corpus came in various formats. The main formats were PDF files, Word files, XML files, text drawn from databases, and text harvested directly from the web. Texts from PDF files were extracted by a special program developed by a member of the *MÍM* team. As a last resort, we used optical character recognition software that is used for extracting text from scanned paper documents (ABBYY FineReader: `http://finereader.abbyy.com/`). Some texts came in Word documents which are easy to convert to text. The parliamentary speeches were delivered as XML files from a database at Alþingi. Text and metadata were extracted automatically with a program developed by a member of the *MÍM* team. Text from the *Morgunblaðið* database is easy to handle and contains metadata that can be extracted automatically and then removed before the text is morphosyntactically tagged and included in the corpus. Text harvested from the web is usually quite clean.

The importance of the cleaning phase should be emphasized. The quality of the text will influence later phases of the corpus building, i.e. sentence segmentation and tokenisation, which in turn influence the quality of the morphosyntactic tagging.

Texts from printed books and periodicals that are delivered either as PDF files or Word files usually contain hyphenation. Those texts were run through a program that joined the two parts of a word that had been split between lines. Various other measures had to be taken, either with automatic or semi-automatic means. We removed manually long quotations in a foreign language, long quotations from Old Icelandic texts and from new texts that we did not have permission to use, as well as footnotes, tables of content, indexes, reference lists, poems, tables and pictures. All texts were run through a cleaning program that standardizes quotation marks, both single and double, and hyphens.

The text files obtained for the corpus were either encoded using UTF-8 or ISO-8859-1 character encoding. It was decided that all texts in *MÍM* should be converted to UTF-8 encoding. However, in the tagging process (Section 3.6.) one of the taggers used requires text in ISO-8859-1 character encoding and the current version of the software used for searching the corpus also requires text in ISO-8859-1 character encoding. As a consequence all characters that are not a part of the ISO-8859-1 character set had to be substituted with simplified versions. Although long texts in foreign languages and Old Icelandic were removed there still remain names and short quotations. As an example of characters that had to be replaced the character *œ* was substituted with the character *ö* from the modern Icelandic alphabet and the the Greek character η was replaced with the character sequence *eta*.

3.6. Annotating the text

The annotation phase consists of sentence segmentation, tokenisation, morphosyntactic tagging and lemmatisation. After morphosyntactic tagging and lemmatisation, the texts, together with the relevant metadata, are transferred into TEI-conformant XML format with special programs developed by the *MÍM* team.

The procedure and software used for sentence segmentation, tokenisation, morphosyntactic tagging and lemmatisation has been explained by (Loftsson et al., 2010) in their work on the *GOLD* corpus (see Section 5.).

The tagset used was developed for the *IFD* corpus (see Section 2.). The automatic morphosyntactic tagging accuracy has been estimated as 88.1-95.1%, depending on text type (Loftsson et al., 2010).

3.7. Metadata

All texts in the corpus are accompanied by metadata. For published texts, the metadata comprises bibliographic data like title, name of author(s), age and gender of author(s), name of editor(s) (if applicable), publisher, date and place of publishing. For other texts, metadata is recorded to identify the text. For spoken data, various information on the recorded sessions and the speakers is registered. Most of the metadata had to be manually created, but metadata on files from the newspaper *Morgunblaðið* and on parliamentary speeches was created automatically. The metadata is shown for each text example retrieved through the search interface and is a part of the downloadable texts in TEI-conformant XML format. Individual texts can be selected for search through the search interface and also classified by source which reflects approximately the classification in Table 1. The texts will also be searchable by the target age group (adults, teenagers, children).

4. Availability and use of *MÍM*

4.1. Availability

As mentioned in Section 1., the corpus was originally made to be used in LT projects. However, it soon became obvious that a web-based search interface to the corpus was necessary to enable researchers, teachers, students and lexicographers to search the tagged corpus. The Norwegian search interface *Glossa* (Johannessen et al., 2008), which in turn uses the *IMS Corpus Workbench*[10] as a search engine, is being adapted to be used with the *MÍM* corpus.

An experimental search interface is already operating where about 17.7 million words of the corpus texts are available for search (see Section 1.). In the summer of 2012, all the corpus texts will be searchable. The corpus will also be available in TEI-conformant XML format in the summer of 2012, through download from a special webpage where prospective users register and agree to the licensing terms.

As a part of the project META-NORD[11], the Icelandic META-NORD team has established a special website (`http://www.malfong.is/`) where Icelandic Language Resources can be identified and located. Information on the *MÍM* corpus will be available there, as well as links to webpages for search and download of the corpus material.

Most of the published texts have been made accessible for search in their entirety (without annotation) in the Text collection of the AMI[12], where the outcome of the search is presented in KWIC format.

4.2. Uses of the corpus

The search interface is already being used in teaching Icelandic at the University of Iceland. The texts have been made available to the same projects as *Íslenskur orðasjóður* has been used for, as mentioned in Section 2.

The texts in the corpus are being used to augment the vocabulary in the *Database of Modern Icelandic Inflection* (*DMII*) (Bjarnadóttir, 2012). This database is available for search[13] and for download[14] for use in LT projects.

In the future, automatic lookup in the corpus will be possible, both from the Nordic *ISLEX*[15] database (Sigurðardóttir et al., 2008) and from the *DMII*. The user would then be given a chance to retrieve text examples from the corpus containing the word(s) he has looked up in the respective database. There is also a possibility of offering information on the frequency of particular word forms found in electronic databases based on the frequency in the *MÍM* corpus.

5. Related projects

The *MÍM* project has been carried out over a number of years. Various other projects have been worked on at the same time by the *MÍM* project group. Four will be mentioned here.

The first is a corpus of about 1 million running words which has been sampled from *MÍM*. This corpus which we call *GOLD* (Loftsson et al., 2010) is intended as a reliable standard for the development of LT tools. Tagging and lemmatisation of this subcorpus will be manually corrected.[16] This corpus will augment the *IFD* corpus (see Section 2.) which has been used for training statistical taggers and developing LT tools. The *GOLD* corpus is nearly twice the size of the *IFD* corpus and the texts are more varied. The *GOLD* corpus will be made available through the official site for Icelandic LT Resources, (`http://www.malfong.is`), for search, for download, and as training and test sets for the training of statistical taggers.

The second project is a separate corpus of about 500,000 words of spoken language, described in Section 3.4. This corpus is intended for theoretical and practical purposes relating to the spoken language.

The third is a project where about 1.7 million words of old Icelandic texts in normalized spelling have been tagged with morphosyntactic tags and lemmatised (Rögnvaldsson and Helgadóttir, 2011). Accuracy of the tagging was estimated as 92.7%. These texts are available (`http://www.malfong.is`) for search and download for use in linguistic research and LT projects.

The fourth is an experimental project, carried out in the summer of 2011, to add semantic analysis to the morphosyntactic tagging in the *MÍM* corpus, using the semantic analysis and classification of the vocabulary of *Íslenskt orðanet*[17] (Jónsson, 2010). *Íslenskt orðanet* is a database tracing semantic relations based on a large collection of word combinations. As a result, various links between lexical, grammatical and semantic features in the text examples of the corpus were established and users equipped with new and varied search choices.

6. Conclusion

As pointed out in the introduction, the *MÍM* corpus was built to serve two purposes. Firstly, it can be used in LT projects and, secondly, for language research. The part of the corpus open for search has already proved to be useful. The texts cannot be downloaded yet, but they have been made available to researchers, e.g. to a project where a Database of Semantic Relations is being created and in a project to develop context sensitive spelling correction for

[10] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

[11] `http://www.meta-nord.eu/`

[12] `http://www.arnastofnun.is/page/arnastofnun_gagnasafn_textasafn`

[13] `http://bin.arnastofnun.is/`

[14] `http://ordid.is/gogn/`

[15] `http://www.islex.hi.is/`

[16] As of February 2011 about 90% of the morphosyntactic tags have been manually corrected.

[17] `http://www.ordanet.is/`

Icelandic. Various spin-offs of the corpus project that will serve the LT community have been identified. The *MÍM* corpus is unique in the context of Icelandic LT, as it is the only large tagged corpus in Icelandic. Since permission for the use of texts in the corpus was secured from all copyright holders, and since researchers can obtain the texts and use them in LT projects despite some restrictions, the availability of the *MÍM* corpus will be better than is usually the case of corpora.

It is our wish that this work will inspire those wishing to develop a similar resource for less-resourced languages.

7. Acknowledgements

This project has been financed by the Language Technology Program of the Icelandic Ministry of Education, Science and Culture which ran during the years 2000–2004, the Icelandic Research fund through the project Viable Language Technology Beyond English (no 090662011), the University of Iceland Research Fund, the Student Innovation Fund and META-NORD (funded by the ICT PSP Programme, as part of the Competitiveness and Innovation Framework Programme under grant agreement no 270899). The authors would like to thank the many young students and researchers who have worked on various aspects of the project from time to time. Thanks are also due to the developers of *Glossa*, especially Anders Nøklestad. We would also like to thank the anonymous reviewers for helpful comments. Last, but not least, thanks are due to the numerous authors, writers, editors and others who have very kindly given permission for the use of their text and the publishers who made digital copies available. Without their cooperation this project could not have been completed.

8. References

G. Aston and L. Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

K. Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *Proceedings of "Language Technology for Normalization of Less-Resourced Languages", workshop at the* 8^{th} *International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

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

E. Hallsteinsdóttir, T. Eckart, D. Biemann, and M. Richter. 2007. Íslenskur orðasjóður – Building a Large Icelandic Corpus. In *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, Tartu, Estonia.

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

J. H. Jónsson. 2010. Lemmatisation of Multi-word Lexical Units: Motivation and Benefits. In H. Bergenholtz, S. Nielsen, and S. Tarp, editors, *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pages 165–194. Bern: Peter Lang.

H. Loftsson, J. H. Yngvason, S. Helgadóttir, and E. Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

A. B. Nikulásdóttir and M. Whelpton. 2010. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

R. Ólafsson, E. Rögnvaldsson, and Þ. Sigurðsson. 1999. Tungutækni [Language Technology]. Skýrsla starfshóps [Committee Report]. Menntamálaráðuneytið [Ministry of Education, Science and Culture]. Reykjavik, Iceland.

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

A. Przepiórkowski, R. L Górski, M. Łaziński, and P. Pęzik. 2010. Recent Developments in the National Corpus of Polish. In *Proceedings of LREC 2010*, Valetta, Malta.

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

E. Rögnvaldsson, H. Loftsson, K. Bjarnadóttir, S. Helgadóttir, A. B. Nikulásdóttir, M. Whelpton, and A. K. Ingason. 2009. Icelandic Language Resources and Technology: Status and Prospects. In R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Odense, Denmark.

E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. 2011. Creating a Dual-Purpose Treebank. *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

A. Sigurðardóttir, A. H. Hannesdóttir, H. Jansson, H. Jónsdóttir, L. Trap-Jensen, and Þ. Úlfarsdóttir. 2008. ISLEX – an Icelandic-Scandinavian Multilingual Online Dictionary. In *Proceedings of the XIII Euralex International Congress*, Barcelona.

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