The Tagged Icelandic Corpus (MÍM)

Sigrún Helgadóttir, Kristín Bjarnadóttir, Ásta Svavarsdóttir
The Árni Magnússon Institute for Icelandic Studies, Iceland (AMI)

Eiríkur Rögnvaldsson
Department of Icelandic
University of Iceland, Iceland

Hrafn Loftsson
School of Computer Science
Reykjavik University, Iceland

Introduction

We describe the development of a new, synchronic, balanced tagged Icelandic corpus consisting of about 25 million tokens developed at The Árni Magnússon Institute for Icelandic Studies (AMI). The corpus is intended for use in Language Technology projects and linguistic research.

Creating the MÍM Corpus

Text collection

Selection criteria:

- A balanced or representative text collection.
- Texts produced during the years 2000–2010.
- Texts that are electronically available.
- Original writings in Icelandic.
- Permission available for copyrighted text.

Permission clearance

- About 88% of texts in the corpus are protected by copyright.
- Cooperation secured from the *Writer's Union of Iceland*, the *Association of Non-fiction and Educational Writers in Iceland* and the *Icelandic Publishers' Association*.
- All copyright owners signed a special declaration and agreed that their material may be used free of licensing charges.
- The operator (AMI) agrees that only 80% of each published text is included in the corpus and that copies of the *MÍM* corpus are only made available under the terms of a special license agreement.

Text extraction and cleaning

- Text was extracted from publishers' files (e.g. *pdf*, *Word*, *XML*).
- All texts were converted to UTF-8 character encoding.
- Hyphenated words that were split between two lines were joined. Line breaks were inserted after lines (headings) that do not end with an end-of-sentence marker.

Annotating the text

Sentence segmentation and tokenisation

- The text was split into individual sentences and each sentence into individual tokens, one token per line using the *IceNLP* toolkit (http://icenlp.sourceforge.net/).

Part-of-speech tagging

- The text was tagged with four individual taggers: one rule-based tagger (*IceTagger*) and three data-driven taggers (*TnT*, *fnTBL*, *MXPOST*). A tagger combination was applied with *CombiTagger* (http://combitagger.sourceforge.net/) using majority voting as a combination method.
- This method is especially suitable when tagging a corpus, i.e. when effectiveness is more important than efficiency. The tagging accuracy has been estimated to lie between 89% and 92.5%.

Lemmatisation

- The text was lemmatized with the program *Lemmald* which is a part of the *IceNLP* toolkit.

Metadata

- All texts in the corpus are accompanied by metadata.

Texts in MÍM by source

Source	%
Printed newspapers	27.9
Printed books	22.3
Printed periodicals	8.7
Blogs	7.6
The Icelandic Web of Science (http://www.why.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

Written texts
Obtained from digital copies of printed publications and the web.

Spoken texts
Obtained through four different projects consisting of monologues, interviews and spontaneous conversations between adults of both sexes and with different back-grounds.

Availability

Search interface (http://www.malfong.is/mim/).

- We use the Norwegian search interface *Glossa*, which in turn uses the *IMS Corpus Workbench* as a search engine, with the *MÍM* corpus. The corpus is freely open for search.

Download (http://www.malfong.is/mim/)

- All texts will be available for download in TEI-conformant XML-format together with annotation and metadata for use in LT projects and for linguistic research under a special license agreement. 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.

Conclusion

- A new corpus, *MÍM*, consisting of about 25 million tokens of Icelandic text has been created for use in Language Technology projects and linguistic research.
- The corpus contains a varied selection of contemporary Icelandic texts written during the years 2000-2010.
- Permission clearance was obtained for all copyrighted texts.
- The corpus is available for search through a web interface and will soon be available for downloading in TEI-conformant XML- format.

LREC 2012: SaLTMIL-AfLaT Workshop. Istanbul 22.05.2012