An Icelandic Gigaword Corpus

Steinþór Steingrímsson, Sigrún Helgadóttir & Eiríkur Rögnvaldsson

The paper describes work in progress to compile an Icelandic Gigaword Corpus (IGC). The initial aim of the project was to compile a large corpus of contemporary texts with at least a billion running words, with the minimum amount of work and resources. Thus we focussed on material not protected by copyright and sources which could provide us with large chunks of text for each cleared permission. The two main sources considered were therefore official texts and texts from news media. Only digitally available texts are included in the corpus, and formats that can pose problems are not processed. The corpus texts are morpho-syntactically tagged and provided with metadata. Processes have been set up for continuous text collection, cleaning and annotation. The corpus will be made available for search and download with permissive licenses. The first version of the corpus will be released by the end of 2017. Texts will be added continually and a new version published every year.

1. Introduction

The lack of a very large Icelandic text corpus has been evident for some time. The compilation of such a corpus has therefore been considered a top priority in order to further Language Technology (LT) in Iceland (Anna Björk Nikulásdóttir et al. 2017). Large text corpora are e.g. necessary for the design of language models that are used in building a variety of LT tools such as speech recognizers, spell and grammar checkers and automatic machine translation. With the increased importance of machine learning methods such as neural networks in LT, the importance of large text corpora and other textual resources has increased considerably.

The aim of the corpus project is to compile as large a corpus as possible with the minimum amount of work and resources. We want the corpus to be attractive for use in LT projects as well as for other research and study. In planning the project it was decided to aim for the following goals:

- The IGC will contain more than a billion running words, morpho-syntactically tagged and lemmatized and provided with metadata.
- Only digitally available texts will be included in the IGC. Formats that may pose a difficulty will not be processed.
- The IGC will be open and constantly expanding.
- A closed version will be published every year.
- The IGC will be accessible through an online concordance search tool.

- Trend data from the IGC will be searchable in an n-gram viewer.
- The IGC will be made available for download with a permissive license.

In Section 2 the compilation of the MIM corpus (Sigrún Helgadóttir et al. 2012) is described where the intention was to create a "balanced" and a "representative" text collection. In order to achieve representativity and balance text was sampled from many genres and often a very small chunk of text was acquired for each license. However, there are several problems connected with trying to achieve representativeness in a corpus. For the first, what should it be representative of? And because it can be hard to determine where a variety of language ends and another begins, any corpus is 'virtually by definition biased to a greater or a lesser extent' (Nelson 2010).

One of the design goals for the IGC is for it to be open, that it will be constantly expanding, but closed versions will be published every year to make it possible for researchers to verify others' results. Furthermore, in order to accomplish our goal of more than a billion words we need to build a collection of texts from sources who have available material that is not protected by copyright or where it is possible to get big chunks of text for each license secured. The two main sources considered are therefore official texts and texts from news media. Only digitally available texts will be included in the corpus and formats that are difficult to process, like pdf documents, will not be used. This design makes it even harder to consider representatitiveness. The corpus will therefore be biased towards journalistic and official texts, but more detailed description of the corpus texts is given in section 3.2.

The corpus texts are morphosyntactically tagged and provided with metadata. Processing pipelines are set up for continuous text collection, text cleaning and annotation where the processing tools will be continually updated.

This paper is structured as follows. In Section 2 we describe briefly existing Icelandic corpora. In Section 3 an account is given of the creation of the IGC. Availability of the corpus is discussed in Section 4 and in Section 5 we sum up and conclude the paper.

2. Icelandic Corpora

In this section existing Icelandic corpora are listed and described briefly, to explain their shortcomings and hence the need for a new corpus.

A small corpus was compiled at the Institute of Lexicography[1] for the making of the Icelandic Frequency Dictionary (IFD), Íslensk orðtíðnibók, published in 1991 (Jörgen Pind et al. 1991). The IFD corpus[2] consists of just over half a mil-

[1] Now a part of the Árni Magnússon Institute for Icelandic Studies.

[2] Available at <http://www.malfong.is>.

lion running words. The corpus has a heavy literary bias as about 80% of the texts stem from fiction. The corpus is annotated with morphosyntactic tags and lemmata. Tagging and lemmatization was manually corrected and hence the corpus has been used as a gold standard for training part-of-speech (PoS) taggers, lemmatizers and parsers. It can be stated that the IFD corpus has laid the ground for most work on PoS tagging, lemmatization and parsing that has been performed on Icelandic during the last 15 years.

The Tagged Icelandic Corpus (MIM) was released in the spring of 2013, both for search[3] and download.[4] This corpus contains 25 million running words from various genres dating from the first decade of the 21st century (Sigrún Helgadóttir et al. 2012). The corpus was intended for use in LT projects and for linguistic research. About 86% of the texts are protected by copyright, the remainder being official text (parliamentary speeches, legal text, adjudications and text from government websites). The largest proportion of the text, just less than 24%, comes from published books containing both fiction and non-fiction. The second largest portion, about 22%, derives from newspapers, mostly printed newspapers. The corpus is annotated with morphosyntactic tags and lemmata.

To enable the use of the corpus in LT projects it was considered important to secure copyright clearance for the texts to be used. All owners of copyrighted text signed a special declaration and agreed that their material may be used free of licensing charges.

MIM-GOLD is a corpus of about 1 million running words which was sampled from the MIM corpus (Hrafn Loftsson et al. 2010; Sigrún Helgadóttir et al. 2012; Steinþór Steingrímsson et al. 2015). The corpus is intended as a reliable standard for the development of LT tools. Tagging of this subcorpus has been manually corrected. MIM-GOLD will augment the IFD corpus for training statistical taggers and developing LT tools. The MIM-GOLD corpus is nearly twice the size of the IFD corpus and the texts are more varied, less than 25% of the texts in MIM-GOLD are literary texts compared to about 80% of the texts in the IFD corpus.

Training and testing using the Average Perceptron Tagger *Stagger* (Östling 2012) on MIM-GOLD after two correction phases has already been described (Steinþór Steingrímsson et al. 2015). The result showed that there were still errors in the tagging that needed to be corrected. Work on locating and correcting these errors was completed in the fall of 2017.

The Icelandic Parsed Historical Corpus (IcePaHC)[5] is a diachronic treebank that contains about one million running words from every century between the 12th and the 21st centuries, inclusive (Eiríkur Rögnvaldsson et al. 2011). The texts are annotated for phrase structure, PoS-tagged and lemmatized. 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 in copyright.

---

3 *Mörkuð íslensk málheild*: <http://mim.arnastofnun.is>.

4 At <http://www.malfong.is>.

5 <http://www.linguist.is/icelandic_treebank/>.

Íslenskur orðasjóður[6] is an Icelandic corpus of more than 550 million running words collected from all domains ending in *.is* in 2005 and 2010 (approx. 33 million sentences). Moreover, additional newspaper texts (2 million sentences) and the Icelandic Wikipedia are included. The web texts were cleaned substantially before their inclusion in the corpus.

Although the corpora mentioned in this section have been useful in LT and language research they do not fulfill the requirements that present day LT makes to language resources as regards size and quality. Therefore it was considered necessary to embark on the project of compiling the IGC.

3. Creating the corpus

In Section 1 the aims of the corpus project were described, the primary aim being to compile as large a corpus as possible, at least a billion words, with the minimum amount of work and resources. In this section we will give an account of permissions clearance, collecting the texts and the cleaning and annotation process.

3.1 Permission clearance and licensing

One of the design considerations for the IGC was to make the corpus available with a permissive license, such as a Creative Commons license.[7] Work on permission clearance for the first version of the corpus concluded in early 2017. We cleared permission from 19 content providers but found that Creative Commons licensing is not widely known in Iceland so eventually it was necessary to use the license used for the compilation of the MIM corpus for a substantial part af the texts. Although some of the copyright protected texts in the IGC will be made available with a CC license a great part will be tied to the special license developed for the MIM corpus. Together with text not protected by copyright we have access to more than 40 different text sources. The texts include general and local news from print and the web, transcribed television and radio news, commentary on politics and current affairs and texts on scientific matters. Furthermore, we collect parliamentary speeches, adjudications from courts and a selection of recent fiction and non-fiction from The Árni Magnússon Institute's text collection.

3.2 Collecting texts

A pragmatic approach to text collection was adopted. Texts requiring a minimum of cleaning and processing and texts accompanied by relevant metadata are preferred. This applies to texts obtained from databases of text owners and text har-

6 <http://wortschatz.uni-leipzig.de/ws_ice/>.

7 Cf. <https://creativecommons.org/>.

vested from the web. Texts in MS Word document format, in Excel spreadsheets or in XML format have also been accepted.

Texts not protected by copyright will be collected from official sources, the biggest of which is the Icelandic parliament, providing parliamentary speeches dating back to 1940 in XML format, containing all relevant metadata. The speeches are transcribed at Alþingi and have been extensively proofread. We also harvest legal text and adjudications from official websites.

Text has been acquired from all the largest newspaper publishers in Iceland, and a number of smaller ones have given permission for use of their text both from online and printed sources. The corpus collection includes the Icelandic Wikipedia, the University of Iceland's Science Web, The Árni Magnússon Institute's text collection (fiction and non-fiction, from recent decades), translations of EEA documents and other smaller sources.

Text genre	**Sources**	**Word count**
Newspaper articles	*Morgunblaðið*, *Vísir*, *DV* and various other smaller news sources	745,708,958
Parliamentary speeches	Alþingi	210,580,253
Adjudications	Supreme court and district courts of Iceland	88,351,996
Transcribed radio/television news	RÚV and 365	54,129,051
Sports news	Fótbolti.net and 433.is	45,992,991
Current affair blogs	Jónas.is, Andríki.is and other smaller sources.	13,030,217
Informational articles	Wikipedia and Science Web	10,738,060
Gossip/entertainment	Bleikt.is	5,316,675
Total		1,173,848,201

Table 1: Retrieved texts as of August 2017.

Table 1 lists text genres and word count for texts that have been retrieved in August 2017. At that point the majority of texts in the first version of the IGC have been processed. Unprocessed sources are listed in table 2.

Text source	Estimated word count
EEA translations	20,000,000
Newspaper articles (6 smaller news sources)	30,000,000
Legal text	5,000,000
The Árni Magnússon Institute's text collection	70,000,000
Total	**125,000,000**

Table 2: Texts to be included in the first version, not retrieved in August 2017.

3.3 Text cleaning and annotation

Texts in the corpus can be divided into written texts and transcribed spoken text. Transcribed spoken text includes parliamentary speeches and news from the main radio and television stations in Iceland.

Procedures have been devised for automatic editing and cleaning of the text, annotation and metadata extraction. There is no manual post-editing.

The annotation phase consists of sentence segmentation, tokenization, morphosyntactic tagging and lemmatization. After morphosyntactic tagging and lemmatization, the texts, together with the relevant metadata, are transferred into TEI-conformant XML format (TEI Consortium 2017). N-grams (n up to 5) are also created for use with the n-gram viewer and for distribution.

Sentence segmentation and tokenization is performed with the same procedures as were used for the MIM corpus (Sigrún Helgadóttir et al. 2012). *IceStagger* (Hrafn Loftsson & Östling 2013) is used for tagging the IGC, initially trained on the IFD corpus but will be retrained and rerun when MIM-GOLD is available.

A new tool is currently being developed for lemmatizing Icelandic text. This tool will be used for lemmatizing the IGC and first results indicate a great improvement over the tool used to lemmatize the MIM corpus. A thorough analysis and comparison of the two systems remains to be done.

A pipeline for harvesting, cleaning and annotating the corpus texts has been developed. Individual tools in the pipeline will be continually updated to produce a more precise and reliable annotation with each new version of the corpus.

4. Availability and use

The main object of the corpus is for use in LT projects. For other uses, such as linguistics research, teaching, lexicography or other studies the data will be searchable in a web-based concordance tool. The Swedish platform KORP[8] (Borin et al. 2012) which in turn uses the IMS Corpus Workbench[9] (Evert & Hardie 2011) as a search engine is being adapted to be used for the corpus. Users of the search interface can take advantage of the annotation of the texts when specifying search criteria. Texts will be added continually to the searchable corpus.

The corpus texts will be made available for download in the TEI-conformant XML format (TEI Consortium 2017). As mentioned in Section 1 some of the corpus texts are not protected by copyright, some can be distributed with relatively open CC licenses and some texts will be made downloadable with the special license developed for the MIM corpus. This situation will be reflected in the download procedures. The corpus can be downloaded through the Icelandic LT resources website *Málföng*.[10]

[8] <https://spraakbanken.gu.se/korp/>.

[9] <http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html>.

[10] <http://www.malfong.is/>.

The corpus texts will also be searchable through an n-gram viewer based on NB N-gram viewer (Birkenes et al. 2015).

To aid developers of LT tools the corpus website will allow download of the n-grams (n up to 5) used for the n-gram viewer.

5. Conclusion and further work

The new Icelandic Gigaword Corpus will be a valuable resource for builders of LT tools for Icelandic. It will also be useful for researchers, lexicographers, teachers, journalists and others working with or researching the Icelandic language.

The compilation of the corpus will be an ongoing process although closed versions, which will not be changed, will be published yearly. Official texts will be added continually as well as texts protected by copyright, as long as permission for their use has been secured. The tools in the corpus pipeline will also be upgraded following the development of better tools or versions and the corpus texts reannotated to reflect improved precision and reliability of the tools.

References

Anna Björk Nikulásdóttir, Jón Guðnason & Steinþór Steingrímsson (2017): *Máltækni fyrir íslensku 2018–2022: verkáætlun.* Reykjavík: Mennta- og menningarmálaráðuneytið.

Birkenes, Magnus B., Lars G. Johnsen, Arne M. Lindstad & Johanne Ostad (2015): From digital library to n-grams: NB N-gram. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA-2015),* NEALT Proceedings Series Vol. 23. Vilnius, Lithuania, 293–295.

Borin, Lars, Markus Forsberg & Johan Roxendal (2012): Korp – the corpus infrastructure of Språkbanken. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012),* Istanbul, 474–478. <http://www.lrec-conf.org/proceedings/lrec2012/pdf /248_Paper.pdf> (Retrieved September 10, 2017).

Eiríkur Rögnvaldsson, Anton K. Ingason, Einar F. Sigurðsson & Joel Wallenberg (2011): Creating a Dual-Purpose Treebank. In: *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

Evert, Stefan & Andrew Hardie (2011): Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In: *Proceedings of the Corpus Linguistics 2011 conference*. Birmingham, UK: University of Birmingham. <https://www.birmingham.ac.uk /documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf> (Retrieved September 10, 2017).

Hrafn Loftsson & Robert Östling (2013): Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA-2013)*. NEALT Proceedings Series 16. Oslo. <http://www.ep.liu.se/ecp/085/013/ecp1385013.pdf> (Retrieved September 10, 2017).

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir & Eiríkur 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 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta. <https://www.ru.is/~hrafn/papers/corpusTagging.final.pdf> (Retrieved September 10, 2017).

Jörgen Pind, Friðrik Magnússon & Stefán Briem (1991): *Íslensk orðtíðnibók* [The Icelandic Frequency Dictionary]. Reykjavík: Orðabók Háskólans.

Nelson, M. (2010): Building a written corpus. In: A. O'Keeffe & M. McCarthy (Eds.): *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, 53–65.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir & Hrafn Loftsson (2012): The Tagged Icelandic Corpus (MIM). In: *Proceedings of the workshop "Language Technology for Normalization of Less-Resourced Languages"– SaLTMiL 8 – AfLaT2012 at the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, 67–72. <http://aflat.org/files/saltmil8-aflat2012.pdf> (Retrieved September 10, 2017).

Steinþór Steingrímsson, Sigrún Helgadóttir & Eiríkur Rögnvaldsson (2015): Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA-2015)*. NEALT Proceedings Series Vol. 23. Vilnius, Lithuania, 287–291. <https://aclanthology.info/ papers/W15-1838/w15-1838> (Retrieved September 10, 2017).

TEI Consortium, eds. (2017): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 3.2.0. Last updated on 10th July 2017. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> (Retrieved September 10, 2017).

Östling, Robert (2013): Stagger: An Open-Source Part of Speech Tagger for Swedish. In: *Northern European Journal of Language Technology*, 2013, Vol. 3, 1–18. Linköping: Linköping University Electronic Press. <http://www.nejlt.ep.liu.se/2013/v3/a01/nejlt13v3a1.pdf> (Retrieved September 10, 2017).

Steinþór Steingrímsson
Language Technologist
steinthor.steingrimsson@arnastofnun.is

Sigrún Helgadóttir
Language Technologist
sigrun.helgadottir@arnastofnun.is

The Árni Magnússon Institute for Icelandic Studies
Laugavegi 13
101 Reykjavík, Iceland

Eiríkur Rögnvaldsson
Professor
The University of Iceland
Faculty of Icelandic and Comparative
Cultural Studies
Árnagarði við Suðurgötu
101 Reykjavík, Iceland
eirikur@hi.is