Risamálheild: A Very Large Icelandic Text Corpus

Steinþór Steingrímsson[1], Sigrún Helgadóttir[1], Eiríkur Rögnvaldsson[2], Starkaður Barkarson[1], Jón Guðnason[3]
[1]The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland
[2]The University of Iceland, Reykjavík, Iceland
[3]Reykjavik University, Reykjavík, Iceland
steinst@hi.is, sigruhel@hi.is, eirikur@hi.is, stb36@hi.is, jg@ru.is

Abstract
We present Risamálheild, the Icelandic Gigaword Corpus (*IGC*), a corpus containing more than one billion running words from mostly contemporary texts. The work was carried out with minimal amount of work and resources, focusing on material that is 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 be problematic are not processed. The corpus texts are morphosyntactically tagged and provided with metadata. Processes have been set up for continuous text collection, cleaning and annotation. The corpus is available for search and download with permissive licenses. The dataset is intended to be clearly versioned with the first version released in early 2018. Texts will be collected continually and a new version published every year.

Keywords: text corpora, Icelandic

1. Introduction

The lack of a very large Icelandic text corpus has been evident for some time. Data oriented methods have increasingly come to dominate the field of NLP and this has led to the need for more data and bigger datasets in order to achieve better performance. In the last few years, with machine learning methods such as neural networks reaching preeminence in various areas of Language Technology (LT) the importance of large text corpora and other textual resources has increased considerably. The compilation of a corpus such as the one described here has therefore been considered a top priority in order to further LT in Iceland (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 systems.

The aim of the *IGC* project was to compile as large a corpus as possible with the minimum amount of work and resources. The corpus should be clearly versioned in order to facilitate reproducible experiments. The design should make it easy to compare NLP algorithms on contemporary Icelandic and serve as a resource for a wide range of linguistic research, research in the field of culturomics (Michel et al. 2011) and for the interested public. The corpus should be attractive for use in LT projects as well as for other research and study. Therefore we aimed for the following goals:

- The *IGC* will contain more than a billion running words, morphosyntactically 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 (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 only a very small chunk of text was acquired for each license. There are several problems connected with trying to achieve representativity 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). As the goal of our current project was to create as large a corpus as possible of contemporary texts in a language spoken by less than 350 thousand people, instead of emphasizing on representativity we aimed for as much coverage as possible and providing extensive meta-data so that users of the corpus can construct their own subcorpora as needed.

A primary design goal for the *IGC* was for it to be open, and that it will be constantly expanding. To make it possible for researchers to verify findings made using the corpus, static versions will be published every year, containing all texts collected up to that point. Furthermore, in order to accomplish our goal of more than a billion words we built a collection of texts from sources where it is possible to acquire 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 were therefore official text and text from news media. Only digitally available texts are included in the corpus and formats that may pose problems, like pdf documents, were not processed. This results in the corpus as a whole being biased towards journalistic and official texts, but more detailed description of the corpus texts is given in section 3.2.

The 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.

The 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*, in Section 4 availability of the corpus is discussed and in Section 5 we conclude.

2. Icelandic Corpora

Existing Icelandic corpora will be listed and described briefly in this section 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 (Pind et al. 1991). The *IFD* corpus[2] consists of just over half a million running words. The corpus has a heavy literary bias as about 80% of the texts are 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 (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 text, just less than 24%, comes from published books containing both fiction and non-fiction. The second largest portion, about 22%, is taken from newspapers, mostly from printed newspapers. The corpus is annotated with morphosyntactic tags and lemmata and each text segment contains metadata. 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 (Loftsson et al. 2010; Helgadóttir et al. 2012; 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* can augment the *IFD* corpus for training stochastic 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 (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 fall 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 (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.

Íslenskur orðasjóður[6] is an Icelandic corpus of more than 550 million running words collected from all domains ending in *.is* during the autumns of 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 inclusion in the corpus.

Although the corpora mentioned in this section have been useful in LT and language research they lack in size and/or coverage to fulfill the requirements that present day LT makes. 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, text collection 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] However, Creative Commons licensing does not seem to be widely known in Iceland so eventually it was necessary to use the same license as was developed for the *MIM* corpus texts for some of the texts in the *IGC*. Some of the copyright protected texts in the *IGC* will be made available with a CC BY license but a great part will be tied to a modified version of the *MIM* corpus license. Work on permission clearance for the first version of the corpus concluded in early 2017. We cleared permission from 20 content providers. 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 printed media and the web, transcribed television and radio news, commentary on politics and current affairs and texts on scientific matters. Furthermore we collect parliamentary

1 Now a part of the Árni Magnússon Institute for Icelandic Studies.
2 Available at http://www.malfong.is
3 http://mim.arnastofnun.is
4 http://www.malfong.is
5 http://www.linguist.is/icelandic_treebank/
6 http://wortschatz.uni-leipzig.de/ws_ice/
7 https://creativecommons.org/

Text Genre	Word Count	No. of documents	Time period
Newspaper Articles	796,526,434	3,029,985	1998-2017
Parliamentary Speeches	210,699,883	380,557	1911-2017
Adjudications	92,696,289	23,634	1999-2017
Published Books	5,729,543	120	1980-2008
Transcribed Radio/Television News	54,129,050	313,749	2004-2017
Sports News Websites	47,431,733	280,838	2002-2017
Regulations	26,038,153	12,038	1275-2017
Current Affair Blogs and articles	10,511,776	43,678	1973-2017
Informational Articles	10,796,107	55,091	2000-2017
Lifestyle	4,027,506	14,671	2010-2017
Total	1,261,026,503	4,154,528	

Table 1: Retrieved texts for the IGC 2017

speeches, adjudications from courts and a selection of recent fiction and non-fiction from The Árni Magnússon Institute's text collection.

As a consequence the downloadable corpus is divided into two parts: *IGC1* and *IGC2*. *IGC1* contains texts that can be used with a special license developed for *MIM*. 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. The license granted to the licensee is non-transferable. *IGC2* contains official texts and texts that can be used with a CC BY license. All copyright holders have agreed that their material may be used free of licensing charges. Copyright owners that did not accept the CC BY license signed a special declaration developed for *MIM* with necessary amendments for the *IGC1*.

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 were preferred. This applied to texts obtained from databases of text owners and text harvested from the web. Text in MS Word documents, in Excel spreadsheets or in XML format was also accepted. All documents in the corpus are provided with extensive meta-data, but they can be categorized into 10 genres as shown in Table 1.

3.2.1 Newspaper Articles

Text was acquired from the largest newspaper publishers and news websites in Iceland and a number of smaller publishers, 16 sources in total. Documents from 1998 were acquired from the largest source, *Morgunblaðið*, and from 2004 from the second largest, *Vísir.is*. Documents from these two sources contain more than 75% of the running words for all documents in this category.

3.2.2 Parliamentary Speeches

Texts not protected by copyright were collected from official sources, the biggest of which is Alþingi, the Icelandic parliament, providing parliamentary speeches dating back to 1911 in XML format, containing all relevant meta-data. The speeches have been transcribed and extensively proofread. Although the oldest speeches are from 1911 the bulk of the texts are fairly recent, as documents from the first 30 years (1911-1940) include about 6 million words but documents from the last 30 years (1988-2017) have 120 million words, more than half of this subcorpus.

3.2.3 Adjudications

Adjudications were harvested from the official websites of the Supreme Court of Iceland and the eight district courts of Iceland. Around half the documents come from the Supreme Court, dating from 1998-2017. The other half is from the district courts and date from 2006-2017.

3.2.4 Published Books

Published books from The Árni Magnússon Institute's text collection were incorporated into the corpus. Only books published since 1980 were included. They include fiction and non-fiction and vary considerably in length.

3.2.5 Transcribed Radio/Television News

We received transcribed documents of all news programs from 2006-early 2017 from the two biggest broadcasting companies in Iceland, RUV, the Icelandic National Broadcasting Service and 365 the biggest privately owned media company in Iceland. These transcripts include both scripted news, read by reporters, and transcribed interviews aired on the news programs.

3.2.6 Sports News Websites

Articles from two sport news websites, dedicated to football news, are included in the corpus. In the newspaper category sports news can also be found, but we do not separate the sports articles from those sources, but rather keep them with other articles from the same sources.

3.2.7 Law and Regulations

Icelandic law as of September 2017 is included in the corpus. The oldest documents date back to the 13th century but a majority is fairly recent, with more than 60% of the documents dated in the 21st century.

3.2.8 Current Affair Blogs and Articles

Articles on current affairs from three sources who have all been publishing articles for more than 10 years are included in the corpus. One of the three sources includes opinion pieces published in newspapers since 1973.

3.2.9 Informational Articles

The corpus collection also includes the Icelandic *Wikipedia* and the University of Iceland's *Science Web*. They include informational articles on various topics. Approximately 60% of this material comes from *Wikipedia* and 40% from the *Science Web*.

3.2.10 Lifestyle

This category includes articles from a web site concerned with famous people and lifestyle.

Table 1 lists the ten text genres and word count for texts retrieved for the first version of the *IGC*. In total more than 4 million documents were collected containing 1.26 billion words. About 57% of the texts are a part of *IGC1* or available with a special license and about 43% are a part of *IGC2*, available with CC BY license.

3.3 Text cleaning and annotation

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

A pipeline for harvesting, cleaning and annotating the corpus texts was 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.

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). Each document collected for the corpus is distributed in one file, which is comprised of a header, containing metadata and a body which includes the document text, lemmas and morphosyntactic tags.

N-grams (n up to 5) are also created for use with the n-gram viewer and for distribution.

```
<text>
  <body>
    <div1>
      <p n="1">
        <s n="1">
          <w lemma="hver" type="fshen">Hvað</w>
          <w lemma="vera" type="sfg3en">er</w>
          <w lemma="hægur" type="lhensf">hægt</w>
          <w lemma="að" type="cn">að</w>
          <w lemma="lesa" type="sng">lesa</w>
          <w lemma="úr" type="aþ">úr</w>
          <w lemma="þessi" type="favebþ">þessari</w>
          <w lemma="mynd" type="nveþ">mynd</w>
          <c type="punctuation">?</c>
        </s>
      </p>
      <p n="2">
        <s n="1">
          <w lemma="blár" type="lkfnvf">Bláu</w>
          <w lemma="litur" type="nkfng">litirnir</w>
          <w lemma="sýna" type="sfg3fn">sýna</w>
          <w lemma="hvar" type="aa">hvar</w>
          <w lemma="kljósandi" type="nkfn">kljósendur</w>
          <w lemma="liggja" type="sfg3fn">liggja</w>
          <w lemma="þétt" type="aae">þéttast</w>
          <c type="punctuation">.</c>
        </s>
      </p>
    </div1>
  </body>
</text>
```

Figure 1: Text body in an XML file from the corpus.

Figure 1 shows the text body in one TEI-conformant XML document. The text is divided into numbered paragraphs and within each paragraph there are numbered sentences. Each line within the sentences contain different elements for words and for punctuation. The elements for words have lemma and type elements, the type element contains the morphosyntactic tag.

3.4 Tagset

Sentence segmentation and tokenization is performed with the IceNLP toolkit (Loftsson and Rögnvaldsson 2007), the same procedures as were used for the *MIM* corpus (Helgadóttir et al. 2012). *IceStagger* (Loftsson & Östling 2013) is used for tagging the *IGC*. A corpus made by concatenating the *IFD* corpus and the *MIM-GOLD* corpus was used to train IceStagger. Dictionaries used when tagging were augmented with the dictionary of The Database of Modern Icelandic Inflection, *BÍN* (Bjarnadóttir, 2012).

The tagset is a revised version of the tagset used for the *IFD* corpus. A tag for abbreviations has been added and another for e-mail and web addresses. There are more than 670 possible morphosyntactic tags in the tagset, and 559 are found in the corpus. More than 50% of the words in the corpus are tagged with only 16 of the most frequent tags. Examples of tags are shown in Table 2, which lists the 5 most frequent tags in the *IGC*. A complete description of the tagset is available at málföng.is, where the corpus can be downloaded.[8]

Tag	Description	Count	% of Total
aa	Adverb – does not govern case	97,761,983	7.79%
c	Conjunction	93,105,617	7.41%
aþ	Preposition – governs dative	91,594,602	7.29%
ao	Preposition – governs accusative	53,114,369	4.23%
sfg3en	Verb – indicative, active, 3rd person, singular, nominative	48,700,790	3.88%

Table 2: Most frequent tags in the *IGC*

3.5 Lemmatization

A new tool is being developed for lemmatizing Icelandic text. A pre-release version of this tool was used for lemmatizing the *IGC* as 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 carried out.

3.6 Metadata

All texts in the corpus are accompanied by metadata. For published texts, the metadata comprises bibliographic data like title, name of author(s), name of editor(s) (if applicable), publisher, date and place of publishing. If it was published on the web the URL is included. For other texts, metadata is recorded to identify the text. For spoken data, various information on the recorded sessions and the speakers is registered. The metadata is shown for each text example retrieved through the search interface and is

8 http://malfong.is/files/rmh_tagset_files_en.pdf

a part of the downloadable texts in TEI conformant XML format. Texts can be selected for search through the search interface classified by publishing date, author and source, which reflects approximately the classification in Table 1.

4. Availability and use

The main object of building the corpus is to make it available for use in LT projects. For other uses, such as linguistics research, teaching, lexicography or other studies the data will be available in a web-based concordance tool on the website malheildir.arnastofnun.is. The Swedish platform KORP[9] (Borin et al. 2012) which in turn uses the *IMS* Corpus Workbench[10] (Evert & Hardie 2011) as a search engine was adapted to be used with 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 as they become available.

The corpus texts are available for download in the TEI-conformant XML format (TEI Consortium 2017). As mentioned in Section 3.1 the corpus has been divided into two parts, *IGC1* and *IGC2* for download where *IGC1* is made available with a special license developed for the *MIM* corpus and *IGC2* with CC BY license. This situation is reflected in the download procedures. The corpus is available for download through the Icelandic LT resources website *Málföng*.[11]

The corpus texts are also available through an n-gram viewer based on NB N-gram viewer (Birkenes et al. 2015). The n-gram viewer is accessible on n.arnastofnun.is.

To aid developers of LT tools the corpus website allows 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 is a valuable resource for builders of LT tools for Icelandic. It is also useful for linguists, 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 will be published yearly. Official texts will be added continually and texts protected by copyright as long as permission for the use of the text has been secured. The tools in the corpus pipeline will also be upgraded as better tools or versions become available and the corpus texts reannotated to reflect improved precision and reliability of the tools.

6. Acknowledgements

The corpus was compiled during the years 2015 to 2017 at the Árni Magnússon Institute for Icelandic Studies and was funded mostly by the Infrastructure Fund (no. 151110-0031, project manager Eiríkur Rögnvaldsson), the Contribution Grants Fund (Mótframlagasjóður) at the University of Iceland and by the Icelandic Ministry of Education and Culture. The authors would like to thank Gunnar Thor Örnólfsson for setting up the search interface and Kristján Rúnarsson for helping with text collection. Thanks are also due to the developers of Korp, especially Lars Borin. We would also like to thank the anonymous reviewers for helpful comments. Last, but not least, thanks are due to publishers and editors who gave permission for the use of their text and made digital copies available. Without the cooperation of all these people this project could not have been completed.

9 https://spraakbanken.gu.se/korp/

10 http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html

11 http://www.malfong.is /

7. Bibliographical References

Birkenes, M., Johnsen, L., Lindstad, A. and Ostad, J. (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, pages 293–295, Vilnius, Lithuania.

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. 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), pages 67–72, Istanbul, Turkey.

Borin, L., Forsberg, M. and Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pages 474–478, Istanbul, Turkey.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, Birmingham, UK.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K. and Loftsson, H. (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), pages 67–72, Istanbul, Turkey

Loftsson, H., Yngvason, J., Helgadóttir, S. and Rögnvaldsson, E. (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, Malta.

Loftsson, H. and Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In Proceedings of Interspeech 2007, Special Session: "Speech and language technology for less-resourced languages", Antwerp, Belgium.

Loftsson, H. and Östling, R. (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, Norway.

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., The Google Books Team, Pickett, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. In *Science* 331(6014): 176-182.

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

Nikulásdóttir, A., Guðnason, J. and Steingrímsson, S. (2017). Máltækni fyrir íslensku 2018-2022: verkáætlun. [Language Technology for Icelandic 2018-2022: Strategic Plan] Mennta- og menningarmálaráðuneytið. Reykjavík.

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

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

Steingrímsson, S., Helgadóttir, S. and Rögnvaldsson, E. (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, pages 287–291, Vilnius, Lithuania.

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/> (September 2017).

Östling, R. (2012). Stagger: A modern POS tagger for Swedish. In: Proceedings of the Swedish Language Technology Conference, SLTC, Lund, Sweden.