Proceedings of ParlaCLARIN II Workshop, pages 11–17
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

IGC-Parl: Icelandic Corpus of Parliamentary Proceedings

Steinþór Steingrímsson, Starkaður Barkarson, Gunnar Thor Örnólfsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
steinthor.steingrimsson, starkadur.barkarson, gunnar.thor.ornolfsson@arnastofnun.is

Abstract

We describe the acquisition, annotation and encoding of the corpus of the Althingi parliamentary proceedings. The first version of the corpus includes speeches from 1911-2019. It comprises 404 thousand speeches and just under 219 million words. The corpus has been automatically part-of-speech tagged and lemmatised. It is annotated with extensive metadata about the speeches, speakers and political parties, including speech topic, whether the speaker is in the government coalition or opposition, age and gender of speaker at the time of delivery, references to sound and video recordings and more. The corpus is encoded in accordance with the Text Encoding Initiative (TEI) Guidelines and conforms to the Parla-CLARIN schema. We plan to update the corpus annually and its major versions will be archived in the CLARIN.IS repository. It is available for download and search using the KORP concordance tool. Furthermore, information on word frequency are accessible in a custom made web application and an n-gram viewer.

Keywords: corpora, parliamentary, Icelandic

1. Introduction

Parliamentary records, reports, written questions and inquiries, legal documents, and transcriptions of debates are a rich source of data for research in various disciplines. Not only do they enable new lines of research in fields as diverse as linguistics, political science, sociology, economic history, gender studies and information retrieval, they can also be an important source of data for natural language processing. This has led to a number of projects aiming to compile, analyze and enrich parliamentary records, some of which are listed in Section 2.

Transcriptions of speeches from debates in Althingi, the Icelandic parliament, have previously been made available as part of the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018), but as parliamentary data differ in many ways from other data in the IGC we have compiled a special parliamentary corpus, IGC-Parl, enriched with meta-data pertaining to parliamentary corpora. This includes information about age, gender and role of the parliamentarian delivering the speech, when they were active in parliament, party affiliation, whether their political party was part of a ruling coalition or in opposition, topics of the speech and more. This project is the first step in compiling a number of different corpora with rich, relevant metadata, from the different text classes in the IGC. Others are for example: news media, books, adjudications, social media, etc. As with the IGC we plan to update the corpus annually and make it available through a variety of means as detailed in Section 5. In subsequent versions we also plan to add other parliamentary records, starting with inquiries, resolutions and bills as discussed in section 7.

In order for IGC-Parl to be in line with comparable corpora from other countries we aspire to adhere to the Parla-CLARIN scheme[1]. This is in accordance with the TEI-standard which is the standard the IGC follows. For further discussion on encoding and annotation of the corpus, refer to Section 4.

[1] https://clarin-eric.github.io/parla-clarin

Icelandic parliamentary speeches have been used for linguistic research, most recently in (Stefánsdóttir and Ingason, 2019) where change in stylistic fronting in the speeches of one long-standing parliamentarian is used to support the hypothesis that a person's language can vary in accordance to change in social status.

By making parliamentary data available with rich metadata we facilitate replicability of results as well as further research in this field. In Section 6 we give a few examples of possible research directions the corpus data and accompanying metadata enable.

2. Related work

Parliamentary corpora have been available to NLP researchers for a long time. In the early days of statistical machine translation (SMT), the Canadian Hansard Corpus (Roukos and Melamed, 1995), consisting of debates from the Canadian Parliament parallel in English and Canadian French, was used by machine translation researchers to advance their field. Europarl (Koehn, 2005), a multilingual parallel corpus containing the proceedings of the European parliament has also proved to be useful for SMT, and the DCEP corpus (Hajlaoui et al., 2014) adds a variety of document types published by the European parliament, enabling new lines of research on the European parliament data.

Recently there has been increased interest in parliamentary corpora and compiling of corpora has been undertaken in a number of countries. Examples of monolingual corpora are the Hansard corpus, a collection of parliamentary records of the British Parliament from 1803-2005 (Alexander and Davies, 2015), the ParlAT beta corpus (Wissik and Pirker, 2018) containing Austrian parliamentary proceedings from 1996-2017, presented at the ParlaCLARIN workshop in 2018 (Fišer et al., 2018) along with a Slovenian Corpus (Andrej Pančur and Erjavec, 2018), a Polish one (Ogrodniczuk, 2018) and a corpus of the Grand National Assembly of Turkey (Onur Gungor and Çağıl Sönmez, 2018). As of February 2020, the CLARIN ERIC infrastructure offers access to 22 parliamentary corpora. This includes corpora in almost all of the languages spoken in CLARIN ERIC

member and observer countries, including all the Nordic languages except Icelandic.[2]

As the number of parliamentary corpora increases, the need for a common format and interoperability grows accordingly. A common schema could facilitate comparison and analysis of topics across parliamentary data from different countries. A standard format for parliamentary data is being proposed by CLARIN. At a workshop in May 2019, Tomaž Erjavec & Andrej Pančur introduced the proposed scheme, Parla-CLARIN (Erjavec and Pančur, 2019).

3. Building the Corpus

The corpus covers speeches delivered from 1911 to mid 2019 in the Parliament of Iceland, Althingi, and are found on Althingi's website (www.althingi.is). Many of the older speeches are missing from the website despite still being listed, as discussed further in Section 6. Speeches predating 1991 are tagged only with a date, while those delivered after that time are tagged with a timestamp.

A short biography of each speaker is available on the website along with information about what political party and constituency they belonged to and what role they had in different periods.

Since 2001 debates concerning specific issues have been grouped by topic (e.g. Industries: Fisheries) so it is possible to link each speech to one or more topics.

A schedule of each session from 1995 onwards is available, but earlier speeches can only be grouped by date. This does not always cohere with sessions since one session can span more than one calendar day.

All the speeches on the website are available in HTML files. Older speeches, for which there are no available sound recordings, have been OCR-read from Alþingistíðindi, parliamentary records that have been published by the Icelandic Parliament since 1875. These speeches have been manually corrected. Refer to Section 6 for further discussion.

3.1. From IGC to IceCor-Parl

The first version of The Icelandic Gigaword Corpus (IGC) was published in 2018 and the second one in January 2020. It contains almost 1,400 million words from different sources, mainly official texts (e.g. parliamentary speeches as far back as 1911, law texts, adjudications) and texts from news media. By the end of 2021 we intend to split the corpus into various sub-corpora, contained in separate files, each with its own metadata structure. In the current version of the IGC each speech is contained in a separate file with information about the name of the speaker, date of delivery and the title. IGC-Parl contains much more detailed information about each speech and speaker.

We started by scraping the website and entering information about each speaker into a database: his or her id used on the website, full name, date of birth, and gender. For each period information was gathered about the person's status in the parliament, to which political party and constituency they belonged and if they held a position as a minister in the government. We also collected information about which political parties were in the government from 1944.

Information about all issues discussed, for each parliament, was inserted into the database and speeches linked to topics where possible.

```
<teiCorpus>
 <teiHeader>
   <profileDesc>
    <particDesc>
     <listPerson>
      <!-- List of all speekers -->
     </listPerson>
     <listOrg>
      <!-- list of all political parties -->
     </listOrg>
    </particDesc>
   </profileDesc>
 </teiHeader>

 <TEI> <!-- TEI element for each day/session

  <teiHeader>
   <!-- information about time and setting -->
  </teiHeader>

  <TEI> <!-- TEI element for each speech
   <teiHeader>
    <sourceDesc>
     <biblStruct>
      <analytic>
       <title></title>
       <author>
        <!-- Information about the speaker -->
       </author>
       <date>
        <!-- date and time of delivery -->
       </date>
      </analytic>
      <ref> <!-- linkt to speech --> </ref>
     </biblStruct>
     <recordingStmt>
      <!-- information about and link to media files -->
     </recordingStmt>
    </sourceDesc>
    <profileDesc>
     <textClass>
      <catRef>
       <!-- list of topics -->
      <catRef>
     </textClass>
    </profileDesc>
   </teiHeader>
  </TEI>

 </TEI>

</teiCorpus>
```

Figure 1: The TEI encoding structure of IGC-Parl.

4. Annotation

For annotation we opted for TEI rather than other XML schemas for encoding parliamentary proceedings, such as Political Mashups (Gielissen and Marx, 2009), Parliamentary Metadata Language (Gartner, 2014) or the Akoma Ntoso[3]. Iceland joined CLARIN ERIC in 2020 after having been an observer since 2018. CLARIN has advocated for the use of TEI and has published the Parla-CLARIN schema[4] for use for annotation of political proceedings. TEI has been used for annotation of other Icelandic corpora and the IGC is all annotated in TEI.

[2] https://www.clarin.eu/resource-families/parliamentary-corpora

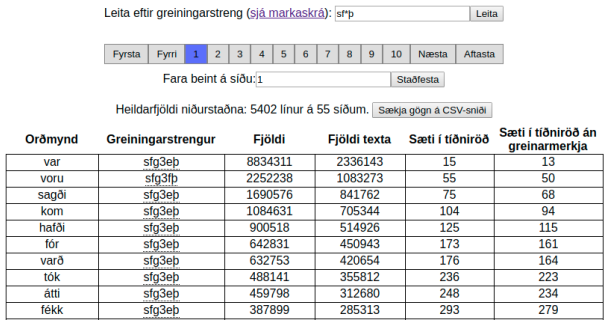
[3] http://www.akomantoso.org

[4] https://clarin-eric.github.io/parla-clarin

4.1. Encoding into XML

We mostly follow the proposed schema, Parla-CLARIN. The root element <teiCorpus> contains a header with metadata for the entire corpus. The <particDesc> element contains a list of all the speakers in the <listPerson> element, their name, gender, date of birth, their party affiliation and roles in different periods. It also contains a list of all political parties, congresses, governments and constituencies in <listOrg> elements. The header is followed by teiCorpus-elements. Each teiCorpus-element contains data for one date (but not session due to lack of information as mentioned above) - a header with metadata and one TEI-element for each speech. Each TEI-element contains another header with metadata and the text element for each speech. The metadata contains detailed information about the speaker in the <author> element, where age, gender, party affiliation and role at the time of delivery are each in one <note> element. If the speaker belonged to the government when the speech was delivered another <note> will indicate that. The number and title of the issue discussed are in the <title> element, and date and time of delivery in the <date> element. Related topics and categories are listed in the <textClass> element in the <profileDesc> element while the link to the written speech on Althingi's website is located in a <ref> element nested in the <biblScope> element. Sound files for each speech have been available since 2006 and video files since 2009. Information and links to those files are in the <recordingStmt> element. The structure is shown in Figure 1.

The Parla-CLARIN schema proposes two layers, a <teiCorpus> containing a header and <TEI> elements that each contains one session, sitting or day. We have three layers since we keep all speeches in a separate TEI-element. The reason for this was twofold. Firstly we wanted to facilitate a search where a user of the corpus would for example only want to look at speeches by women or by a specific speaker. Secondly we want to keep a certain conformity between the sub-corpora of IGC and one of the general rules is that a text by one author/speaker is contained in one separate TEI-element.


Leita eftir greiningarstreng (sjá markaskrá): sf*þ Leita

Fyrsta Fyrri 1 2 3 4 5 6 7 8 9 10 Næsta Aftasta

Fara beint á síðu: 1 Staðfesta

Heildarfjöldi niðurstaðna: 5402 línur á 55 síðum. Sækja gögn á CSV-sniði

Orðmynd	Greiningarstrengur	Fjöldi	Fjöldi texta	Sæti í tíðniröð	Sæti í tíðniröð án greinarmerkja
var	sfg3eþ	8834311	2336143	15	13
voru	sfg3fþ	2252238	1083273	55	50
sagði	sfg3eþ	1690576	841762	75	68
kom	sfg3eþ	1084631	705344	104	94
hafði	sfg3eþ	900518	514926	125	115
fór	sfg3eþ	642831	450943	173	161
varð	sfg3eþ	632753	420654	176	164
tók	sfg3eþ	488141	355812	236	223
átti	sfg3eþ	459798	312680	248	234
fékk	sfg3eþ	387899	285313	293	279

Figure 2: The word frequency database. An example of a search using wildcards. The search term *sf*þ* corresponds with verbs in the indicative mood and past tense. The columns of the results table are respectively: Word form, PoS-tag, Count, Number of texts containing word, Frequency ranking, Frequency ranking excluding punctuation tokens.

The speeches have been tokenised and both lemmas and PoS-tags are listed with each token. Sentences are marked up using the <s> element, words with the <w> element and punctuation symbols with the <c> element. The base form of words is given in the lemma attribute while the tag is given in the type attribute.

Alongside the main TEI-file another XML-file resides that contains a list of all the speakers and a link to all their speeches to facilitate search by speakers.

4.2. Tokenisation, POS-tagging and lemmatisation

All linguistic annotation is carried out automatically with no manual correction. *Tokenizer*, developed by Miðeind (Þorsteinsson, 2020) was used to divide the text into sentences and running words. Morphosyntactic tagging was performed with *ABLTagger* (Steingrímsson et al., 2019) and lemmatisation with the lemmatiser *Nefnir* (Ingólfsdóttir et al., 2019). These tools have been shown to achieve state-of-the-art results for Icelandic, with *ABLTagger* reaching over 95% accuracy on 10-fold validation using a gold standard corpus, and *Nefnir* has been shown to reach almost 97% accuracy when lemmatising words that have previously been tagged automatically. It should be noted that none of these tests have been carried out on texts from parliamentary speeches and thus the exact accuracy for this corpus is not known. The tagset used for tagging IGC is almost the same tagset that was developed for compiling the Icelandic Frequency Dictionary (IFD) (Pind et al., 1991), with only a few changes. A corpus made by concatenating the IFD corpus and the MIM-GOLD corpus (Loftsson et al., 2010) was used to train the tagger. Lexical data from The Database of Icelandic Inflections (DIM) (Bjarnadóttir et al., 2019) was used to augment the tagger for increased accuracy.

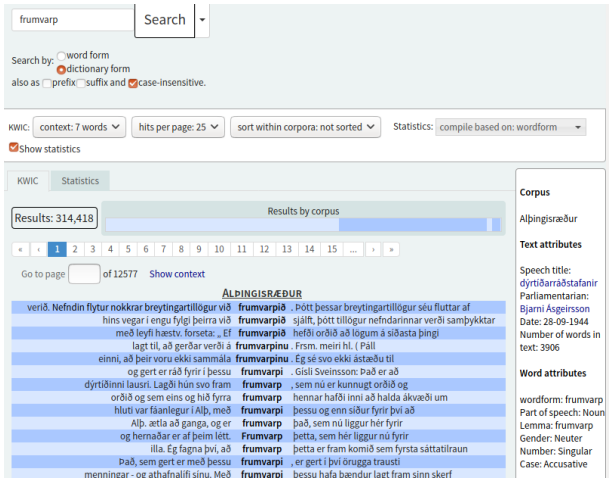


Figure 3: Search results as shown in the Korp-based concordance search tool. The text that matches the search term or pattern is shown in bold, with its context on either side displayed. The right-hand sidebar displays metadata about the text from which each result is sourced.

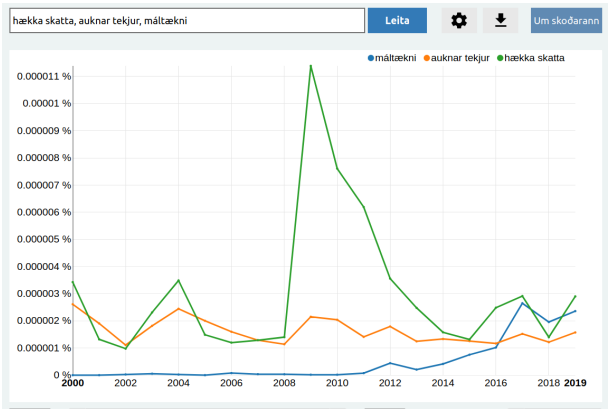


Figure 4: n-gram statistics displayed by the n-gram viewer. The search terms roughly correspond to "raise taxes", "increased income", and "language technology". The large spike is for "raise taxes" in 2009, in the wake of the financial collapse of 2008.

5. Availability and Maintenance

IGC-Parl is available for download on CLARIN.is, the official website of the Icelandic branch of CLARIN ERIC. It is published under a CC-BY 4.0 license and may therefore be used for both academic and industry purposes. We aim to publish a new version annually, adding more material and employing more sophisticated and accurate methods of tokenisation, PoS-tagging and lemmatisation as they become available. Furthermore, we will adapt existing Icelandic corpus tools to support IGC-Parl.

Various tools have recently been developed or adapted to make Icelandic text corpora more accessible to researchers and the general populace. These tools include a concordance search tool, a searchable word frequency database, and an n-gram viewer. They are described in detail in Steingrímsson et al. (2020). Here we will give a brief description of the tools through which IGC-Parl will be made accessible.

The concordance tool is powered by the Swedish Language Institute Språkbanken's Korp (Borin et al., 2018) and offers a concordance search using a rich query language[5] (Evert, 2005), enabling users to search by all linguistic features denoted in Section 4.2.

The word frequency database enables users to search by words, canonical forms (lemmas) and fine-grained morphosyntactic PoS-tags. It also supports wildcard searches for each of these search terms, enabling users to see frequency statistics on all words or tags which match a pattern. An example of a wildcard search is shown in Figure 2.

Analysing trends in word occurrence by year can prove useful, for example in research into language change and neologisms. We make the data available in an n-gram viewer, which enables the user to view frequency trends for n-grams up to a length of 3, for both lemmas and word forms as they appear in the text. The user can configure the viewer in a number of ways, for instance to give absolute or relative numbers for each year or display the results as a cumulative or non-cumulative curve. Results can be downloaded, either as SVG graphics or in a comma separated text file. The n-gram viewer is based on the NB n-gram viewer. (Breder Birkenes et al., 2015) All n-grams will be made available for download. Trigrams of length $<= 3$ are available in their entirety, but with 4- and 5-grams, only those which occur more than 3 times in the corpora are included in the download files.

[5]The CQP query language: `http://cwb.sourceforge.net/`

6. Quantitative Analysis

In total, the corpus contains 218,889,307 tokens in 404,401 speeches, given by 987 different speakers. The Althingi parliamentary records were printed from 1845. They include parliamentary documents and speeches. In recent years, the Althingi has been working on digitising old documents and they intend to go as far back as 1875. Currently, all speeches since 1937 have been made available as well as some of the speeches from 1911 to 1936, which were published on the Althingi website as part of celebrating certain events, e.g. women's suffrage, enacted in 1915. As evident in Figure 6 our corpus reflects this. A reasonable amount of speeches from 1913–1915 are included but there is a gap in the speech collection until 1937, when we finally have a complete set of speeches for each year. As we intend to publish new versions of the corpus annually, which include the most recent data, older speeches will also be included in future versions as they become available.

As the IGC-Parl corpus has been packaged with rich metadata, interesting information about the parliament can easily be congregated. In the following examples we have done so with Python scripts that read through the data. Figure 6 shows the number of words spoken each year in parliament. The columns show the division of word count between male and female parliamentarians. The graph shows us that it isn't until recent years that women have come close to being represented as well as men. By having a gender flag for the parliamentarians, we can study various gender related issues, look into whether men or women are more likely to speak on certain topics, or compare linguistic characteristics or speech length, to name a few. Similar analyses could be done on other parliaments and the comparison could give us valuable insights into gender biases in various countries, at least as they are reflected in parliament.

Enthusiastic followers of the Icelandic parliamentary debates would probably agree that a prominent trait of the Icelandic parliament is that the opposition traditionally speaks more than the ruling coalition, even though there is no tradition for minority coalition governance in Iceland and the

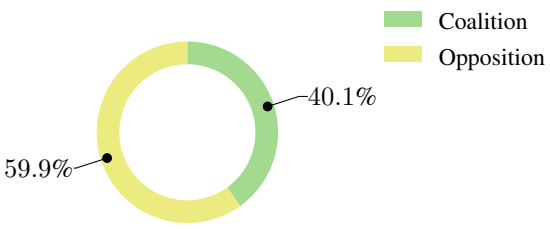


Figure 5: Words spoken by opposition vs. ruling parties

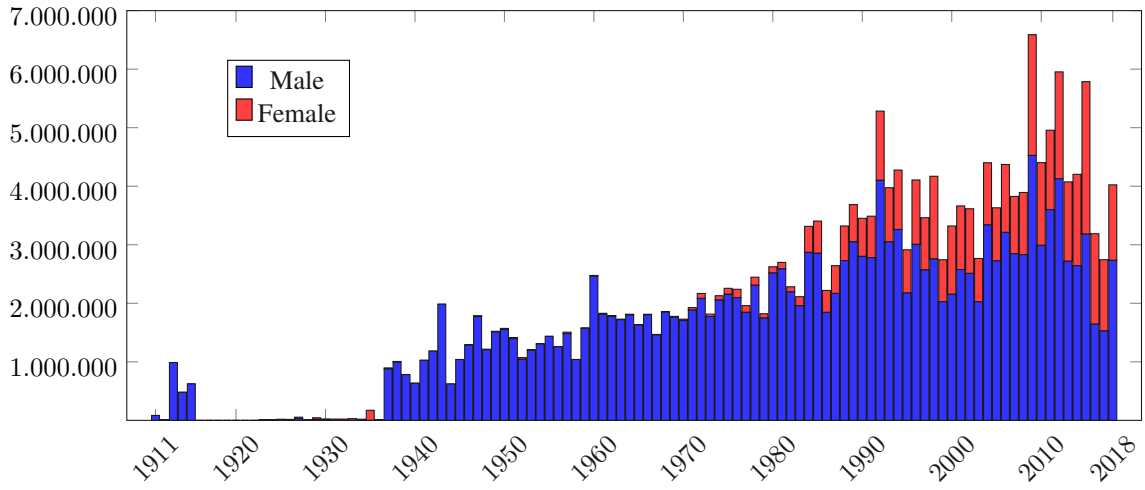


Figure 6: Number of words spoken by year.

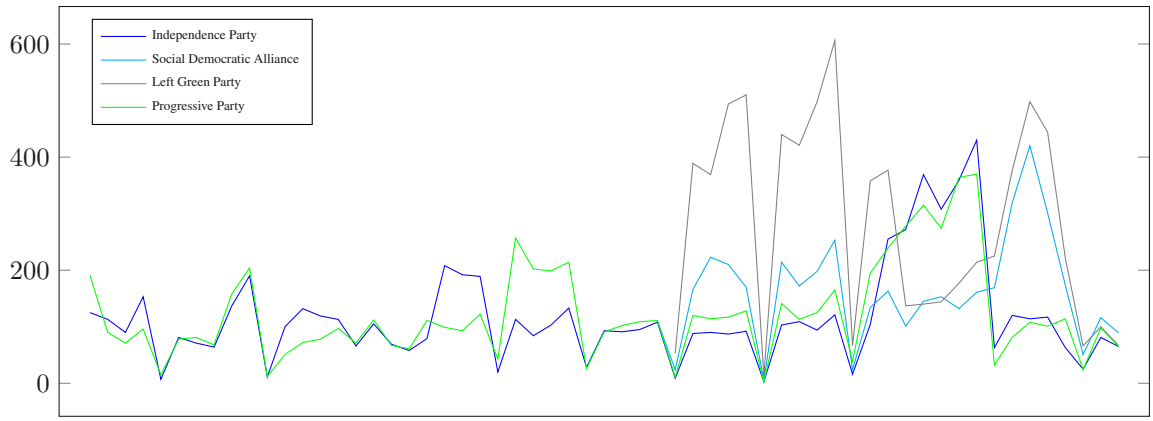


Figure 7: Words spoken on average per day per parliamentarian per parliamentary term from 1970 to 2018

coalition has almost always held the majority. We can test this characteristic of the parliament by using the information the corpus gives us on which parties are in power at which time. We use this information to count speeches and total words spoken by both the opposition and coalition from 1937 when the data is first complete. Figure 5 shows the results and verifies our suspicion.

Figure 7 shows another aspect of this same thing. The graph shows words spoken on average per day, by the average parliamentarian in each of the four largest parties. Each point in the graph stands for a legislative period, lasting from a few weeks and up to five months. The number of meetings in each legislative period varies, with occasional summer legislative periods that have relatively few meetings compared to number of days, explaining the points in the graph where the average for all parties is very low. The graph starts in 1970 and ends in 2018.

The Independence Party and the Progressive Party usually speak less than the Left Greens and the Social Democrats. This may be because they are more often a part of the ruling coalition. An exception to that in 2009–2013 can be seen clearly in the graph as the parliamentarians of these parties had never before spoken as much. When they came back to power they went back to their old ways.

These are only a few examples of how the data can be analysed. The rich metadata provides abundant research paths for academics and other users to discover.

7. Current and Future Work

While we have published the first version of IGC-Parl[6], an annotated corpus of parliamentary speeches enriched with a variety of relevant metadata, work on Icelandic corpus compilation is ongoing. In subsequent versions of IGC-Parl we want to improve on the metadata even further. Speeches from 1991 onwards have a topic classification, but the topics are only listed in Icelandic. We want to translate the topic listings to English in order to make them more useful for non-Icelandic speakers. A new named entity recogniser is being built for Icelandic. We will use that to add named entity annotation to future versions of the corpus. As we do not know the PoS-tagging and lemmatisation accuracy for this particular corpus we plan to manually create a test set that will allow us to measure the accuracy of the automatic tools. As this corpus is somewhat different from the data the PoS-tagger is trained on, we want to investigate whether there are any systemic errors made in the tagging process. If so, we want to try to alleviate them in later versions of the corpus by adjusting the tagger.

[6] http://hdl.handle.net/20.500.12537/14

The frequency database tool described in Section 5 will get a more user friendly design and will be optimised for speed. And while the corpus can currently be queried in an n-gram viewer we have plans to allow for more detailed querying, for instance by party affiliation.

Furthermore we intend to add other parliamentary data to the corpus, starting with inquiries and replies to them, resolutions and bills and amendments to them.

As we are trying to adhere to the Parla-CLARIN schema, we will follow the advancement of the scheme and in future versions amend our encoding to be as close to the CLARIN standard for parliamentary corpora as we see fit.

8. Conclusion and Acknowledgements

We have described the compilation of IGC-Parl, the metadata and the composition of the data. The work is funded by the Language Technology Program for Icelandic 2019–2023 (Nikulásdóttir et al., 2020).

9. Bibliographical References

Andrej Pančur, M. and Erjavec, T. (2018). SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, NODALIDA 2019, Turku, Finland.

Borin, L., Forsberg, M., and Roxendal, J. (2018). Korp the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 474–478.

Breder Birkenes, M., Johnsen, L. G., Lindstad, A. M., 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, Vilnius, Lithuania.

Erjavec, T. and Pančur, A. (2019). Introduction to the proposed annotation scheme. ParlaFormat Workshop.

Evert, S. (2005). The CQP query language tutorial. *IMS Stuttgart. CWB version*, 2.

Darja Fišer, et al., editors. (2018). *Proceedings of LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora.*, Paris, France, May. European Language Resources Association (ELRA).

Gartner, R. (2014). A metadata infrastructure for the analysis of parliamentary proceedings. In *Big Humanities Data, The Second IEEE Big Data 2014 Workshop*, Bethesda, Maryland, USA.

Gielissen, T. and Marx, M. (2009). Exemplification of Parliamentary Debates. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop*, DIR 2010, pages 19–25.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). DCEP -digital corpus of the European parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, NODALIDA 2019, Turku, Finland.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In Francis M. Tyers Sarasola, Kepa et al., editors, *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, Valetta, Malta.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Ogrodniczuk, M. (2018). Polish Parliamentary Corpus. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Onur Gungor, M. T. and Çağıl Sönmez. (2018). A Corpus of Grand National Assembly of Turkish Parliament's Transcripts. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Þorsteinsson, V. (2020). Tokenizer for icelandic text. CLARIN-IS, Stofnun Árna Magnússonar.

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

Roukos, Salim, D. G. and Melamed, D. (1995). Hansard French/English LDC95T20.

Stefánsdóttir, L. B. and Ingason, A. K. (2019). Lifespan Change and Style Shift in the Icelandic Gigaword Corpus. In K. Simov et al., editors, *Proceedings of CLARIN Annual Conference 2019*, pages 138–141.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Steingrímsson, S., Kárason, Ö., and Loftsson, H. (2019). Augmenting a BiLSTM tagger with a morphological lex-

icon and a lexical category identification step. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2019, Varna, Bulgaria.

Steingrímsson, S., Barkarson, S., and Örnólfsson, G. T. (2020). Facilitating Corpus Usage: Making Icelandic Corpora More Accessible for Researchers and Language Users. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

10. Language Resource References

Marc Alexander and Mark Davies. (2015). *Hansard Corpus 1803-2005*. Available online at http://www.hansard-corpus.org.

Tanja Wissik and Hannes Pirker. (2018). *ParlAT Corpus*. Austrian Centre for Digital Humanities.