Proceedings of ParlaCLARIN II Workshop, pages 44–50
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Parsing Icelandic Alþingi Transcripts: Parliamentary Speeches as a Genre

Kristján Rúnarsson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
Árnagarði við Suðurgötu, IS-102
krunars@hi.is, einar.freyr.sigurdsson@arnastofnun.is

Abstract

We introduce a corpus of transcripts from Alþingi, the Icelandic parliament. The corpus is syntactically parsed for phrase structure according to the annotation scheme of the Icelandic Parsed Historical Corpus (IcePaHC). This addition to IcePaHC makes it more diverse with respect to text types and we argue that having a syntactically parsed corpus facilitates research on different types of texts. We furthermore argue that the speech corpus can be treated somewhat like spoken language even though the transcripts differ in various ways from daily spoken language. We also compare this text type to other types and argue that this genre can shed light on their properties. Finally, we show how this addition to IcePaHC has helped us identify and solve issues with our parsing scheme.

Keywords: parliamentary corpus, parliamentary transcripts, text types

1. Introduction

In this paper we discuss a corpus of Icelandic Alþingi parliamentary speeches, syntactically parsed for phrase structure.[1] The corpus, which contains approx. 60,000 words, is parsed in accordance with the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011).[2] This new addition to IcePaHC makes it more diverse with respect to genres of texts. We chose unprepared speeches to make the parliamentary speech corpus more coherent and closer to reflecting actual spoken language.

In this paper, we first of all argue that having a syntactically parsed corpus facilitates research on different text types. It is therefore crucial for us to have the parliamentary speeches parsed in the same way as the other 1 million words found in IcePaHC.

Secondly, we focus on the properties of the unprepared speeches we chose for the corpus. We argue that they can be treated like spoken language in important ways, though they differ in various ways from "regular" spoken language.

Thirdly, we argue that the text type under discussion can shed light on other text types. For example, long clauses containing many words seem to be one of the characteristics not only of the parliamentary speeches but also of religious texts, whereas clauses in narratives tend to be much shorter. Finding common traits in the speeches and the religious texts may help us discover the defining characteristics of these two genres.

Furthermore, we will show examples of how this new addition to IcePaHC has helped us identify and solve issues with our parsing scheme.

The paper is structured as follows: In Section 2 we give a brief description of IcePaHC. Section 3 discusses the parliamentary transcripts and looks at parliamentary speeches as a text type as opposed to other genres. Section 4 discusses how the addition of parliamentary transcripts has impacted the annotation scheme of IcePaHC. Section 5 concludes the paper.

2. IcePaHC

The Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011; Rögnvaldsson et al., 2012) is a collection of parsed texts containing 1 million running words from the 12th through the 21st centuries. It is annotated according to a scheme based on that of the Penn Parsed Corpora of Historical English (https://www.ling.upenn.edu/hist-corpora/) (Kroch and Taylor, 2000; Kroch et al., 2004) using the annotation tool Annotald (Ecay et al., 2018), after preprocessing, including lemmatization and preliminary parsing, using IceNLP (http://icenlp.sourceforge.net/) — see Loftsson (2008), Loftsson and Rögnvaldsson (2007) and Ingason et al. (2008) — as well as various scripts developed specifically for IcePaHC.

IcePaHC has been designed to capture the Icelandic language in various contexts with regard to time period and subject matter. The texts have been selected so as to be presumably written each mainly by a single author and the length of the excerpts has been decided so that they are short enough that many diverse texts could be included, while still providing adequate coverage of the authors' internal grammar.

IcePaHC aims to include texts in each of several genres (narratives, religion, biographies, science, law) from every century from the 12th century to the present, but currently includes mainly narratives and religious texts. There is also a need for still more types of texts from different authors and times dealing with diverse subjects. The new additions to IcePaHC are genres not previously included, namely parliamentary transcripts and news articles. This paper focuses on the parliamentary transcripts and discusses their importance as a text type.

[1] The creation of the parsed corpus of parliamentary speeches is part of a bigger project named "Universal Treebanking" (Einar Freyr Sigurðsson PI), funded by the Strategic Research and Development Programme for Language Technology 2019–2020 in which IcePaHC is also being converted to a Universal Dependencies scheme.

[2] The parsed Icelandic Alþingi parliamentary speech corpus is available along with the rest of IcePaHC at https://github.com/antonkarl/icecorpus.

3. The Parliamentary Transcripts

3.1. The Nature of the Texts and Their Selection

The parliamentary transcripts are from a small set of only four speakers that have been chosen so as to represent both male and female speakers of different generations: Steingrímur J. Sigfússon (b. 1955), Þorgerður Katrín Gunnarsdóttir (b. 1965), Helgi Hrafn Gunnarsson (b. 1980) and Björt Ólafsdóttir (b. 1983).[3] An important secondary consideration was the existence of enough material from each speaker from a similar time, in this case between 2011 and 2015.

The transcripts, which were extracted from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018), were chosen from among responses rather than prepared speeches, so as to better represent spontaneous speech. The transcripts have, however, been edited by parliamentary secretaries for publication, so the text we have to work with is not pure speech. This is a drawback, especially if the intention is to examine in detail the structure of spoken language as opposed to written language, e.g., getting accurate statistics about the relative prevalence of specific features. Nevertheless, it has been evident in the annotation process that certain features mainly associated with spoken language appear frequently in the transcripts despite the apparent tendency of the editing process to make them more concise and regularly structured and adhere more closely to the norms of formal written language.

It may also be noted that published novels such as are included in IcePaHC's narratives category also go through an editing process with similar aims and tendencies to that of parliamentary speeches before being published, so they too do not perfectly represent the speaker's idiolect.

There are also other circumstances in favor of the parliamentary speeches including ease of access to both text and original audio, lack of copyright restrictions and individual authorship (indisputable in the case of responses, apart from editor changes), which aligns very well with the design goals of IcePaHC.

3.2. Comparison with Other Text Types

It was our belief that there would be important differences between the parliamentary transcripts and the existing IcePaHC corpus, and that constructions might be found there that are not found, or are significantly less common, in other more formal text types.

As expected, the parliamentary transcripts differ from previously added IcePaHC texts in several ways. Disfluencies, fragment answers (i.e., shortened answers to questions), resumptive elements, clefts and arguments shared by conjoined clauses (instances of which might be analysed as right node raising) are some of the phenomena which occur frequently in the parliamentary transcripts.

Adding new text types may also help us understand the nature of other text types, because there are many linguistic factors that could conceivably be affected by the genre. Several factors may be unique for a text type, while others might be shared with other text types.

For example, as discussed in Section 3.4, we see that there is a notable similarity between the parliamentary speeches and religious texts, in contrast to narrative texts. Such findings may reveal something of the nature of these texts, and comparisons of this kind can spark new research, e.g., in sociolinguistics.

In Sections 3.3–3.7 we discuss various linguistic features of the parliamentary transcripts.

3.3. Disfluencies

We can expect to find disfluencies of various sorts – such as breaks, false starts and repetitions – to a much higher degree in spoken language than in written texts that are carefully planned and thought through. These include breaks where a sentence or a phrase breaks off or is not finished. An example from our Alþingi corpus is shown below.

(1) Ég veit ekki alveg hvernig ætti að vinna þetta tiltekna frumvarp frekar vegna þess að **það er svo**, – nú vantar mig aftur íslenska orðið fyrir „brutal“ –
'I don't know exactly how this particular bill should be further worked on because **it is so [BREAK]** – now I need again the Icelandic word for "brutal"'

The speaker in this example breaks off when describing the bill as he cannot remember the Icelandic word for English *brutal*. Such breaks are marked specially in the parsing scheme and can therefore be easily found.

The unprepared speeches in our corpus, being spontaneous and not written beforehand, do in fact contain a higher total number of breaks than all the rest of IcePaHC. Even though the parliamentary speech corpus only contains around 60,000 running words, as opposed to the 1 million words of IcePaHC, it has 14 breaks whereas IcePaHC has only 5. It is possible that there is some inconsistency in parsing between the speeches and the rest of IcePaHC but this nonetheless suggests that breaks are more frequently found in the speeches due to their nature as spoken language.

3.4. Clause Length

Matthíasson (1959) argues that parliamentary speeches tend to contain exceptionally many subordinate clauses as a result of their nature, with, e.g., the speeches often being spontaneous.[4] Matthíasson (1959, 206) furthermore claims that increased frequency of subordinate clauses results in longer matrix clauses. It is quite straightforward to investigate the length of clauses with our parsed corpus and we can compare the speeches with other genres, namely narrative and religious texts. When we look at the relative frequency of the three text types (see Figure 1), it turns out that the proportion of short clauses, with four to eight words, is much higher in the narratives. Religious texts and the parliamentary speeches exhibit a similar proportion of longer clauses, on the other hand, as opposed to narrative texts whose longer clauses are proportionally less frequent, as can be clearly seen in Figure 1.

[3] Helgi Hrafn Gunnarsson and Steingrímur J. Sigfússon's speeches were selected as their language had been investigated before; see Stefánsdóttir (2016) and Stefánsdóttir and Ingason (2018).

[4] Verifying Matthíasson's claim should now be possible as different types of subordinate clauses (complement clauses, relative clauses, adverbial clauses, etc.) are all parsed in the corpus.

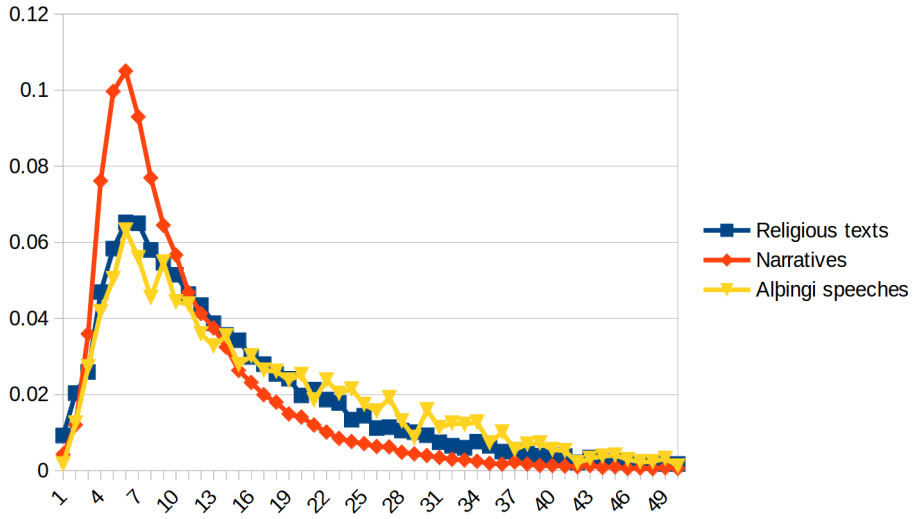

Figure 1: Relative frequency distribution of lengths (in words) of root clauses in different text types (religious texts, narrative texts, parliamentary speeches).

3.5. Resumptive Elements

The resumptive element *þá* 'then' is frequently found in the parliamentary transcripts immediately following left-dislocated subordinate clauses headed by *ef* 'if', *þegar* 'when', *þótt* 'even', *þó (að)* 'even', *þrátt (fyrir að)* 'despite', etc.

(2) en ef fólk vill fara á hausinn **þá** er það
but if people wants.to go bankrupt **then** is that
væntanlega möguleiki líka.
presumably a.possibility too

Some such use of the resumptive *þá* has been linked to spoken language (Thráinsson, 2005, 578) and we have in fact noticed that it is sometimes deleted in the transcripts, presumably because it is frowned upon to an extent. This use is, nonetheless, quite frequently found in our parliamentary speech corpus as well as in other texts in IcePaHC from all periods.
However, we find a certain use of resumptive elements in the transcripts that we do not in other texts in IcePaHC. In some cases, resumptive *þá* is immediately preceded by *að*.[5] *Að* can be many things in Icelandic syntax, such as a preposition, an infinitival marker or a complementizer. Without going into details, it is presumably a complementizer in the *að þá* construction.

(3) Vandinn er auðvitað sá að þegar menn tala um hin ósnortnu víðerni, sem því miður gerast nú ansi fágæt og Ísland býr yfir sumum þeirra, sennilega stærstu ósnortnu víðernum í Evrópu, a.m.k. í Vestur-Evrópu, **að þá** er skilgreiningin sú að þar [...]
'The problem is of course that when people talk about the untouched wilderness, which unfortunately are now becoming quite rare and Iceland has got some of them, probably the biggest untouched wilderness in Europe, at least in Western Europe, **that then** the definition is that there...'

[5] For a syntactic analysis, see Jónsson (2019).

Thráinsson (2005, 578) mentions the use of resumptive *að þá*, taking it to be even more connected to spoken language use than *þá*. The use of resumptive *þá* and *að þá* merits further research but for now it suffices to point out that the use of resumptive *að þá* found in the parliamentary speeches is indicative of the spoken language trait of this particular text type.

3.6. Topic Expressions

Topic-introducing expressions seem to be relatively frequent in the parliamentary speeches as opposed to other text types in IcePaHC. These are expressions starting with words like *varðandi* 'regarding', *að því er varðar* 'regarding', *hvað varðar* 'as regards', *hvað viðkemur* 'as regards'. To give an example, there are 42 instances of matrix clauses starting with the topic introducer *varðandi* in our parliamentary corpus, as in (4), but none in other IcePaHC texts.[6]

(4) **Varðandi aukna kostnaðarþátttöku sjúklinga** líst mér að sjálfsögðu illa á hana.
'**Regarding an increased cost participation of patients**, I do not, of course, like it.'

This is something that needs further investigation. That is, are such topic introducers more frequent in spoken language than written texts? This shows a clear need for more parsed transcripts of spoken language. We therefore leave this for future research.

3.7. Words Indicative of Informal Register

When we try to figure out the properties of a certain text type, it is worth looking at the individual words used as well as the syntax. For that purpose, we do not need to rely on a syntactically parsed corpus as we can search for particular words in other corpora such as the Icelandic Gigaword Corpus (IGC; https://malheildir.arnastofnun.is/) (Steingrímsson et al., 2018). Svavarsdóttir (2007, 38–39) looks at word use

[6] It should be noted that we would not expect to find this particular expression in older texts in IcePaHC.

in three corpora of different types; she discusses words that are found in spoken language dialogues and to some degree in what she calls informal texts (diaries, etc.) but not, or to a much lesser degree, in rather formal, written texts (newspaper texts). Looking through her list of words, we note that she mentions, for example, *ókei* 'okay', which is neither found in her formal nor informal text corpus. There are, however, 16 instances of *ókei* in the spoken language corpus she reports on. Searching IGC, we find several instances of the word *ókei* in parliamentary transcripts, which without a doubt does not belong to a formal register.

Note that while we are arguing that the Icelandic parliamentary speeches share various properties with other types of spoken language, we are not arguing that parliamentary speeches are like any other type of (informal) spoken language. Members of parliament sometimes use loanwords from other languages, like English, which they often ask the audience to excuse (by adding a phrase like *svo ég sletti* 'so I use a foreign word/expression'); this may be indicative of a somewhat formal setting. We will not look further into this for now.

4. Development of the Annotation Scheme

The prevalence of certain features has prompted a deeper look into the way syntactic structure is analysed and annotated in IcePaHC, both shedding light on old issues that had never been definitively settled during earlier work on IcePaHC and bringing new issues to our attention.

The situation of IcePaHC is peculiar in that its annotation scheme (http://linguist.is/icelandic_treebank/) is derived from one developed for historical English texts (https://www.ling.upenn.edu/hist-corpora/) (Kroch and Taylor, 2000; Kroch et al., 2004). While the fact that it has been developed for Early Modern and Middle English rather than just contemporary English has made the annotation scheme more suitable for Icelandic, there are still important features of Icelandic that affect the practicability of specific analytical choices that have been retained from the Penn Parsed Corpora of Historical English (PPCHE) scheme.[7]

In particular, Icelandic is a highly inflected language, much more so than English, especially with regard to case. While the English of PPCHE, especially Middle English, does have some limited case inflection, it is a language in transition and it is not always clear to what degree inflection exists and case has generally not been annotated. By contrast, Icelandic shows a clear distinction between the cases and cases have been annotated in IcePaHC.

This affects the analysis of presumed instances of right node raising and that of comparative phrases which have been presumed to contain a prepositional phrase.

4.1. Right Node Raising

One issue where Icelandic does not seem to conform to the scheme is that of right node raising. An English example from Postal (1974, 126) is shown in (5).

(5) Jack may be—and Tony certainly is—a werewolf.

Here, the NP *a werewolf* applies to the two matrix clauses, i.e., *Jack may be a werewolf* and *Tony certainly is a werewolf*.[8]

The parsing scheme employed in IcePaHC has presumed that the second clause in right node raising is parenthetical and that its rightmost element is raised so as to appear in the appropriate place in the encompassing prior clause. This analysis has been inherited from PPCHE. While it might be practical for English, it causes problems when applied to Icelandic, because the evidence clearly shows that it is the second clause which governs the case of the shared constituent and not the former. If the shared phrase has been moved (with argument movement), it would be expected to acquire its case from the governor of the place it was moved to, and if it has not been moved it should likewise retain the appropriate case for its position. An analysis involving right node raising is shown below where the dashed line rectangle marks the parenthetical clause (IP-MAT-PRN).

```
( (IP-MAT (NP-SBJ (PRO-N Við)) 'we'
  (BEDI vorum) 'were'
  (PP (P í) 'in'
    (NP (ADJ-D miklu) 'much'
      (N-D sambandi) 'contact'
      (PP (P við) 'with'
        (IP-MAT-PRN (CONJ og) 'and'
          (NP-SBJ *con*)
          (VBDI fengum) 'got'
          (NP-OB1 (NP (ADJ-A góða) 'good'
            (N-A leiðsögn)) 'guidance'
            (CONJP (CONJ eða) 'or'
              (NP (ADJ-A góða) 'good'
                (N-A áminningu))) 'reminder'
            (CP-REL *ICH*-1))
          (PP (P frá-frá))) 'from'
        (NP (NPR-D Sambandi) 'association'
          (NP-POS (ADJ-G íslenskra)
                          'Icelandic-GEN'
            (NS-G sveitarfélaga))
                  'municipalities-GEN'
        ))))
  (CP-REL-1 (WNP-2 0)
    (C sem) 'which'
      (IP-SUB (NP-SBJ *T*-2)
        (NEG ekki) 'not'
        (BEDI var) 'was'
        (VAN svarað))) 'answered'
  (, ,-,)))
```

In this example, one and the same NP, *Sambandi íslenskra sveitarfélaga*, applies to two clauses; it is simultaneously, in a way, the object of two prepositions, *við* and *frá*, and the question is how best to account for that within the scheme. The NP headed by *Sambandi* is in the dative case as the preposition *frá* assigns dative to its complement, but according to the analysis above it ends up in a PP with the

[7] More information about the annotation scheme for the Penn Historical Corpora may be found in the annotation manual at https://ling.upenn.edu/~beatrice/annotation/.

[8] Without going into details of the original account in Postal (1974), right node raising "places a double of the sequence [i.e. the phrase which is identical in both clauses] on the right, by Chomsky adjunction, and deletes all original occurrences" (p. 126).

preposition *við* 'with', which should govern the accusative case. Furthermore, there is an extraposed relative clause (CP-REL) belonging to the parenthetical clause (IP-MAT-PRN), but it ends up having to be raised to the outer main clause because it appears after the raised NP.
We therefore came up with a different scheme which appears to fit the Icelandic pattern better:

```
( (IP-MAT (IP-MAT (NP-SBJ (PRO-N Við))
      (BEDI vorum)
      (PP (P í)
        (NP (ADJ-D miklu)
        (N-D sambandi)
        (PP (P við)))))
    (CONJP (CONJ og)
      (IP-MAT (NP-SBJ *con*)
        (VBDI fengum)
        (NP-OB1 (NP (ADJ-A góða)
          (N-A leiðsögn))
          (CONJP (CONJ eða)
            (NP (ADJ-A góða)
              (N-A áminningu)))
          (CP-REL *ICH*-1))
        (PP (P frá)
          (NP (NPR-D Sambandi)
            (NP-POS (ADJ-G íslenskra)
              (NS-G sveitarfélaga))))
        (CP-REL-1 (WNP-2 0)
          (C sem)
          (IP-SUB (NP-SBJ *T*-2)
            (NEG ekki)
            (BEDI var)
            (VAN svarað)))))
  (, ,-,)))
```

Here, the shared NP appears in a PP with a preposition governing the correct case, and instead of a parenthetical clause, the two matrix clauses are conjoined in the most usual way, using a conjunction phrase.

4.2. Resumptive NPs in Comparative Clauses

Another issue that has been identified as a problem in IcePaHC is the treatment of comparative constructions of the form *<COMP ADJ/ADV> than ...*, *so/as <ADJ/ADV> as ...*, etc., also inherited from PPCHE. These constructions are parsed as adjectival or adverbial phrases containing a prepositional phrase, where the preposition is the word *than* or *as* (in Icelandic *en*, *og*) that immediately follows the head adjective/adverb. In case a subordinate clause with a gap corresponding in function to the head follows, the complement of the preposition is a complementizer phrase containing the subordinate clause IP, as in the following example:

```
(ADJP (ADJR-N helgari) 'holier'
  (PP (P en) 'than'
    (CP-CMP (WADJP-2 0)
      (C 0)
      (IP-SUB (ADJP *T*-2)
        (NP-SBJ (OTHERS-N aðrir) 'other'
          (ADJ-N helgir) 'holy'
          (NS-N menn)))))) 'men'
```

While this two-layer PP/CP combination might seem odd it is in line with the treatment of various other subordinate clause types in IcePaHC and the Penn Parsed Corpora of Historical English. However, when, instead of gap, an overt phrase corresponding to the antecedent is used, the structure is simplified, as shown in the following example, where the pronoun *sig* corresponds to the prior NP *engan betri vin*:

```
(IP-SUB (NP-SBJ (NPR-N Gróa)) 'Gróa'
  (VBDS ætti) 'had'
  (NP-OB1 (Q-A engan) 'no'
    (ADJR-A betri) 'better'
    (N-A vin) 'friend'
    (PP *ICH*-4))
  (ADVP-LOC (ADV hér) 'here'
    (PP (P á) 'on'
      (NP (N-D jörðu)))) 'earth'
  (PP-4 (P en) 'than'
    (NP (PRO-A sig)))) 'him/her'
```

From an English-speaking, or caseless, point of view, this seems to work rather well and even to confirm the appropriateness of calling the traditionally termed subordinating conjunctions prepositions, which may otherwise seem idiosyncratic. For Icelandic, however, the grammatical case of the supposed prepositional complement shows that it cannot be a direct complement to the preposition, since a preposition governs a specific case, but the phrases in question do not take their case from any preposition, but rather agree in case with their antecedents.
This prompted us to include the CP-CMP in such constructions as well, allowing for an IP therein where the phrase could fill the same role as its antecedent. That raised another issue: how does the WH-phrase in the CP connect to the subordinate clause? The following shows an attempt at this, using a dummy adverbial phrase (ADVP) that has no clear semantic role or connection to the antecedent:

```
( (IP-IMP (VBPI Nýtum) 'let's utilize'
  (ADVP (ADV þá)) 'then'
  (NP-OB1 (N-A tíma$) 'time'
    (D-A $nn)) 'the'
  (PP (RP fram) 'forward'
    (P að) 'to'
    (NP (ADJS-D næstu) 'next'
    (NS-D þingkosningum)
           'parliamentary elections'
    (, ,-,)
    (CP-REL (WNP-1 0)
      (C sem) 'which'
      (IP-SUB (NP-SBJ *T*-1)
        (RDPI verða) 'will be'
        (ADVP (ADV vonandi)) 'hopefully'
        (ADVP (ADVR fyrr) 'sooner'
        (PP (P en) 'than'
          (CP-CMP (WADVP-2 0)
            (C 0)
            (IP-SUB (ADVP *T*-2)
              (ADVP (ADVR síðar))) 'later'
          )))))))
    (. .-.)))
```

A possible solution to the problem was found in another CP construction – the relative clause. The following example shows a resumptive NP, *spítalann* 'the hospital', being used in lieu of a gap in a relative clause.

```
( (IP-MAT (ADVP (ADVS Helst)) 'chiefly'
  (BEPI er) 'is'
  (NP-SBJ (PRO-N það)) 'it'
  (NP-PRD
    (NPR-N Landspítali$) 'National Hospital'
    (D-N $nn)) 'the'
  (CP-CLF (WNP-1 0)
    (C sem) 'which'
    (IP-SUB (NP-SBJ (PRO-N við)) 'we'
      (VBPI sjáum) 'see'
      (CP-THT (C að) 'that'
        (IP-SUB (NP-SBJ *exp*)
          (NP-ADV (NP-ADV (N-A ár) 'year'
            (PP (P frá) 'from'
              (NP (N-D ári))))) 'year'
          (, ,-,)
          (BEPI er) 'is'
          (NEG ekki) 'not'
          (DAN gert) 'done'
          (ADVP
            (ADV nægjanlega) 'adequately'
            (ADV vel)) 'well'
          (PP (P við) 'to'
            (NP-RSP-1
              (N-A spítala$) 'hospital'
              (D-A $nn))))))) 'the'
  (. .-.)))
```

This analysis is based on that of the PPCHE; it is also used particularly in the Penn audio-aligned corpora (Tortora et al., 2017; Tortora et al., 2020) and has been used in IcePaHC before. The WH-phrase is generally considered to have been moved from the IP and in a relative clause corresponds to the antecedent that the relative clause speaks about. In comparative clauses there is also a comparative phrase that is an antecedent to the CP. Using the same method of connecting the phrase that corresponds to the antecedent to the WH-phrase neatly ties together the treatment of different types of CP in gapped and ungapped variants.

```
( (IP-IMP (VBPI Nýtum) 'let's utilize'
  (ADVP (ADV þá)) 'then'
  (NP-OB1 (N-A tíma$) 'time'
    (D-A $nn)) 'the'
  (PP (RP fram) 'forward'
    (P að) 'to'
    (NP (ADJS-D næstu) 'next'
    (NS-D þingkosningum)
        'parliament elections'
    (, ,-,)
    (CP-REL (WNP-1 0)
      (C sem) 'which'
      (IP-SUB (NP-SBJ *T*-1)
        (RDPI verða) 'will be'
        (ADVP (ADV vonandi)) 'hopefully'
        (ADVP (ADVR fyrr) 'sooner'
        (PP (P en) 'than'
          (CP-CMP (WADVP-2 0)
            (C 0)
            (IP-SUB (ADVP-RSP-2
                      (ADVR síðar))) 'later'
        )))))))
  (. .-.)))
```

5. Conclusion

We have established that the Icelandic Parsed Historical Corpus benefits from the addition of parliamentary transcripts by demonstrating their unique qualities while also showing their potential relationships to other types of text. Furthermore, we have found that adding new types of texts inspires us to improve our analysis in unanticipated ways. We are currently working on a further addition to the treebank, and it is our hope that even more text types will be added in the future so that it represents as good a cross-section of the language as possible.

6. Acknowledgements

This project is funded by the Strategic Research and Development Programme for Language Technology, grant no. 180020-5301.

We would also like to thank the three anonymous reviewers for their comments on the paper.

7. Bibliographical References

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In Bengt Nordström et al., editors, Advances in Natural Language Processing. 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25–27, 2008 Proceedings, pages 205–216, Berlin. Springer.

Jónsson, J. G. (2019). The XP-*þá*-construction and V2. In Ken Ramshøj Christensen, et al., editors, The Sign of the V: Papers in Honour of Sten Vikner, pages 341–360, Aarhus University. Department of English School of Communication and Culture.

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An incremental finite-state parser for Icelandic. In Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007), pages 128–135, Tartu, Estonia, May. University of Tartu, Estonia.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31:47–72.

Matthíasson, H. (1959). Setningaform og stíll. Bókaútgáfa Menningarsjóðs, Reykjavík.

Postal, P. M. (1974). On Raising. One Rule of English Grammar and its Theoretical Implications. MIT Press, Cambridge, MA.

Stefánsdóttir, L. B. and Ingason, A. K. (2018). A high definition study of syntactic lifespan change. *University of Pennsylvania Working Papers in Linguistics*, 24.1:169–178. https://repository.upenn.edu/pwpl/vol24/iss1/20/.

Stefánsdóttir, L. B. (2016). Breytingar á framburði. Með hliðsjón af félagslegum þáttum. B.A. thesis, University of Iceland, Reykjavík. http://hdl.handle.net/1946/24333.

Svavarsdóttir, Á. (2007). Talmál og málheildir — talmál og orðabækur. *Orð og tunga*, 9:25–50.

Thráinsson, H. (2005). Setningar. Handbók um setningafræði. Íslensk tunga III. Almenna bókafélagið, Reykjavík.

8. Language Resource References

Ecay, A., Beck, J., and Ingason, A. K. (2018). Annotald. Version 1.3.10. http://annotald.github.io.

Kroch, A. S. and Taylor, A. (2000). Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition. Size: 1.3 million words.

Kroch, A. S., Santorini, B., and Delfs, L. (2004). Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. Size: 1.8 million words.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey.

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.

Tortora, C., Santorini, B., Blanchette, F., and Diertani, C. (2017). The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE), version 0.1. www.aapcappe.org. Size: 1.8 million words.

Tortora, C., Cutler, C., Haddican, B., Newman, M., Santorini, B., and Diertani, C. (2020). Corpus of New York City English (CUNY-CoNYCE). https://conyce.commons.gc.cuny.edu/.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank.