CLARIN Annual Conference Proceedings

2024

Edited by

Vincent Vandeghinste and Thalassia Kontino

15 – 17 October 2024
Barcelona, Spain

Please cite as:
CLARIN Annual Conference Proceedings, 2024. ISSN 2773-2177 (online).
Eds. Vincent Vandeghinste and Thalassia Kontino.
Barcelona, Spain, 2024.

IcePaHC 2024.03 – A Significant Treebank Upgrade

Joel C. Wallenberg
University of York
joel.wallenberg@york.ac.uk

Anton Karl Ingason
University of Iceland
antoni@hi.is

Einar Freyr Sigurðsson
Árni Magnússon Institute for Icelandic Studies
einar.freyr.sigurdsson@arnastofnun.is

Eiríkur Rögnvaldsson
University of Iceland
eirikur@hi.is

Abstract

The version of the Icelandic Parsed Historical Corpus (IcePaHC) that was released in 2011 and later made available through CLARIN has facilitated a wide range of studies, both in terms of theoretical linguistics and Natural Language Processing. Here, we discuss how IcePaHC has been used throughout the years and present the first major update to IcePaHC in 13 years, version 2024.03, involving thousands of corrections that allow for more precise research than before. The current version is released under a Creative Commons Attribution license (CC BY).

1 Introduction

The Icelandic Parsed Historical Corpus (IcePaHC) (Rögnvaldsson et al., 2011, 2012; Wallenberg et al., 2011) is a manually annotated phrase structure treebank that contains approximately 1 million words of parsed text that has been sampled from historical data.[1] The treebank contains texts from every century from the 12th century to the 21st century, mostly from narratives and religious texts that are spread more or less evenly across this period.

When IcePaHC 0.9 was released in 2011, it was the first major treebank for Icelandic and thus it paved the way for various types of studies that had not been feasible before. However, a treebank is a complicated dataset that must be maintained because there are no practical methods available that make sure that a treebank is completely free from errors. Despite various manual and automatic methods to minimize errors, it remains an ongoing process to fix mistakes in the data, and thus the GitHub repository for IcePaHC (https://github.com/antonkarl/icecorpus/) has evolved continuously over the last 13 years, and especially during 2023 and 2024 as the authors of the corpus have made a systematic effort to correct as many errors as possible. The outcome of these efforts is the release of a new major update of the treebank to CLARIN, version 2024.03, now available for download (Wallenberg et al., 2024).

This paper is organized as follows. Section 2 reviews some background on the IcePaHC treebank and related resources. In Section 3 we discuss how the corpus has been used in linguistics research and in Section 4 we go over some studies that have been carried out using IcePaHC that involve Natural Language Processing (NLP). Section 5 discusses the new version and Section 6 concludes.

2 Background

The IcePaHC corpus has its roots in a research program that goes back to the annotation of the Penn Treebank, the first major phrase structure treebank (Marcus et al., 1993). The Penn Treebank was, in turn, followed by the Penn Parsed Corpora of Historical English, also developed at the University of Pennsylvania (Kroch & Taylor, 2000b; Kroch et al., 2004). This second iteration of treebank development involved some improvements to the annotation scheme, including a more flat structure in cases that involved ambiguity, such as in PP-attachment and the ordering of elements within the verb phrase. This is important in order to not have the annotation involve too many decisions that are simply based on a

[1] We thank three anonymous reviewers for their comments on this paper.

convention rather than a reliable analysis and also because during historical change, the theoretically appropriate analysis is not always clear as grammar competition may be ongoing in the speech community (Kroch, 1989, 2001; Kroch & Taylor, 2000a).

This tradition of annotating historical corpora was continued and extended to the Icelandic language in the project Viable Language Technology beyond English – Icelandic as a test case whose PI was Eiríkur Rögnvaldsson. This served a dual function as the treebank was intended both for linguistic research and development of Natural Language Processing tools (Rögnvaldsson, 2010). The IcePaHC project used a semi-automatic method, for example by running the shallow parser IceParser (Loftsson & Rögnvaldsson, 2007) and structure-modifying search queries in CorpusSearch (Randall, 2005) before manual correction and further manual annotation of the phrase structure was carried out.

3 Uses in Linguistics

The IcePaHC treebank has been used for several linguistics studies. For example, Ingason et al. (2013) used the frequencies of certain types of passive constructions to model the evolutionary trajectory of the Icelandic New Passive and make predictions about its spread, even into the future. In this case, it was the Variational Model of Language Acquisition (Yang, 2002) that provided a theoretical foundation for a predictive analysis but the treebank was the source of the empirical counts that were used in the analysis.

Wallenberg et al. (2021) used IcePaHC to study the relationship between information smoothness in the sense of Information Theory and the trajectory of diachronic change. In this case, word orders that serve as a diagnostics for certain constructions were extracted along with features that characterized the environment of each example and the text of each sentence was submitted to a function that calculated an indicator of information smoothness/density. The study found that Information Theoretic factors have a significant effect on how historical change evolves over time.

Various other topics have been studied, including expletives and cataphora (Booth, 2018, 2019), and verb second vs. verb first word orders (Booth & Beck, 2021). Schätzle (2018) furthermore studied dative subjects in the history of Icelandic with an emphasis on data visualization.

4 Uses in Natural Language Processing

The most prominent task for treebanks in Natural Language Processing is the possibility of training data-driven parsers on the manually annotated trees such that a system becomes available for automatic parsing of the same type of trees, given any arbitrary text. IcePaHC was used along with the Faroese Parsed Historical Corpus (FarPaHC) (Ingason et al., 2012) for an experiment that made use of the fact that the two insular Scandinavian languages have a similar syntax. The experiment involved training a parser on a mixture of the two languages and found that parsing accuracy in Faroese can be improved by adding Icelandic data to the training dataset (Ingason et al., 2014).

A pipeline for parsing Icelandic text was trained and made available by Jökulsdóttir et al. (2019). This CLARIN resource focused on providing the basic infrastructure needed for setting up parsing pipelines for Icelandic and it included a configuration of the Berkeley parser that had been trained on IcePaHC. This pipeline setup was used to facilitate the first release of a neural parsing pipeline for Icelandic in Arnardóttir and Ingason (2020), a system that got an F1 score of 84.74% and was also trained on IcePaHC – and furthermore it was supported by the application of a multilingual BERT model. Although IcePaHC is a phrase structure treebank, it has also served as the foundation for development of dependency parsing for Icelandic due to the development of conversions from phrase structure to Universal Dependencies (UD) (Arnardóttir et al., 2020; Arnardóttir et al., 2023). Both IcePaHC and FarPaHC have been converted to a UD format. Furthermore, IcePaHC served as a model for the parsing of parliament speeches (Rúnarsson & Sigurðsson, 2020) and sports-news texts[2] with the parsing subsequently being converted to a UD format. The newest versions of all three UD treebanks, UD_Icelandic-IcePaHC (Arnardóttir, Hafsteinsson, Sigurðsson, Jónsdóttir, et al., 2024), UD_Icelandic-Modern (Rúnarsson et al., 2024) and

[2]The parsed texts have not been published as of yet but the parsing, carried out by Kristján Rúnarsson, can be found at our GitHub repository: https://github.com/antonkarl/icecorpus/.

UD_Faroese-FarPaHC (Arnardóttir, Hafsteinsson, Sigurðsson, Ingason, et al., 2024), are found in the latest release of Universal Dependencies.

5 The New Version

Building on the success of IcePaHC and its open access release on CLARIN, we now present a new major upgrade of the resource. The treebank contains the same texts but thousands of corrections have been made to the annotation, resulting in a more accurate resource for use in both linguistics and language technology. The new and updated version has already been made available on CLARIN and it is called IcePaHC 2024.03. This means that instead of using version numbers like 0.3, 0.5, 0.9, like we have done in the past, the version number now follows a system where the release date is used as the basis of the numbering, March 2024 in this case.

Various types of corrections have been made to both syntactic structure, Part-of-Speech tags and lemmatization. This last point is particularly significant since thousands of corrections involve correcting lemmas in the corpus. This is particularly useful when designing queries that target the dictionary form of a word. Structure correction has also made use of the fact that the treebank is in open access and therefore enjoys regular feedback from its users. We are grateful for the emails we have received about aspects of the annotation that needed to be reconsidered and we have done so on many occasions throughout the years.

6 Conclusion

We have described the IcePaHC corpus, its many uses, and the new significant upgrade that has now been released. While several language resources have been made available in the last few years (Nikulásdóttir et al., 2022), Icelandic remains a low-resource language (Rehm & Way, 2023) and therefore every step counts along the path towards a more robust Language Technology ecosystem for the language. The release of IcePaHC 0.9 was a major step in 2011 and facilitated diverse research in the following years. Now that a new version has been made available with more precise annotation, we remain optimistic that the tradition of studying Icelandic phrase structure computationally continues to be a fruitful enterprise.

References

Arnardóttir, Þ., Hafsteinsson, H., Jasonarson, A., Ingason, A., & Steingrímsson, S. (2023). Evaluating a Universal Dependencies Conversion Pipeline for Icelandic. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 698–704.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., & Steingrímsson, S. (2020). A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 16–25.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Ingason, A. K., Rögnvaldsson, E., & Wallenberg, J. C. (2024). UD_Faroese-FarPaHC. In Universal Dependencies 2.14.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Jónsdóttir, H., Bjarnadóttir, K., Ingason, A. K., Rúnarsson, K., Steingrímsson, S., Wallenberg, J. C., & Rögnvaldsson, E. (2024). UD_Icelandic-IcePaHC. In Universal Dependencies 2.14. http://hdl.handle.net/11234/1-5502

Arnardóttir, Þ., & Ingason, A. K. (2020). A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser. *Proceedings of CLARIN Annual Conference*, 48–51.

Booth, H. (2018). *Expletives and Clause Structure. Syntactic Change in Icelandic* [Doctoral dissertation, The University of Manchester].

Booth, H. (2019). Cataphora, expletives and impersonal constructions in the history of Icelandic. *Nordic Journal of Linguistics*, *42*(2), 139–164.

Booth, H., & Beck, C. (2021). Verb-second and verb-first in the history of Icelandic. *Journal of Historical Syntax*, *5*(28), 1–53.

Ingason, A. K., Legate, J. A., & Yang, C. (2013). The Evolutionary Trajectory of the Icelandic New Passive. *University of Pennsylvania Working Papers in Linguistics*, *19*(2), 91–100.

Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Wallenberg, J. C. (2014). Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 91–95.

Ingason, A. K., Sigurðsson, E. F., Rögnvaldsson, E., & Wallenberg, J. C. (2012). Faroese Parsed Historical Corpus (FarPaHC) 0.1 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/92

Jökulsdóttir, T. F., Ingason, A. K., & Sigurðsson, E. F. (2019). A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus. *Proceedings of CLARIN Annual Conference*, 138–141.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*, 199–244.

Kroch, A. S. (2001). Syntactic Change. In M. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory* (pp. 698–729).

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

Kroch, A. S., & Taylor, A. (2000a). Verb-Object Order in Early Middle English. *Diachronic Syntax: Models and Mechanisms*, 132–163.

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

Loftsson, H., & Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. *Proceedings of the 16th Nordic Conference of Computational Linguistics*, 128–135.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

Nikulásdóttir, A. B., Arnardóttir, Þ., Barkarson, S., Guðnason, J., Gunnarsson, Þ. D., Ingason, A. K., Jónsson, H. P., Loftsson, H., Óladóttir, H., Rögnvaldsson, E., Sigurðsson, E. F., Sigurgeirsson, A. Þ., Snæbjarnarson, V., Steingrímsson, S., & Örnólfsson, G. T. (2022). Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. *Selected Papers from the CLARIN Annual Conference 2021*, 109–125.

Randall, B. (2005). CorpusSearch 2 User's Guide. University of Pennsylvania.

Rehm, G., & Way, A. (Eds.). (2023). *European Language Equality – A Strategic Agenda for Digital Language Equality*.

Rúnarsson, K., & Sigurðsson, E. F. (2020). Parsing Icelandic Alþingi Transcripts: Parliamentary Speeches as a Genre. *Proceedings of the Second ParlaCLARIN Workshop*, 44–50.

Rúnarsson, K., Arnardóttir, Þ., Hafsteinsson, H., Barkarson, S., Jónsdóttir, H., Steingrímsson, S., & Sigurðsson, E. F. (2024). UD_Icelandic-Modern. In Universal Dependencies 2.14.

Rögnvaldsson, E. (2010). Icelandic language technology: An overview. *Language, Languages and New Technologies: ICT in the Service of Languages*, 187–195.

Rögnvaldsson, E., Ingason, A. K., & Sigurðsson, E. F. (2011). Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). *Language Variation Infrastructure*, 97–112.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1977–1984.

Schätzle, C. (2018). *Dative subjects: Historical change visualized* [Doctoral dissertation, University of Konstanz].

Wallenberg, J. C., Bailes, R., Cuskley, C., & Ingason, A. K. (2021). Smooth Signals and Syntactic Change. *Languages*, *6*(2), 60.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC) 0.9 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/62

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2024). Icelandic Parsed Historical Corpus (IcePaHC) 2024.03 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/325

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press.