# Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian

**Anton Karl Ingason,[1] Hrafn Loftsson,[2] Eiríkur Rögnvaldsson,[3]**
**Einar Freyr Sigurðsson[1] and Joel Wallenberg[4]**
[1]University of Pennsylvania, [2]Reykjavík University, [3]University of Iceland, [4]Newcastle University
[1]Department of Linguistics, Philadelphia, PA 19104-6305,
[2]School of Computer Science, Menntavegi 1, IS-101 Reykjavík, Iceland,
[3]Faculty of Icelandic and Comparative Cultural Studies, Árnagarði við Suðurgötu, IS-101 Reykjavík, Iceland,
[4]School of English Language, Literature, and Linguistics, Newcastle University, Newcastle Upon Tyne, UK, NE1 7RU
[1]{ingason,einarsig}@ling.upenn.edu, [2]hrafn@ru.is, [3]eirikur@hi.is, [4]joel.wallenberg@ncl.ac.uk

**Abstract**

This paper presents ongoing work that aims to improve machine parsing of Faroese using a combination of Faroese and Icelandic training data. We show that even if we only have a relatively small parsed corpus of one language, namely 53,000 words of Faroese, we can obtain better results by adding information about phrase structure from a closely related language which has a similar syntax. Our experiment uses the Berkeley parser. We demonstrate that the addition of Icelandic data without any other modification to the experimental setup results in an f-measure improvement from 75.44% to 78.05% in Faroese and an improvement in part-of-speech tagging accuracy from 88.86% to 90.40%.

**Keywords:** parsing, Faroese, Icelandic

## 1. Introduction

Automatic phrase structure (constituent) parsing is by now available for a number of languages. The state-of-the-art systems are statistical, based on Probabilistic Context-Free Grammars (PCFGs), and trained on parsed corpora, e.g., the Penn Treebank (Marcus et al., 1993) in the case of English. Corpus construction is very time consuming and therefore it remains an important task to develop methods for rapid deployment of parsers for new languages.

In this paper, we describe work in progress on two new parsers for the Insular Scandinavian languages: IceParsald for Icelandic and FarParsald for Faroese. IceParsald is based on a traditional model where the development of a parser is preceded by a full scale corpus building project. In contrast, FarParsald is rapidly deployed by constructing a small corpus and partially employing existing training data from a related language, Icelandic.

We show that our approach improves parse quality for Faroese compared to only using our limited Faroese data. In addition to our methodological experiments, our systems fill a gap in the set of basic language resources for these languages. Full phrase structure parsers have not been available for either of the two languages until now.

Furthermore, our systems are based exclusively on GPL-based free and open source resources, which allows us to make them freely available for future work in academia and beyond.

## 2. Related work

For the current type of a task, we need two types of resources. First, we need general purpose software that is capable of inducing a grammar from appropriately parsed input data, i.e., from a treebank. Second, we need language specific resources, i.e., treebanks for Icelandic and Faroese.

Since the Penn Treebank, considerable attention has been paid to the development of parsing methods/software for training parsers on treebanks (see esp. Collins, 1999; Klein and Manning, 2003; Bikel, 2004; Petrov et al., 2006; Petrov and Klein, 2007). For our purpose, it is important that the software is robust and is freely available under a license that facilitates future work on derived resources and experiments. For these reasons, we decided to use the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007), an implementation of the so-called unlexicalised parsing method (Klein and Manning, 2003).

Efforts in recent years have made available a number of resources for processing Icelandic (e.g., Loftsson, 2008; Dredze and Wallenberg, 2008; Ingason et al., 2008; Loftsson et al., 2009; Rögnvaldsson et al., 2009). This includes IceParser (Loftsson and Rögnvaldsson, 2007), a shallow parser for the language, but until now, a full phrase structure parser has not been developed. Some progress has also been made in developing resources for Faroese, e.g., Trosterud (2009); Antonsen et al. (2010), but the availability of language processing software for Faroese is still quite limited. Our contribution of parsed corpora, which are then trained using a state-of-the-art parsing method, is therefore a significant addition to the set of tools available for the two languages.

## 3. Parsed corpora for training

Our training corpora, IcePaHC for Icelandic (Wallenberg et al., 2011) and FarPaHC for Faroese (Ingason et al., 2012), use the Penn Parsed Corpora of Historical English (PPCHE) annotation scheme (e.g., Santorini 2010). IcePaHC contains 1 million words compared to 1.2 million words in the Penn Parsed Corpus of Middle English (Kroch and Taylor, 2000) and 1.7 million words in the Penn Parsed Corpus of Early Modern English (Kroch et al., 2004). It consists of 60 texts ranging from the $12^{th}$ century to the present – most of them are either narratives (39 texts; ≈ 675,000 words) or religious texts (13 texts; ≈ 235,000 words).

FarPaHC, on the other hand, is much smaller. It consists of 53,000 words from three religious texts, one from the $19^{th}$ century (1823) and two from the $20^{th}$ century (1928 and 1936).

The Insular Scandinavian languages differ from English in, e.g., that they both have rich case morphology (nominative, accusative, dative, genitive) which is, of course, annotated in the corpora. Finite verbs agree with subjects not only in number (as in English) but also in person. The increased morphological richness does of course affect the distribution of evidence available to a parser during training.

## 4. Parsing Icelandic and Faroese

The goals of the current project to parse Icelandic and Faroese are twofold. First, we want to bring full phrase structure parsing to the two languages. Second, we want to demonstrate that having a parsed corpus of a closely related language can improve parsing quality of a language for which only a small corpus is available.

For the Icelandic treebank, IcePaHC, it took 3 annotators 2 years to annotate 1 million words for full phrase structure. This process of building IcePaHC is described in Rögnvaldsson et al. (2011a,b, 2012). The development approach of the Icelandic parser, IceParsald, can therefore be viewed as traditional, with a costly annotation phase. The most significant speed improvement in this project was due to the development of new annotation software, Annotald (Beck et al., 2011), which continues to be actively developed and used in various other projects. IceParsald is the result of training the Berkeley parser on IcePaHC.

Methodologically, the main contribution of the current paper is the less traditional approach to our Faroese parser, FarParsald. The Faroese treebank, FarPaHC, which underlies FarParsald, is a data set which was parsed by one annotator in 3 months (excluding time that was used for coping with the peculiarities of some historical text sources). This short annotation phase for Faroese yielded 53,000 annotated words.

A treebank of this size is quite limited in the context of training a parser. However, making use of the fact that Icelandic and Faroese are closely related languages, we were able to improve the quality of the Faroese automatic parses by training on a combination of Icelandic and Faroese data. The parser has access to the most important patterns in the distribution of terminal nodes from the Faroese data. Adding Icelandic data to the input gives a more accurate view of the syntax of Insular Scandinavian, which is in most respects the same for the two languages.

We observed a positive trend for precision, recall and f-measure as measured by EVALB (Collins, 1997), when adding Icelandic data. Compared to using just FarPaHC, adding 10% and then 20% of IcePaHC results in better parses. The mean f-measure in a 10-fold cross validation improves from 75.44% to 77.22% by adding 10% of IcePaHC, and to 78.05% with 20% of IcePaHC included (the last split was 4-fold) (see Figure 1). The released FarParsald will be based on the combination of FarPaHC and all the Icelandic data.

We should note that the overall quality of the parsing may seem unsatisfying at this point. The software is originally developed for English and nothing has been done in the current experiment to optimize its configuration for Faroese. Also, the annotation scheme in the historical treebanks gives relatively flat structures that sometimes do not include extra levels of bracketing that is unlikely to be misbracketed by a parser. Such structures may artificially increase the number of correct brackets in experiments that assume the Penn treebank annotation scheme. We hope to explore these issues in more detail in future work but for the present purpose, all else being equal, the main result is that adding Icelandic data to the Faroese treebank has a positive effect on the results.

Our experiments also show that even if the most direct evidence for the distribution of terminal nodes in Faroese comes from FarPaHC, the constraints on phrase structure enforced by the added Icelandic data does in fact improve part-of-speech (PoS) tagging accuracy of the Faroese terminals. Using just the Faroese data resulted in 88.86% tagging accuracy, but adding Icelandic data improved the outcome to 90.06% with 10% of IcePaHC added and 90.40% with 20% of IcePaHC added (see Figure 2).

For completeness, we note that it should be quite obvious that using only the Icelandic data does not give good results for Faroese because the information about Faroese terminal nodes comes from FarPaHC. To confirm this, we trained a parser on 30% of IcePaHC and did not include any data from FarPaHC. The resulting parser gives an f-score around 40% and a tagging accuracy around 50%. This confirms that it is in fact the combination of data from the two languages that is important.

## 5. Further work

There is still considerable potential for improving the parsing of Faroese within the general approach of the current work. In this small initial experiment we have used only a subset of IcePaHC to demonstrate the effect we are interested in. Furthermore, even if our combined corpora significantly expand FarPaHC, they are relatively small compared to the Penn Treebank. In future work on the eventual release of FarParsald we plan to use the full IcePaHC corpus and, building on the success of the current experiment, we plan to investigate the effects of adding data from even more sources.

Another avenue we intend to explore in future work is the application of more types of parsing software on the Faroese and Icelandic data and a more careful configuration of the setup to suit the intricacies of these languages and the type of an annotation scheme that is used for the respective treebanks. Whereas the purpose of this paper is to contribute to a discussion on training on multiple languages, the ultimate future goal of this research project will be to make available useful parsers for the two languages.

## 6. Conclusion

We have described work in progress on statistical parsers for Icelandic and Faroese. While the current results are

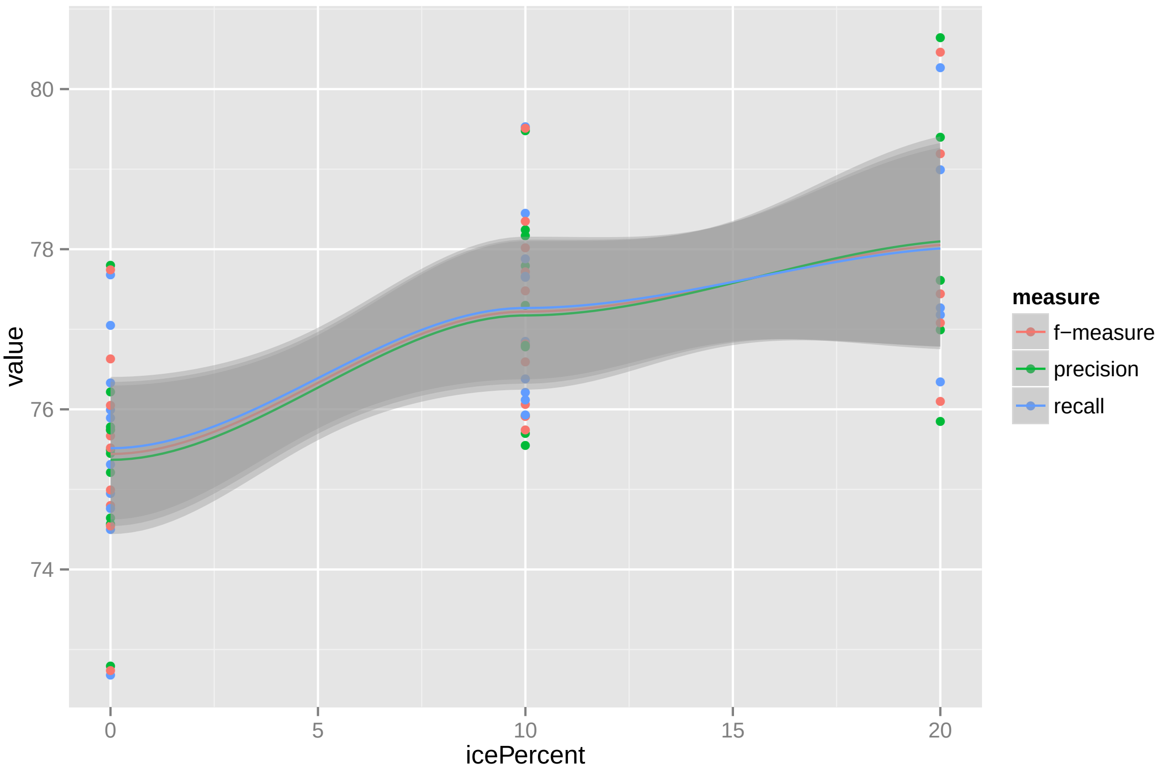

Figure 1: Parser evaluation for FarParsald: icePercent indicates how much of IcePaHC was added to the Faroese training data.

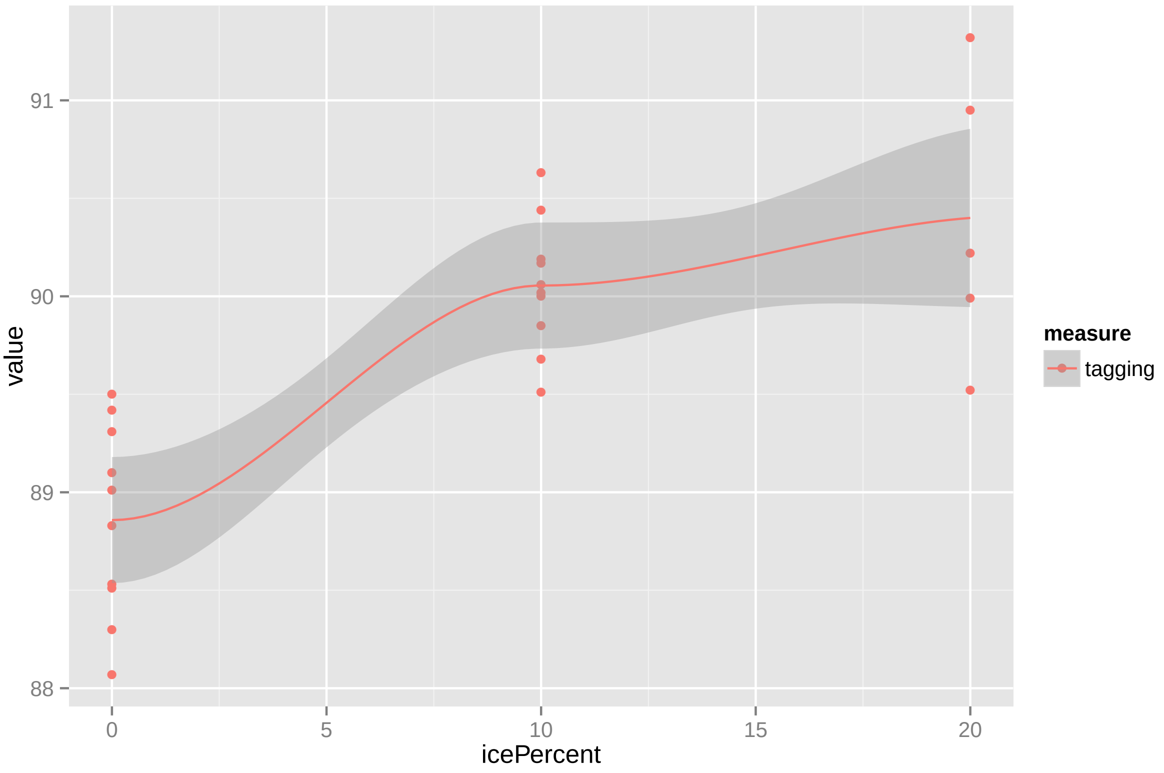

Figure 2: Evaluation of tagging accuracy: icePercent indicates how much of IcePaHC was added to the Faroese training data.

somewhat preliminary and call for more detailed evaluation, they are promising, and they show that having data from a related language, Icelandic, does improve the quality of parsing and PoS tagging of Faroese. In addition to our methodological experiments, we make available parsed corpora and parsers for two languages for which such resources were not previously available.

## 7. Acknowledgements

The construction of IcePaHC was supported by the Icelandic Research Fund (Rannsóknasjóður), grant no 090662011, Viable Language Technology beyond English — Icelandic as a test case; the U.S. National Science Foundation (NSF) International Research Fellowship Program (IRFP), grant #OISE-0853114, Evolution of Language Systems: a comparative study of grammatical change in Icelandic and English; the University of Iceland Research Fund (Rannsóknasjóður Háskóla Íslands), grant Icelandic Diachronic Treebank (Sögulegur íslenskur trjábanki); and the EU ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no 270899 (META-NORD). The Faroese Parsed Historical Corpus was funded by the University of Iceland Research Fund (Rannsóknasjóður Háskóla Íslands), grant Faroese treebank (Frumgerð færeysks trjábanka).

We would like to thank members of the Treebanks Lab at the University of Pennsylvania for helpful comments and discussions. Thanks to anonymous reviewers for useful comments.

## 8. References

Antonsen, L., Trosterud, T., and Wiechetek, L. (2010). Reusing Grammatical Resources for New Languages. In *Proceedings of the $7^{th}$ International Conference on Language Resources and Evaluation*, LREC, Valetta, Malta.

Beck, J. E., Ecay, A., and Ingason, A. K. (2011). Annotald. Version 11.11. [Software for treebank annotation.] http://github.com/janabeck/annotald.

Bikel, D. M. (2004). *On the parameter space of generative lexicalized statistical parsing models*. Ph.D. thesis, University of Pennsylvania.

Collins, M. (1997). Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the $35^{th}$ Annual Meeting of the ACL and Eighth Conference of the EACL*, Madrid, Spain.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Dredze, M. and Wallenberg, J. (2008). Icelandic Data Driven Part of Speech Tagging. In *Proceedings of the $46^{th}$ Annual Meeting of the ACL*. Columbus, OH, USA.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In Raante, A. and Nordström, B., editors, *Advances in Natural Language Processing, $6^{th}$ International Conference on NLP, GoTAL 2008, Proceedings*. Gothenburg, Sweden.

Ingason, A. K., Rögnvaldsson, E., Sigurðsson, E. F., and Wallenberg, J. C. (2012). Faroese Parsed Historical Corpus (IcePaHC). Version 0.1. http://www.linguist.is/farpahc.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the $41^{st}$ Annual Meeting of the ACL*, Sapporo, Japan.

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

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

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M., editors, *Proceedings of the $16^{th}$ Nordic Conference of Computational Linguistics*, NODALIDA. Tartu, Estonia.

Loftsson, H., Kramarczyk, I., Helgadóttir, S., and Rögnvaldsson, E. (2009). Improving the PoS tagging accuracy of Icelandic text. In Jokinen, K. and Bick, E., editors, *Proceedings of the $17^{th}$ Nordic Conference of Computational Linguistics*, NODALIDA. Odense, Denmark.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

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

Petrov, S. and Klein, D. (2007). Improved Inference for Unlexicalized Parsing. In *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL (NAACL-HLT)*, Rochester, NY, USA.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the $21^{st}$ International Conference on Computational Linguistics and the $44^{th}$ Annual Meeting of the ACL*, Sydney, Australia.

Rögnvaldsson, E., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Whelpton, M., Nikulásdóttir, A. B., and Ingason, A. K. (2009). Icelandic Language Resources and Technology: Status and Prospects. In Domeij, R., Koskenniemi, K., Krauwer, S., Maegaard, B., Rögnvaldsson, E., and de Smedt, K., editors, *Proceedings of the $17^{th}$ Nordic Conference of Computational Linguistics, workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, NODALIDA. Odense, Denmark.

Rögnvaldsson, E., Ingason, A. K., and Sigurðsson, E. F. (2011a). Coping with variation in the Icelandic Parsed Historical Corpus (IcePaHC). In Johannessen, J. B., editor, *Language Variation Infrastructure. Papers on selected projects. Oslo Studies in Language 3.2*. University of Oslo.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2011b). Creating a dual-purpose treebank. *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In N. Calzolari and K. Choukri and

T. Declerck and M. U. Doğan and B. Maegaard and J. Mariani and J. Odijk and S. Piperidis, editor, *Proceedings of the* $8^{th}$ *International Conference on Language Resources and Evaluation*, LREC. Istanbul, Turkey.

Santorini, B. (2010). Annotation manual for the Penn historical corpora and the PCEEC. University of Pennsylvania, Philadelphia. http://www.ling.upenn.edu/hist-corpora/annotation/index.html.

Trosterud, T. (2009). A constraint grammar for Faroese. In Bick, E., Hagen, K., Müürisep, K., and Trosterud, T., editors, *Proceedings of the* $17^{th}$ *Nordic Conference of Computational Linguistics, workshop Constraint Grammar and robust parsing*, NODALIDA. Odense, Denmark.

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.