The evolution of relative clauses in the IcePaHC treebank

Anton Karl Ingason and **Johanna Mechler**
University of Iceland
Sæmundargötu 2
102 Reykjavík, Iceland
antoni@hi.is, mechler@hi.is

Abstract

We examine how the elements that introduce relative clauses, namely relative complementizers and relative pronouns, evolve over the history of Icelandic using the phrase structure analysis of the IcePaHC treebank. The rate of these elements changes over time and, in the case of relative pronouns, is subject to effects of genre and the type of gap in the relative clause in question. Our paper is a digital humanities study of historical linguistics which would not be possible without a parsed corpus that spans all centuries involved in the change. We relate our findings to studies on the Constant Rate Effect by analyzing these effects in detail.

1 Introduction

In this paper, we report on a study that analyzes the historical evolution of the elements that introduce relative clauses in Icelandic. Two types of elements are involved. First, we have relative complementizers, the elements *sem* and *er*, similar to English *that*. Second, we have relative pronouns, words that start with *hv-*, similar to English *wh-* words. Examples of the relative complementizers are given in (1) and (2) and a relative pronoun is shown in (3).

(1) stelpan **sem** fór burt
girl.the that went away
'the girl that went away'

(2) stelpan **er** fór burt
girl.the that went away
'the girl that went away'

(3) stelpan við **hverja** hann talaði
girl.the with whom he talked
'the girl with whom he talked'

Historically, the relative complementizer used to be *er*, but over time, this element has been replaced with *sem*, which has the same syntactic distribution and function. These elements are very common in historical texts, *er* is dominant in the earliest texts and *sem* in modern texts. In contrast, relative pronouns like *hverja* 'whom' in example (3) have never been very common, but for a while they gained some popularity among writers who produced Icelandic texts. The present study examines this historical development.

As thousands of examples need to be analyzed in order to uncover the relevant facts, a parsed corpus (a treebank) is essential. That is a collection of texts that has been annotated in terms of syntactic structure. We use the IcePaHC treebank for this purpose (Wallenberg et al., 2011). This means that different uses of the words *sem* and *er* are disambiguated. For example, *sem* can be a relative complementizer or a comparative complementizer and *er* can either be a relative complementizer or an inflected form of the verb 'to be' in Icelandic.

The paper is organized as follows: The background section introduces relative complementizers *sem* and *er* as historically competing forms and provides methodological details on the treebank. We extract all relevant examples of the environments in question and report on the findings drawn from these in the sections on relative complementizers and relative pronouns. We show that the change from *er* to *sem* in the history of Icelandic follows a very regular S-shaped curve and discuss how changes in the introduction of relative clauses relate to the Constant Rate Effect (Kroch, 1989), an important property of syntactic change in the languages of the world. The main findings are then summarized in the conclusion.

2 Background

While some traditional texts categorize the Icelandic words *sem* and *er* as relative pronouns, Þráinsson (1980) argues that they are in fact relative complementizers. This is because they do not pattern with pronouns in their formal properties. Unlike Icelandic pronouns, they do not manifest

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 202–208
May 3-4, 2025 ©2025 Association for Computational Linguistics

case declension and they cannot be the complements of prepositions. They also always appear at the beginning of subordinate clauses. This is in contrast with actual relative pronouns, which also appear in Icelandic, that start with *hv*, such as *hver* 'who', much like English *wh-* words. Of these elements that introduce relative clauses, the complementizers are much more common. There are some more examples of elements introducing relative clauses but the other types are comparatively rare. These include clauses that begin with the form *sá*, typically used as demonstrative, and it has also been noted that the Icelandic author, Halldór Laxness, sometimes uses *og*, typically a coordinating conjunction 'and', to start his relative clauses (Rögnvaldsson, 1983).

The older Icelandic texts use *er* as a relative complementizer (sometimes written *es*). This is a frozen form of what used to be a pronoun historically (Matthíasson, 1959, 10). The form *sem* then evolves from the comparative particle *sem*, but *sem* as a relative complementizer is not attested in the oldest written sources, i.e., runes from before the year 1000 (Matthíasson, 1959, 79–85). The change from *er* to *sem* is quite interesting from the point of view of historical linguistics because the entire change is attested in the historical record, unlike some changes that as linguists we only get to observe once the change is underway.

It has long been noted that when two linguistic forms compete for use, the transition from one to the other may follow an S-shaped curve if the rate of use is plotted against a time axis. This type of a historical change has been derived from certain hypotheses about how children acquire language (Yang, 2002). One well-known example of a syntactic change that follows an S-shaped curve is the rise of *do*-support in the history of English (Kroch, 1989). In his analysis of *do*-support, Kroch proposes a Constant Rate Effect for historical change, such that when a change applies in more than one syntactic context, the rate of change is the same across contexts, even though the rate of use is different depending on context. We revisit S-curves and the Constant Rate Effect below, as testing these hypotheses/ effects sheds light on the theoretical implications of our study.

3 The Icelandic Parsed Historical Corpus

The Icelandic Parsed Historical Corpus, IcePaHC (Wallenberg et al., 2011; Rögnvaldsson et al., 2011; Rögnvaldsson et al., 2011, 2012), is a manually annotated phrase structure treebank, developed in the tradition of the Penn Parsed Corpora of Historical English (PPCHE) (Kroch and Taylor, 2000; Kroch et al., 2004). While the Penn treebank (Marcus et al., 1993) was the first major treebank to be developed and remains the best known such resource, various lessons were learned during its development and some of these led to changes in the annotation scheme for the historical corpora, notably including a more flat phrase structure for constructions where structural ambiguity makes consistent and informative annotation challenging. The Icelandic treebank builds on this experience by adopting an annotation scheme which is in most respects identical to the PPCHE scheme, only adjusting it in minor ways where Icelandic requires additional information. The modifications include more morphological information at the PoS-tag level, such as the annotation of morphosyntactic case features.

IcePaHC consists of one million words of text, all of which have been manually annotated. This includes samples from 61 texts and in the corpus distribution, a plain text version of each text, along with a version that is PoS-tagged and lemmatized, and finally, and most importantly, a version that has been annotated for phrase structure according to the PPCHE guidelines. Since this is a historical corpus, an even distribution of samples from all centuries is emphasized and the corpus contains texts from the 12th century to the 21st century inclusive.

The texts come from five genres. Most of the samples are narratives or religious texts, and these two genres are found for almost all centuries. The corpus also contains biographies, legal text, and scientific text. IcePaHC has been used for a variety of research projects, both in linguistics as well as Natural Language Processing, and it has been widely cited in such work. For example, IcePaHC has been used to predict historical change in the case of the so-called New Passive (or New Impersonal Construction) (Ingason et al., 2012) and it has also been used to train phrase structure parsers (Ingason et al., 2014; Jökulsdóttir et al., 2019; Arnardóttir and Ingason, 2020).

To extract the examples from the treebank, we used the Parsed Corpus Query Language, PaCQL. (Ingason, 2016), and we performed all of our quantitative analysis in R (R Core Team, 2023). The publication of the IcePaHC treebank was a milestone in the ongoing effort to build Language Tech-

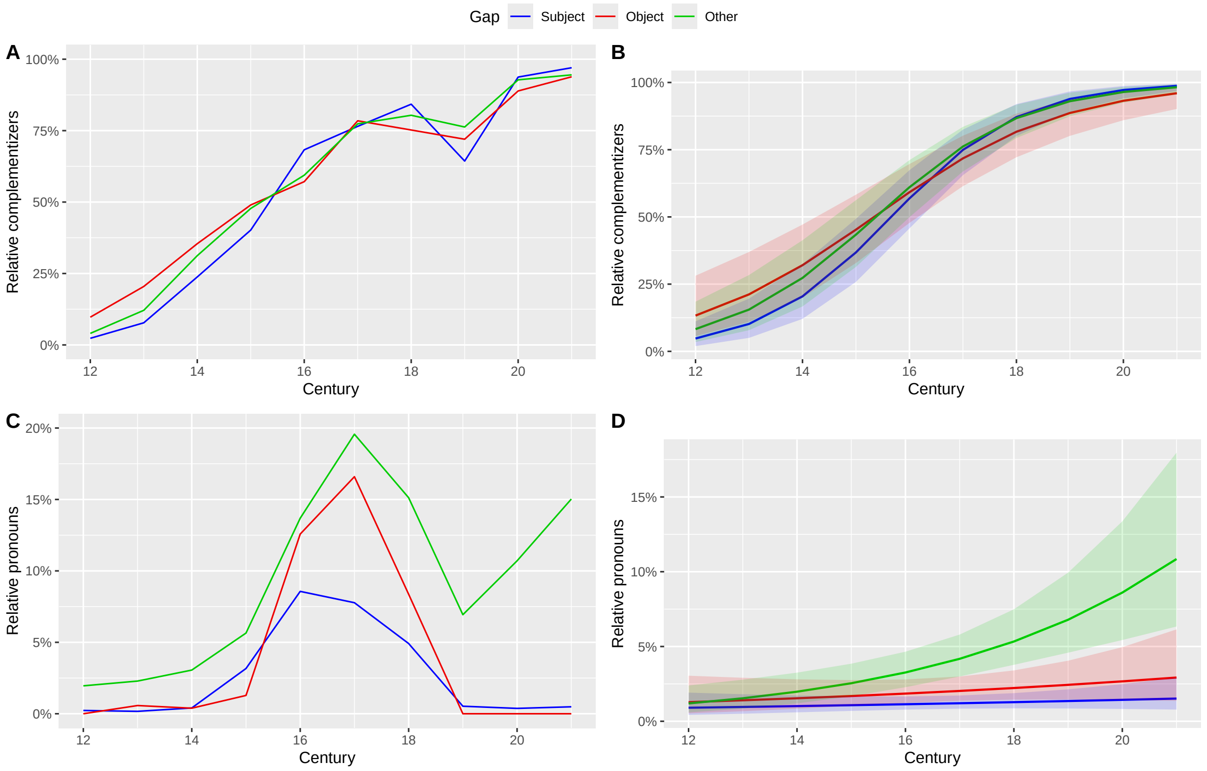

Figure 1: (A) The empirical rate of *sem* over time, by gap type in relative clauses. (B) The predicted probabilities of *sem* over time with an interaction between century and gap type. (C) The rate of relative pronouns over time, by gap type in relative clauses. (D) The predicted probabilities of relative pronouns over time with an interaction between century and gap type.

Table 1: Statistical results for the mixed-effects regression model for relative complementizers *(left)* and relative pronouns *(right)*.

Mixed-Effects Regression Model: Relative Complementizers *(left)* and Relative Pronouns *(right)*

	Relative Complementizers				*Relative Pronouns*			
Predictors	**Odds Ratios**	**std. Error**	**Statistic**	***p***	**Odds Ratios**	**std. Error**	**Statistic**	***p***
(Intercept)	0.00	0.00	-8.58	**<.001**	0.00	0.01	-4.65	**<.001**
century	2.27	0.20	9.09	**<.001**	1.06	0.07	0.86	0.391
gap [object]	66.48	36.42	7.66	**<.001**	0.93	1.18	-0.06	0.953
gap [other]	6.27	2.99	3.85	**<.001**	0.12	0.11	-2.29	**0.022**
cen × gap [object]	0.77	0.03	-7.55	**<.001**	1.04	0.08	0.47	0.636
cen × gap [other]	0.90	0.03	-3.49	**<.001**	1.22	0.07	3.68	**<.001**
genre [rel]					2.41	0.77	2.74	**0.006**
genre [bio]					6.13	2.72	4.09	**<.001**
genre [law]					1.01	1.18	0.01	0.992
genre [sci]					0.48	0.48	-0.73	0.463
Random Effects								
σ^2	3.29				3.29			
τ_{00}	2.84				0.77			
ICC	0.46				0.19			
N $_{\text{text-id}}$	61				61			
Observations	10206				12140			
Marginal R^2	0.395				0.183			
Conditional R^2	0.675				0.338			

nology resources for the Icelandic language. These efforts facilitate not only practical development outside academia, but also studies like the present one within the realm of the Digital Humanities. While IcePaHC was one of the early outputs that support the Digital Humanities in the context of the Icelandic language, the work on further resources continues, as evidenced by the more recent Language Technology Programme of the Icelandic government (Nikulásdóttir et al., 2020).

4 Relative complementizers

Let us first consider the evolution of relative complementizer over time, i.e. how *sem* replaces *er* as the form used for this purpose in Icelandic. Figure 1 shows how these forms evolve based on our data from the IcePaHC corpus. First consider part (A) of the figure. This shows the empirical rate of *sem* for each century of written text, as a proportion of total *sem* + *er* clauses, split up by the type of subject gap in the clause. Distinctions between types of subject gap are demonstrated by the examples in (4) and (5).

(4) The girl [that _ chased the boy]

(5) The girl [that the boy chased _]

These examples show that the empty slot in the relative clause can correspond to constituents that have a different grammatical status. This is interesting because previous research has found that subjects are more accessible in processing than objects and objects are more accessible than obliques (Lau and Tanaka, 2021). Being accessible in this context means that for comprehension purposes, less accessible elements suffer from lower accuracy, longer processing time, and greater working memory burden. For production, less accessible objects result in slower responses, more errors and more omissions or substitutions. Additionally, in both child and second language acquisition, they are characterized by later acquisition and greater avoidance.

The first thing to notice about Figure 1 (A) is that the empirical rate of *sem* over time follows a very regular curve. This is interesting because even the well-known S-curve from Kroch (1989) that describes the rise of *do*-support in the history of English is quite wiggly. The only century that appears to deviate from a regular rise is the 19th century and it turns out that this exception has a straightforward explanation. The corpus contains two texts from the 19th century, *Sagan af Heljarslóðarorrustu* and *Hellismanna saga*, both of which manifest a low rate of *sem* because they are intentionally written in an archaic style. These two texts contribute substantially to the overall rate for the 19th century. Apart from this, the curve is remarkably regular.

Furthermore, if we look at Figure 1 (B), we see the predicted probabilities of *sem* over the same centuries, again split by gap type, and in this case based on the output of a mixed-effects regression model. The model is built with usage of *sem* as the response variable and the predictors century, gap type, as well as an interaction between century and gap type; text-ID was added as random effect. All of these predictors are highly significant as shown in Table 1. It is not surprising that century is significant as this predictor tracks the historical change we are investigating. It is more surprising that adding the century * gap interaction improves the model because if a Constant Rate Effect (Kroch, 1989; Fruehwald et al., 2013) was present, adding the interaction should not improve the model fit as the change spreads at the same rate in all grammatical contexts. However, if we look at Figure 1 (B), we find that during the initial period when *er* is more common than *sem*, *er* is more likely to be selected in relative clauses with a subject gap. This effect reverses during the later period; when *sem* is more common than *er*, *sem* is more likely to be selected in clauses with a subject gap. We hypothesize that there are processing reasons for this effect; somehow the faster processed subject gap clauses are associated with the selection of the most frequent variant of the complementizer. Perhaps, this is related to the more frequent variant of the complementizer also being subject to faster access from memory. Such effects might matter when planning sentences, even though this is written text and not spoken language. We nevertheless emphasize that further interpretation of this effect requires more research and likely also comparisons with other similar phenomena, which, to our knowledge, does not exist currently.

The IcePaHC corpus contains metadata about the text genre (e.g., narrative or religious text), as mentioned above. Unexpectedly, genre was not significant in the model selection process for relative complementizers. This suggests that other factors such as century or gap type were better suited to explain the observed variation in the data set. We considered genre because religious texts might be expected to be more conservative than narratives;

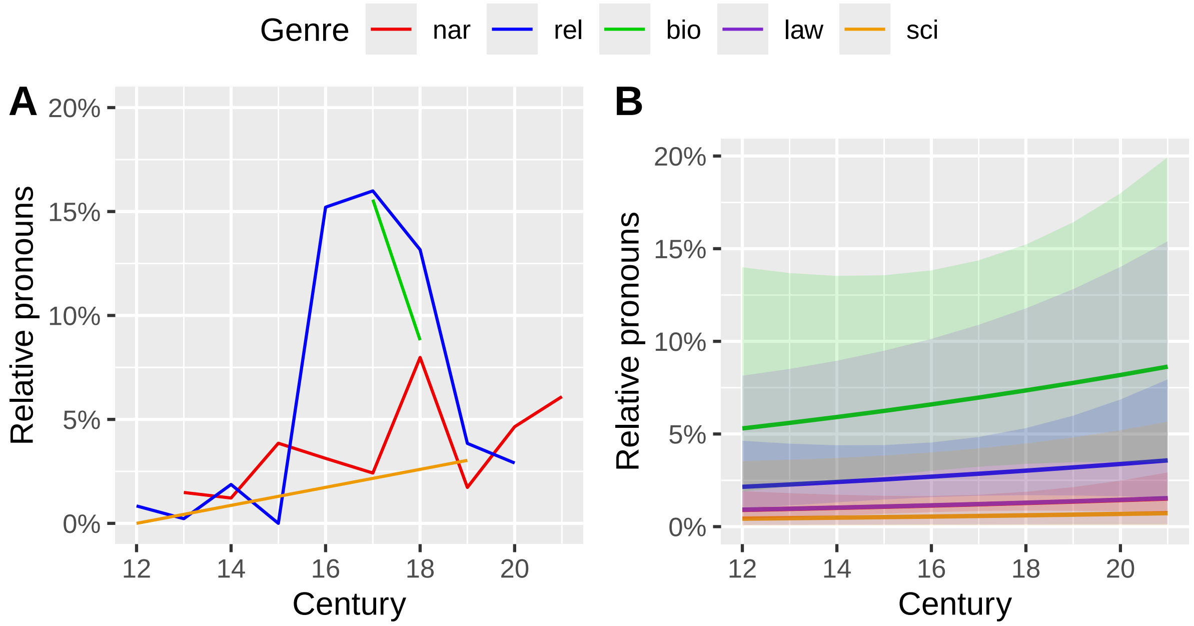

Figure 2: (A) The empirical rate of relative pronouns over time by genre. (Note that there is only one legal text so there is no graph for that genre.) (B) The predicted probabilities of relative pronouns over time by genre. (Genre types: nar = narrative, rel = religious, bio = biographical, law = legal, and sci = scientific texts.)

those are the two genres for which we have most data. However, such an effect is not found.

5 Relative pronouns

As outlined, in addition to relative complementizers, another way to introduce relative clauses in Icelandic is relative pronouns. Figure 1 (C) shows the rate of relative pronouns over time, as a proportion of all relative clauses, split up by the type of gap (subject, object, and other). It reveals a relatively low rate of relative pronoun use across all centuries considered here. The highest rates of relative pronouns can be found between the 16th and late 18th centuries, with a peak of about 20% in the 17th century. Thus, relative complementizers, the alternative to relative pronouns, remain the most common form historically.

This distribution also affects the type of gap. For subject gaps, relative complementizers are more readily available since they are the most common form, so they are chosen more frequently. Inversely, relative pronouns are less likely to be used with subject gaps (see Figure 1 (D); interaction Table 1). For object gaps, using relative pronouns is more likely, as they have slightly longer processing times. Relative pronouns are most commonly used with another argument, e.g., with prepositional phrases. Here, a possible translation effect needs to be considered since more texts during this time were translated from German to Icelandic. It might be the case that in German texts this type was the most common form, so logically this trend would transfer to Icelandic relative pronoun use.

The regression model for relative pronouns (response variable) includes century, type of gap, and genre as fixed effects, an interaction between century and type of gap, and finally text-ID as random effect (see Table 1). Regarding the Constant Rate Effect, as was the case for relative complementizers, adding the interaction between century and type of gap improves the model fit.

Further analysis reveals that besides century and the type of gap, genre is also an important factor in conditioning relative pronoun use (see Table 1). Narrative texts, which serve as response/default level here, are significantly different from religious and biographical texts (Figure 2), but we also lack extensive data on these two text types. Interestingly, legal and scientific texts are not significantly different. In scientific texts, it is also less likely to find relative pronouns than in any other type of text according to the mixed-effects regression model.

6 Conclusion

In this paper, we have shown that using the phrase structure analysis of the IcePaHC treebank provides valuable insights into the diachronic evolution of Icelandic relative clauses. From the 12th century up until the 21st century, relative complementizers have been more common than relative pronouns. The choice of complementizer is conditioned by the type of gap in relation to frequency, e.g., *sem* is selected more frequently for subject gaps when *sem* is the most common form and *er* is selected more frequently for subject gaps when *er* is more common. For relative pronouns, the analysis reveals a genre effect, which is not present for relative complementizers. We find that the relative pronouns are used most often in biographies and religious texts in the 16th to 18th century and they are especially frequent in clauses whose gap is not an argument, i.e., not a subject or an object, but rather something else. In sum, we add new evidence to an ever-growing body of research on Icelandic using Language Technology resources. The findings of this study further inform future work on the Constant Rate Effect, providing another test case for this effect.

Limitations

Regarding the limitations of this paper, it is possible that other predictors affect the distribution of relative complementizers and prounouns that could not be considered in the analysis here. While they are very rare, there are also some other elements that introduce relative clauses that were not taken into account in the analysis. Further, the analysis is based on written language, and spoken language might be more nuanced (although we believe that written language is appropriate for studying this type of change). Lastly, we rely on the annotation provided in the IcePaHC corpus, which might contain errors; however, we checked several examples, and overall, the corpus proves very accurate.

Acknowledgments

We would like to thank the reviewers for helpful comments that contributed to making this a better paper.

References

Þórunn Arnardóttir and Anton Karl Ingason. 2020. A neural parsing pipeline for Icelandic using the Berkeley neural parser. In *Proceedings of CLARIN Annual Conference*, pages 48–51.

Josef Fruehwald, Jonathan Gress-Wright, and Joel Wallenberg. 2013. Phonological rule change: The Constant Rate Effect. In *NELS 40: Proceedings of the 40th Annual Meeting of the North East Linguistic Society*, volume 1, pages 219–230. GLSA Publications.

Anton Karl Ingason. 2016. PaCQL: A new type of treebank search for the digital humanities. Handrit. *Italian Journal of Computational Linguistics*, 2(2):51–66.

Anton Karl Ingason, Julie Anne Legate, and Charles Yang. 2012. The evolutionary trajectory of the Icelandic New Passive. *University of Pennsylvania Working Papers in Linguistics*, 19(2):11.

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel Wallenberg. 2014. Rapid deployment of phrase structure parsing for related languages: A case study of Insular Scandinavian. In *LREC*, pages 91–95. Citeseer.

Tinna Frímann Jökulsdóttir, Anton Karl Ingason, and Einar Freyr Sigurðsson. 2019. A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus. In *Proceedings of CLARIN Annual Conference*, pages 138–141.

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

Anthony S. Kroch, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. Size: 1.8 million words.

Anthony S. Kroch and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition. Size: 1.3 million words.

Elaine Lau and Nozomi Tanaka. 2021. The subject advantage in relative clauses: A review. *Glossa: a journal of general linguistics*, 6(1).

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Haraldur Matthíasson. 1959. *Setningaform og stíll*. Bókaútgáfa Menningarsjóðs.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for Icelandic 2019-2023.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Eiríkur Rögnvaldsson, Anton Karl Ingason, and Einar Freyr Sigurðsson. 2011. Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). In *Language Variation Infrastructure*, pages 97–112.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984.

Eiríkur Rögnvaldsson. 1983. "Tilvísunartengingin" OG í bókum Halldórs Laxness. *Mímir*, 30:8–18.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel C. Wallenberg. 2011. Creating a dual-purpose treebank. *Journal for Language Technology and Computational Linguistics*, 2(26):141–152.

Joel Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9.

Charles Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford.

Höskuldur Þráinsson. 1980. Tilvísunarfornöfn? *Íslenskt mál*, pages 53–96.