# Stability and change in Icelandic corpora: The case of Stylistic Fronting

**Anton Karl Ingason**[1] **and Johanna Mechler**[1]

[1]University of Iceland, Iceland

Corresponding author: Anton Karl Ingason , antoni@hi.is

## Abstract

We use Icelandic corpora, the Icelandic Gigaword Corpus, the Icelandic Parsed Historical Corpus, and the Newspaper Corpus, in order to investigate the history of a syntactic construction, Stylistic Fronting (SF). SF has long been noted to be associated with formal style and it has received considerable attention in the theoretical syntax literature, but less has been said in the literature about its history throughout the centuries and modern times. We find that use of SF remained stable from the 12th century to the 20th century, but its rate declines in the (late) 20th and 21st century. Our analysis furthermore shows that use of SF is only significantly connected to genre in the most recent data, suggesting that the link between SF and style may be a relatively modern innovation. Finally, we test our data for the Constant Rate Effect, revealing that it is present for some grammatical contexts of SF. Our paper is a digital humanities study of historical linguistics which would not be possible without parsed corpora that together span all centuries involved in the change.

## Keywords

historical linguistics; digital humanities; syntax; Icelandic; Constant Rate Effect

## I INTRODUCTION

Stylistic Fronting (SF) is a word order phenomenon in Icelandic that has been studied in detail in the theoretical syntactic literature and various proposals have been put forth for its analysis [Maling, 1980, Holmberg, 2000, 2006, Thráinsson, 2007, Angantýsson, 2017, Ingason and Wood, 2017]. This construction involves the movement of a word or a phrase into a subject gap. This may take place in both subordinate and main clauses. While some work has looked at SF quantitatively [Wood, 2011] and in terms of lifespan change in modern times [Stefánsdóttir and Ingason, 2018], the overall historical evolution of SF remains understudied.

In this paper, we aim to contribute to the knowledge of how SF evolved throughout the centuries and in modern times and we would also like to shed light on the history of this construction's association with formal style. With our approach from a digital humanities perspective, we make use of the Icelandic Parsed Historical Corpus [Wallenberg et al., 2011] and the Icelandic Gigaword Corpus [Steingrímsson et al., 2018]. As for the Icelandic Gigaword Corpus, our investigation includes samples from the Icelandic Parliament Corpus [Steingrímsson et al., 2020] and the Icelandic Newspaper Corpus, subsets of this corpus. We make use of these corpora in order to examine the historical stability of SF until the 20th century and its decline in the late 20th and 21st century. In the final part of this paper, we also show how the type of finite auxiliary affects the rate of SF and how this predictor relates to the Constant Rate Effect [Kroch, 1989] over time.

## II BACKGROUND

SF appears in various grammatical contexts which have a distinct quantitative distribution as found by Wood [2011]. Here, we control for some of these contexts by only focusing on SF in relative clauses with a subject gap where a finite auxiliary and a non-finite main verb appear at the beginning of the clause in either of the two possible word orders. Where SF has not applied (1), the auxiliary precedes the non-finite main verb. Where SF has applied (2), the non-finite main verb precedes the auxiliary.

(1) Varðandi það [$_{CP}$ sem **var** **sagt** hér] ...
regarding it that **was** **said** here
'Regarding what was said here ...'

(2) Varðandi það [$_{CP}$ sem **sagt** **var** hér] ...
regarding it that **said** **was** here
'Regarding what was said here ...'

Quantitatively, Wood [2011] reveals that the rate of SF depends on the category of the fronted constituent. Considering analytical options, Holmberg [2000] finds that SF may apply where there is a phonological subject gap. Stefánsdóttir and Ingason's [2018] work suggests that the use of SF as a formality marker evolves across the lifespan, depending on personal and political events. Thus, research on SF illustrates its grammatical contexts and (in)stability in use across the individual lifespan. We are the first, to our knowledge, that add historical evidence across multiple centuries to explore the evolution of this linguistic variable in detail.

In our data, SF patterns together with three types of finite verbs, namely modal verbs, *be*, and *have*. If SF changes across time, we can test whether it changes at the same rate in all three environments of finite verbs. Hereby, we refer to the so-called Constant Rate Effect, proposed by Kroch [1989]. Kroch [1989], in his examination of *do*-support in English, introduces the Constant Rate Effect as a theory for historical linguistic change. According to this concept, when a change occurs in multiple syntactic environments, the rate of change remains consistent across these various contexts. This holds true even if the actual rate of use might differ in each context. Following the historical analysis of SF, we will re-examine the Constant Rate Effect, relating it to our findings on SF use in Icelandic.

It is important to contribute evidence to the study of the Constant Rate Effect because this is a theoretical construct that has a central place in the study of historical linguistics. Support for the Constant Rate Effect has been found in more syntactic phenomena than those originally studied by Kroch, e.g. in Pintzuk [1995], and evidence has been found that it may also hold for phonological change in Fruehwald et al. [2013]. Furthermore, building on Yang [2000] and Yang [2002], Kauhanen and Walkden [2018] propose a model of how the Constant Rate Effect can be derived from a model of language acquisition. This means that strengthening the empirical basis on which this area of research is built is valuable for the field.

## III METHODS AND DATA

Our analysis relies on three corpora. The first one is the *Icelandic Parsed Historical Corpus* (IcePaHC; $n$=1,232) [Wallenberg et al., 2011]; we use this corpus to explore the linguistic stability of SF throughout the centuries, as the corpus covers the time span from 1150–2008. IcePaHC [Wallenberg et al., 2011, Rögnvaldsson et al., 2011, Rögnvaldsson et al., 2011, 2012], is a phrase structure treebank with manual annotations, designed in alignment with the principles of the Penn Parsed Corpora of Historical English (PPCHE) [Kroch and Taylor, 2000, Kroch et al., 2004]. The creation of the English historical corpora revealed new insights into annotation schemes for historical corpora, for instance, adopting a flatter phrase structure in instances where structural ambiguity complicates the provision of consistent, informative annotations. The Icelandic treebank builds upon these lessons by following an annotation framework that is fundamentally aligned with the PPCHE, with only minor modifications to account for the unique linguistic properties of Icelandic. These modifications include the addition of more detailed morphological features at the part-of-speech tagging level, such as the inclusion of morphosyntactic case annotations. For extracting examples from the treebank, we utilized the Parsed Corpus Query Language (PaCQL) [Ingason, 2016], while all quantitative analyses were conducted in R [R Core Team, 2023].

IcePaHC comprises one million words of manually annotated text, encompassing samples from 61 different sources. Each text in the corpus is provided in three versions: a plain text format, a version with part-of-speech tagging and lemmatization, and notably, a version annotated for phrase structure following the PPCHE guidelines. As a historical corpus, particular attention is given to ensuring an even distribution of samples across all centuries, as mentioned above. The texts in the IcePaHC span five different genres. The majority of the samples consist of narratives and religious texts, both of which are present across nearly all centuries. In addition, the corpus includes genres such as biographies, legal documents, and scientific writing.

The release of the IcePaHC treebank marked a significant achievement in the continued effort to develop Language Technology resources for Icelandic. These initiatives not only promote practical applications beyond academia but also support research in fields such as the Digital Humanities, as demonstrated by the present study. IcePaHC has been employed in various research initiatives, not only within the field of linguistics but also in Natural Language Processing, and it has been extensively cited in related studies. For instance, the corpus has been instrumental in modeling historical changes, such as the development of the so-called New Passive (or New Impersonal Construction) [Ingason et al., 2012], as well as to train phrase structure parsers [Ingason et al., 2014, Jökulsdóttir et al., 2019, Arnardóttir and Ingason, 2020]. Although IcePaHC was among the initial resources to contribute to the Digital Humanities within the Icelandic language context, work on additional resources is ongoing, as highlighted by the more recent Language Technology Programme, initiated by the Icelandic government [Nikulásdóttir et al., 2020].

However, IcePaHC has fewer data points for more recent years, especially for the 21st century. This is why we add parliament speeches from the *Icelandic Parliament Corpus* (IPC; $n$=514,465) [Steingrímsson et al., 2020] and newspaper articles from the *Icelandic Newspaper Corpus* (NPC; $n$=2,146,487) to our data set, which include data from 1909–2021 and 1998–2021 respectively. As mentioned, the two corpora are subsets of the *Icelandic Gigaword Corpus* [Steingrímsson et al., 2018]. For both corpora, Python scripts, using PoS-tags and lemmas from the corpus, were applied in data processing to extract relevant examples of SF.

Together, the three corpora allow for a large scale and fine-grained diachronic analysis of SF (total $n$=2,662,184). Data visualization and analysis was performed in R [R Core Team, 2023]. The best model fit was determined based on the AIC, and the make-up of each model is described in more detail in the relevant sections below.

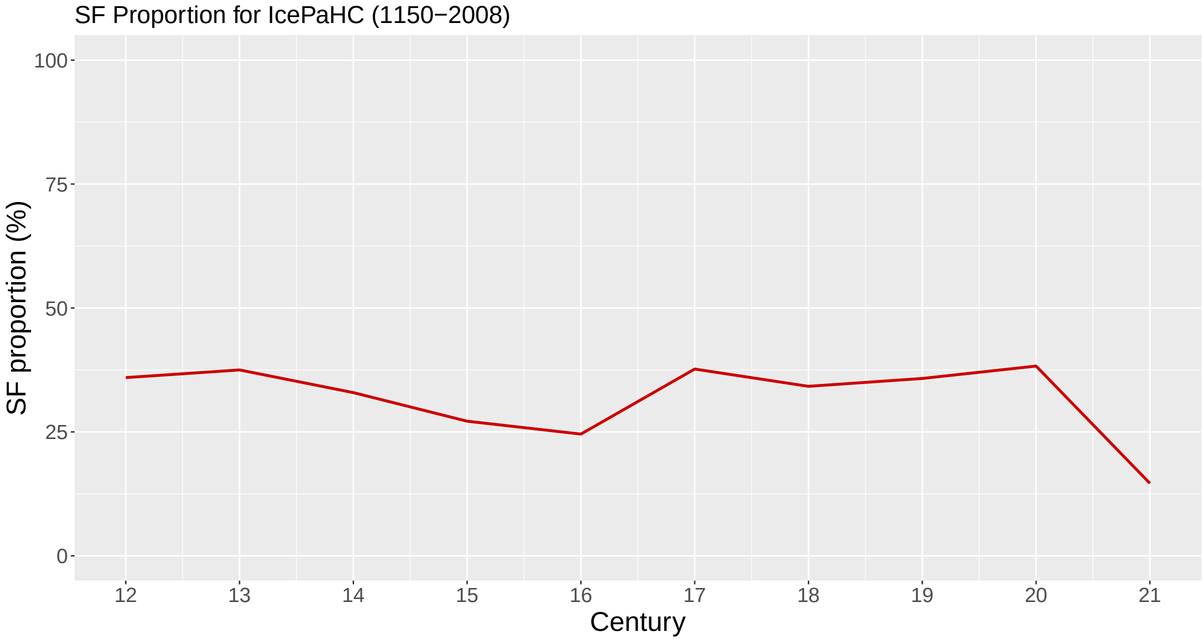

Figure 1: The rate of SF in the history of Icelandic over time.

## IV HISTORICAL STABILITY

Considering the rate of SF in IcePaHC in Figure 1, we find that SF remains relatively stable from the 12th to the 21st century. Only from the 20th to 21st century we see a decline in SF use, and we will further explore the more recent development of SF in the following section using the IPC and NPC.

The remarkable stability of SF is also reflected in the results of the statistical analysis. Fitting a mixed-effects regression model with century, genre, and finite verb as fixed effects and text-ID as a random effect, we find that none of the factors are significant, suggesting that SF did not change over this period (see Table 1). For the effect of finite verb, '*be*' was used as reference level; for genre, it was 'narrative' texts. There are furthermore no significant differences between the levels of predictors. Based on these findings, we conclude that the rate of SF is overall notably stable, even across different text genres and different linguistic contexts, such as the three types of finite verbs.

Table 1: Regression model for IcePaHC with SF as response variable.

**Mixed-Effects Regression Model: IcePaHC**

| **Predictors** | **Odds Ratios** | **Std. Error** | **Statistic** | ***p*** | **Random Effects** | |
|---|---|---|---|---|---|---|
| (Intercept) | 0.59 | 0.07 | -4.27 | **<.001** | $\sigma^2$ | 3.29 |
| century | 0.96 | 0.08 | -0.50 | 0.616 | $\tau_{00\ \text{text-ID}}$ | 0.13 |
| finite verb [have] | 0.90 | 0.13 | -0.72 | 0.473 | ICC | 0.04 |
| finite verb [modal] | 0.72 | 0.13 | -1.88 | 0.060 | $N_{\text{text-ID}}$ | 61 |
| genre [biographical] | 1.05 | 0.29 | 0.18 | 0.854 | | |
| genre [religious] | 0.78 | 0.15 | -1.27 | 0.206 | Observations | 1232 |
| genre [legal] | 0.76 | 0.43 | -0.48 | 0.635 | Marginal $R^2$ | 0.016 |
| genre [scientific] | 0.22 | 0.18 | -1.84 | 0.065 | Conditional $R^2$ | 0.054 |

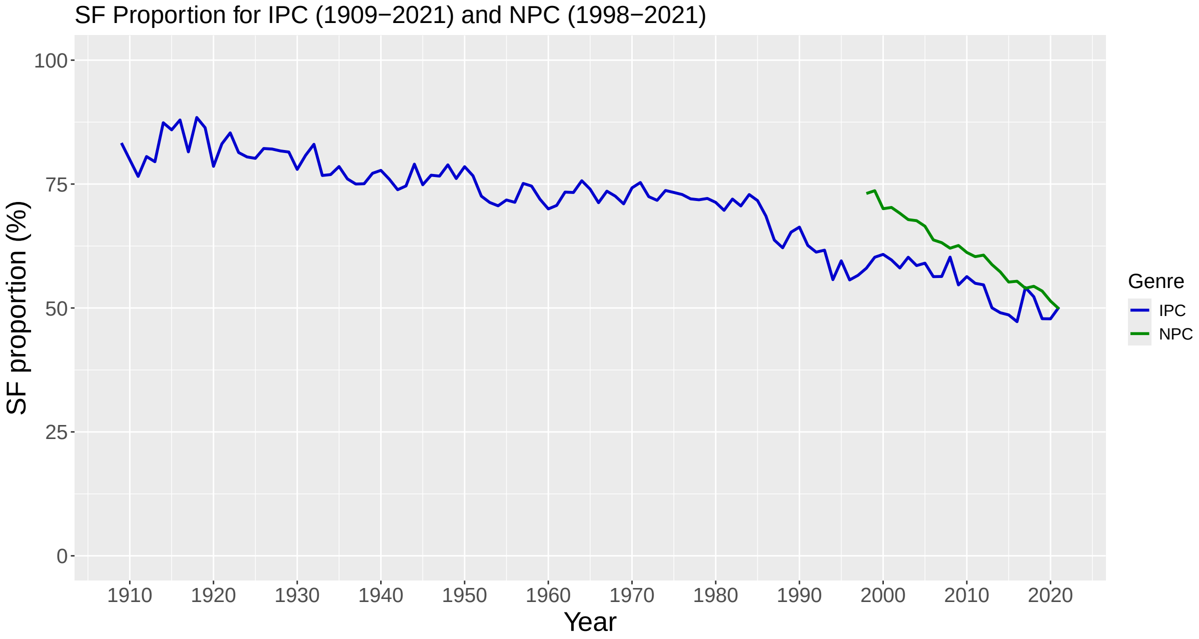

Figure 2: The rate of SF in the history of Icelandic in the 20th and 21st century, by genre (IPC = Icelandic Parliament Corpus, NPC = Newspaper Corpus).

## V GENRE EFFECT WITH SF DECLINE IN THE LATE 20TH AND 21ST CENTURY

If we take a closer look at the 20th and 21st century, for which we have abundant data in the IPC and NPC, we find that use of SF declines gradually over this period (see Figure 2). For this more recent time period, we define the IPC and the NPC as two genres, with the IPC covering speech data in a parliament setting, and the NPC including written texts in the form of newspaper articles (as described in Section III). In general, SF use in the NPC is much higher than in IcePaHC, but it decreases in more recent decades, falling below the SF rate in the IPC (see Figure 2). Both NPC and IPC end at about 50% SF use in the final years included in the data set.

The mixed-effects regression models for parliament speeches (IPC) was built with year, role in parliament, party status, and type of finite verb as fixed effects and non-finite verb and person/ member of parliament as random effects (see Table 2). For newspaper articles (NPC), the model includes year and type of finite verb as fixed effects, an interaction between year and type of finite verb, as well as non-finite verb and author as random effects (see Table 2). Results for both models find that (among other factors) year is highly significant, indicating that SF changes over this period. Additionally, further analysis suggests that there is a significant difference between the two genres, which implies a genre effect in the more recent data.

We hypothesize that a historical change is indeed taking place, with SF use decreasing over time, but an alternative interpretation is that these two social spaces, the parliament and newspapers, are becoming more informal over time. As a formality marker is subject to style shift [Labov, 1972], a decline in formality may lead to a decline in use, independent of any historical development. It is also possible that both explanations contribute to the evolution of SF in this period.

Table 2: Regression model for IPC and NPC with SF as response variable (fv = finite verb, nfv = non-finite verb).

**Mixed-Effects Regression Model: IPC**

| **Predictors** | **Odds Ratios** | **std. Error** | **Statistic** | ***p*** | **Random Effects** | |
|---|---|---|---|---|---|---|
| (Intercept) | 2.88 | 0.11 | 28.34 | **<.001** | $\sigma^2$ | 3.29 |
| year | 0.54 | 0.01 | -39.38 | **<.001** | $\tau_{00\ \mathrm{nfv}}$ | 0.84 |
| fv [have] | 0.27 | 0.00 | -157.97 | **<.001** | $\mathrm{N}_{\mathrm{nfv}}$ | 2795 |
| fv [modal] | 0.12 | 0.00 | -110.82 | **<.001** | $\tau_{00\ \mathrm{person}}$ | 0.33 |
| party status [minority] | 0.86 | 0.01 | -13.89 | **<.001** | $\mathrm{N}_{\mathrm{person}}$ | 899 |
| role [minister] | 1.26 | 0.02 | 17.18 | **<.001** | ICC | 0.26 |
| role [replacement] | 1.06 | 0.03 | 2.11 | **0.034** | Observations | 514,465 |
| | | | | | Marginal $\mathrm{R}^2$ | 0.149 |
| | | | | | Conditional $\mathrm{R}^2$ | 0.372 |

**Mixed-Effects Regression Model: NPC**

| **Predictors** | **Odds Ratios** | **std. Error** | **Statistic** | ***p*** | **Random Effects** | |
|---|---|---|---|---|---|---|
| (Intercept) | 1.31 | 0.03 | 11.05 | **<.001** | $\sigma^2$ | 3.29 |
| year | 0.78 | 0.00 | -91.69 | **<.001** | $\tau_{00\ \mathrm{nfv}}$ | 1.22 |
| fv [have] | 0.33 | 0.00 | -269.19 | **<.001** | $\mathrm{N}_{\mathrm{nfv}}$ | 5853 |
| fv [modal] | 0.12 | 0.00 | -237.61 | **<.001** | $\tau_{00\ \mathrm{author}}$ | 0.93 |
| fv [have] x year | 0.95 | 0.00 | -15.94 | **<.001** | $\mathrm{N}_{\mathrm{author}}$ | 8376 |
| fv [modal] x year | 0.92 | 0.01 | -10.33 | **<.001** | ICC | 0.40 |
| | | | | | Observations | 214,6487 |
| | | | | | Marginal $\mathrm{R}^2$ | 0.081 |
| | | | | | Conditional $\mathrm{R}^2$ | 0.444 |

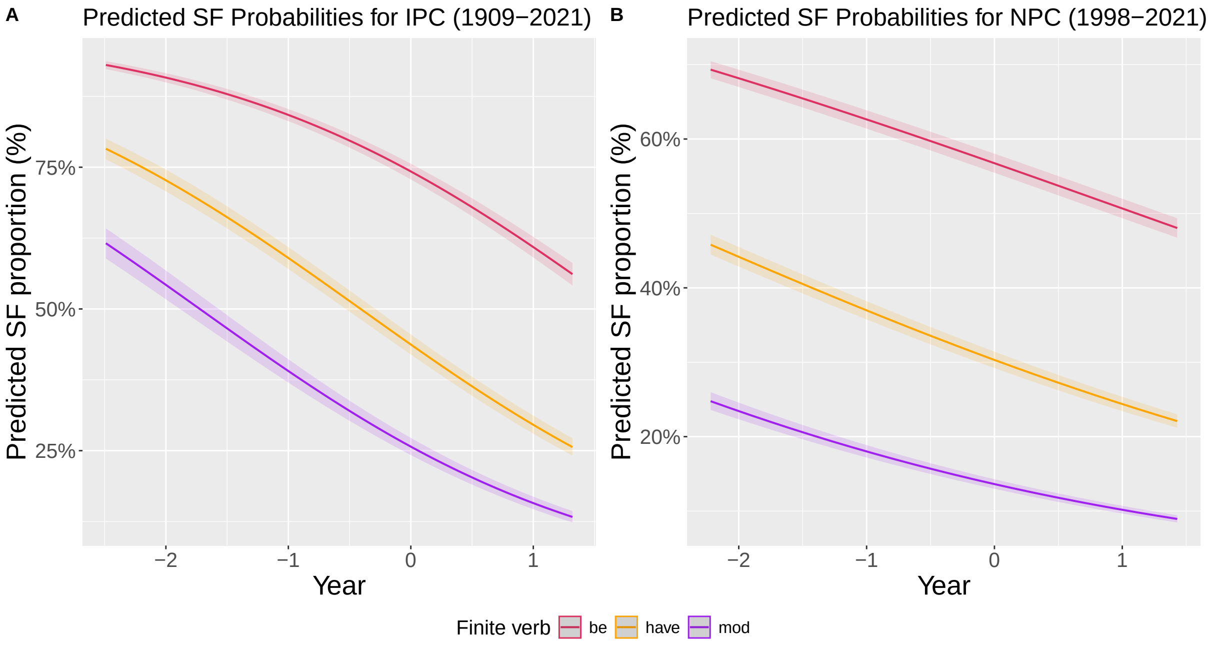

Figure 3: The predicted probabilities of SF for three types of finite verbs over time, by genre (A: IPC = Icelandic Parliament Corpus, B: NPC = Newspaper Corpus). (Note that year is scaled in the models and plots respectively.)

## VI TESTING FOR THE CONSTANT RATE EFFECT

Both regression models for IPC and NPC suggest a change across time (see Table 2), so the question arises whether SF changes at the same rate in different grammatical context, i.e. whether the Constant Rate Effect applies. Here we test the Constant Rate Effect for three types of finite verbs, by adding an interaction of year and finite verb to the regression model for IPC and NPC separately. For parliament speeches (IPC), adding the finite verb-year interaction does not significantly improve the model, which can be interpreted as evidence for all three types of finite verbs changing at the same rate (so we find a Constant Rate Effect). Figure 3 (A) also supports this finding, showing all three types of finite verbs declining at the same rate, while the individual rate of SF remains different for each type. For the newspaper articles (NPC), adding the interaction significantly improves the model fit, which suggests that the three types of finite verbs do *not* change at the same rate. However, Figure 3 (B) illustrates that the three types of finite verbs are predicted to pattern very similarly. Therefore, the fact that including the interaction in the model improves the model fit might be due to underfitting the model – not due to the absence of the Constant Rate Effect. Future analysis will reveal more in this regard. Taken together, these results add important new evidence on the Constant Rate Effect for a syntactic variable in Icelandic.

## VII CONCLUSION

Our digital humanities study reveals that, historically, SF remains stable across time, but more recent data based on corpora from the 20th and 21st century suggests a decline in SF use during this period. Considering the type of finite verb as a grammatical context, we looked for a Constant Rate Effect but could only confirm it for the IPC. However, the lack of a Constant Rate Effect in the NPC might be due to the nature of the model. Crucially, our study contributes significant evidence for this effect in Icelandic syntax, underscoring the value of digital humanities approaches in uncovering patterns in historical language data. Previous literature has reported that SF is associated with formal style; in the historical data (IcePaHC) we do not find a significant effect of genre, but in the more recent data (IPC, NPC) we do, suggesting that the association of SF with style might be a more modern innovation. Only the triangulation of extensive data from three annotated corpora allowed us to draw such conclusions, highlighting the importance of these types of digital humanities studies.

## LIMITATIONS

Focusing on one Icelandic variable is both a strength of this paper, as it investigates a non-dominant language, but also a limitation, making generalizations difficult. Furthermore, the more recent data includes two genres only; future research might consider additional situational and stylistic contexts to substantiate the findings outlined above. Lastly, the NPC lacks comparable predictors to include in the model to avoid underfitting the model; with a more complex model, we might be able to increase the confidence in the claims made in relation to the Constant Rate Effect.

## ACKNOWLEDGMENTS

This project is supported by a grant from the European Research Council (ERC), project ID 101117824.

## References

Ásgrímur Angantýsson. Stylistic Fronting and related constructions in the insular Scandinavian languages. In Höskuldur Þráinsson, Caroline Heycock, and Zakaris Svabo, editors, *Syntactic Variation in Insular Scandinavian. Studies in Germanic Linguistics*, pages 277–306. John Benjamins Publishing Company, Netherlands, 2017.

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

Josef Fruehwald, Jonathan Gress-Wright, and Wallenberg Joel. Phonological rule change: The constant rate effect. In *NELS 40*, pages 219–230. GLSA Publications, 2013.

Anders Holmberg. Scandinavian Stylistic Fronting: How any category can become an expletive. *Linguistic Inquiry*, 31(3):445–483, 2000.

Anders Holmberg. Stylistic Fronting. *The Blackwell Companion to Syntax*, pages 532–565, 2006.

Anton Karl Ingason. PaCQL: A new type of treebank search for the digital humanities. Handrit. *Italian Journal of Computational Linguistics*, 2(2):51–66, 2016. doi: 10.4000/ijcol.391.

Anton Karl Ingason and Jim Wood. Clause bounded movement: Stylistic Fronting and phase theory. *Linguistic Inquiry*, 3(48):513–527, 2017.

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

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel C. Wallenberg. Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 91–95, Reykjavik, Iceland, 2014.

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

Henri Kauhanen and George Walkden. Deriving the constant rate effect. *Natural Language & Linguistic Theory*, 36(2):483–521, 2018.

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

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

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

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, 1972.

Joan Maling. Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál*, 2:175–193, 1980.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. Language technology programme for Icelandic 2019-2023, 2020. URL `https://aclanthology.org/2020.lrec-1.418/`.

Susan Pintzuk. Variation and change in Old English clause structure. *Language Variation and Change*, 7(2): 229–260, 1995.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2023. URL `http://www.R-project.org`.

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

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

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

Lilja Björk Stefánsdóttir and Anton Karl Ingason. A high definition study of syntactic lifespan change. *U. Penn Working Papers in Linguistics*, 24(1):1–10, 2018.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Steinþór Steingrímsson, Starkaður Barkarson, and Gunnar Thor Örnólfsson. IGC-Parl: Icelandic Corpus of Parliamentary Proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 11–17, 2020.

Höskuldur Thráinsson. *The Syntax of Icelandic*. Cambridge University Press, Cambridge, 2007.

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

Jim Wood. Stylistic Fronting in spoken Icelandic relatives. *Nordic Journal of Linguistics*, 34(1):29–60, 2011.

Charles Yang. Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250, 2000.

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