Using the Icelandic Gigaword Corpus to Explain Lifespan Change

Lilja Björk Stefánsdóttir
University of Iceland
lbs@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

In this paper, we demonstrate research on syntactic lifespan change in the speech of an Icelandic MP, using data from the Icelandic Gigaword Corpus. Our study exemplifies how advances in language technology infrastructure can play an important role in linguistic studies, providing comprehensive linguistic data that would have been impossible to acquire only a few years ago.

1 Introduction

In this paper, we describe a case study where we use the Icelandic Gigaword Corpus (Steingrímsson et al., 2018) in order to examine syntactic lifespan change in the speech of an Icelandic MP, with an emphasis on the period before, during, and after the Icelandic economic crash of 2008. Our study exemplifies the benefits of using a CLARIN language resource for big data humanities.

We examine the variable use of the syntactic process of Stylistic Fronting (SF), a stylistic indicator associated with formal speech in Icelandic, throughout the career of an Icelandic MP, Bjarni Benediktsson. We observe Benediktsson having a high rate of SF usage during his first years as an MP, a pattern that is disrupted in the year 2008, where the rate suddenly drops following the Icelandic economic crash of 2008. Benediktsson's rate of SF continues to drop in the years following the economic crash, an interesting development as it occurs at the same time as his party, The Independence Party, is in a very weak position, having lost their seat in government and facing the lowest support in the history of the party. We attribute this decline to a dramatic change in Benediktsson's Linguistic Market Value (LMV) in the sense of D. Sankoff and Laberge (1978); greater contextual importance of language correlates positively with the use of more formal speech variants. This temporary change is then reversed in 2012 when the rate of SF increases again, during a time where the aftermath of the economic crash is mostly over and Benediktsson's status, as well as his party's status within the parliament is getting stronger.

The findings from our study demonstrate how a fine-grained view of syntactic lifespan change yields insights about status-associated usage as interrelated aspects of the social dimension of language. Our findings also provide evidence of the importance of a high-definition approach (Stefánsdóttir & Ingason, 2018, 2019), that is, using comprehensive and continuous linguistic data derived from corpora, due to the complex and fluctuating nature of individual lifespan change.

2 Background

In the past two decades, there has been considerable growth in research on changes in how individuals speak over the course of their lives, generally referred to as lifespan change. Most studies on lifespan change are not high-definition studies, meaning they are commonly based on comparing two periods in the individuals's life. In some cases, samples from several periods are combined into one and then compared to another perio, like in MacKenzie's (2017) study on David Attenborough's speech. Sankoff's (2004) study on phonological variation among members of the film series "7 Up" is unusually fine-grained because the same individual was examined five times at seven-year intervals, therefore relying on a higher time resolution than most studies on individual lifespan change. Yet, we cannot see how the

changes develop accurately because seven years passed between points in time when the same individual was observed. The study by Arnaud (1998) is similar to the present one as it uses a corpus as a source for tracking syntactic change. Although the study certainly does have a higher time resolution than many comparable studies, it nevertheless leaves room for enhanced detail, for example, the study is exclusively based on written texts, and its quantitative findings are not based on a well-defined envelope of variation but rather a more coarse density measure. Furthermore, the data are grouped into 5-year periods, yielding a maximum of 11 readings per speaker. Therefore, we believe that there is much to be learned from our present study, which looks at a **continuous year-by-year account** of well-understood linguistic variable that spans several years and is derived from spoken language transcriptions, thus having a very high time resolution and a robust connection with spoken Icelandic.

Two of our previous studies (Stefánsdóttir and Ingason, 2018, 2024) on Icelandic MPs during the financial crash are examples of the benefits of this kind of methodology. The former Icelandic Minister of Finance, Steingrímur Sigfússon, showed an upward shift in his use of SF during the economic crash, which we attributed to a dramatic change in his Linguistic Market Value (LMV), as he was in a position of great responsibility as a Minister of finance during an economic crisis. Another MP, Þorgerður Gunnarsdóttir, showed a different reaction to the crisis as the rate of SF dropped during the period, especially between the years 2008–2009, when the rate went from 78% to 50%. This change is interesting because Gunnarsdóttir had been a minister in the so-called crash-government, that is, the government that had been serving during and in the years before the economic crash, which eventually collapsed in early 2009 after the biggest public protests in Icelandic political history. We attributed this drop to a change in her LMV, as she had gone from being a minister in the government to an MP in the opposition as well as being a member of a party that suddenly had become very unpopular.

All of our studies on Icelandic MPs are high-definition studies, meaning that they rely on a very high time resolution, with the number of readings matching the number of years the MPs have been in office. For Sigfússon and Gunnarsdóttir, we had a total of 38 and 20 readings, respectively, giving us a clear overview of changes and their development with no gaps between readings.

2.1 The Stylistic Fronting variable

Stylistic Fronting is an optional movement process, found in Icelandic, of a word or a phrase into a subject gap (Angantýsson, 2017; Maling, 1980; Thráinsson, 2007; Wood, 2011).

(1) *Bækur* *[$_{CP}$ sem* *{eru* *lesnar* *(No SF)* *//* *lesnar* *eru* *(SF)}* *til* *skemmtunar]* *eru* *bestar.*
books [$_{CP}$ that are read // read are for entertainment] are best
'Books that are read for entertainment are the best ones.'

The contrast between the word orders with and without SF illustrates the optionality. The relative clause has a subject gap and thus SF can apply and move the non-finite main verb in front of the finite auxiliary. SF has no effect on truth-conditional meaning, and its only clear meaning component is a sociolinguistic one; the movement is associated with formal style. SF is found in both main clauses and subordinate clauses, as long as the subject is not phonologically overt. The phonological subject gap condition holds for relative clauses with extracted subjects, as above, and impersonal main clauses where there is no overt subject. Although full phrases can be stylistically fronted, we only focus on the cononical case here, as we limit the scope of the study to word orders involving the complementizer *sem* that introduces Icelandic relative clauses (e.g., by excluding frontable elements other than non-finite main verbs) and finite auxiliaries and non-finite main verbs in either of the two possible orders.

We do this to control for factors that can condition the use of SF, building on findings from Wood (2011) who showed that prosodic factors and syntactic category can affect the rate of SF. Obviously, this does not include all cases of SF but allows us to extract a well-defined envelope of variation with high accuracy (where SF application and non-application are accounted for).

3 Detecting patterns in the Icelandic Gigaword Corpus

The Icelandic Gigaword Corpus consists of 2429 million running words of text, and a part of the corpus is parliament speeches. This resource allows for a high-definition time resolution; thus, we can observe

gradual changes over time where the time axis is continuous. We wrote a Python script that analyzed a part of the corpus, the parliament speeches by Benediktsson between 2003–2021, and we extracted sequences with a relative complementizer followed by a finite verb and a non-finite one in either of the two possible word orders. The corpus includes audio files and a transcription, making observations for accuracy possible as we can listen to each audio file and verify the transcription. The patterns that we search for are very reliable as confirmed by our manual checks. As mentioned, we only collected subject relatives with a potential for SF of a non-finite main verb. This provided us with 2729 tokens of the SF variable, with each token coded for SF application and the year of the speech.

4 Lifespan change in Benediktsson's speeches

Figure 1 shows Benediktsson's use of SF across his career. In the early years, Benediktsson's rate of SF is relevatively high, with the average of 91.16% in the years 2003–2007. This pattern is not an unexpected one since Benediktsson was a new MP at the time and the situational effect of his surroundings and new status is likely to have caused him to be become more aware of his language use, which positively correlates with a frequent use of formal variants such as Stylistic Fronting, according to Labov's (1972) attention-paid-to-speech model. In addition to situational effects, the high rate of SF during this period can also be interpreted in terms of the so-called Linguistic Marketplace (D. Sankoff and Laberge, 1978), a long term style predictor that explains the relationship between a speakers linguistic behavior and their Linguistic Market Value (LMV). Individual's LMV is highly connected to their social status, which is shaped by personal and professional experience over the years, and a high LMV correlates positively with the probability of using formal variants. Benediktsson's LMV was high during this period, both as an MP but also as an MP for a party that was in government and therefore, in a strong position within the parliament, resulting in his speech becoming formal.

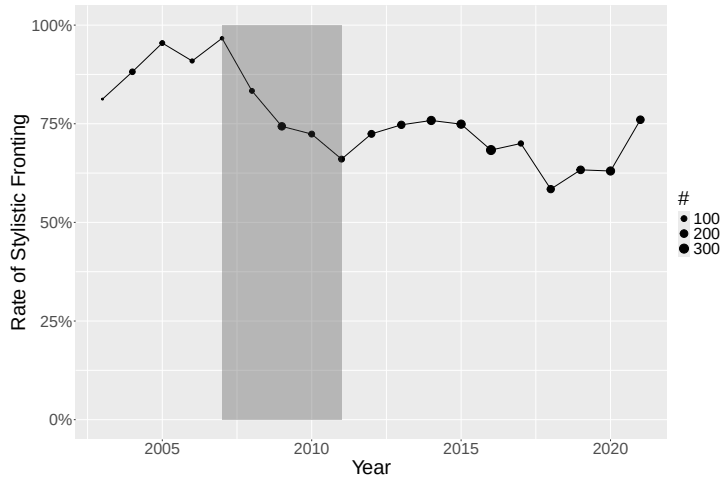

Figure 1: Evolution of SF in Benediktsson's career.

However, this trend is reversed in 2008 when the rate of SF suddenly drops, going from 96.7% in 2007 to 83.3% in 2008. The rate continues to drop in the following years (2007–11 highlighted), reaching its lowest rate in 2011 at 66.0%. This decline in the use of SF is interesting because it occurs almost simultaneously with the Icelandic economic crash of 2008 and during the crash's aftermath which lasted till the year 2011. The economic crash took a big toll on Benediktsson's party, The Independence Party, as it had been a part of the so-called crash-government, which collapsed in early 2009, and lost a lot of support from its voters with many looking at the party as the culprit for the economic crash. We analyze this change in Benediktsson's linguistic behaviour as a reflection of a dramatic change in his LMV, due to his weak political position at the time.

This temporary change is again reversed in 2012, when the rate of SF increases again, when the aftermath of the financial crash is mostly over and Benediktsson, and his party, start to prepare for the upcoming elections, which took place a year later. The 2013 elections resulted in the Independence Party regaining their status as Iceland's biggest political party and Benediktsson took on the role of Minister

of Finance in a right-wing government. Perhaps unsurprisingly, the rate of SF continues to be relatively high in the following years, reflecting Benediktsson's high LMV as a person in a position of power.

An analysis of Benediktsson's use of SF in the period before, during and after the financial crash in Iceland confirms that the various nuances of individual lifespan change can only be studied in a large digitized corpus. If we had, for example, only the first data point and the last, many crucial aspects of the development would have gone missing from the picture, no matter how carefully the data would have been collected. Note that analysis of lifespan change has to be evaluated on a case-by-case basis and, for example, in our study of Ásmundur Daðason (Stefánsdóttir & Ingason, Forthcoming), we believe that aspects of identity are more important than LMV. Also note that it is not trivial to decide how to evaluate statistical signficance in such a data set. Some pairwise differences between years are significant but we seek tools that evaluate differences in wiggly long-term trends. We leave a more accurate analysis for future work.

5 Summary

In this paper we used a CLARIN resource, a corpus containing the speeches of the Icelandic parliament, to analyze the formality levels of one politician, Benediktsson, over time. This builds on our previous work on other Icelandic Members of Parliament and continues to demonstrate the value of open access parliament data for the digital humanities. We found that fluctuations in the use of SF by Benediktsson reflects explanations given by the literature on style shift, as it connects with his attention-paid-to-speech and Linguistic Market Value over time.

References

Angantýsson, Á. (2017). Stylistic fronting and related constructions in the insular scandinavian languages. *Syntactic Variation in Insular Scandinavian*, *1*, 277.

Arnaud, R. (1998). The development of the progressive in 19th century English: A quantitative survey. *Language Variation and Change*, *10*(02), 123–152.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

MacKenzie, L. (2017). Frequency effects over the lifespan: A case study of Attenborough's r's. *Linguistics Vanguard*.

Maling, J. (1980). Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál*, *2*, 175–193.

Sankoff, D., & Laberge, S. (1978). The linguistic market and the statistical explanation of variability. In *Linguistic variation: Models and methods*. Academic Press.

Sankoff, G. (2004). Adolescents, young adults and the critical period: Two case studies from 'Seven up'. *Sociolinguistic variation: Critical reflections*, 121–139.

Stefánsdóttir, L. B., & Ingason, A. K. (2018). A high definition study of syntactic lifespan change. *U. Penn Working Papers in Linguistics*, *24*(1), 1–10.

Stefánsdóttir, L. B., & Ingason, A. K. (2019). Lifespan change and style shift in the Icelandic Gigaword Corpus. *Proceedings of CLARIN Annual Conference 2019*, 138–141.

Stefánsdóttir, L. B., & Ingason, A. K. (2024). *Lífsleiðarbreytingar á Alþingi* [The Annual Humanities Conference, University of Iceland. Reykjavík 8–9 October.].

Stefánsdóttir, L. B., & Ingason, A. K. (Forthcoming). Wiggly lifespan change in a crisis – contrasting reactive and proactive identity construction. *U. Penn Working Papers in Linguistics*.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., & Guðnason, J. (2018, May). Risamálheild: A very large Icelandic text corpus. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1690

Thráinsson, H. (2007). *The syntax of Icelandic*. Cambridge University Press.

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