CLARIN Annual Conference 2019

PROCEEDINGS

Edited by

Kiril Simov, Maria Eskevich

30 September – 2 October 2019
Leipzig, Germany

Please cite as:
Proceedings of CLARIN Annual Conference 2019. Eds. K. Simov and M. Eskevich.
Leipzig, Germany: CLARIN, 2019.

Lifespan Change and Style Shift in the Icelandic Gigaword Corpus

Lilja Björk Stefánsdóttir
University of Iceland
Reykjavík, Iceland
lbs11@hi.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

We demonstrate research on syntactic lifespan change and style shift in Icelandic that is made possible by recent advances in Language Technology infrastructure for Icelandic. Our project extracts data from the Icelandic Gigaword corpus and allows us to shed light on how social meaning shapes the linguistic performance of speakers using big data methods that would not have been feasible for us to use without a corpus of this type.

1 Introduction

In this paper, we describe a case study where we use the recently constructed Icelandic Gigaword Corpus (Risamálheild) (Steingrímsson et al., 2018) in order to examine syntactic lifespan change and style shift in the speech of an Icelandic speaker, former minister of finance, Steingrímur J. Sigfússon. Our study exemplifies how a general-purpose resource for Language Technology can facilitate big data research in the digital humanities if it is carefully curated and made freely available to researchers. Without recent advances in Language Technology infrastructure for Icelandic, our study would have been a prohibitively daunting task, showing that important progress is being made, even in the case of a less-resourced language like Icelandic.

In recent years, studies of language variation and change have increasingly paid attention to linguistic change across the lifespan of an individual. This is interesting because it is widely believed that a critical period for language acquisition constrains the malleability of linguistic abilities (Lenneberg, 1967) and, empirically, the organization of language is indeed rather stable in the adult brain. It is therefore important to improve our understanding of what can change in the language of adults and how. Most current studies on lifespan change have a limited time resolution, typically looking at only 2–3 periods in the speaker's life (see for example Harrington, 2006, Sankoff and Blondeau 2007, Kwon 2017, MacKenzie 2017, but also Arnaud 1998 and Sankoff 2004).We argue that an improved time resolution is critical for studies of this type as well as a focus on qualitative detail when interpreting quantitative findings.

We examine the variable use of the syntactic process of Stylistic Fronting (SF) throughout the career of an Icelandic politician, Steingrímur J. Sigfússon. We reveal an age grading pattern (e.g., Labov 1994; Wagner 2012) toward less formal usage that is disrupted by a spike in use of the formal variant following the Icelandic economic crash of 2008 when Sigfússon, the leader of the Left-Green Movement, becomes the Minister of Finance and becomes publicly responsible for the fate of the Icelandic economy. We attribute the spike to a dramatic change in his Linguistic Market Value (LMV) in the sense of (Sankoff and Laberge, 1978). This temporary change is then reversed when the left wing government loses its majority in the 2013 election and Steingrímur stops being a Minister in the government as well as the leader of his party. The findings demonstrate how a fine-grained view of syntactic lifespan change yields insights about age-associated usage and status-associated usage as interrelated aspects of the social dimension of language. We also examine the stylistic dimension and suggest that style shift reflects a situational LMV.

2 The Stylistic Fronting variable

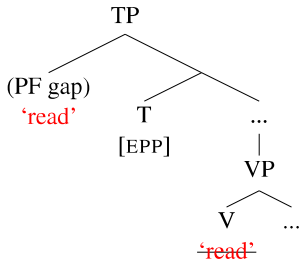
SF is an optional movement in Icelandic of some category to the apparent subject position of a finite clause which does not have an overt subject (Maling, 1980). The envelope of variation includes subject relatives, embedded subject questions and various impersonal constructions. Although the (phrasal) subject position, e.g., Spec,TP, seems to be the target of SF, the moved category is often a head, canonically a non-finite main verb that moves in front of a finite auxiliary like the passive participle in (1). The second word order variant shown in the curly bracket is a non-SF counterpart. According to one analysis (Holmberg, 2006), SF is one way to satisfy the PF part of the EPP requirement.

The example in (1) demonstrates the variation associated with SF. In a relative clause with a subject gap, the non-finite verb *lesnar* can be moved into the subject gap.

(1) *Bækur [$_{CP}$ sem {lesnar eru / eru lesnar } til skemmtunar] eru bestar.*
books [$_{CP}$ that {read are / are read } for entertainment] are best
'Books that are read for entertainment are the best ones.'

Structurally, the movement looks as in (2). The T head has some kind of a subject requirement, whose precise formulation is beyond the scope of this paper, but crucially the absence of an overt subject, even in cases where a covert subject is likely to be assumed by many analysts, allows for SF. The PF gap allows any SF-suitable element to move into an apparent subject position, e.g., Spec,TP.

(2)



3 Detecting patterns in the Icelandic Gigaword Corpus

The Icelandic Gigaword corpus consists of about 1300 million running words of text and a part of the corpus are parliament speeches. Since the subject of our study is a member of the Icelandic parliament, Alþingi, and due to the fact that Sigfússon is the parliament member who has spoken the most words at the Icelandic parliament (about 6 million words), we make use of the biggest collection of spoken language from the same speaker available to us in order to study the nature of lifespan change and style shift in spoken language. We wrote a Python script that analyzed part of the Gigaword corpus, the parliament speeches given by Sigfússon between 1990-2013, 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. It should be noted that the corpus is not parsed for syntactic structure. Nevertheless, the patterns that we search for are very reliable as confirmed by our manual checks of the extracted data.

To control for various contextual factors (Wood, 2011) we only collected subject relatives with a potential for SF of a non-finite verb. This provided us with 8005 tokens of the SF variable. Each token was coded for SF application, speaker's age and type of speech (prepared/response).

4 Lifespan change

Use of SF across Sigfússon's career is shown in Fig. 1. There is a downward trend in the use of SF from age 35 onwards. We interpret this as an age-associated pattern (age grading) resulting from a reduced pressure to conform to the formality demands of the parliament as he gains seniority. However, his role as a leading opposition voice increases his LMV when the economy fails in 2007-2008 and the trend is reversed. When he becomes Minister of Finance (green period in Fig. 1) his SF use rises sharply.

When he stops being a Minister in government in 2013 and steps down as the leader of the Left-Green Movement, returning to being a common opposition MP, his SF returns to pre-economic-crash levels.

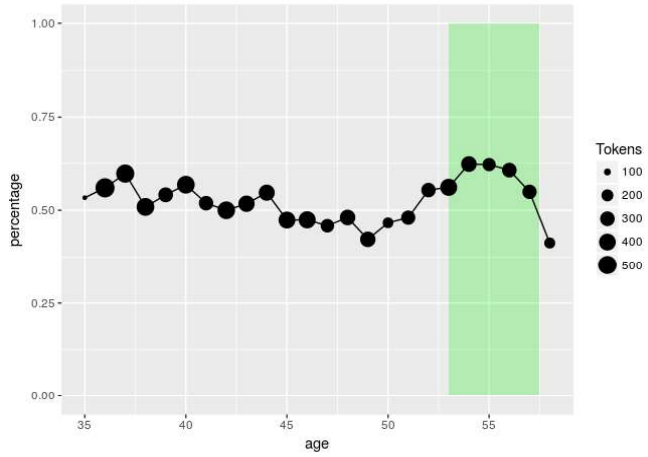


Figure 1: Evolution of SF use in Sigfusson's career.

The main point to be taken from these facts is that various nuances in the development can only be studied in a large digitized corpus. If we had, for example, only the first data point and the last data point, it would not matter how carefully the data were collected and curated; many crucial aspects of the development would simply be missing from the picture.

(3) **Main point about the methodology:**
While community-wide usage evolution is often regular and gradual, individual lifespan change responds rapidly to idiosyncratic sociolinguistic pressures – demanding a high-definition approach.

5 Style shift

In Fig. 2, we split the data between prepared speeches (more SF) and responses (less SF), visualized as a locally weighted regression. Consider now the audience design theory (Bell, 1984) which was constructed to amend some issues with the (Labov, 1972) notion of attention-paid-to-speech. The audience design theory is not the most obvious explanation for a style shift in 8005 tokens given from the same podium in the same room, all of which have the Icelandic parliament as an audience. Of course it is possible to say that preparation simply yields a fundamentally different type of language, but this is nevertheless as close to a fixed audience type as one can imagine.

Figure 2: Style shift between prepared speeches (FALSE/red) and responses (TRUE/green)

In the presentation we will discuss our view that usage probability at time p_t is a function of a base probability p_0, the LMV of the individual and the LMV of the current situation, thus reducing the social and stylistic dimension of variation to two interrelated aspects of the linguistic market. This view can capture LMV via attention-to-speech as well as adaptation to the audience.

(4) $p_t = p_0 + \text{LMV(individual)} + \text{LMV(situation)}$

What we mean be LMV via attention-to-speech is the fact that language which is prepared on paper is likely to be viewed as something that should strive to meet a higher standard with respect to the social evaluation of language. We believe that unifying social properties of the individual and the social properties of the speech situation is a feasible theoretical reduction which is a useful null hypothesis until proven wrong.

6 Summary and future work

Our findings add to much ongoing work on lifespan change and because of the wealth of data that are available in Sigfússon's speeches we get a high definition view of syntactic change across the lifespan. The findings reveal an interplay of age-associated and status-associated factors. In our presentation, we will also discuss how this study demonstrates that qualitative detail is important when interpreting quantitative findings; the computational power that the digital humanities have made available for researchers complements rather than replaces well established methods that focus on attention to detail and context.

References

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

Bell, Alan. 1984. Language style as audience design. *Language in society*, 13:145–204.

Harrington, Jonathan. 2006. An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, 34:439–457.

Holmberg, Anders. 2006. Stylistic fronting. *The Blackwell companion to syntax*, pp. 532–565.

Kwon, Soohyun. 2014. Noam Chomsky's vowel changes across the lifespan. *Selected papers from NWAV 42, U. Penn Working Papers in Linguistics*, 20:91–100.

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

Lenneberg, Eric. 1967. The biological foundations of language. *Hospital Practice*, 2:59–67.

MacKenzie, Laurel. 2017. Frequency effects over the lifespan: A case study of Attenborough's r's. *Linguistics Vanguard*, 3.1 (2017).

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

Sankoff, Gillian. 2004. Adolescents, young adults and the critical period: Two case studies from 'Seven Up'. *Sociolinguistic Variation: Critical Reflections*, ed. Carmen Fought, 121–139. Oxford University Press, New York.

Sankoff, Gillian, and Helene Blondeau. 2007. Language change across the lifespan: /r/ in Montréal French. *Language*, 83:560–588.

Sankoff, David, and Suzanne Laberge. 1978. *The linguistic market and the statistical explanation of variability. Linguistic variation: Models and methods*. Academic Press, New York.

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jon Gudnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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