

Word of the year 1919: Conveying the media's favorite annual linguistic parlor game to a different era

Steinþór Steingrímsson^{1,*}, Einar Freyr Sigurðsson¹, Starkaður Barkarson¹,
Atli Jasonarson¹ and Ágústa Þorbergsdóttir¹

¹The Árni Magnússon Institute for Icelandic Studies

Abstract

In Iceland, the word of the year is chosen annually, both by the Icelandic National Broadcasting Service and the Árni Magnússon Institute for Icelandic Studies (AMI). We explore the possibility of doing the same but for a year more than 100 years ago. We try using the same methods as AMI does for our times. This approach has various limitations, which we discuss, and raises many questions, such as how much texts from journals and periodicals reflect the actual word use of the time.

Keywords

word-of-the-year, linguistics, neologism, corpus linguistics, corpora

1. Introduction

In the last few decades, the Word of the Year has become a popular tradition with various dictionaries, linguistics societies and media outlets. In 1971, Gesellschaft für deutsche Sprache in Germany was one of the first to start this tradition.¹ In 1990, the American Dialect Society was the first to follow in the English-speaking world. Now, the word of the year is chosen for many or most of Europe's national languages, in some cases by multiple bodies.

In Iceland, the Icelandic National Broadcasting Service, RÚV, has listed candidates for the word of the year and allowed visitors to its website to vote on it since 2015. The Árni Magnússon Institute for Icelandic Studies (AMI) participated in generating the candidate lists for the first three years but has since 2018 published its own list and selected the word of the year from that list. For generating the candidate lists, AMI uses the Icelandic Gigaword corpus (IGC) (Steingrímsson et al. 2018; Barkarson, Steingrímsson, and Hafsteinsdóttir 2022). The last version of IGC, from 2022 (Barkarson et al. 2022), contains more than 2,4 million running words, mainly from news media, official documents (such as parliamentary speeches) and social media. The

The 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024), Reykjavík, Iceland, May 29–31, 2024


*Corresponding author.

✉ steinthor.steingrimsson@arnastofnun.is (S. Steingrímsson); enar.freyr.sigurdsson@arnastofnun.is (E. F. Sigurðsson); starkadur.barkarson@arnastofnun.is (S. Barkarson); atli.jasonarson@arnastofnun.is (A. Jasonarson); agusta.thorbergsdottir@arnastofnun.is (Á. Þorbergsdóttir)

🌐 <https://steinst.is/> (S. Steingrímsson)

🆔 0000-0002-9776-9507 (S. Steingrímsson)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

¹<https://gfd.s.de/aktionen/wort-des-jahres/#>

candidate lists generated are of three types: New words in the given year that have never been seen before (in the corpus), words that appear at least twice as often in the given year as in any other year, and words that appear more often in the given year than in the previous four years in total. The words on these lists are often ones that have been used for a short while related to one event or few related events and tend to be forgotten a short while later. As the word of the year is selected at the end of the year, a selection by vote tends to be biased towards words related to events happening in the last months of the year. To prevent such an imbalance, the AMI inspects the automatically generated lists and selects 10–12 candidates for possible words of the year. When the candidates are selected it is of course important that the words have some connection with hot topics in the previous year, as the words should preferably have some descriptive value, describing the year in question. It is therefore instrumental that the words are used by people from different levels of society. The word selected as word of the year from this shortlist is usually one that has gained a lot of popularity and is instantly recognized by most people following the media in Iceland.²

In this paper, instead of applying our methods to newly written texts to find the word that best describes a year that is ending, we try to use the same approaches on written texts from the past. Doing that, we find that we not only expose the limitations of the approaches we commonly use to select a word of the year, but it can also give surprising insights into how what will later be regarded as the major events of the time are viewed and discussed.

For our experiment, we looked at the time period 1914–1920, and decided on selecting a word of the year for one of these years, 1919. We did not select this period by chance. We wanted to know how dominating major events such as World War I and the Spanish flu were in the media at the time. When *sóttkví* ‘quarantine’ was selected as the word of the year 2020 (Þorbergsdóttir and Steingrímsson 2021), it was presumably a factor in the selection process that COVID-19 affected everyone, including the people selecting the word of the year. When we go more than 100 years back, the selection process may become a bit different, as we, the authors of this paper, did not experience the time period ourselves. Also by selecting a period at the very end of our corpus, which goes back to the beginning of the 19th century, we can still go relatively far back in time and yet apply the comparison methods that are used to select the word of the year in the 21st century.

The paper is structured as follows: In Section 2 we describe the data we use and the generated word lists. Section 3 gives our results and Section 4 discusses the limitations of our approach.

2. The data we use

To extract information on which words stand out with respect to their frequency, we created a corpus of texts from 367 newspapers and periodicals from the period 1801 to 1920. The texts were collected from *Tímarit.is*, a digital library where millions of scanned and OCR-read pages are made available on the Internet in digital format. *Tímarit.is* is a collaborative project between the National and University Library of Iceland, the National Library of the Faroe Islands and the National and Public Library of Greenland. In order to reduce the number of OCR errors

²See, for example, Þorbergsdóttir and Barkarson (2023) and Þorbergsdóttir, Barkarson, and Steingrímsson (2024) for discussion on the word of the years 2022 and 2023, respectively, chosen by AMI, and the methods used.

in the corpus we applied the method described in Jasonarson et al. (2023), post-processing the text using a neural model trained to identify spelling errors most likely produced by the OCR process.

The corpus consists of 239,136,590 running words. The amount of texts grows with the years; all the texts from the 19th century add up to only about 76 million words, or less than 32% of the total amount. 68% of the texts belong to the first two decades of the 20th century.

After PoS-tagging and lemmatizing the corpus, statistical information about each noun, adjective and verb was saved into a database table. The table contains information about how often each lemma appears in the corpus, grouped by the source (newspaper or periodical) and year. The next step was to run a script that read information from the database and created three different lists. To take an example of a single year, 1919, the first list contained all the words that appeared in the year 1919 but never in earlier texts, 6,937 in total. The second list contained words that appeared at least half as often in 1919 as in all previous years, 3,405 in total. The third list contained words that appeared at least six times in the year 1919 but less than six times in all the previous years, 747 in total. This is a variation of the criteria used for selecting present day Word of the Year. For the previous period we compare the frequency of a word in a given year to a longer span of previous years, as the smaller amount of data otherwise allows for more noise in the candidate lists.

3. Word lists

When we look at the word lists generated for the years 1914–1920, where the methods described in Section 2 are used, we might expect them to be dominated by words that are related to World War I, from 1914 to 1918. That is not the case, however, even though it is clear from the lists that the war had a large effect on the news reported in Icelandic journals at the time. Some notable words from our generated lists whose use can be linked to the war are *fríðarráðstefna* ‘peace conference’, *gagnáhlaup* ‘counterattack’, *kafbátur* ‘submarine’, *skotgryffa* ‘trench’, *stríðs(vá)trygging* ‘war insurance’ and *vélbyssa* ‘machine gun’. There are also various compounds with *ófriður* as the first part (or contained in the first part) or as the head of the whole compound, such as *heimsófriður* ‘world war’, *ófriðarsmælki* ‘war trifles’ and *ófriðarþjóð* ‘war nation’.

The Spanish flu (or the great influenza epidemic) raged in Iceland during the last months of 1918 and in 1919. We cannot really say that we see its significance in our lists, even though it had tremendous effect in Iceland. One way people talked about it was *sóttin mikla* ‘the great illness’. The frequency of *sótt*, compared to the years and decades before, seems to be, nevertheless, insignificant.

What sticks out, however, in 1917–1919 – using the methods described in Section 2 – are words that refer to Bolsheviks and Bolshevism and other related terms. The words are, of course, of Russian origin and are new in Icelandic discourse in the beginning of the 20th century. In 1917, *bolséviki/bolséviki/bolsheviki* ‘Bolshevik’ is a new word according to our corpus and is used 38 times. In 1918 it occurs 215 times and 591 times in 1919. This term is found in many different Icelandic journals from the time, both published in Iceland and Canada.

Not counted in these numbers are variations of the word where the spelling may suggest

an Icelandic preaspiration, like *bolschevikkar* 'Bolsheviks'; where *-kk-* could be a reflection of [hk] in pronunciation (as opposed to [k] in *bolshevikar*). Interestingly, Bolsheviks are also sometimes referred to as *Bolshevíkingar* 'Bolshevikings', or even *Bolsvíkingar*, with the head of the compound being the Icelandic *víkingur* 'viking'. Furthermore, *bolséviki* is also found in many different compounds from the time, such as *bolsévikastjórn* 'Bolshevik leadership', *bolsévikaflokkur* 'Bolshevik party' and *bolshevikikenning* 'Bolshevik theory/doctrine'.

Finally, there is also some frequency of *Bolsévismi* 'Bolshevism' (with variation in spelling) and there are also examples of Bolsheviks being referred to as *Bolsévistar* (lit. 'Bolshevists') in those years.

Based on the discussion above, we propose to select a word of the year 1919 – that word being *bolséviki* (and related terms).

4. Limitations

Our experiment is based on newspapers and journals from the chosen period. As discussed in Section 2 the text is OCR-ed. While the text seems to contain relatively few errors, these errors may impact which words fulfil the criteria for getting on the candidate lists. Furthermore, when describing new concepts or objects, the different publications tend to use different words and spelling and standardisation is slower than it is in the media now, as demonstrated by the many different forms of our selected word of the year 1919, which later was standardised as *bolséviki* or *bolséviki*.³

The newspapers and journals contain advertisements and column headers that are repeated day after day. The words printed there, commonly product names or product descriptions, are quite prevalent on the automatically generated lists and need to be filtered out before we compile the candidate lists. We do that manually. Another aspect of working with newspapers and journals is that we of course do not have any spoken language in our data. While most of the texts used when selecting word of the year for recent years is also from written texts, the data we use there does contain transcribed news from radio and television. By only looking at written texts we may thus be missing words commonly used for a short while in the period we are looking at, but we have no way of knowing that.

When, for example, we were selecting the word of the year at the end of last year, all of the people in the selection group had lived through the year, followed news and participated in society. We cannot say the same when we select the word of the year for 1919. There were still almost 50 years until the first of us was born so we can never really feel or understand the mood of the times. Our ideas about the period could unknowingly affect the results when we work with the data. We may be more likely to find different orthographies of words that we expected to find, while missing discussion on other things that we did not expect. We tried to keep this in mind when manually inspecting the automatically produced lists. On the other hand, these are not all disadvantages. When we select words from the present, we may have strong personal opinions about some words, like temporary "linguistic fads" – and these opinions can affect the

³Both versions are found in *Íslensk nútímamálsorðabók* (the dictionary of modern Icelandic; Jónsdóttir and Úlfarsdóttir) and *Íslensk stafsetningarorðabók* (the Icelandic spelling dictionary; Sigtryggsson) gives *bolséviki*.

choice. It is perhaps less likely that we have strong feelings about word use in the past, in times we did not live.

Finally, when looking at one calendar year at a time, notable events that happen at the end of a year can be less conspicuous as the discussion is often divided between two years. An example of this is the discussion on the Spanish flu, referred to in Section 3. When the influenza epidemic was at its worst in Reykjavík in November 1918, the Icelandic newspapers were not published for more than a week (cf., e.g., Bjarnason 2020, 89). A flyer distributed that week reports that the majority of Reykjavík's inhabitants are ill with the disease. In other parts of the country the Spanish flu was not as prevalent, and as most of the newspapers were published in Reykjavík they may not have found any need for in-depth reporting on it or discussing it in great detail when the flu had blown over.

5. Conclusions

We have described an experiment in selecting a word of the year for a different era, applying the methodology used today. We inspect a number of years over the period 1914–1920 and find that while the methodology gives us convincing candidate lists for each year, it has limitations that have to be kept in mind. These include the effects of OCR, orthography that is not well standardised, missing language registers as well as the editorial practices being different from what they are today, possibly giving a narrow view of the topics of the day. Working around this, we created lists of words for each year and selected the one that sticks the most out, *bolséviki* in the year 1919.

Acknowledgments

We would like to thank three anonymous reviewers for their comments and feedback on the paper and Ásta Svavarsdóttir for discussions on aspects of this work.

References

- Barkarson, Starkaður, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022. "Icelandic Gigaword Corpus (IGC-2022) - annotated version." CLARIN-IS. <http://hdl.handle.net/20.500.12537/254>.
- Barkarson, Starkaður, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. "Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al., 2371–2381. Marseille. European Language Resources Association. <https://aclanthology.org/2022.lrec-1.254>.
- Bjarnason, Gunnar. 2020. *Spænska veikin*. Reykjavík: Mál og menning.

- Jasonarson, Atli, Steinþór Steingrímsson, Einar Sigurðsson, Árni Magnússon, and Finnur Ingimundarson. 2023. "Generating Errors: OCR Post-Processing for Icelandic." In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, edited by Tanel Alumäe and Mark Fishel, 286–291. Tórshavn. University of Tartu Library. <https://aclanthology.org/2023.nodalida-1.29>.
- Jónsdóttir, Halldóra, and Þórdís Úlfarsdóttir, eds. *Íslensk nútímamálsorðabók*. Reykjavík: The Árni Magnússon Institute for Icelandic Studies. <https://islenskordabok.arnastofnun.is/>.
- Sigtryggsson, Jóhannes B., ed. *Íslensk stafsetningarorðabók*. Reykjavík: The Árni Magnússon Institute for Icelandic Studies. <https://stafsetning.arnastofnun.is>.
- Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. "Risamálheild: A Very Large Icelandic Text Corpus." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 4361–4366. Miyazaki. European Language Resources Association. <https://aclanthology.org/L18-1690>.
- Tímarit.is*. Landsbókasafn Íslands – Háskólabókasafn. <https://timarit.is>.
- Þorbergsdóttir, Ágústa, and Starkaður Barkarson. 2023. "Orð ársins 2022: Innrás." *Hugrás*, January 11, 2023. <https://hugras.is/2023/01/ord-arsins-2022-innras/>.
- Þorbergsdóttir, Ágústa, Starkaður Barkarson, and Steinþór Steingrímsson. 2024. "Orð ársins 2023: Gervigreind(in)." *Hugrás*, January 5, 2024. <https://hugras.is/2024/01/ord-arsins-2023-gervigreindin/>.
- Þorbergsdóttir, Ágústa, and Steinþór Steingrímsson. 2021. "Orð ársins 2020: Sóttkví." *Hugrás*, January 20, 2021. <https://hugras.is/2021/01/ord-arsins-2020-sottkvi/>.