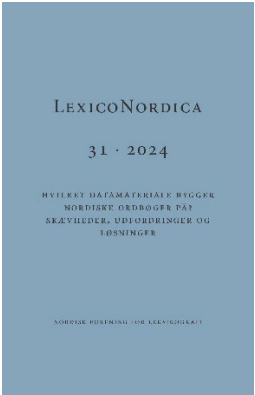


LexicoNordica

Titel:	Representativeness and biases in Icelandic corpora	
Forfatter:	Einar Freyr Sigurðsson & Steinþór Steingrímsson	
Kilde:	LexicoNordica 31, 2024, s. 201-224	
URL:	https://tidsskrift.dk/lexn/issue/archive	

© 2024 LexicoNordica og forfatterne

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Representativeness and biases in Icelandic corpora

Einar Freyr Sigurðsson & Steinþór Steingrímsson

All language data are inherently biased, as collection methods, availability of texts and recordings, and the views of the collectors will always affect the process and its results. We examine how bias is manifested in NLP tools trained on corpora and how that could be used to detect biases in Icelandic corpora. We also look at how sports coverage seems to exhibit something that we call male-by-default bias, as an example of a bias that might be hard to detect using automatic approaches. Finally, we suggest how metadata could be enriched to perform better analyses.

1. Introduction

In recent years there has been substantial growth in available language resources for use in Icelandic language technology and linguistic research.¹ The Icelandic Gigaword Corpus (IGC) is the largest of these resources. Its latest version comprises approximately 2.5 billion words. *Tímarit.is* is a collection of all major newspapers, magazines and periodicals published in Icelandic from the 19th century to the present day, digitized using OCR (Optical Character Recognition) and made available online. For both corpora, the focus is on quantity rather than on balance or on being representative of the language as a whole. By *representativeness* of a corpus, we refer to the relation between the corpus and the language it is being used to represent (Hunston 2008).

When we do research using corpora, we are faced with questions of potential biases regarding what is represented and to what

1 We would like to thank two anonymous reviewers and the editors for very helpful comments on this paper.

degree. Some things may be overrepresented, while others may be underrepresented. These biases can, however, often be difficult to detect. The data contained in the two corpora mentioned above stem from different sources and represent different registers and genres. It can be argued that a certain dataset is in some way representative of a certain type of Icelandic, due to its origins and how it was collected. Nonetheless, all language data are inherently biased to some extent, as collection methods, availability of texts and recordings, and the views of the collectors will always affect the process and its results. When language data are used for research, the researcher must be aware of these limitations.

In this paper, we propose two research questions:

1. How can we use existing corpora to find ingrained biases, such as gender biases?
2. What kind of metadata is needed to facilitate research on biases and representativeness?

We seek to answer these questions from the viewpoint of Icelandic corpora, discuss potential biases in these corpora with respect to representativeness, and discuss possible approaches for answering these questions.

We examine how bias is manifested in NLP (Natural Language Processing) tools trained on corpora and how that could be used to detect biases in Icelandic corpora. We also look at how sports coverage seems to exhibit something that we call male-by-default bias, as an example of a bias that might be hard to detect using automatic approaches. Furthermore, we suggest how metadata could be enriched in order to better analyse where and how biases and other specific types of artifacts present themselves in the data.

2. A note on biases

Before we go any further, it is crucial to state what we mean when we refer to biases. It may be helpful to look at dictionary definitions of the word *bias*. Among other things, *Merriam-Webster* mentions ‘prejudice’ and ‘deviation’ under its definitions of the noun *bias*. One definition of the noun in *Collins English Dictionary* talks about bias being “a tendency to prefer one person or thing to another, and to favour that person or thing”, and includes the example “Bias against women permeates every level of the judicial system”. Finally, *Cambridge Dictionary* talks about *bias* as “the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment” and also as “the fact of a collection of data containing more information that supports a particular opinion than you would expect to find if the collection had been made by chance”.

The definitions picked out from these dictionaries are in line with our thinking of the term and what we are referring to when we use it in this paper, where *bias* entails that there is something skewed, where, for example, the sample does not represent the population.

Many different types of bias have been detected and some are well known and have a name of their own. The entry for *bias* in the *Cambridge Dictionary* gives three examples of biases, *publication bias* (negative results are not reported), *selection bias* (e.g., in social studies) and *survivorship bias* (the tendency for failed samples to be excluded from performance studies). Other commonly known biases include *confirmation bias* and *authority bias*. When we discuss various biases in this paper, we name them for the sake of clarity.

When we work with language, whether it is in the context of linguistic research, when building language technology tools or as

lexicographers, we are constantly faced with questions of what and whose language is represented.

Linguists sometimes give their own linguistic judgments when they are studying their native language. From their own acceptability judgments, they may overgeneralize and claim that a certain phenomenon or construction is ungrammatical in the language they speak. If the linguists do not consult other speakers, who might disagree with their judgments, we may witness something we can call *my-grammar-as-the-default bias*.

We might encounter something similar if we were, for example, working on enhancing a dictionary by finding and selecting words not found in previous versions. We might be biased towards accepting new words that we are familiar with rather than words we have not heard before. We could refer to this as *my-dialect-as-the-default bias*.

Furthermore, representativeness and biases are clearly a factor when we look at prescriptivism. Prescriptivist directions on language use often give examples of actual language use that is to be avoided, for one reason or another. When Böðvarsson (1992:21) states that the plural of *Ástrali* ‘Australian’ is *Ástralar* ‘Australians’, he also adds that the plural *Ástralir* is incorrect. He would probably not have mentioned this unless he knew or believed *Ástralir* is a form that people actually use. In the light of this, it is interesting that when we search for “Ástralir” and “Ástralar” in the IGC, we get 2,465 results for the “incorrect” form *Ástralir*, while we only get 1,888 hits for the “correct” form *Ástralar*, or 57% vs. 43%, respectively. This indicates that the prescriptively favoured variants do not necessarily represent the language or grammar of the majority of speakers – or how the language is actually spoken.

Prescriptivism often also invokes authority bias – if directions on how to speak properly do not conform with our own grammar, it can matter who it is that considers this or that to be improper. An interesting situation may currently be developing in Icelandic

discourse: In recent years, professor emeritus Eiríkur Rögnvaldsson, a well-known linguist in Iceland, has objected to how prescriptivist directions are often given and contradicted in various ways by actual language use – see, e.g., Rögnvaldsson (2022). On the one end we have the prescriptivist authorities that have been followed throughout the years, whereas on the other we have another authority objecting to saying that linguistic phenomena like “dative sickness” are incorrect or improper language.

Still, the question arises why we should be concerned about biases in corpora. For one thing, some biases we encounter can be harmful. Nonetheless, it is not always clear in what way. Blodgett et al. (2020) surveyed 146 papers on biases in NLP systems. It was generally unclear in these papers “what kinds of system behaviors are harmful, in what ways, to whom, and why” (Blodgett et al. 2020:5454). It is not immediately clear whether something like the my-grammar-as-the-default bias can cause harm, but it might be harmful if certain features of someone’s grammar or dialect are stigmatized; individuals may, for example, become insecure when expressing themselves. Such biases could also “lead to feelings of invisibility and marginalization”, to use Friðriksdóttir & Einarsson’s (2024:7596) words when discussing potential harms caused by gender biases.

In fact, in some cases it is more obvious how gender biases could be harmful. Such biases have frequently been discussed in recent years in relation to language technology. Some of this research, like the aforementioned study by Friðriksdóttir & Einarsson (2024), focuses on Icelandic. Friðriksdóttir & Einarsson conducted an experiment in which they got large language models to predict the pronoun in a sentence on the form *he/she/they is/are a(n) <occupation term>*. The goal was to see whether large language models “merely echo the gender distribution within respective professions” or if “they exhibit biases aligned with their grammatical genders” (Friðriksdóttir & Einarsson 2024:7597). In this

context, they define gender bias “as the tendency of these models to generate or perpetuate gender stereotypes” (Friðriksdóttir & Einarsson 2024:7596). They furthermore state that a gender bias “can reinforce harmful societal norms, such as by influencing individuals’ perceptions regarding the careers or roles accessible to them based on their gender” (ibid.).

We will not be discussing Friðriksdóttir & Einarsson’s (2024) study further. However, we look at a different study on gender bias, by Sólmundsdóttir et al. (2021, 2022), in machine translation from English to Icelandic in section 3.2, and in section 3.3 we show examples of how we can detect biases and imbalances in the IGC and *Tímarit.is* that relate to gender or sex.

3. Detecting biases and imbalance in Icelandic corpora

The largest resources, both in terms of mere quantity and time-span, for studying written Icelandic are the IGC and *Tímarit.is*. Therefore, when examining potential biases and imbalances in Icelandic corpora and how we can detect them, we work with these two corpora. In this section we also discuss how biases present themselves in NLP tasks such as machine translation (MT) and word embeddings, and how such tasks can be used to help identifying biases in corpora.

3.1 Representativeness and balance

The IGC is an ongoing corpus project, with new versions of the corpus published on a regular basis. The first edition, published in 2017, contained 1.3 billion running words. The latest edition, published in 2022, has over 2.5 billion running words in eight sub-corpora: journals, news, social media, parliamentary proceedings,

adjudications, laws, published books, and Wikipedia (see description in Barkarson et al. 2022). The corpus has become fundamental for linguistic research, dictionary work, and language technology for Icelandic. The data collection has focused on collecting recently published data from sources that allow for the data to be distributed with permissive licenses. All the data are collected in digital format from the source or publishers; no OCR is carried out when the corpus is compiled. This entails that the corpus comprises mostly recent texts, with the vast majority having been published after the year 2000. *Tímarit.is*, on the other hand, contains only OCR-processed texts. The accessibility also differs, the IGC is PoS-tagged and lemmatized, searchable through Korp, a KWIC-system for studying corpora, as well as being downloadable, while *Tímarit.is* is only available through a string-based search engine and is neither tagged nor lemmatized.

The goal of the IGC project is to build as large a corpus as possible of contemporary texts in a language spoken by less than 400 thousand people and, instead of emphasizing representativeness, the aim was to achieve as much coverage as possible and to provide extensive metadata so that users of the corpus can construct their own subcorpora as needed. Steingrímsson et al. (2018:4361) point out in the original IGC paper that trying to achieve representativeness in a text corpus can be problematic. First of all, what should it be representative of? It can be difficult to determine whose language and perspective is represented, and then again how accurately it is represented. And because it can be hard to determine where a variety of language ends and another begins, any corpus is virtually by definition biased to a greater or a lesser extent. In their discussion on balance and representativeness, Beelen et al. (2022) raise the question of what type of representativeness is desirable, and they describe how problematic it can be to achieve a balance of perspectives without downplaying any of them.

Finally, to know the strengths and limitations of the corpus

better, the user needs to understand which type of language the corpus represents. Rich metadata on the origins, classification, and analysis of the text are essential for reaching this understanding and to discern which biases can be expected, although experimentation is also needed to reveal them.

3.2 Machine translation

Machine translation (MT) systems rely on large amounts of data for training. Neural machine translation (NMT), which has been the dominant paradigm in MT since 2016–2017, needs large parallel corpora in order to achieve acceptable translation capability, while the more recent large language models (LLMs) are trained on billions of words of monolingual texts. In both cases, the texts are likely to reflect views and opinions of those who wrote or published the texts, and these may or may not be appropriate for the MT systems being trained.

Sólmundsdóttir et al. (2021, 2022) detected difference in gender use in Google translations from English to Icelandic. The authors point out that technology could maintain “societal inequalities and outdated views” (Sólmundsdóttir et al. 2022:3113) and come to the conclusion that “the results show a pattern which corresponds to certain societal ideas about gender and gender roles” (Sólmundsdóttir et al. 2021:199). Furthermore, a large amount of data will not necessarily guarantee its diversity, see, e.g., the discussion in Bender et al. (2021).

English does not exhibit gender on adjectives and past participles, whereas Icelandic does. It is therefore interesting to see how sentences like *I am <ADJECTIVE>* are translated. In their research, using Google Translate, Sólmundsdóttir et al. (2021, 2022) observed a peculiar gender bias when translating predicative sentences from English to Icelandic. They compiled a list of adjectives generally used to describe people and classified them in two categories: 1) words that describe personality traits, such as

strong, weak, clever, or stupid, and 2) words that describe appearance, such as *beautiful, ugly, fat, or thin*.

The authors found that adjectives describing positive personality traits appeared more often in the masculine form, while negative ones were more likely to appear in the feminine form. They examined 262 adjectives that describe people's personal traits. 156 of them were translated in the masculine, whereas 65 were translated in the feminine – with the rest being translated in the neuter, as uninflected adjectives or as syncretic for masculine and feminine (Sólmundsdóttir et al. 2021:189). Interestingly, 59% of the adjectives used in the masculine were considered by the authors to describe positive features, whereas only 23% of the adjectives used in the feminine were positive (Sólmundsdóttir et al. 2021:189). Two examples are shown below, where *strong* is translated as the masculine *sterkur* and *weak* as the feminine *veik*:

- (1) English: I am *strong*.
Icelandic: Ég er *sterkur*. (masculine; positive)
 - (2) English: I am *weak*.
Icelandic: Ég er *veik*. (feminine; negative)
- (Sólmundsdóttir et al. 2021:190)

The study also looked at 67 adjectives that describe people's look or appearance, such as *beautiful* and *handsome*. Of these, 31 adjectives were translated into the masculine gender, whereas 15 were translated into the feminine. Here, however, the ratio of positive adjectives is much higher among the feminine usage: 67% of the adjectives in the feminine were positive as opposed to 23% of the adjectives that were used in the masculine (Sólmundsdóttir et al. 2021:191–192).

Moreover, the research tested adjectives in predicative sentences that describe the speaker's ability to carry out certain tasks. Two examples are shown in the following:

- (3) English: I am *good* at electrical work.
 Icelandic: Ég er *góður* í rafmagnsvinnu. (masculine)
- (4) English: I am *good* at cooking.
 Icelandic: Ég er *dugleg* að elda. (feminine)
 (Sólmundsdóttir et al. 2021:193)

In this part of the research, Sólmundsdóttir et al. focused on sentences that describe people's ability in craft and industry, on the one hand, and housekeeping, on the other. When adjectives in sentences like *I am good at electrical work* (ability in craft and industry) were translated, they were used in the masculine in 12 out of 15 cases. Furthermore, different adjectives seemed to be used, depending on the gender: *dugleg* for the feminine, *góður* for the masculine, even though they were used to translate the same English adjective, *good*. When adjectives in sentences describing housekeeping were translated, they were in the feminine in 18 out of 21 examples (Sólmundsdóttir et al. 2021:192–193).

The results for sentences containing adjectives which describe people's appearance (e.g., *beautiful*, *handsome*) are in some ways opposite to the results for adjectives that describe people's personality traits (e.g., *strong*, *weak*). The authors state that this shows a pattern which corresponds to societal ideas about gender. In the light of how MT systems are developed, this bias must reflect the data the systems are trained on. When the systems are faced with ambiguity, they generate the most likely translation with the likelihood derived from the training data. This is thus an example of an MT system perpetuating a societal bias presented in corpora used to build these systems.

LLMs trained on massive monolingual data sets in multiple languages rather than parallel corpora, have been shown to exhibit similar tendencies. Vanmassenhove (2024) ran a small experiment where she evaluated ChatGPT (based on GPT-3.5) in terms of translating ambiguous words with respect to gender from

English to Italian. Her findings indicate a strong male bias which becomes even more prevalent when asked to generate alternatives. She concludes with a call to raise awareness about these issues and taking proactive steps to address them.

Going back to our discussion on detecting bias in corpora, the case of MT shows that understanding what different parts of text corpora represent, and what biases we are likely to find in them, can be crucial when developing NLP systems which derive their model of the language from data, whether we want to mitigate gender bias or other undesirable artifacts.

3.3 Bias and imbalance detection

In previous subsections, we have discussed how biases and imbalances can have undesirable effects in different NLP tasks, where models based on text corpora are employed. In this subsection we look at how data in corpora can be imbalanced and biased with respect to gender, and how these biases are not necessarily easily detected.

3.3.1 *I am* ... (counting linguistic phenomena)

Inspired by Sólmundsdóttir et al.'s (2021, 2022) work, we decided to look for examples that are somewhat similar to the ones they discuss (e.g., *I am good at housekeeping*).² We searched for sentences that start with *ég er* 'I am' immediately followed by a "strongly" inflected adjective or past participle³ in masculine, feminine and neuter, which in turn is followed by an infinitival marker and

2 We have, however, not looked at the distribution of the gender of adjectives with respect to, e.g., positive and negative personality traits. We leave that for future research.

3 By excluding weak inflection we exclude various examples that are syncretic for feminine and masculine, and we also exclude many examples that are not applicable, such as *Ég er meira að segja* ... 'I even am ...'

then a verb in the infinitive.⁴ We show examples for each gender use below that we found using our search queries in the IGC.

- (5) *Ég er glaður* að hafa gert það.
 I am glad.MASC.SG to have done that
 ‘I’m glad I did it.’
- (6) *Ég er búin* að reyna allt.
 I am done.FEM.SG to try everything
 ‘I have tried everything.’
- (7) *ég er búíð* að biðjast afsökunar :)
 I am done.NEUT.SG to ask apology
 ‘I have apologized.’

With this search we find examples where the predicate consists of a single word – an adjective or a participle – as it is immediately followed by an infinitival clause; therefore, the adjective or participle could not be part of a larger noun phrase (which would determine its gender, as in *Ég er glaður nemandi* ‘I’m a glad/happy student’, where the masculine noun *nemandi* determines the grammatical gender of the adjective). This gives us examples where an adjectival or participial predicate agrees with *ég* ‘I’ (which itself does not show gender features). The gender of the adjective or participle then indicates the gender or sex of the speaker. We take the speaker in (5) to be a male speaker, the one in (6) to be a female speaker, while speakers in examples like (7) could be non-binary or genderqueer using neuter when referring to themselves.

We expected utterances by male speakers to be the most frequent out of the three as we expected male speakers to be the most

4 The search query for feminine gender was as follows: <sentence> [word = “Ég” %c] [word = “er” %c] [(pos = “I” | hattu = “þ”) & lob = “s” & kyn = “v”] [word = “að” %c] [hattu = “n”]. By replacing *kyn* = “v” with *kyn* = “k” or *kyn* = “h” in the query, we get the search queries we used for the masculine and neuter, respectively.

dominant in the overall discussion. That was not the case, however.⁵

- (8) a. Masculine: 31,731 results
 b. Feminine: 75,263 results
 c. Neuter: 742 results

The fact that there are twice as many examples of feminine agreement with *ég* 'I' comes as a surprise if we take these results at face value. Why are there so many utterances that suggest female speakers as opposed to male speakers? Does the IGC generally represent much more female speakers than male speakers? Or do female speakers use this construction more than male speakers? We could ask many other questions to get a clearer picture – showing that we need to take a closer look at what is behind these numbers.

	Feminine	Masculine
Journals	20	14
News	5,519	8,392
Social media	69,134	21,412
Parliamentary proceedings	494	1,772
Adjudications	4	16
Laws	5	1
Books	87	124
Wikipedia	0	0
All IGC	75,263	31,731

Table 1: Feminine vs. masculine agreement in the IGC.

- 5 It should be noted that the majority of the 742 results for the neuter do not reflect the gender of the speaker, as many of these results are utterances like *Ég er satt að segja ...* 'I am, truth be told, ...' where the gender feature on the adjective – in this case *satt* 'true' – does not come from the subject *ég* 'I'. If the subject were *hann* 'he' or *hún* 'she', the form of the adjective would still be neuter: *Hann/hún er satt að segja ...*

When we look at the distribution with respect to subcorpora within the IGC, see Table 1, we see a different picture than the one painted in (8) above. In four out of eight subcorpora, namely news, parliamentary proceedings, adjudications and books, the masculine is used in the majority of cases. There are no examples of what we looked for in the Wikipedia subcorpus, and only in three subcorpora, namely journals, social media and laws, is there more use of the feminine than the masculine form in the construction we are looking at. Furthermore, out of 75,263 examples of feminine-form use, 69,134 examples are from the social-media subcorpus.

3.3.2 Sports coverage with respect to sex

In a survey looking at sports coverage with respect to sex in Icelandic newspapers from 1 May 1999 to 30 April 2000, around 85% of the coverage was on men's sports, 7% on women's sports, and 8% on sports in general (*Nefnd um konur og fjölmíðla: álit og tillögur* 2001:20). To test whether this has changed, we might try some simple methods, like looking at the frequency of the use of words that indicate which sex is being discussed. Such a word pair is *kvennalandsliðið* 'the women's national team' and *karlalandsliðið* 'the men's national team'.

We looked for these two terms on *Tímarit.is* to see whether the survey's results would be reflected in the use of the two terms. The numbers for the decade 2010–2019 are 1653 (48%) search results for *kvennalandsliðið* vs. 1802 (52%) results for *karlalandsliðið* (these two word forms are each syncretic for nominative and accusative case and have a suffixed definite article; we did not search for results in the dative and the genitive case). Given the numbers in the report mentioned above, this would certainly indicate that there is more discussion on women in sports than before. However, the total results for the two terms are rather interesting, as shown in Table 2.

	Kvennalandsliðið	Karalandsliðið
1950–1959	9 (100%)	0 (0%)
1960–1969	77 (76%)	24 (24%)
1970–1979	138 (69%)	62 (31%)
1980–1989	631 (73%)	234 (27%)
1990–1999	862 (71%)	354 (29%)
2000–2009	1541 (64%)	851 (36%)
2010–2019	1653 (48%)	1802 (52%)

Table 2: *Kvennalandsliðið* vs. *karlalandsliðið* on *Tímarit.is*.

In all the decades prior to 2010–2019, *kvennalandsliðið* is more frequent in the corpus on *Tímarit.is* than *karlalandsliðið*. Various news reports like the one in (9) shed a light on the possible reason for this (English translation ours).

(9) *Landsliðið í handbolta*

Landsliðið í handbolta sem leikur gegn Luxemborg hefur verið valið [...]

The national handball team

The national handball team, which plays against Luxemburg, has been chosen [...]

(*Þjóðviljinn*, 25 November 1975, p. 14)

This is the heading together with the first words of a very short article about the men's national team in handball. However, there is no mention of this being the men's team – it only says “Landsliðið í handbolta” ‘the national handball team’. It cannot be inferred from the context elsewhere on the page that this is the men's national team. That is, ‘the national team’ without mentioning the sex seems to refer to men by default, something we can call *male-by-default bias*.

The results of our little search query on *Tímarit.is* are yet another example where we cannot simply take the results at face

value and they reveal a gender bias in the data. Even though the ratio for *kvennalandsliðið*, compared to *karlalandsliðið*, is lower for 2010–2019 than any other decade, this may in fact reflect a more balanced coverage of men and women in sports – especially if there is less of the male-as-default bias than in previous decades.

3.4 Word embeddings

Word embeddings are vector representations of words calculated from very large data sets. Using the popular Word2Vec model (Mikolov et al. 2013), each word is typically represented by a 300-dimensional vector of real numbers. These vectors capture semantic and syntactic information about the word, based on surrounding words in the training data. Figure 1 shows an example of how four words could be represented in vector space. By observing their position in relation to one another, we find that by simple arithmetic we can calculate semantic relationships; in this case *King* – *Man* + *Woman* = *Queen*. Similarly, we can find the closest equivalents for any word in a corpus by training word embeddings on the corpus and in a similar manner, calculate the distance between words in the vector space.

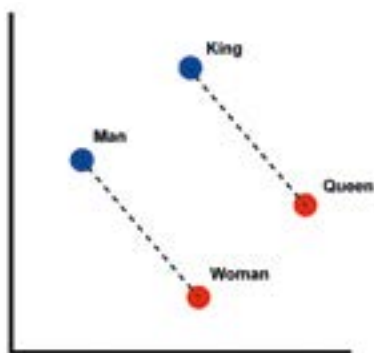


Figure 1: A simplified visual illustration of the relation between four words in vector space. Figure taken from Wikipedia (<en.wikipedia.org/wiki/Word2vec>).

Garg et al. (2018) show that word embeddings can be effective tools to study historical biases and stereotypes. They do this by relating measurements from embeddings trained on 100 years of text data to historical census and survey data, and they find that changes in the embeddings track with demographic and occupation shifts over time.

They use embeddings each trained on a decade of 20th century texts and inspect words from word lists they collate to represent gender and ethnicity, as well as lists of adjectives and occupations. Using the embeddings and word lists they measure the strength of association between a group and neutral words in several experiments. For example, they compute the average embedding distance between words that represent gender, on the one hand, and ethnicity, on the other, as well as words for occupations to estimate the strength of the embeddings to calculate sociological trends over time. They compare the results to historical surveys and show that the embeddings capture both gender and ethnic occupation percentages and consistently reflect historical changes. They also try to quantify ethnic and gender stereotypes by finding the top adjectives associated with different groups over time and the adjectives most biased towards Asians.

1910	1950	1990
irresponsible	disorganized	Inhibited
envious	outrageous	Passive
barbaric	pompous	Dissolute
aggressive	unstable	Haughty
transparent	effeminate	Complacent
monstrous	unprincipled	Forceful
hateful	venomous	Fixed
cruel	disobedient	Active
greedy	predatory	Sensitive
bizarre	boisterous	Hearty

Table 3: Adjectives most biased towards Asians in 1910, 1950 and 1990 in the experiment carried out by Garg et al. (2018).

Table 3 shows how adjectives biased towards Asians have changed over time in the data the embeddings have been trained on. While this may not completely reflect the general attitude of the time, it is at least an indication of how the discussion has changed.

Work such as this shows that machine learning approaches can be used to help us understand bias and representativeness in texts. It can give us some indication of what can be expected if we, or the tools we build, try to generalize from the corpus, but it can also be used for teaching us about the fluidity of the language and how biases may change over time or even between different text sources.

While the IGC mainly comprises texts from the 21st century, the news subcorpora contain information about sources of the texts and for some sources there is also further categorization of the texts. Some of the other sources have information about age and gender of authors, and the parliamentary speeches have party affiliation. Using the word embeddings approach could potentially help us identify biases in the corpus not only over time but also between text categories and in speeches of parliamentarians from different parties.

4. Which metadata are needed?

In the IGC, all information accompanying the texts that were collected is distributed as metadata. All eight subcorpora have publisher, date, and place of publishing, if available. All texts are tagged and lemmatized. For books, journals, and parliamentary proceedings, information on the author is also available. For parliamentary proceedings, gender, year of birth, age when the text was written, and political affiliation are also included. As an example of the usefulness of the metadata for linguistic research, we can mention the study by Stefánsdóttir & Ingason (2018, 2022) on the variable use of stylistic fronting in Icelandic in thousands

of parliament speeches given by parliamentarian Steingrímur J. Sigfússon, where they identify syntactic change across his lifespan related to status-associated factors.

4.1 Generating metadata using NLP tools

Enriching a corpus with more metadata that allow for more fine-grained research could be achieved by analysing the data using NLP tools. *Tímarit.is* does not have any categorization that can be used while searching the corpus, and while the IGC contains some categorization of texts, in particular news from some media outlets, more thorough categorization could be helpful, for example for studies such as the one conducted in section 3.3.2 above. An example of such a categorization could be to analyse news and classify them into fine-grained categories, such as:

(10) *News* → *Sports* → *Handball* → *Men* → ...

Sentiment analysis could be used to investigate which topics are being discussed in a positive or negative fashion, or even how people or groups of people are being discussed in the texts. To make the user able to gather information on certain individuals or groups of people, named entity recognition would make analysis easier.

4.2 Other corpora

For other corpora, for example web-scraped corpora and data in newspaper scanning projects (e.g., *Tímarit.is*), less metadata are available. On the other hand, a lot is often known about the data. A newspaper could be right-leaning or left-leaning, it could be funded by a certain industry or solely by subscriptions. It could be yellow press journalism or business oriented, etc. If we are seek-

ing balance of representation, should we then also take circulation into account? All these types of information can be important when the texts are analysed to identify biases or to understand what or who the texts are likely to represent. By providing these types of additional information, the corpus would make all such endeavours easier for researchers.

Furthermore, some newspapers are read by many, while others are read by few. Should these newspapers be regarded as equal or should researchers using these data use some sort of oversampling or undersampling approaches? Similar principles could apply to books. A book may be popular and be read for decades after publication while another book is only read by 20 people in the first few months after publication and then forgotten. While such information can be useful for some researchers, it has to be acknowledged that corpus publishers do have to prioritize and include what is most likely to be of use, especially in cases where adding the information requires time-consuming manual work.

5. Conclusion

In the introduction above, we outlined the following research questions:

1. Can we use existing corpora to find ingrained biases?
2. What kind of metadata is needed to facilitate research on biases and representativeness?

While we have not made an effort to give concrete and definite answers to these big questions, we have tried to shed a light on the topic with respect to Icelandic and Icelandic corpora.

Firstly, we have, based on discussion on research in machine translation, considered the importance of understanding what dif-

ferent parts of text corpora represent, and what biases we are likely to find in them. Furthermore, we discussed how word embeddings can play a role in detecting biases and how they change over time or differ between text sources. We also pointed out biases that are not necessarily easily detectable and need to be considered.

Secondly, we sketched up possible approaches to enrich the metadata with the goal in mind to facilitate bias and imbalance detection.

A demonstration of how biases are to various extent ingrained in all text corpora and how they can be detected, may help lexicographers, as well as language technologists, understand the limitations of corpora and perhaps facilitate richer data selection that reflects more diverse aspects of the language and language use.

It is a constant problem trying to strike the balance between different sources, quantity and quality of texts, and the perfect balance for one user may not suit all others. Introducing new text sources may introduce biases not prevalent in other sources and while larger corpora may give us more examples to work with, they may also amplify various biases. When enlarging corpora, giving users the tools to analyse them, descriptive metadata and perhaps some analyses may help the users find the balance that suits their needs.

References

Dictionaries and corpora

- Cambridge Dictionary*. <dictionary.cambridge.org> (April 2024).
Collins English Dictionary. <www.collinsdictionary.com> (April 2024).
 Icelandic Gigaword Corpus = Barkarson, Starkaður, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteins-

- dóttir, Finnur Ágúst Ingimundarson & Árni Davíð Magnússon (2022): *Icelandic Gigaword Corpus (IGC-2022) – annotated version*. CLARIN-IS. <hdl.handle.net/20.500.12537/254> (April 2024).
- Merriam-Webster.com*. <www.merriam-webster.com> (April 2024).
- Timarit.is*. Landsbókasafn Íslands – Háskólabókasafn. <timarit.is> (February 2024).

Other references

- Barkarson, Starkaður, Steinþór Steingrímsson & Hildur Hafsteinsdóttir (2022): Evolving large text corpora: Four versions of the Icelandic Gigaword Corpus. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 2371–2381 <aclanthology.org/2022.lrec-1.254>.
- Beelen, Kaspar, Jon Lawrence, Daniel C.S. Wilson & David Beavan (2022): Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. In: *Digital Scholarship in the Humanities* 38, 1–22. <doi.org/10.1093/llc/fqac037>.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell (2021): On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*. New York, NY: Association for Computing Machinery. 610–623. <doi.org/10.1145/3442188.3445922>.
- Blodgett, Su Lin, Solon Barocas, Hal Daumé III & Hanna Wallach (2020): Language (technology) is power: A critical survey of “bias” in NLP. In: Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Associ-

- ation for Computational Linguistics. 5454–5476. <aclanthology.org/2020.acl-main.485>.
- Böðvarsson, Árni (1992): *Íslenskt málfar*. Reykjavík: Almenna bókafélagið.
- Friðriksdóttir, Steinunn Rut & Hafsteinn Einarsson (2024): Gendered Grammar or Ingrained Bias? Exploring Gender Bias in Icelandic Language Models. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, LREC-COLING 2024*. Torino: ELRA and ICCL. 7596–7610. <aclanthology.org/2024.lrec-main.671>.
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky & James Zou (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. In: *Proceedings of the National Academy of Sciences* 115, E3635–E3644. <www.pnas.org/doi/abs/10.1073/pnas.1720347115>.
- Hunston, Susan (2008): Collection strategies and design decisions. In: Anke Lüdeling & Merja Kytö (eds.): *Corpus Linguistics: An International Handbook*, Volume 1. Berlin: De Gruyter. 154–168.
- Mikolov, Tomas, Kai Chen, Gregory S. Corrado & Jeffrey Dean (2013): Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations*. <arxiv.org/abs/1301.3781>.
- Nefnd um konur og fjölmiðla: álit og tillögur* (2001). Reykjavík: Menntamálaráðuneytið. <rafladan.is/handle/10802/6120>.
- Rögnvaldsson, Eiríkur (2022): *Alls konar íslenska. Hundrað þættir um íslenskt mál á 21. öld*. Reykjavík: Mál og menning.
- Stefánsdóttir, Lilja Björk & Anton Karl Ingason (2018): A high definition study of syntactic lifespan change. In: *University of Pennsylvania Working Papers in Linguistics* 24, 169–178.
- Stefánsdóttir, Lilja Björk & Anton Karl Ingason (2022): Einstaklingsbundin lífsleiðarbreyting: Þróun stílfærslu í þingræðum Steingríms J. Sigfússonar. In: *Íslenskt mál og almenn málfræði* 44, 151–178.

- Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson & Jón Guðnason (2018): Risamálheild: A Very Large Icelandic Text Corpus. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki: European Language Resources Association. 4361–4366. <aclanthology.org/L18-1690>.
- Sólmundsdóttir, Agnes, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir & Anton Karl Ingason (2021): Vondar vélþýðingar: Um kynjahalla í íslenskum þýðingum Google Translate. In: *Ritið* 3/21, 177–200. <doi.org/10.33112/ritid.21.3.7>.
- Sólmundsdóttir, Agnes, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir & Anton Karl Ingason (2022): Mean machine translations: On gender bias in Icelandic machine translations. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 3113–3121. <aclanthology.org/2022.lrec-1.333>.
- Vanmassenhove, Eva (2024): Gender bias in machine translation and the era of large language models. In: *arXiv* <arxiv.org/html/2401.10016v1>.

Einar Freyr Sigurðsson
Research Associate Professor, ph.d.
The Árni Magnússon Institute for
Icelandic Studies
Edda, Sæmundargata 5
IS-107 Reykjavík
einar.freyr.sigurdsson@
arnastofnun.is

Steinþór Steingrímsson
Research Assistant Professor, ph.d.
The Árni Magnússon Institute for
Icelandic Studies
Edda, Sæmundargata 5
IS-107 Reykjavík
steinthor.steingrimsson@
arnastofnun.is