A word frequency dictionary of Icelandic child-directed speech

Hinrik Hafsteinsson & Einar Freyr Sigurðsson

In this paper we present the making of a word frequency dictionary of Icelandic child-directed speech. The material consists of transcribed video recordings of two boys during their language acquisition period and their fathers (and, to a lesser degree, their mothers). The older boy, born in 1982, is the father of the younger one, born in 2011, which means that through the video recordings we get access to the spoken language of three generations within the same family. Data containing child-directed speech is important for research in language acquisition. Both the dictionary and the tools made in the project are freely available online, which hopefully facilitates Icelandic language acquisition research.

1. Introduction[1]

We present the making of a word frequency dictionary of Icelandic child-directed speech.[2] The product itself is twofold. On the one hand, it is a dictionary which contains concrete information on children's input in language acquisition. It is available at www.github.com/hinrikur/BKL and can be used in research on child-directed speech. On the other hand, to be able to make the dictionary, we have developed methods and tools (Python scripts) which are, just like the dictionary, freely available online.

When finished, the dictionary will be the product of two transcribed corpora—the Gunnar Corpus (S.J. Sigurðsson 2019) and the Kalli Corpus (Sigurðsson & Árnadóttir 2019)—of approximately 50 hours in total of video recordings of two children, born 30 years apart, during their language acquisition period. The videos, recorded in most part by the children's fathers, contain conversations between the children and their parents. By focusing on what the children hear (child-directed speech) we gain information on the input in language acquisition and are able to, e.g., assess changes in Icelandic that are currently underway.

[1] We thank two anonymous reviewers and the editors for their comments on the paper.
[2] The dictionary is a part of the project "Í beinan karllegg: Skráning talmáls þriggja ættliða" ('Patrilineal Descent: Transcribing Spoken Language of Three Generations') which was funded by The University of Iceland Research Fund in 2019.

2. Why child-directed speech?

The term *child-directed speech* (or motherese or caregiver speech) is often used when highlighting the fact that parents, for example, talk in a particular way to their young children (such as in infant-directed speech; see discussion in, e.g., de Boer 2012). We use the term to focus on the fact that we are investigating not what the children themselves say, but the input they receive from their parents and others, which may differ in various ways from other types of speech or texts.

Even though children's input in language acquisition has been studied in detail cross-linguistically, few studies exist on the input in the acquisition of Icelandic (see discussion in Nowenstein 2014). However, Icelandic language acquisition is well researched and, as a matter of fact, a word frequency dictionary of Icelandic child language does exist (Einarsdóttir et al. 2019). There are also various Icelandic child language corpora (which also contain child-directed speech), most notably the Ari/Kari Corpus (Strömqvist, Ragnarsdóttir et al. 1995), the Birna Corpus (Gíslason et al. 1983), the Dóra Corpus (Pálsdóttir 1979), the Einarsdóttir Corpus (Einarsdóttir 2018), the Eva Corpus and the Fía Corpus (Sigurjónsdóttir 2000, 2007). Nonetheless there is a clear lack of child-directed speech data.

A variety of changes are currently underway in Icelandic syntax. It is important to study language acquisition to better understand these changes, and child-directed speech, which reflects the input in acquisition, is well suited for this. Some changes in Icelandic are more lexically rooted than others, such as dative substitution, where accusative subjects of experiencer verbs are replaced by dative subjects (e.g., Svavarsdóttir 1982, Jónsson & Eyþórsson 2003, Yang 2016 and Nowenstein 2017). Speakers have to memorize, word by word, which experiencer verbs take an accusative subject. Some such verbs, e.g., *dreyma* 'dream', *langa* 'want' and *vanta* 'need', are frequently used in child-directed speech. As other such verbs, e.g., *bresta* 'lack', *hrylla við* 'be horrified by' and *sundla* 'feel dizzy', may be used much less frequently, it is important to explore to what degree they are used. If it turns out that (i) the number of experiencer verbs that take an accusative subject in standard Icelandic is much lower in child-directed speech than the number of experiencer verbs that take a dative subject and (ii) the frequency of many or most of the accusative subject verbs that do occur in child-directed speech is low, then we can better understand the ongoing change of dative substitution.

It should be emphasized that even though a change like dative substitution has been investigated in great detail, linguists have had very limited access to the input of children acquiring Icelandic. We believe that the dictionary, as well as the corpora which it is based on, is a step in the right direction.

Word frequency can potentially also help us understand how dative subjects in passives are acquired. A recent study by Sigurðsson, Nowenstein & Sigurjónsdóttir (2018) suggests that children acquire dative subjects that originate as

indirect objects in passives earlier than dative subjects that originate as direct objects. That is surprising and may indicate that ditransitives, such as *gefa* 'give', *selja* 'sell' and *senda* 'send', are more frequent in the input than dative-taking monotransitives, such as *hjálpa* 'help', *hrinda* 'push' and *kasta* 'throw'. This can be explored further, using our word frequency dictionary.

3. The videos

The video recordings feature two boys, Gunnar (b. 1982) and Kalli (b. 2011). The videos were recorded in most part by the children's fathers and are from the periods 1987 to 1990 (when Gunnar was 5–8 years old) and 2013 to 2018 (when Kalli was 2–7 years old).

Through the transcriptions we have access to the language of three generations within one and the same family, as Gunnar is, in fact, Kalli's father: On the one hand we have Gunnar (as a child) and his father and, on the other hand, Gunnar (as a grown-up) and his son, Kalli. This adds an interesting diachronic dimension to the project. Also, this opens a window into Kalli's and Gunnar's language acquisition. We observe Gunnar as a grown-up serving as the input of Kalli and Gunnar's father as the input of young Gunnar during the boys' language acquisition. It should be noted that the recordings also feature the boys' mothers talking to them even though the fathers conducted the vast majority of the recordings. The dictionary therefore also includes the mothers' speech.

The goal (pending on more funding) is to transcribe 50 hours—10 from Gunnar's acquisition period and 40 from Kalli's acquisition period. We have currently transcribed approximately 6 hours of video recordings.

4. The making of the dictionary

4.1. Building the corpora

The corpora consist of video recordings transcribed using the multimedia annotation tool ELAN (version 5.4). This tool enables detailed text annotation for video and audio files and is used extensively in the fields of conversation and discourse analysis (see, e.g., Brugman & Russel 2004). For the purposes of this project we focused on the basic text annotation functionality in ELAN, where each annotation contains roughly a single utterance by a speaker, and each speaker has her own specific tier in the annotated data. We also opted for a simple transcription scheme, where notation of detailed discourse features is omitted. A screenshot displaying this is shown in Figure 1.

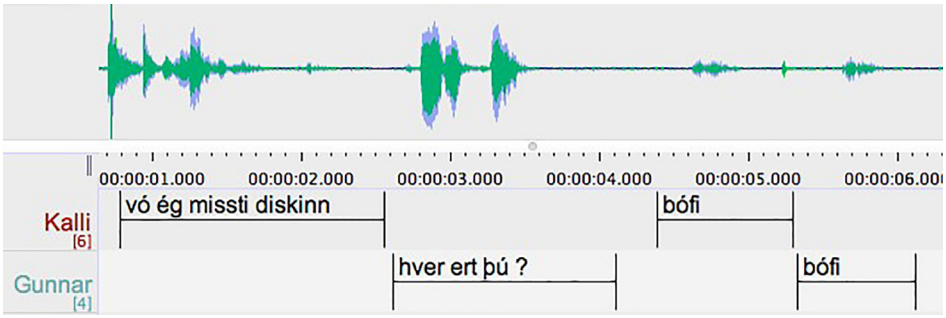

Figure 1: Screenshot from ELAN.

Even though Figure 1 shows only a part of the features ELAN offers, it gives an idea of how the program works. At the top we have a waveform window that displays the sound in the recording at a given time. At the bottom, the speaker-specific tiers are visible, one for each speaker in the recording. At the beginning of this recording (from 2016), Kalli drops a CD he is looking at on the floor and says: *Vó ég missti diskinn* 'Whoa, I dropped the disc'. In the accompanying video recording—which can be played through ELAN in real time during the transcription—Kalli is dressed up with something covering his head and his father, Gunnar, asks him: *Hver ert þú?* 'Who are you?' which Kalli replies to: *Bófi* 'A thug'. Gunnar then repeats Kalli's reply. In this recording, Kalli's and Gunnar's speech does not overlap, i.e., they do not speak at the same time, which the tier-specific annotations clearly capture.

As the goal is to build a dictionary of child-directed speech, we need young Gunnar's and Kalli's input. The transcriptions in ELAN are saved as ELAN Annotation Format (EAF) files, a variant of the XML file schema. When working with these transcriptions we use the Pympi package (Lubbers & Torreira 2018) for Python (and additional scripts written in the project) to extract all utterances made by the adult speakers. This is relatively straightforward, due to the speaker-specific tiers in the annotation data. By counting only the words uttered by young Gunnar's and Kalli's parents, we get a total number of 11,927 words of input in the current data, with 1,910 words uttered by young Gunnar's parents and 10,017 by Kalli's parents.[3]

Once we have extracted the texts from the boys' parents from the ELAN-transcriptions, they need to be processed further before word frequency can be calculated. In our approach we need the lemmas of the words, i.e., the morphologically

[3] This difference in size between the datasets is expected, as the length of transcribed Kalli recordings is currently greater than that of transcribed Gunnar recordings. As mentioned in Section 3, the total length of the recordings of Kalli is much greater than of Gunnar (40 vs. 10 hours), so this relative difference can be expected to remain throughout the project as the data is expanded upon with further transcriptions.

neutral form of each word. To achieve this, the text is first sent to an automatic Part-of-Speech (PoS) tagger. As Icelandic is a language with rich inflection, this greatly eases the lemmatization process, especially regarding morphological ambiguity between word forms. Several freely available PoS taggers exist that can tag Icelandic texts with good results, such as IceTagger (Loftsson 2008), IceStagger (Loftsson & Östling 2013) and, most recently, ABLtagger (Steingrímsson et al. 2019), a neural network-based tagger. We opted for using ABLtagger, as it has been reported to give the best accuracy in tagging Icelandic texts. For the lemmatization itself, we used Nefnir (Daðason 2018), which has been shown to be a very reliable lemmatizer, given that the source text is accurately PoS tagged (Ingólfsdóttir et al. 2019).

To illustrate the output of the lemmatization step, Table 1 shows the PoS tagged and lemmatized output for the sentence in (1):

(1) langar þig að halda á þessu
want you to hold on this
'Do you want to carry this?'

langar	sfg3en	langa
þig	fp2eo	þú
að	cn	að
halda	sng	halda
á	aþ	á
þessu	faheþ	þessi

Table 1: PoS tagged and lemmatized text.

In Table 1 we see three columns. The first one shows the original text, the second column shows the PoS tags and the column on the right shows the lemmas. For example, *langar* 'want' is the first word in the sentence and it is tagged as "sfg3en" which stands for verb-indicative-active-third person-singular-present tense. Its lemma is *langa* (the infinitival form of the verb).

4.2. Extracting word frequency data

With the 11,927 lemmatized words at hand, we counted the occurrences of each lemma using a simple Python script,[4] giving us detailed information on the frequency of each word in the dataset. There are various factors that need to be

[4] As previously noted, all scripts written in this project in order to make the dictionary are available online: www.github.com/hinrikur/BKL.

taken into account in this step. There might, for example, be inaccuracies in the tagging and lemmatization step which have an effect on the end word frequency values. We may also assume that there will be errors in the original text itself, as the annotations have not been proofread in any fashion yet and will inevitably contain some typographical errors. The interactions of such errors in each step of the process may also produce wrong lemmas that skew the final word frequency values. These minor issues must eventually be manually checked and corrected, but leaving them aside for now, we can discuss the word frequency data itself.

The objective in creating the word frequency dictionary is to provide data on the frequency of different words in the input Kalli and Gunnar receive, and thus much-needed information on child-directed speech as a whole. Besides this main objective, the dictionary will enable comparison of the input of two Icelandic children roughly 30 years apart from each other, coupled with the children being father and son. As of now the dictionary achieves both of these objectives to an extent. Due to the relatively small size of the current dataset, especially for Gunnar's input, we will not make any decisive claims based on this data yet. We can, however, make broad observations regarding the content of the boys' input. For example, (2) and (3) below show the ten most frequent verbs in Gunnar's and Kalli's input, respectively—starting with the most frequent verb, which is *vera* in both cases.

(2) vera, segja, fara, koma, sjá, ætla, vilja, eiga, gera, taka
be say go come see intend want own do take

(3) vera, fara, gera, koma, eiga, ætla, halda, geta, verða, segja
be go do come own intend think/hold can become/must say

It is apparent from these excerpts from the dictionary that the most frequent verbs in the two inputs are relatively similar, despite the difference in size of the two datasets.

If these datasets indeed reflect the content of each boy's language input, we would expect them to be broadly compatible with more general data on Icelandic. This seems to be the case when we compare the excerpts in (2) and (3) to the one in (4), which shows the ten most frequent verbs in *Íslensk orðtíðnibók* ('Icelandic Frequency Dictionary', IFD; Pind et al. 1991:612), which is based on a corpus of roughly 590,000 word tokens.

(4) vera, hafa, koma, verða, segja, fara, geta, taka, eiga, gera
be have come become/must say go can take own do

From these examples we can infer that despite the relatively small size of our datasets, and even though the text type we are working with—child-directed speech—is different from the text types that the IFD consists of, our word frequency dictionary does describe its source data in an expected, systematic fashion, just like dictionaries based on larger corpora. Therefore, we are confident that our work could be expanded upon, given a larger dataset and furthermore that our production methods produce applicable results.

5. By linguists, for linguists

The dictionary is a part of the project "Í beinan karllegg: Skráning talmáls þriggja ættliða" ('Patrilineal Descent: Transcribing Spoken Language of Three Generations'), in which the focus is language acquisition. Our hope is that the dictionary will be immediately useful in research on Icelandic language acquisition—that is, it is aimed at linguists.

The second author of this paper (alphabetically ordered), who was awarded the grant, is a linguist whose theoretical interests include language acquisition. Furthermore, the first author holds a B.A. degree in general linguistics and is currently an M.A. student in language technology. The dictionary is therefore not only made for linguists doing research, but also made by linguists.

In addition to this we hope that the dictionary, the scripts written in the project and the transcripts, which we aim at making freely available, will become useful in various language technology projects.

6. Current status and future directions

We now have a corpus of 22,698 words, of which almost 12,000 are child-directed speech. Much more transcribing is needed, however, as we have only transcribed approximately 6 out of 50 hours of video recordings. The word frequency dictionary (the most recent version at each time) and the scripts written specifically in order to make the word frequency dictionary are available online, at www.github.com/hinrikur/BKL. We also aim at making all the transcriptions available online, at www.github.com/einarfs.

The Gunnar Corpus and the Kalli Corpus provide new Icelandic acquisition data, including child-directed speech. In the near future we also wish to gather more recordings, old and new, from more speakers. Hopefully, this is just the start.

References

de Boer, Bart (2012): Infant-directed speech and language evolution. In: Maggie Tallerman & Kathleen R. Gibson (eds.): *The Oxford Handbook of Language Evolution*. Oxford: Oxford University Press, 322–327.

Brugman, Hennie & Albert Russel (2004): Annotating Multimedia / Multi-modal resources with ELAN. In: *Proceedings of LREC 2004.*

Daðason, Jón Friðrik (2018): Nefnir. <www.github.com/jonfd/nefnir>.

Einarsdóttir, Jóhanna Thelma (2018): *Gagnabanki Jóhönnu Einarsdóttur um málsýni* (GJEUM). [Einarsdóttir Corpus.]

Einarsdóttir, Jóhanna Thelma, Anna Lísa Pétursdóttir & Íris Dögg Rúnarsdóttir (2019): *Tíðni orða í tali barna.* Reykjavík: Háskólaútgáfan.

ELAN (Version 5.4) [Computer software] (2018): Nijmegen: Max Planck Institute for Psycholinguistics. <tla.mpi.nl/tools/tla-tools/elan>.

Gíslason, Indriði, Randa Mulford & Ásgeir S. Björnsson (1983): Upp vek þú málið mitt. In: Sigurjón Björnsson (ed.): *Athöfn og orð. Afmælisrit helgað Matthíasi Jónassyni áttræðum*. Reykjavík: Mál og menning, 107–114.

Ingólfsdóttir, Svanhvít Lilja, Hrafn Loftsson, Jón Friðrik Daðason & Kristín Bjarnadóttir (2019): Nefnir: A high accuracy lemmatizer for Icelandic. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics.* Linköping University Electronic Press, 310–315.

Jónsson, Jóhannes Gísli & Þórhallur Eyþórsson (2003): Breytingar á frumlagsfalli í íslensku. In: *Íslenskt mál* 25, 7–40.

Loftsson, Hrafn (2008): Tagging Icelandic text: A linguistic rule-based approach. In: *Nordic Journal of Linguistics* 31, 47–72.

Loftsson, Hrafn & Robert Östling (2013): Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In: *Proceedings of NODALIDA 2013*. Linköping University Electronic Press, 105–119.

Lubbers, Mart & Francisco Torreira (2018): *pympi-ling: a Python module for processing ELAN's EAF and Praat's TextGrid annotation files*. (Version 1.69.) Retrieved from: <pypi.python.org/pypi/pympi-ling>.

Nowenstein, Iris Edda (2014): *Tilbrigði í frumlagsfalli á máltökuskeiði. Þágufallshneigð og innri breytileiki.* M.A. thesis, University of Iceland.

Nowenstein, Iris Edda (2017): Determining the nature of intra-speaker subject case variation. In: Höskuldur Thráinsson, Caroline Heycock, Hjalmar P. Petersen & Zakaris Svabo Hansen (eds.): *Syntactic variation in Insular Scandinavian*. Amsterdam: John Benjamins, 91–112.

Pálsdóttir, Margrét (1979): *Máltaka barns*. B.Ed. thesis, Kennaraháskóli Íslands.

Pind, Jörgen (ed.), Friðrik Magnússon & Stefán Briem (1991): *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.

Sigurðsson, Einar Freyr & Hlíf Árnadóttir (2019): *The Kalli Corpus*.
Sigurðsson, Einar Freyr, Iris Edda Nowenstein & Sigríður Sigurjónsdóttir (2018): *The acquisition of dative subjects in L1 Icelandic*. Talk given at GALANA 8, Bloomington, September 28th, 2018.
Sigurðsson, Sigurður Júlíus (2019): *The Gunnar Corpus*.
Sigurjónsdóttir, Sigríður (2000): *The Eva Corpus*.
Sigurjónsdóttir, Sigríður (2007): *The Fía Corpus*.
Steingrímsson, Steinþór, Örvar Kárason & Hrafn Loftsson (2019): Augmenting a BiLSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In: *Proceedings of Recent Advances in Natural Language Processing*, 1161–1168.
Strömqvist, Sven, Hrafnhildur Ragnarsdóttir et al. (1995): The Inter-Nordic Study of Language Acquisition. In: *Nordic Journal of Linguistics* 18, 3–29.
Svavarsdóttir, Ásta. (1982): „Þágufallssýki.“ Breytingar á fallnotkun í frumlagssæti ópersónulegra setninga. In: *Íslenskt mál* 4, 19–62.
Yang, Charles (2016): *The Price of Linguistic Productivity. How Children Learn to Break the Rules of Language.* Cambridge, MA: The MIT Press.

Hinrik Hafsteinsson
M.A. student
University of Iceland
Sæmundargata 2
IS-102 Reykjavík
hih43@hi.is

Einar Freyr Sigurðsson
research lecturer, Ph.D.
Árni Magnússon Institute for Icelandic Studies
Laugavegur 13
IS-101 Reykjavík
einar.freyr.sigurdsson@arnastofnun.is