Developing Flashcards for Learning Icelandic

Xindan Xu
University of Iceland
Reykjavík, Iceland
xindanxu@hi.is

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

Abstract

This paper describes the process of developing flashcards for the most frequently used words in Icelandic. The process involves utilising currently available open-source online databases, the Tagged Icelandic Corpus, MÍM, and the Database of Modern Icelandic Inflection, BÍN, to extract a list of the most frequently used words, their part-of-speech tags, and inflectional forms. This was combined with newly developed language technology tools for Icelandic to generate phonetic and audio transcriptions of the words. The final product is a combination of printable flashcards and digital flashcards which are easily accessible through smart devices.

1 Introduction

Flashcards are a useful tool for learning. They are frequently used for memorising new words when learning a new language. When combined with spaced repetition, they can produce long-term knowledge retention.

In this project, we created a deck of flashcards that consists of the 4,000 most frequently used words in Icelandic. On the front side of each flashcard, a word is shown along with a sample sentence. On the back of each flashcard, more detailed information about the word is shown, including the following: its English translation, essential morpho-syntactic information (e.g. word class and gender, if applicable), the phonetic transcription, dialectal variation (if applicable), and selected inflectional forms.

The production of this flashcard dataset was made possible due to the recent developments in language technologies for Icelandic. Twenty years ago, this project would have to be carried out manually because Icelandic language technology resources were almost non-existent (Rögnvaldsson et al., 2009). Since 2000, a lot of effort and financial support have been put into developing language technologies for Icelandic. This included building online corpora of texts and sound files, e.g. the Tagged Icelandic Corpus MÍM (Helgadóttir et al., 2012), online dictionaries, e.g. The Database of Modern Icelandic Inflection BÍN (Bjarnadóttir, 2012), and basic tools for natural language processing, e.g. IceTagger (Loftsson, 2008) and Lemmald (Ingason et al., 2008).

By utilising these resources, we have compiled a novel dataset that contains a rich variety of information for selected words. This information was incorporated into flashcards to create a more detailed and effective learning material. We developed two versions of the flashcards: a printable pdf-version and a digital Anki-version that supports media files and is available on multiple platforms. Both versions of the flashcards will be accessible to the public without charge, and the dataset will be published under an open-source license (CC BY 4.0).

2 Flashcards for vocabulary learning

Vocabulary learning is a fundamental aspect of second language acquisition and lasts throughout the learning process. Vocabulary learning involves two scopes: vocabulary size and depth of vocabulary knowledge (Schmitt, 2008). Without sufficient vocabulary size, understanding input and producing satisfactory output in a second language can be frustrating for learners. Furthermore, a lexical item is learned not only by making a form-

Xindan Xu and Anton Karl Ingason 2021. Developing flashcards for learning Icelandic. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 55–61.

meaning connection, but also by understanding how it is used in context (Schmitt, 2008).

Flashcards are a learning tool that facilitates the acquisition of vocabulary. Through the use of high frequency words of a second language, flashcards can help acquire sufficient vocabulary size more effectively. Flashcards can also provide lexical items with context, as well as additional information that aids the depth of vocabulary knowledge, for example, word class, pronunciation and inflectional forms. Furthermore, flashcards can incorporate spaced repetition learning that can produce long-term knowledge retention of the vocabulary. Studies have shown that spaced repetition is one of the most effective learning techniques (Dunlosky et al., 2013; Kang, 2016). This is a learning technique that allows initial study and subsequent reviews to be spaced out over time, and that new and more difficult material is reviewed more often than well-known and easy material.

3 Source of material

Vocabulary and associated morphological information was extracted from two main sources: the Tagged Icelandic Corpus, MÍM (Helgadóttir et al., 2012), and the Database of Mordern Icelandic Inflection, BÍN (Bjarnadóttir, 2012).

3.1 MÍM corpus

The Tagged Icelandic Corpus (hereafter referred to as MÍM) contains approximately 25 million tokens collected from contemporary Icelandic texts during the period 2006–2010. The texts are selected from a variety of sources, including published books, newspapers, Icelandic parliament speeches, legal texts, and student essays. These texts are considered to be representative of the Icelandic society's language usage. The texts are morphosyntactically tagged, lemmatized, and formatted into XML-documents defined by TEI (Text Encoding Initiative). This makes it possible to extract a variety of useful information from the corpus. In this study, we extracted the frequency of headwords and their part-of-speech tags, as well as sample sentences for the selected headwords.

The corpus was tagged and lemmatized automatically using software *IceNLP* (Loftsson, 2019). The accuracy of morphosyntactic tagging was estimated to be 88.1%–95.1% depending on text type (Loftsson et al., 2010). The accuracy of lemmatization was estimated to be approximately 90%. The corpus is available through a special user license.[1]

An example of entries for the headword *ár* (e. *year*) in the MÍM corpus is shown in Listing 1. The inflectional form of the headword is shown between <w> and </w>: *árum* and *ára*. Type shows the POS-tag used for the inflectional form, i.e. "nhfþ" for *árum* and "nhfe" for *ára*.[2]

```
<w lemma="ár" type="nhfþ">árum</w>
 <w lemma="ár" type="nhfe">ára</w>
```

Listing 1: Example from the MÍM Corpus

The first character in the tag always shows the word class, e.g. "n" for "nafnorð" (e. *noun*), "s" for "sagnorð" (e. *verb*). The number of characters used in the tag depends on the word class. In this case, "árum" in the first entry was tagged: noun, neutral, plural and dative, whilst "ára" in the second entry was tagged: noun, neutral, plural and genitive.

3.2 BÍN corpus

The Database of Modern Icelandic Inflection (hereafter referred to as BÍN) consists of more than 270,000 headwords with approximately 5.8 million inflectional forms. Language technology data from the database are distributed under a CC BY-SA 4.0 license and are available at `https://bin.arnastofnun.is/DMII/`. The basic version of the database, Sigrún's format, was used in the development of the flashcards. The data consists of 6 fields: lemma, id, word class, semantic fields, inflectional form, and grammatical tag (see example of *ár* in Figure 1).

4 Data processing

A Python script was used to parse XML-documents and count the frequency of occurrence for each pair of lemma and the first two characters of the tag in the MÍM corpus. The resulting dataset was cleaned and expanded upon by comparison with the BÍN corpus.

Unnecessary tokens in the resulting dataset (e.g. symbols and roman numbers) were filtered out by comparing all the entries with the headword entries in the BÍN corpus. Subsequently, since the MÍM corpus was tagged and lemmatized automatically, it was necessary to double-check the extracted tags

[1] See `http://www.malfong.is/files/userlicense_mim_download_en.pdf`.

[2] See the full list of tagsets used in MÍM corpus: `http://www.malfong.is/files/mim_tagset_files_en.pdf`.

	lemma	ind	cat	bin_tag	word_form	tag
	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>
1	ár	466	hk	alm	ár	NFET
2	ár	466	hk	alm	árið	NFETgr
3	ár	466	hk	alm	ár	ÞFET
4	ár	466	hk	alm	árið	ÞFETgr
5	ár	466	hk	alm	ári	ÞGFET
6	ár	466	hk	alm	árinu	ÞGFETgr
7	ár	466	hk	alm	árs	EFET
8	ár	466	hk	alm	ársins	EFETgr
9	ár	466	hk	alm	ár	NFFT
10	ár	466	hk	alm	árin	NFFTgr
11	ár	466	hk	alm	ár	ÞFFT
12	ár	466	hk	alm	árin	ÞFFTgr
13	ár	466	hk	alm	árum	ÞGFFT
14	ár	466	hk	alm	árunum	ÞGFFTgr
15	ár	466	hk	alm	ára	EFFT
16	ár	466	hk	alm	áranna	EFFTgr

Figure 1: Example of the entry for *ár* in the BÍN corpus.

and make corrections where necessary. For example, prepositions and adverbs share the same tag ("a") in the MÍM corpus, whilst they have separate tags in the BÍN corpus ("fs" for prepositions and "ao" for adverbs). Furthermore, a lemma can be two or more separate words from different word classes. For example, lemma *sig* can be both a neutral noun meaning "subsidence", and a reflexive pronoun referring to oneself. To make sure these instances are tagged correctly, all the headwords and tags extracted from the MÍM corpus were compared against tags in the BÍN corpus. If the tags did not match, the tags from the BÍN corpus were used. Finally, words were ranked by their frequency of occurrence and the top 4,000 were chosen for the project.

As it would not be beneficial to show all the inflectional forms of a headword at once, selected inflectional forms were chosen based on word class. Selected inflectional forms of the chosen words were retrieved from the BÍN corpus. An example entry for the noun *ár* is shown in the Table 1. In this case, the frequency of occurrence of lemma "ár" of class "hk" in the MÍM corpus was 96,849 times. This was ranked 29th amongst all the headwords in the MÍM corpus. Its genitive singular form (EFET) is *árs* and nominative plural form (NFFT) is *ár*.

Lemma	Class	Freq	Rank	W_form	Tag
ár	hk	96,849	29	árs ár	EFET NFFT

Table 1: Example entry for the noun *ár*.

4.1 Phonetic and audio transcription

Phonetic transcriptions of the words were generated using LSTM encoder-decoder sequence-to-sequence models developed by Grammatek ehf. (2021). These models transcribe grapheme to phoneme (g2p) in four pronunciation variants of Icelandic: the standard pronunciation of modern Icelandic, the northern variant (post-aspiration), the southern variant (hv-pronunciation), and the northeast variant (post-aspiration + voiced pronunciation).[3] The R package ipa (Hayes and Alexander, 2020) was used to convert the X-SAMPA phonetic transcription resulting from the g2p models to ipa transcription.

In Icelandic, the pronunciation of a lemma is the same in different word classes. For example, the lemma *tala* can be used as a feminine noun meaning "number, speech", or as a verb meaning "talk, speak". In both instances, pronunciation of the lemma is the same: [tʰa:la]. Accounting for these duplicates, a total of 3,933 unique lemmas (out of 4,000 in total) was used for phonetic transcription.

Audio transcriptions were generated using the Icelandic Dóra voice included in the Amazon Polly text-to-speech service (Amazon Web Services, 2021).

4.2 Translation and sample sentence

Translation of the Icelandic words was carried out semi-automatically. A list of words was translated automatically using the Google Translate web service. However, the translation accuracy turned out to be poor in some cases. Poor translation accuracy mainly occurs when there is minimal difference in written form between two different words. For example, lemma *hár* can be a noun meaning "hair" and an adjective meaning "high". In such cases, Google Translate failed to differentiate the word class and their meanings. Furthermore, Google Translate did not recognise the acute accent in some cases. For example, *dýr* (e. *animal* (no.) and *expensive* (adj.)) and *dyr* (e. *door*) are only distinguished by the acute accent, but they were both translated into "animals" using Google Translate. According to a recent study (Aiken, 2019), Icelandic was among the lowest scoring languages in terms of translation accuracy using Google translate. Therefore, translations were reviewed manually using the Concise Icelandic-English Dictio-

[3] For more information about the regional pronunciation variants of Icelandic, see Rögnvaldsson (2020).

nary (Hólmarsson et al., 2006) as a reference.

The process of selecting sample sentences was also carried out semi-automatically. A python script was used to parse the XML-files from the MÍM corpus and 10 sentences were selected for each headword. Subsequently, sentences were arranged based on their complexity, i.e., length of the sentence and whether there are any uncommon words in the sentence. Finally, the most easily understandable sentence was selected manually for each headword to be shown on the flashcards.

After this step, the data was ready to be used in the production of the flashcards. Table 2 shows a demonstration data-frame with all information excluding the sample sentences and selected inflectional forms.

4.3 Printable and digital flashcards

Both a printable pdf version and a digital version of the flashcards were made in the project. The pdf version of the flashcards was generated using the R package Knitr (Xie, 2021) and the LaTeX-package Flacards (Stuhrmann, 2005). The main difference between the two versions is that the digital version contains audio files of the selected words so that users can listen to their pronunciation; while the physical flashcards contain the phonetic transcriptions in regional variants of Icelandic (if applicable).

Digital flashcards

Digital flashcards were made using the Python library Genanki (Staley, 2021). The script produces an Anki-deck package which can be imported into the Anki-app. Anki is available on multiple platforms and supports different media types in the cards. Another advantage of Anki is the inclusion of spaced repetition, which is considered to be one of the most effective learning techniques (Dunlosky et al., 2013; Kang, 2016).

Basic components of an Anki deck are notes. Each note contains a front (question) and a back (answer) side with information to memorise. The notes in the Genanki library are defined by two components:

1. *models*, which indicate the information to be shown on the card by defining the *fields* and how the card should look like by defining the *templates*.

2. *fields*, which are the actual information to be shown on the card and should correspond to the fields defined by the model.

The difference between the *fields* in the model and the *fields* in the note is that the fields in the model act like a placeholder for the fields of information to be shown, while the fields in the notes are the actual information.

Figure 2 shows an example of the front and back of the Anki flashcard for *ár*. The triangle button which is located next to the phonetic transcription is used to replay the audio of the word. At the bottom of the user interface, the user can choose the interval between repeated viewings. A short interval should be chosen for flashcards that are difficult to memorise so that they are repeated more frequently, whilst a long interval should be chosen for flashcards that are easy to memorise. This process is done to prioritise the flashcards that are harder to learn and thus to improve the overall efficiency of learning. For example, the card would be reviewed immediately by clicking the "again" button, after 1 day by clicking the "Good" button, and after 4 days by clicking "Easy" button. Different interval settings can be selected by the user on their Anki app.

Figure 2: Example of the front and back of the flashcard for *ár* in Anki.

Printable flashcards

Despite all the advantages that Anki offers, some studies also showed that physical flashcards may produce learning outcomes similar to those for digital flashcards (Sage et al., 2020; Nikoopour and Kazemi, 2014). Furthermore, studies have shown that digital flashcards on mobile devices have led to distractions (Sage et al., 2020) and low enjoyment (Hanson and Brown, 2019) amongst students.

lemma	cat	freq	rank	ipa_sd	ipa_north	ipa_northeast	ipa_south	full_cat	eng
vera	so	1,083,582	1	vɛːra	vɛːra	vɛːra	vɛːra	Verb	be
og	st	953,690	2	ɔːɣ	ɔɣ	ɔɣ	ɔɣ	Conjunction	and
í	fs	810,646	3	iː	iː	iː	iː	Preposition	to; in
að	nhm	540,429	5	aːð	aːð	aːð	aːð	Infinitive marker	to
það	pfn	495,273	6	θaːð	θað	θaːð	θaːð	Pronoun	it, that
ekki	ao	209,020	16	ɛhcɪ	ɛhcɪ	ɛhcɪ	ɛhcɪ	Adverb	not
ár	hk	96,849	29	auːr	auːr	auːr	auːr	Noun	year
mikill	lo	75,043	42	mɪːcɪtl̥	mɪːcʰɪtl̥	mɪːcʰɪtl̥	mɪːcʰɪtl̥	Adjective	large, big; much; great
einn	to	50,885	54	eitn̥	eitn̥	eitn̥	eitn̥	Numeral	one
hinn	gr	27,844	94	hɪn	hɪn	hɪn	hɪn	Article	that, the other
nei	uh	6,774	345	neiː	neiː	neiː	neiː	Interjection	no

Table 2: A demonstration data-frame for flashcard production.

The pdf-version of the flashcards is generated by a mother RNW document and eight child RNW documents. The mother RNW document defines the document class *flacards*, reads in the dataset (similar to the one shown in Table 2), and loops through each row to create the respective flashcard. The child RNW documents define different presentations of the cards for different word classes. For example, three inflectional forms were chosen for the word classes *noun* (lemma, genitive singular and nominative plural), *personal pronoun* (lemma, genitive singular and nominative plural), and *verb* (3rd person singular in present tense and past tense, and past participle in neuter singular nominative case). Four child RNW documents were created to accommodate different word classes. Subsequently, four corresponding child RNW documents were created to accommodate the regional pronunciation variants. For each row in the dataset, the mother RNW document selected the child RNW document required to produce the flashcard. For example, the child RNW document for the word class *noun* without pronunciation variant would be selected for the noun *ár*, whilst the child RNW document for adjective with pronunciation variant would be selected for the adjective *mikill* (Figure 3).[4]

The front side of the pdf-version (Figure 3) is the same as the Anki version (Figure 2). On the back side of the pdf-version, regional variants of pronunciation are shown (Figure 3) as opposed to the audio version of the word in the Anki-version (Figure 2). The noun *ár* has the same pronunciation across all regions of Iceland. The adjective *mikill* has regional pronunciation variants in the north and northeast regions of Iceland (Figure 3).

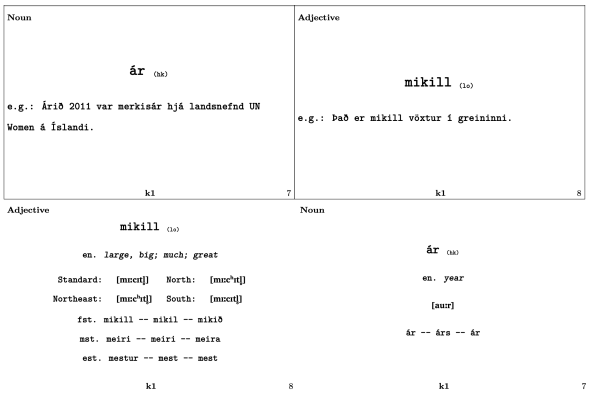

Figure 3: Example of the front and back side of the pdf flashcard for *ár* and *mikill*.

5 Summary and future implementations

In this paper, we have described the process of the production of printable and digital flashcards for the most frequently used words in Icelandic (based on the MÍM corpus). The flashcards dataset will be published under an open-source license which means that it will be freely accessible to the public for use and as a template for further flashcard production.

The flashcards will be useful for anyone who is interested in learning Icelandic, especially at the beginning stage where large quantities of vocabulary need to be acquired. By learning the high frequency words in the language, learners can understand a high percentage of words in common texts such as newspapers and books.

During the production of the flashcards, all steps were carried out automatically except for translation and selecting sample sentences which were both semi-automatic (Table 3). The most time consuming parts are, as expected, the manual steps:

[4]The abbreviations *fst*, *mst* and *est* in Figure 3 refer to *positive degree*, *comparative degree* and *superlative degree* respectively.

double-checking the translation accuracy and selecting sample sentences.

Steps		Efficiency
1	Extract word lists and frequency from MÍM	Automatic
2	Filter out undesirable entries by comparing against lemmas in BÍN corpus	Automatic
3	Extract selected inflectional forms from BÍN	Automatic
4	Phonetic transcription	Automatic
5	Audio transcription	Automatic
6	Translation	Semi-automatic
7	Sample sentences	Semi-automatic
8	Generate printable flashcards	Automatic
9	Generate Anki-flashcards	Automatic

Table 3: Summary steps for the production of flashcards in the project.

A complete list of resources used for the development of the flashcards and their respective licenses are shown in Table 4.

Resource	License
MÍM	Special User License
BÍN	CC BY-SA 4.0 license
g2p-lstm	Apache License 2.0
ipa	MIT | Alexander Rossell Hayes (2020)
Amazon Polly	Creative Commons Attribution-ShareAlike 4.0 International Public License
Genanki	MIT
Knitr	GPL-2 | GPL-3
Flacards	GNU General Public License

Table 4: List of resources used and information about their licences.

In conclusion, we have described the development of a flashcard dataset for leaning Icelandic. The work will serve as a useful template for further development of flashcards as a learning material for Icelandic. For example, a variety of practice decks of the Anki-version can be made so that users can test their learning progress. In Anki, a cloze-deletion field or type-in text field can be implemented into the front of a card. The user's answer will be reviewed automatically and shown in the back (answer) side of the flashcard. This could easily be incorporated into the flashcards so that users can type in the Icelandic words according to the English translation or the phonetic transcription of words with audio display.

Furthermore, the two flashcard decks will serve as a useful resource for the evaluation of flashcards as a learning material, and to ascertain the relative benefits of digital versus physical flashcards for second language learners. We leave that for future work.

Acknowledgements

We would like to thank Anna Björk Nikulásdóttir for the phonetic transcription models, and Atli Jasonarson for writing a script that submitted our word list to Amazon Polly to generate the sound files.

References

Milam Aiken. 2019. An updated evaluation of google translate accuracy. *Studies in Linguistics and Literature*, 3:p253.

Amazon Web Services. 2021. Amazon Polly. `https://aws.amazon.com/polly/`.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection (Beygingarlýsing íslensks nútímamáls). In *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalisation of Less-Resourced Languages"*, pages 13–18.

John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1):4–58. PMID: 26173288.

Grammatek ehf. 2021. g2p-lstm. Phonetic transcription tool. `https://github.com/grammatek/g2p-lstm.git`.

Aroline Hanson and Christina Brown. 2019. Enhancing l2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning*, 33:1–23.

Rossell Hayes and Alexander. 2020. *ipa: convert between phonetic alphabets*. R package version 0.1.0.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 67–72.

Sverrir Hólmarsson, Christopher Sanders, and John Tucker. 2006. Íslensk-ensk orðabók / Concise Icelandic-English Dictionary.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A mixed

method lemmatization algorithm using a hierarchy of linguistic identities (holi). In *Advances in Natural Language Processing*, pages 205–216, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sean H. K. Kang. 2016. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19.

Hrafn Loftsson. 2008. Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

Hrafn Loftsson. 2019. IceNLP natural language processing toolkit. CLARIN-IS, Stofnun Árna Magnússonar.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a pos-tagged corpus using existing tools. In *LREC 2010 Proceedings: Proceedings of the Workshop on "Creation and use of basic lexical resources for less-resourced languages"*.

Jahanbakhsh Nikoopour and Azin Kazemi. 2014. Vocabulary learning through digitized & non-digitized flashcards delivery. *Procedia - Social and Behavioral Sciences*, 98:1366–1373. Proceedings of the International Conference on Current Trends in ELT.

Eiríkur Rögnvaldsson. 2020. *A Short Overview of the Icelandic Sound System Pronunciation Variants and Phonetic Transcription*.

Eiríkur Rögnvaldsson, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna Björk Nikulásdóttir, Matthew Whelpton, and Anton Karl Ingason. 2009. Icelandic language resources and technology: status and prospects. In *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources. Odense, Denmark*.

Kara Sage, Michael Piazzini, IV John Charles Downey, and Sydney Ewing. 2020. Flip it or click it: Equivalent learning of vocabulary from paper, laptop, and smartphone flashcards. *Journal of Educational Technology Systems*, 49(2):145–169.

Norbert Schmitt. 2008. Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3):329–363.

Kerrick Staley. 2021. Genanki 0.10.1. `https://github.com/kerrickstaley/genanki.git`.

Norbert Stuhrmann. 2005. *The flacards class*. CTAN package version 0.1.1b.

Yihui Xie. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.31.