CLARIN Annual Conference Proceedings

2023

Edited by

Krister Lindén, Jyrki Niemi, and Thalassia Kontino

16 – 18 October 2023
Leuven, Belgium

Please cite as:
CLARIN Annual Conference Proceedings, 2023. ISSN 2773-2177 (online).
Eds. Krister Lindén, Jyrki Niemi, and Thalassia Kontino.
Leuven, Belgium, 2023.

A Multilingual Database for Icelandic L2 Flashcards

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

Þórunn Arnardóttir
University of Iceland, Iceland
thar@hi.is

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

Abstract

The IceFlash 4K database is a newly developed multilingual resource for Icelandic vocabulary learning as a second language. It currently contains the 4,000 most frequently used words in Icelandic, with translations in four languages: English, Polish, Chinese and Ukrainian. The IceFlash 4K resource, including the flashcards and developer-friendly raw materials, is published under a CC BY 4.0 license. IceFlash 4K will help learners to learn Icelandic vocabulary more efficiently and it is a useful resource for teachers and language resource developers.

1 Introduction

Studies have shown that intentional learning methods, such as using flashcards for vocabulary learning, are very efficient and can produce great retention of knowledge. In this paper, we present our newly compiled multilingual database for developing flashcards for learning Icelandic, IceFlash 4K. It currently includes the 4,000 most common Icelandic words and their translations into four languages: English, Polish, Chinese and Ukrainian.

This database was used to produce flashcards for Icelandic L2 studies and is currently available in four languages: Icelandic–English, Icelandic–Polish, Icelandic–Chinese and Icelandic–Ukrainian. The original proposal, which was discussed in Xu and Ingason (2021), only included Icelandic–English. Whilst more language versions can be added to the database in the future, these languages were prioritized due to their importance within the Icelandic language learning environment. People of Polish origin comprise the largest immigrant group in Iceland (Hagstofa Íslands, 2021) while people of Ukrainian origin have been immigrating to other countries, following the Russian invasion into Ukraine in February 2022, including to Iceland, so it is important to provide both these groups with language aid to ease their adaptation to the Icelandic society. English is an international communication language and Chinese is one of the most widely spoken languages in the world, and the language which has the greatest number of native speakers. The languages chosen cover a large group of people who do not have access to learning materials. The IceFlash 4K database, along with the produced flashcards, is published under a CC BY 4.0 license and is available at the Icelandic CLARIN repository (Xu et al., 2023) and on GitHub.[1] The flashcards are available as a printable PDF version and a digital Anki version.[2] Currently, the English version of the Anki flashcards has been downloaded more than 1,000 times.

The paper is structured as follows. Section 2 discusses relevant research and resources. Section 3 describes the database and how it was created, while Section 4 details the evaluation process. Finally, we conclude in Section 5.

2 Theoretical background

Vocabulary learning is an important aspect in second language acquisition, as well as when acquiring a native language. Nation (2001, p. 60) found that the vocabulary size of native adults who have English as

[1] https://github.com/antonkarl/iceFlash4K

[2] See https://docs.ankiweb.net/background.html.

their native language is approximately 20,000 words, and that this size is very difficult to attain for those who are learning the language as a second language. He claims that one practical approach would be to learn the most common words first in the language. He believed that with good knowledge of the first 1,000 most common words, individuals could achieve a good understanding of about 81% of the written language and 85% of the spoken language. A knowledge of the first 4,000 words allows individuals to understand about 98% of the written language and 96% of the spoken language. Furthermore, the 98% scope is considered optimal for students to understand a second language without further assistance (Nation, 2001, p. 79).

Unlike vocabulary acquisition by native language learners, which is a gradual process of building up their vocabulary from language acquisition, learning a second language requires the same process but over a much shorter period of time in order to become proficient in a new language. Additionally, a long-term retention of the vocabulary knowledge is needed so that it can be called out immediately if necessary. According to Ebbinghaus (1913), our memory weakens over time, and it happens dramatically after being introduced to new learning materials. He conducted an experiment on himself to see how memory works and why people forget. He explained the process of forgetting by means of a line of data points which later became “Ebbinghaus’s forgetting curve". Other researchers have also experimented with different models to interpret forgetting (see e.g. Murre and Dros (2015)). Ebbinghaus (1913) described that when revisions are repeated and spread over a number of time intervals, the rate of forgetting can be delayed or slowed. This phenomenon is known as spacing effect.

Programs specially designed for digital flashcards generally utilize algorithms based on the concept of the forgetting curve and the spacing effect. One such program is called Anki (Damien Elmes, 2023), in which users can make their own multimedia flashcards. Anki flashcards consist of a question (reviewing material) on the front side and an answer (material to be learned) on the back side. During the reviewing process, both active recall testing and spaced repetition are utilized in the program, so that the learning materials are reviewed with different time intervals based on how well the user has learned them. Research (Nakata, 2016; Nation & Hunston, 2013; Webb et al., 2020) suggests that flashcards, especially when applied with spaced repetition, can be very efficient in enhancing vocabulary learning and creating long-term retention of the vocabulary knowledge.

3 A multilingual database

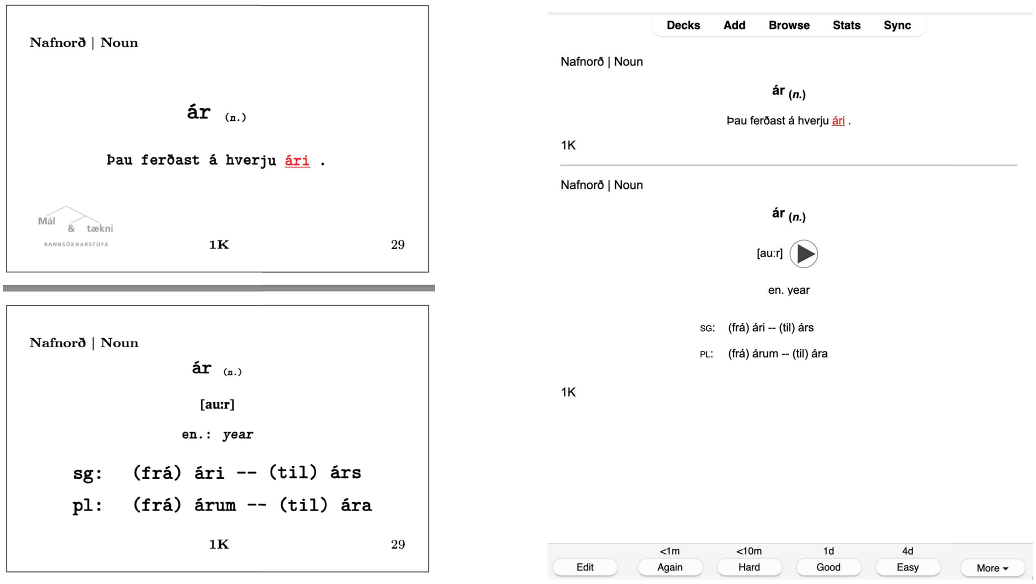
The IceFlash 4K database was designed for producing flashcards for Icelandic L2 studies, consisting of information that is most useful for Icelandic L2 learning. A word frequency list was collected from the Tagged Icelandic Corpus (MÍM; Helgadóttir et al., 2012) and the 4,000 most frequently used words in the corpus were selected as a base vocabulary list for using in the flashcards. The main database is in the form of four tsv files, one for each language version. Each line in the tsv files contains a word and a variety of information about it, including the word category, its frequency in the corpus, a sample sentence which shows the word’s usage in context, the phonetic transcription and the name of the corresponding audio file (which is included with the Anki flashcards), a translation of the most common meaning of the word and selected inflectional forms.[3] Figure 1 shows an example of an Icelandic–English flashcard, both in the PDF version and the desktop version of Anki. The flashcard shows the noun *ár* ’year’ with various pieces of information, including an example sentence and inflectional forms.

Various similar decks exist on AnkiWeb for other languages, but there are always some differences. For example, the most popular deck for French is quite similar[4] but for that deck, only some of the words include sample sentences and the information about inflection is somewhat less detailed.

After the original publication (Xu & Ingason, 2021), in which only an English translation was available, three additional languages were added to the collection: Polish, Chinese and Ukrainian. Furthermore, internal and external reviews have been carried out to evaluate both the content and functionality of the flashcards and the full database has been released using the infrastructure of CLARIN. Results

[3] Note that audio transcriptions of the vocabulary items are not available on the project GitHub page due to the size of the files.

[4] See the French deck on Ankiweb here: https://ankiweb.net/shared/info/893324022.



Nafnorð | Noun

ár (n.)

Þau ferðast á hverju ári .

1K 29

Nafnorð | Noun

ár (n.)

[au:r]

en.: year

sg: (frá) ári -- (til) árs

pl: (frá) árum -- (til) ára

1K 29

Nafnorð | Noun

ár (n.)

Þau ferðast á hverju ári .

1K

Nafnorð | Noun

ár (n.)

[au:r]

en. year

SG: (frá) ári -- (til) árs

PL: (frá) árum -- (til) ára

1K

Figure 1: Flashcards for the noun *ár* 'year' in the PDF version (left) and the desktop version of Anki (right).

from the evaluations will be discussed further in Section 4.

The flashcards are created in nine main steps, apart from further improvements after the evaluations. The majority of the process was carried out automatically. See Xu and Ingason (2021) for a detailed account of data collection, processing and production of the flashcards.

The current flashcards do not contain visual stimuli. It is relatively difficult to find or produce images of certain word categories such as demonstrative pronouns, reflexive pronouns, certain verbs, and abstract nouns. Nevertheless, it would be helpful to add visual stimuli for some cards in future implementations.

4 Initial evaluation

Before the database and the flashcards were released to the public, internal evaluations and external user testing and interviews were carried out to ensure the quality of the resources.

4.1 Evaluation of phonetic transcription, translation and sample sentences

The phonetic transcriptions were originally generated automatically using the g2p-lstm model (Nikulásdóttir, 2020). Upon initial inspection, a few errors were found in the model's output, particularly when it comes to certain pronunciation rules in Icelandic, such as aspiration, preaspiration and *t*-insertion between consonant clusters, in combination with prefixed or compound words. As a result, three resources were combined to produce the final phonetic transcription of the words in the project: the Icelandic Pronunciation Dictionary (Rögnvaldsson, 2015), the Icelandic Pronunciation Dictionary for language technology (Nikulásdóttir, 2021) and an automatic transcription model (Nikulásdóttir, 2020). A total of 526 words were manually reviewed and corrected during this process, with reference to Rögnvaldsson (2020).

Translations of the Icelandic words were carried out in two steps. First, the list of words was automatically translated through the Google Translate web service. The resulting translation accuracy was poor in some cases, see Xu and Ingason (2021) for details. As a result, manual review and translation was carried out on all of the language versions, while translations from Google Translate as well as sample sentences were used as a reference. For the English version, translations were reviewed and corrected manually by the authors and a proofreader. For the Polish and Chinese version, translations from Icelandic were done manually by native speakers, while for the Ukrainian version, a native speaker of Ukrainian, who speaks both English and Polish, manually translated the Icelandic words using English and Polish as intermediate languages.[5]

The sample sentences were collected from the Tagged Icelandic Corpus (MÍM), in which the majority of texts are from published books and online news, comprising approximately 50% of all the texts in the corpus (Helgadóttir et al., 2012). As a result, some sample sentences were considered too difficult for the students who are in the beginner levels of Icelandic studies, especially for the most common words in the first 1,000 tier. This was observed by the teachers of Icelandic as a Second Language in the University of Iceland. In response to this, sample sentences for the first 700 words were completely recreated using short sentences with easily understood vocabulary. This was carried out by a native speaker of Icelandic.

4.2 User testing of Anki flashcards

User testing of the Anki flashcards was carried out prior to the official release of the flashcards. A mini version of the Icelandic–English flashcards was created for this purpose, which consists of 200 words, with a random 5% of each 1,000 words tier. The mini version was consequently sent to volunteered participants, who were asked to use the test flashcards in Anki continuously for two weeks. After two weeks, the participants were asked to report back on their experience using the flashcards through an online form with 5-points likert scale items (Likert, 1932). The online form was completely anonymous and it was not possible to track answers back to individual participants. Apart from some background information, such as age group, native language, length of study of the Icelandic language etc., the form consists of 5 items about the functionality of Anki flashcards and 4 items about participants' Icelandic learning experience.

Survey questions	Strongly disagree (%)	Disagree (%)	Neutral (%)	Agree (%)	Strongly agree (%)
1. I find Icelandic easy to learn.	20	40	30	10	0
2. Knowing Icelandic is important to me in my personal life.	10	10	20	20	40
3. Knowing Icelandic is important to me in my work environment.	10	10	30	10	40
4. I enjoy learning languages through smart devices.	0	0	20	30	50
5. It was very easy to set up the flashcards on my smart device.	0	10	10	40	40
6. I find the flashcards very useful for learning new Icelandic vocabulary.	0	0	40	20	40
7. I find it very useful to have the audio transcriptions of the words.	0	10	0	20	70
8. I find it very useful to have the phonetic transcriptions of the words.	0	10	20	10	60
9. I find the inflectional forms very helpful.	0	11	0	33	56

Table 1: Feedback from initial user testing of Anki flashcards.

A total of 10 volunteered participants finished both testing and reporting on the feedback. Although this is not a big feedback dataset, it gives a general idea of the functionality of the flashcards that we created as well as recommendations on future improvements. Results from this testing is shown in Table 1. The majority of participants gave positive feedback on the different aspects of functionality of the flashcards (see items 5–9). For the Icelandic learning experience, the majority of the participants (60%) considers Icelandic not to be easy to learn (see item 1) and 60% considers that knowing Icelandic is important in their personal life (see item 2). Furthermore, we have received some anecdotal feedback from users who have had a positive experience using the resource.

5 Conclusion

We have presented the IceFlash 4K database, a newly developed multilingual resource for Icelandic vocabulary learning as a second language. It currently contains the 4,000 most frequently used words in Icelandic, with translations in four languages: English, Polish, Chinese and Ukrainian. By learning the

[5] Note that the authors are aware of the disadvantage of this method and an external evaluation will be carried out on the Ukrainian translation of the word list.

high frequency words, learners can understand a large portion of common texts such as newspapers and books. Furthermore, the words were put into four tiers, each containing one thousand words. We believe that this encourages learners to carry on learning and feel the sense of accomplishment when they finish the respective tier. Last but not least, both printable and digital flashcards were made so that learners can choose which format suits them best.

The database is available at the Icelandic CLARIN repository (Xu et al., 2023) under a CC BY 4.0 license. The importance of the currently described database not only lies upon its multilingual application of flashcards, but also its possibility for further language development and applications in other fields of study.

Acknowledgments

We thank the Áslaug Hafliðadóttir Memorial Fund for partially supporting our project, as well as the anonymous reviewers for their comments.

References

Damien Elmes, A. H., AMBOSS MD Inc. (2023, April 5). *Anki* (Version 2.1.57). https://apps.ankiweb.net

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. Teachers College Press. https://doi.org/10.1037/10011-000

Hagstofa Íslands. (2021). Innflytjendur 15,5% íbúa landsins (immigrants are 15.5% of the country's inhabitants) [Retrieved: 2022-08-28]. https://hagstofa.is/utgafur/frettasafn/mannfjoldi/mannfjoldi-eftir-bakgrunni-1-januar-2021/

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., & Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, 67–72.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLOS ONE*, *10*(7), 1–23. https://doi.org/10.1371/journal.pone.0120644

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289. https://doi.org/doi:10.1515/iral-2015-0022

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P., & Hunston, S. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nikulásdóttir, A. B. (2020). Models for automatic g2p for Icelandic (20.10) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/84

Nikulásdóttir, A. B. (2021). Icelandic pronunciation dictionary for language technology 21.10 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/154

Rögnvaldsson, E. (2015). Pronunciation dictionary for Icelandic [Retrieved: 2022-08-22]. http://www.malfong.is/index.php?lang=en&pg=framburdu

Rögnvaldsson, E. (2020). Icelandic pronunciation [CLARIN-IS]. http://hdl.handle.net/20.500.12537/82

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How Effective Are Intentional Vocabulary-Learning Activities? A Meta-Analysis. *The Modern Language Journal*, *104*, 715–738.

Xu, X., & Ingason, A. K. (2021). Developing Flashcards for learning Icelandic. *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, 55–61. https://aclanthology.org/2021.nlp4call-1.5.

Xu, X., Ingason, A. K., Kolka, V. T., Kovalova, A., & Kristínardóttir, I. (2023). Multilingual flashcards with 4,000 most common Icelandic words (IceFlash4K) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/308