Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2311–2316
Marseille, 20-25 June 2022

Samrómur: Crowd-sourcing large amounts of data

Staffan Hedström, David Erik Mollberg, Ragnheiður Þórhallsdóttir, Jón Guðnason
Reykjavik University, Menntavegi 1, 102 Reykjavik
Tiro, Laugavegur 163 Reykjavík
{staffanh, ragnheidurth, jg}@ru.is
{david.mollberg}@tiro.is

Abstract

This contribution describes the collection of a large and diverse corpus for speech recognition and similar tools using crowd-sourced donations. We have built a collection platform inspired by Mozilla Common Voice and specialized it to our needs. We discuss the importance of engaging the community and motivating it to contribute, in our case through competitions. Given the incentive and a platform to easily read in large amounts of utterances, we have observed four cases of speakers freely donating over 10 thousand utterances. We have also seen that women are keener to participate in these events throughout all age groups. Manually verifying a large corpus is a monumental task and we attempt to automatically verify parts of the data using tools like Marosijo and the Montreal Forced Aligner. The method proved helpful, especially for detecting invalid utterances and halving the work needed from crowd-sourced verification.

Keywords: Speech corpora, Icelandic, Crowd Sourcing

1. Introduction

The collection effort is part of the Icelandic national program for language technology (LT), a five-year program launched in October 2019 (Nikulásdóttir et al., 2020). One of the program's goals was to collect a large and diverse collection of Icelandic speech data and make it readily available for use in automatic speech recognition (ASR) and similar fields. The initial goal was to collect 530 thousand utterances (Nikulásdóttir et al., 2017). That goal has been exceeded, and we have collected 1.5 million utterances that we estimate are about 2,250 hours of recorded speech.

Samrómur is the largest prompted speech collection effort for Icelandic so far and verifying the data is just as a monumental task as the collection itself. Therefore the data will be released in batches so that it can become of use as soon as possible. The first corpus to be released from the Samrómur collection was published on OpenSLR[1] and it is also being added to the Linguistic Data Consortium (LDC). That subset contained 100 thousand utterances, or around 114 hours, of manually reviewed utterances for people aged 18 and up.

The next pending subset from the Samrómur Collection will contain 137 thousand utterances, or 131 hours, of speech from children aged between 4-17. That will be the first Icelandic corpus intended for ASR with children's speech. A third corpus containing 17 thousand utterances focused on queries from various sources will also be released on OpenSLR and LDC. As more and more data gets verified, additional coprora will but composed and released.

Another recently released corpus that is part of the LT program is Talrómur (Sigurgeirsson et al., 2021). Although intended for text-to-speech (TTS), the corpus contains 213 hours of high-quality recordings from 8 speakers that have diversity in age, speaking style, dialect, and prosody and can prove useful for ASR and TTS developers alike.

Other notable Icelandic ASR corpora are the "Althingi Parliamentary Speech Corpus" (Helgadóttir et al., 2017), which consists of 542 hours of parliamentary speech with transcripts that have been automatically aligned. Previously, two efforts have been made to collect Icelandic data for speech recognition from the general public. "Icelandic Speech Recognition Project Hjal" (Rögnvaldsson, 2003) which had the primary goal of collecting sufficient material to train a speaker-independent isolated word recognition system, and the Malrómur corpus (Steingrímsson et al., 2017) (Guðnason et al., 2017) published in 2017, which consists of 136 hours of manually evaluated speech utterances with correct transcriptions, similar to Samrómur.

Crowd-sourcing has proved to be an excellent tool for reaching speakers of all ages and genders. In a small nation such as Iceland, which has a rich cultural connection with its language, we have gathered momentum from the standpoint of preserving the language. We chose to engage the community by setting up competitions to achieve this result—two competitions aimed at primary schools and one towards workplaces. The main contribution of this paper is an overview of the collection platform, the results so far and what we have learned from crowd-sourcing data for the last two years.

2. The collection platform

The collection platform is called *Samrómur* [2] and was initially a fork of The Mozilla Common Voice [3] project, which is an open-source platform for crowd-sourcing

[1] https://www.openslr.org/112/

[2] www.samromur.is

[3] https://commonvoice.mozilla.org/

the recording of speech utterances. The platform was previously described in a publication in 2020 (Mollberg et al., 2020), but in the summer of 2020, we began developing our own platform. While still being heavily inspired by the Mozilla Common Voice project, this allowed us to customize the platform further and specialize it to fit our purposes. In figure 1 we show the process for contributing voice recordings on *Samrómur*, there we show that after a participant has chosen to contribute, they have to select how much they want to contribute. They are given three options for how many utterances to donate, small-10, medium-20 and large-50. After a participant has completed their contribution amount and submitted their donation, they are thanked and asked if they want to continue contributing, help with other types of donations (e.g. verification). In a previous iteration of the platform, there was only a single option of donating five utterances at a time, we noticed participants got content with their contribution quickly. The addition of the option of larger contribution packages and the possibility of easily being able to continue contributing with just a mouse click encouraged participants to contribute more, especially during competitions. Once the participants have read the the prompts they can listen and review there contribution and re-read prompts if necessary.
The platform is hosted on Amazon Web Services and is using an Elastic Load Balancing system. All metadata is stored in a Relational Database Service using MySQL and all the utterances are stored in a S3 bucket. This setup has allowed us to handle the heavy load that comes during the competitions. All donated utterances are saved in Waveform Audio format with sample rate of at least 16 kHz.

2.1. Collecting children's data

One of the goals we were working towards was collecting data from children. To collect this data, parents or guardians need to give consent for their children so that they can participate. If a participant selects under 18, additional information is displayed when entering their demographic information. The participant needs to enter their national ID number [4] and an email of the parent or guardian. The parent or guardian will then receive a confirmation email where the parent or guardian can click a link in the email to confirm that consent is given. To smooth out the process, especially for competitions, we also developed an API to allow schools to send in a list of National IDs together with guardians email, which can be used to send the confirmation email to all parents automatically. All data collection on Samrómur is GDPR compliant.

2.2. Prompts and Text Processing

The sentences, or prompts, that participants read were scraped from various sources, as well as some synthesized with scripts. Below is the list of used sources.

- The MIM corpus (Helgadóttir et al., 2012)
- The Icelandic Gigaword corpus (Steingrímsson et al., 2018)
- The Icelandic Web of Science
- The Icelandic Wikipedia page
- A variety of novels (freely donated by the authors)
- A list of Icelandic places, towns and cities
- Synthesized sentences using common queries found in chat-bots, Google, a call center

The process for gathering the prompts is well described in (Mollberg et al., 2020). The same process was used to extend the number of prompts available to 379,695. The prompts were sourced and filtered to be appropriate for different age groups shown in the table 1. To remove prompts that could include profanity or inappropriate language for minors, all prompts where searched for words that could be found in an extensive list of bad Icelandic words. Any prompts containing such words were filtered out. To create a diverse corpus the prompts were first presented to the participants in an order to make sure that every prompt has at least 1 utterances. Once that was achieved, we started to collect up to 10 utterances for each prompt.

Age group	Sentence length	Max word length
10 and under	2-8	8
11-15	6-10	17
16 and older	5-15	35

Table 1: Rules for how the prompts were divided for different age groups. At age 16 and up, the user is expected to be a fully proficient reader.

2.3. Competitions

Three different competitions were organized during the collection period from November 2019 to December 2021. The goal of the competitions was to increase the awareness of the platform and encourage public participation, channel the competitive spirit into useful data. Our focus during the marketing of these competitions was to emphasize the importance of being able to use the Icelandic language in our day to day lives with our ever-evolving technology. As with many other languages, Icelandic is constantly getting more and more influence from other languages with interactions with computers, cell phones, smart homes, etc.. Therefore, an essential part in maintaining the language is to provide the material needed for Language Technology like ASR. Although having these highly aspirational goals is helpful, the simplest message of simply winning the competition was often the most effective.
Two of the competitions were aimed at primary schools. We marketed the competitions via Facebook, Twitter and the Icelandic President also helped with

[4] A 10 digit number where the first six digits show the day, month and year of your birth

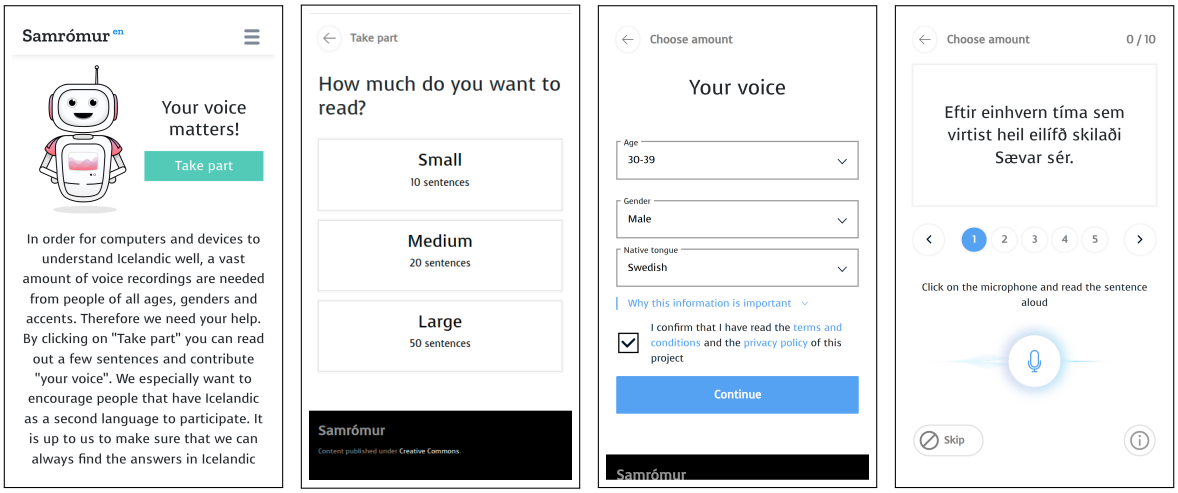

Figure 1: The layout for the recording process with Samrómur. The first image shows where users are prompted to contribute, the second image shows three different contribution sizes for the user to choose (small - 10, medium - 20, large - 50). The third picture shows the page where participants are asked to input demographic information. The fourth image shows where participants are prompted to read out-loud a sentence from the script. Note: the platforms standard language is in Icelandic but it is also available in English which is displayed here.

televised announcements. Direct contact was made with the schools to allow them to prepare and encourage classes to participate. While aimed at schools, everyone was allowed to participate, parents, friends, relatives. All participants needed to choose a school to contribute from during the competition periods. Each competition ran for a week at a time.

The last competition was a collaboration with one of Iceland's largest telephone companies, Síminn and was set as a competition between companies and institutions in Iceland. The company assisted greatly with ads online, on television, at bus stops and created marketing material with locally famous actors, influencers and public figures. The main focus of the ad campaign was to "Save the language". Workplaces had a week to sign up and then the competition ran for a week. The schools/workplaces were split into size categories, and the top ones in each was promised a prize.

3. Verification

Because of the nature of crowd-sourced data, some sort of verification is needed to verify that the utterances align with their prompt. For this purpose, users on the platform can also choose to assist with verifying utterances. Verifiers are presented with a text prompt and the corresponding recording. They must listen to the utterance and give it a positive vote if the text and audio match and a negative vote if not. For an utterance to be classified as valid two positive votes from different users are required. Likewise, two negative votes classify an utterance to be invalid. Users well versed in verification (research personnel, trained students) were given user accounts with super votes. A single super vote was enough to categorize an utterance as valid or invalid. Even though the participation from the public was in the range of tens of thousands of votes, those vote pale in comparison to the gathered data. Therefore, concentrated manual efforts and automatic verification were implemented.

3.1. Automatic verification

Hiring workers to verify a large corpus is both expensive and time consuming. Therefore, we attempted to verify part of the data automatically and to give us a guide for manual verification. A Kaldi based forced alignment tool developed by Language and Voice lab in the University of Reykjavik, commonly called Marosijo (Guðnason et al., 2017) was used to verify the utterances. It outputs a score for each utterance from 0.0 to 1.0. The lower the score is, the more likely a recording is to be bad. The higher it is, then it is more likely to be good. High scoring utterances had a considerable amount of false positives (high scores despite not being good utterances) of around 20%. These utterances were mostly good but often had the start or the end of the audio missing or included mispronunciations. Low scoring utterances had a similar percentage of false negatives (low scores despite being good). These utterances were good but were sometimes scoring low because of loud background noises (which were deemed good for ASR training). Attempts were made to reduce the number of false positives/negatives by using the Montreal Forced Aligner (MFA) tool (McAuliffe et al., 2017). These attempts allowed us to reduce false positives in the 0.01-0.3 range where the utterance could not be aligned with the MFA. In light of the risk of false positives/negatives, it was decided to apply either super or normal votes depending on their score and the false positive/negative rate as seen table 2 below. The result of these votes in the corpus can be seen in table 4. A full documentation and tools on how to apply the automatic verification process

Score range	MFA	Vote type
0.90-1.00		Positive
0.301-0.899		-
0.01-0.3	aligned	Negative
0.01-0.3	unalignable	Negative Super
0		Negative Super

Table 2: Ruleset of how to apply the results from Marosijo scores and MFA alignment as votes

Utterances	Speakers
10+	57.1%
20+	43.2%
50+	23.5%
100+	12.6%
1000+	1.1%

Table 3: The percentage of speakers reaching different levels of donated utterances.

described here can be found on our github page [5].

3.2. Manual verification

During summer 2021, 12 students were hired for 10 weeks to manually verify specific sets of recordings. The sets focused on specific age groups, and utterances that were likely to be invalid were filtered out. However, the verification was not their only task, but we estimate that around 400 man-hours were spent verifying. The students were first instructed in what makes an utterance valid or invalid. The students were then given accounts with super votes to increase the yield of their verification.

4. Results

4.1. Competitions

The competitions yielded an astonishing amount of utterances. In the primary first school competition held in May 2020, 144 thousand utterances were collected. The second school competition held in January 2021 resulted in a total of 790 thousand utterances. Finally, the workplace competition held in November 2021 produced 360 thousand utterances. In total 84% of utterances have been collected during competitions. The accumulation of the data is shown in Figure t2.

4.2. Collected data

The data collection has been ongoing for just over two years. The total amount of collected utterances was 1.5 million (ca 2250 hours) from roughly 20 thousand speakers. The exact number of distinct speakers is unknown as the same speaker using different devices is counted twice if they don't log into their user account. On average, each speaker read 70 utterances. In table 3 we see the percentages of speakers reaching different contribution levels. Four incredible speakers donated more than 10 thousand utterances.

In total, 66% of the donations were from female speakers. Despite marketing efforts to even out this difference between the genders, women were more generous in donating their voice as can be seen in figure 3. Children (contributors under the age of 18) contributed 45% of the collection and adults 55%.

[5] https://github.com/cadia-lvl/samromur-tools/tree/master/QualityCheckPostProcess

4.3. Verification of data

In total 451,861 utterances have been completely verified. Of those, 296,590 utterances are valid (the utterance match the prompt), which means that the split between valid and invalid utterances is 65.6% valid and 34.4% invalid.

4.3.1. Automatic

The automatic verification processed 759 thousand utterances and generated over 500 thousand votes. The split and resulting validations are shown in Table 4.

	Amount	Result
Positive votes	435,550	4,070 validated
Negative votes	15,386	109 invalidated
Negative super votes	60,363	60,354 invalidated

Table 4: The votes generated with automatic verification and their resulting valid-/invalidation of utterances

4.3.2. Manual

For 10 weeks the 12 students verified a total of 192,819 utterances during 400 man-hours. Yielding in 128,827 valid utterances and 63,992 invalid utterances.

5. Discussion

During the first competition it was only possible to contribute 5 utterances at a time therefore we introduced the user selected contribution amounts. Being able to contribute 50 at a time smoothed out the contribution process a lot for the speakers contributing a lot of utterances and we believe that explains the results of table 3. The competitive element introduces some downsides. There were incidents of cheating where students would not read the prompt properly to quickly add more to their schools score. We also observed a lot of students honestly trying to read the prompt well, but in their haste to contribute to their schools score, mispronounced words or stopped the recording before completing the prompt, we believe that this mostly explains the high rate of invalid utterances. And since the schools got very enthusiastic we've also seen gender and age biases as seen in figure 3. We have to keep in mind though, that error rate of 34% might not be representative for the entire collection, the automatic verification process only fully invalidated utterances but indicated that many of the now unverified utterances should be valid so this number might improve over time

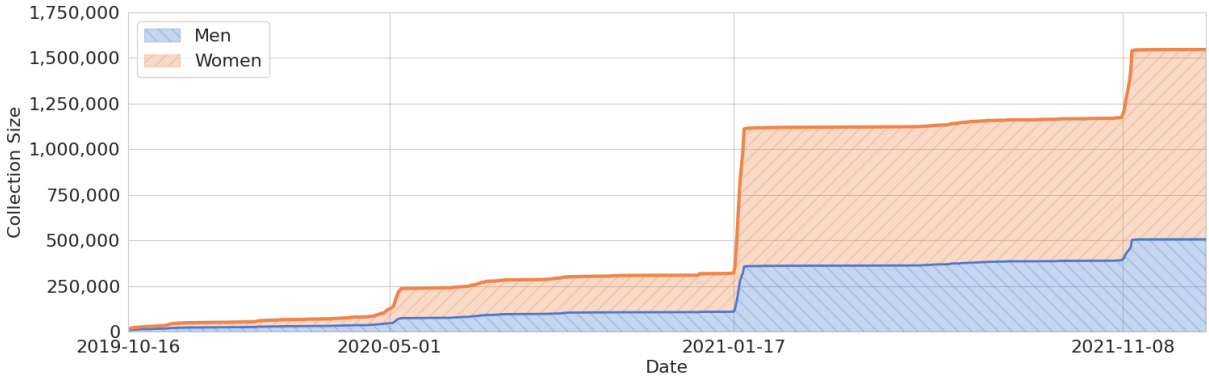

Figure 2: Contribution amounts per day. Heavy increases can be seen on the dates of the competitions.

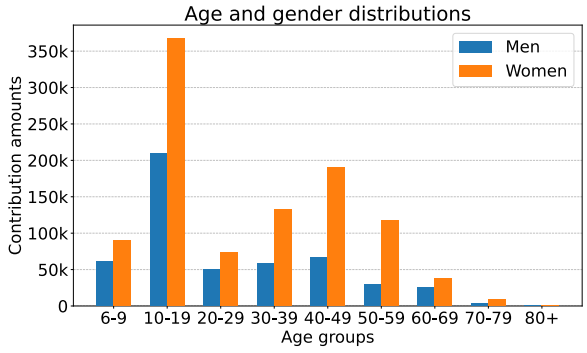

Figure 3: The age and gender distribution for the collection.

when more utterances are verified manually. We have also observed that the second two competitions yielded a lot more data than the first, we speculate that is because the Samrómur brand became more established and familiar to the community and that the first competition really succeeded in increasing awareness and public interest as well as improvements made on the platform itself.

6. Conclusions and further work

The initial goal of the competitions was to increase awareness of the platform and thereby collect more utterances. While that might have been successful, the competitions themselves stood for the largest part of collected utterances by a large margin. Channeling the competitive spirit and the promise of prizes yielded huge successes for this collection and considering that 84% of the 1.5 million donated utterances come from competitions proves that it is a highly effective way of gathering a large amount of data, with the right marketing and social media presence. While the competitions were successful in the collection of utterances getting crowd-sourced verification was more challenging, and considering the higher error rate during the competitions we would not recommend organizing competitions in verification. The automatic verification process used here was very useful, but even though it generated a lot of votes, many utterances still need manual verification to get usable data. Hiring staff to manually verify utterances was expensive but after training gives trustworthy results. In the future we need to evaluate the best way of verifying large amounts of data, either through hired staff, alternative automatic verification or how to increase the crowd-sourced contributions. There are considerations to release unverified data as this might still be useful for ASR research and development, this would allow for collections of this type to release large corpora quicker and using the Marosijo scores, users would get some indication on the dependability of the utterances.

7. Bibliographical References

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gudnason, J. (2020). Samrómur: Crowd-sourcing data collection for Icelandic speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3463–3467, Marseille, France, May. European Language Resources Association.

Nikulásdóttir, A. B., Guðnason, J., and Steingrímsson, S. (2017). *Language Technology for Icelandic 2018-2022: Project Plan*. Mennta og menningarmálaráðuneytið, Reykjavik, Iceland.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, LREC 2020, Marseille, France.

8. Language Resource References

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. (2017). Building asr cor-

pora using eyra. In *INTERSPEECH*, pages 2173–2177.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Helgadóttir, S., Svavarsdóttir, A., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages — SaLTMiL 8 – AfLaT*, Istanbul, Turkey.

Rögnvaldsson, E. (2003). The icelandic speech recognition project hjal. *Nordisk Sprogteknologi. Årbog*, pages 239–242.

Sigurgeirsson, A., Gunnarsson, Þ., Örnólfsson, G., Magnúsdóttir, E., Þórhallsdóttir, R., Jónsson, S., and Guðnason, J. (2021). Talrómur: A large Icelandic TTS corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 440–444, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Steingrímsson, S., Guðnason, J., Helgadóttir, S., and Rögnvaldsson, E. (2017). Málrómur: A manually verified corpus of recorded icelandic speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 237–240.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.