Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, Jón Guðnason
Reykjavik University
Menntavegi 1, 102 Reykjavik
{de14, olafurhj, sunnevath, steinthor18, eydishm, jg}@ru.is

Abstract

This contribution describes an ongoing project of speech data collection, using the web application Samrómur which is built upon Common Voice, Mozilla Foundation's web platform for open-source voice collection. The goal of the project is to build a large-scale speech corpus for Automatic Speech Recognition (ASR) for Icelandic. Upon completion, Samrómur will be the largest open speech corpus for Icelandic collected from the public domain. We discuss the methods used for the crowd-sourcing effort and show the importance of marketing and good media coverage when launching a crowd-sourcing campaign. Preliminary results exceed our expectations, and in one month we collected data that we had estimated would take three months to obtain. Furthermore, our initial dataset of around 45 thousand utterances has good demographic coverage, is gender-balanced and with proper age distribution. We also report on the task of validating the recordings, which we have not promoted, but have had numerous hours invested by volunteers.

Keywords: Speech corpora, Icelandic, Crowd-sourcing

1. Introduction

We describe a work in progress aimed at collecting speech data employing crowd-sourcing. The work started in mid-2019 and a crowd-sourcing platform was launched in fall the same year. It has been very well received and by the end of the first month we had surpassed the data collection goals set for the first three months of the project.

Rapid development in recent years in voice-controlled consumer products has made the apparent stagnation of such solutions for Icelandic all the more noticeable.

This project started out as an effort to translate the Common Voice platform[1] to Icelandic, in collaboration with Deloitte Iceland, a private company. The effort soon got incorporated into a larger program, a new national program for language technology (LT), a five-year national LT program launched in October (Nikulásdóttir et al., 2020). It is funded by the government, managed by Almannarómur, an NGO, and executed by a consortium of universities, research institutes and private enterprise. From the start all facets of this crowd-sourcing effort to collect speech data have been collaborative, which has been a vital part of its success so far.

A report (Nikulásdóttir et al., 2017), written at the preparation stage for the national LT program, describes necessary actions and requirements for building a well working LT infrastructure in Iceland and sets out goals for various core projects of the program. One of these core projects concerns building better speech recognition systems. To this end, goals are set for collecting speech data for training. The goals proposed in the report include collecting 200,000 utterances from adult speakers, 200,000 utterances from youth speakers and 100,000 utterances from individuals with Icelandic as their second language. For the first milestone of our project, set three months after the launch of the collection effort, we set out to collect 30,000 utterances of adult speakers. By utilizing a crowd-sourcing effort and a well thought-out marketing scheme, we achieved that goal one month into the effort. This work has two main contributions. It shows that a carefully structured and well-publicized crowd-sourcing campaign can obtain great results. Furthermore, it gives us a large dataset for training speech recognition models that will continue to grow while the crowd-sourcing website remains up and running.

[1] https://voice.mozilla.org/en

2. Previous Speech Corpora in Icelandic

Previously, two efforts have been made to collect Icelandic data for speech recognition from the general public. The first was the *Hjal* project, which launched in 2003. The goal of *Hjal* was to collect sufficient material to train a speaker-independent isolated word recognition system. Recordings were collected by recruiting volunteers to call a certain phone number and repeat words or phrases. Almost 3,000 people participated, resulting in over 90,000 audio files, which were then transcribed (Rögnvaldsson, 2004).

The second effort was instigated by Google, when the company decided to add Icelandic to its list of languages with speech recognition support. In cooperation with Reykjavik University and the Icelandic Centre for Language Technology, the *Almannarómur* data collection project was launched (Guðnason et al., 2012). Data was collected by asking people to donate their voice by reading prompts into Android G1 phones. They were asked to read for as long as they could, up to a maximum of 30 minutes and/or 350 utterances. This data was subsequently manually inspected by evaluators listening to all the segments, resulting in a speech corpus totaling in 136 hours of correctly transcribed speech segments (Steingrímsson et al., 2017). Another speech corpus available in Icelandic, albeit built using other methods, is the Althingi Parliamentary Speech Corpus (Helgadóttir et al., 2017). In the Icelandic parliament, Althingi, all performed speeches are transcribed manually and published as text on Althingi's web page. While working on building an automatic speech recognition system specifically for the uses of Althingi, a speech corpus suitable for the training the parliamentary Automatic Speech Recognition (ASR) system was build. Text and audio data

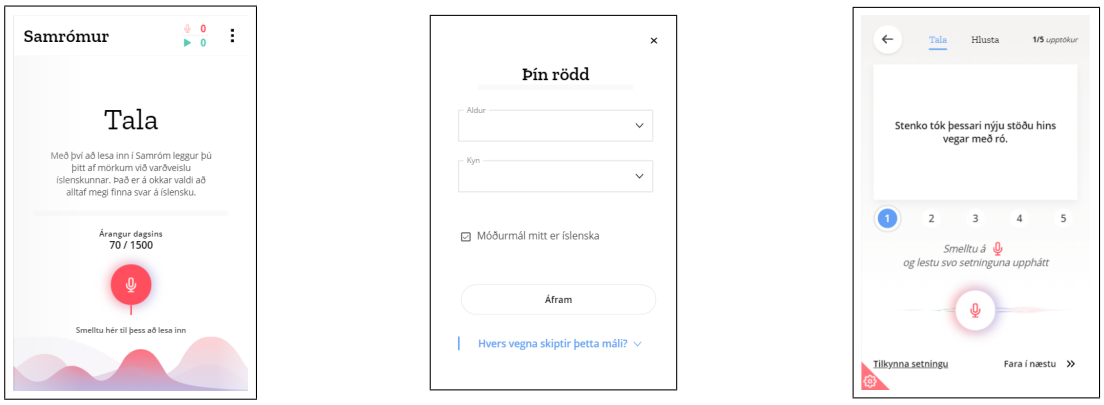

Figure 1: The layout for the recording process with Samrómur. The first image show where users are prompted to participate by donating their voice, the second image shows the card where participants are asked to input demographic information and the third image shows where participants are prompted to read out-loud a sentence from the script.

of manually transcribed speeches were processed to build an aligned, segmented corpus. This resulted in a corpus of 542 hours of speech.

All these three corpora are distributed under an open CC BY 4.0 license on CLARIN[2].

3. Speech Collection Platforms

A few open-source platforms for building ASR corpora have been published, such as the one introduced by Hughes et al. (2010). More mobile collection platforms such as Eyra (Guðnason et al., 2017; Petursson et al., 2016), aimed at enabling under-resourced languages to collect ASR data, and Woefzela (De Vries et al., 2011), aimed for collections in the developing world, have also been reported. These platforms are built for collections on mobile phones with a laborious commitment needed from collectors. Some solutions have introduced apps that require installations by the participants (Leemann et al., 2018) which is a step that might repell participants to commit to a crowd-sourcing collection.

As stated on the official web site, the Common Voice data collection initiative by Mozilla was set up with the aim to *teach machines how real people speak*. The Common Voice platform can be accessed on most browsers on desktop, through mobile with some browsers and through an app for unsupported browsers. It includes quality control through a validation task performed by the participants. This makes true crowd-sourcing possible and easily accessible without time-consuming manual collections and ensures quality control of the data.

Language	Total hours	No. of speakers	Gender	
			Male	Female
English	1,087	39,577	47%	11%
German	350	5,007	68%	10%
French	184	3,005	70%	9%

Table 1: The three languages with the most collected utterances using the Common Voice initiative.

The Common Voice initiative publishes information on the status of collections for the languages using the platform. In Table 1 the languages with the most data collected are listed. These are much bigger speech communities than Iceland. However, our initial results indicate we might well reach that list.

4. Collection Settings

To meet the data requirements needed to build ASR solutions, crowd-sourcing can be an ideal method. It allows for easy collection of vast amounts of voice recordings from numerous different speakers. It may also have its drawbacks, however, a common one being the over-representation of young individuals, as indicated by our initial collection, seen in Figure 4, and (Leemann et al., 2018).

4.1. The Platform

The initiative as introduced in Iceland was given the name *Samrómur*. It is built upon The Mozilla Common Voice project, which is an open-source platform for crowd-sourcing the recording of speech utterances. Using an open, ready-made tool that has been tried and tested is both cost-efficient and helps us avoid possible design pitfalls that might hamper use and thus result in fewer voice samples collected. When implementing our collection platform, we made some adaptations to Common Voice, most notably by simplifying the interface; we also ask users for their demographic information more decisively. Layout examples are shown in Figure 1. We initially direct all users to reading prompts for recording, instead of having equal emphasis on recording and validation. This may result in a lag in validation, but as seen in Figure 2, it has turned out that a considerable amount of validation has been finished. Extensive work was done to obtain full platform compatibility, a greater emphasis was put on collecting demographic information from participants and the recording format was changed from MP3 to WAV. The code for *Samrómur* is available on GitHub[3].

An unknown factor in the process was the traffic due to simultaneous recordings during peak hours, such as right after the launch. Therefore, the website is hosted on European AWS EC2 instances that are optimized for scalability and the data is stored on AWS S3.

[2] https://clarin.is/

[3] https://github.com/aime-island/raddvefur

A considerable amount of work went into making the consent form for participants, in order to make sure the collection was in line with the General Data Protection Regulation (GDPR). All individuals under the age of 18 are required to have consent from a guardian to participate for GDPR compliance (European Commission, 2018).

4.2. Using Samrómur

When participants enter the *Samrómur* website, as seen in the first image in Figure 1, they can choose to speak and record utterances or listen to and verify utterances recorded by other participants. When participants choose to record an utterance into Samrómur for the first time, they are asked to provide optional demographic information: gender, age and their native language. Then they are prompted to read five random sentences from the script, one at a time. They can review and re-record any or all of the sentences, stop the recording process at any time or report a sentence. After completing five sentences, participants are optionally asked to review the recordings. Finally, they are prompted to give their consent to the terms of the project before submitting. Upon finishing, they are prompted to optionally record more utterances. The validation procedure is very similar; participants get five utterances at a time to listen to and have to verify that the utterance is valid by pressing a thumbs-up button, or a thumbs-down button if the recording is invalid. Short instructions of what makes a good recording appear with every recording.

4.3. Validation

For an utterance to be considered valid, two separate participants have to listen to the recording and verify that it matches the given sentence. Likewise, if two separate participants mark the sentence as unfit, it will be marked as invalid. All recordings, also the invalid ones, are stored and marked appropriately.

4.4. Script and Text Processing

While there are two large text corpora available for Icelandic, published under permissive licenses, the MIM corpus (Helgadóttir et al., 2012) and the Icelandic Gigaword Corpus (Steingrímsson et al., 2018), it is not straightforward to adapt them for use in this kind of data collection. The prompts have to be grammatically correct and they cannot be overly complicated, to ensure they will not be problematic for the readers. As the two corpora mentioned above are composed of a variety of texts, choosing a random sample from them would often result in prompts that are not appropriate for our purposes. For this reason, and to try to please and entertain our users in order to retain them, we gathered sentences from sources we deemed interesting, the Icelandic Web of Science[4], the Icelandic part of Wikipedia and a variety of novels that were freely donated by the writers. Furthermore, we generated a collection of sentences, by inserting Icelandic names, to a list of common queries, both to a call center that answers questions about telephone numbers and what kind of service is provided by various companies, and to a generic chatbot that crawls news texts for answers. The corpus contained roughly a hundred thousand sentences, but is to be expanded as needed. Each sentence is only read once.

Allowed characters in the script were limited to the Icelandic alphabet, and a few symbols, numbers and abbreviations were expanded as best possible. In the collection, the minimum length for a sentence is three words and the maximum fourteen words. Random samples from the script were reviewed but each of the prompts shown to users was not manually verified. The code for the text preprocessing is available on GitHub[5].

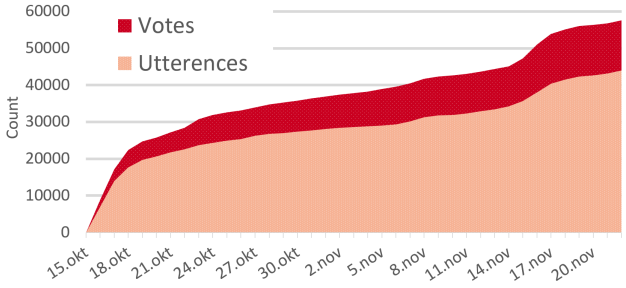

Figure 2: The cumulative count of votes and utterances. Each utterance can have more than one vote as it needs two positive votes to be considered valid and two negative votes to be considered invalid.

4.5. Launch

Samrómur launched at an event that contained short lectures on the status in Icelandic LT and upcoming projects. A number of high-profile individuals were asked to formally open the platform by recording the first sentences in the data collection. Amongst them were the President of Iceland and the Minister of Education, Science and Culture. This generated a substantial amount of media coverage and contributed to creating a positive connection between the public and the data collection effort. A month later, another publicity launch was performed on Icelandic Language Day, when literary writers were asked to read into *Samrómur* while on a live broadcast.

5. Results

We present the results of the data collection for the first phase, spanning the period from launch on October 16, 2019 to November 22, 2019. During this period, two publicity events and some smaller pushes were instigated to boost participation.

5.1. First collection period

Good publicity is an essential part of a successful crowdsourcing collection. Figure 3 gives an example of this as the impact of the first launch obviously gave great results for the first day but diminishes dramatically and is down to a few utterances per day after the third day. Evidently, the launch was still present in the minds of people, since one tweet was all it took to boost the collection noticeably in the first days of November. The third publicity push,

[4] https://www.visindavefur.is/

[5] https://github.com/aime-island/text-extractor

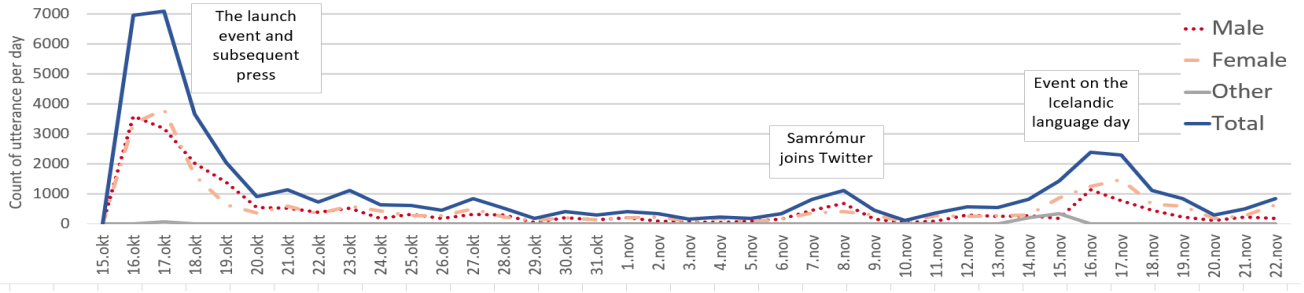

Figure 3: The count of utterances recorded each day from first launch until November 22, 2019.

during Icelandic Language Day, gives a further peak in the data collection.

The overall results for the first period of collections are listed in Table 2. All aspects of the collection exceeded our highest expectations. The number of contributed utterances, validated recordings and number of participants give us great hope for the continuation of the project. The table shows that the participants preferred the task of validation to that of donating speech (note that for each validated utterance, two validations are required). That is perhaps understandable since the task of listening and judging recordings from other participants is much less effort than speaking out loud. This however suits our purposes well, since we need more validations than recorded utterances.

As for the results of the validations, 53% of the recordings have been validated, out of those, 87.45% have been approved by two validators and 12.55% have been rejected by two validators. From the rejected it still remains a possibility to go over the pile to receive an even higher acceptance of utterances.

Speech		**Validation**	
Utterances	45,667	Validated	24,326
Total hours	71	Total hours	40
Speakers	5,718	Validators	3,917
Av. pr. speaker	7.98	Av. pr. validator	14.7

Table 2: Collection results from the speech collection on the left and validation of speech recordings on the right.

5.2. Participant Demographics

It is optional for users to provide demographic information when participating in the project. Nonetheless, in the first phase of the project, 96% of participants have given their demographic information.

Figure 4 gives insight into the demographics, as represented by the the information provided by users. The balance of male and female participants has been close to even from the start of the effort. This is very different from the collections for other languages reported in the Common Voice project, where all of them show much less participation from females as opposed to males. For example, in the English collection, the ratio of participants confirming they are male is 47%, whereas only 11% of participants confirm they are female, see Table 1. At the same time, the reporting of demographic information is much lower compared to the *Samrómur* collection, which may lead to the conclusion that females are more reluctant to give their demographic information. Nevertheless, given these numbers, almost all of the 42% not providing information on gender in the English collection would need to be female to match our numbers. The number of contributions per

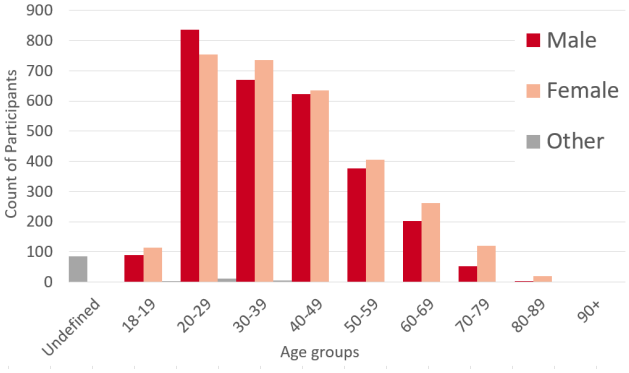

Figure 4: Age groups and gender of all participants in the collection

person is largely one or two rounds of five recordings, i.e. 93% of participants contribute 10 or fewer sentences to the collection. As the data we collect indicates the number of devices used for recording rather than the number of users, the individuals behind this percentage might be even fewer than the number in Table 3.

Utterances	**Speakers**
5 or fewer	4,908
6 to 10	411
11 to 15	136
16 to 20	76
21 to 25	40
26 to 30	33
31 to 40	29
41 or more	85

Table 3: Number of utterances read by participants in increments of five for the period from the first launch October 16, 2019 until November 22, 2019. The average donation per speaker is roughly eight utterances.

6. Availability

There are, however, a number of participants that have contributed vast amounts of utterances. Eighty-one users have

recorded 41 utterances or more. Out of these, the highest score from one individual is 1,242 utterances, with 670 utterances for the second place and 563 for the third. Hopefully more participants will be willing to imitate the effort of these superusers in the next stages of the collection.

The speech corpus is published on the Icelandic LT website *Málföng* under a permissive license (CC BY 4.0). The recordings are made available for download as recorded, in 16-bit, 48kHz WAV-format, accompanied by relevant metadata: hash key that connects the recording to the speaker, duration of the recording, age, gender and native language of speaker if available, prompt text and the type of recording (validated, invalidated, graveyard). They will also be uploaded to a CLARIN repository and integrated into the CLARIN Virtual Language Observatory[6].

The first version of the corpus will be released in March of 2020 and then periodically while the collection effort is ongoing. Version numbers will constitute of year and month of release.

7. Conclusion and further work

The Icelandic LT program is split into multiple milestones and the first one has been reached ahead of schedule, so preliminary work for the next milestones has begun. For the data collected so far we are particularly pleased with the gender balance amongst participants. But presumably we get disproportionately many samples from young adults, which may be because the engagement with devices used for the collection is less prominent amongst older people.

For this first phase of the collection reported here, participation is limited to individuals over the age of 18. In the following stages a collection aimed at minors under the age of 18 will be initiated in collaboration with primary and comprehensive schools, where the plan is to have students participate, e.g. during Icelandic language classes. Special efforts will be made to reach underrepresented groups, such as individuals with Icelandic as their second language. This effort is particularly important so that the ASR solutions based on the data will not discriminate against this growing group of the Icelandic population.

After monitoring the setup of the web so far, there is one significant change that would have been useful to include from the beginning. At the moment, the average user donates roughly eight utterances, which is a rather low count, and any measures to increase that count would be very valuable. In an effort to remedy this, the first reading session will be the same as before but upon completing that, the participants will be prompted with a choice of ex. 15, 20 or 30 utterances in the next session, instead of only 5. This might have a positive anchoring bias effect.

8. Bibliographical References

De Vries, N. J., Badenhorst, J., Davel, M. H., Barnard, E., and De Waal, A. (2011). Woefzela - An Open-Source Platform for ASR Data Collection in the Developing World. In *Proceedings of Interspeech*, Interspeech 2011, Florence, Italy.

European Commission. (2018). General data protection regulation (GDPR) – official legal text. Web page. `https://gdpr-info.eu/`.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. (2012). Almannarómur: An Open Icelandic Speech Corpus. In *Proceedings of the Third International Workshop on Spoken Language Technologies for Under-Resourced Languages*, SLTU 2012, Cape Town, South Africa.

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. (2017). Building ASR Corpora Using Eyra. In *Proceedings of Interspeech 2017*, Stockholm, Sweden.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an ASR Corpus Using Althingi's Parliamentary Speeches. In *Proceedings of Interspeech 2017*, Stockholm, Sweden.

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.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., and LeBeau, M. (2010). Building Transcribed Speech Corpora Quickly and Cheaply for Many Languages. In *Proceedings of Interspeech 2010*, Makuhari, Chiba, Japan.

Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*, 5:1–17.

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

Petursson, M., Klüpfel, S., and Gudnason, J. (2016). Eyra – speech data acquisition system for many languages. *Procedia Computer Science*, 81:53–60.

Rögnvaldsson, E. (2004). The Icelandic Speech Recognition Project *Hjal*. pages 239–242. Museum Tusculanums Forlag, University of Copenhagen, Denmark.

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, Gothenburg, Sweden, May. Association for Computational Linguistics.

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.

[6] `https://www.clarin.eu/content/virtual-language-observatory-vlo`