Samrómur: Crowd-sourcing large amounts of data

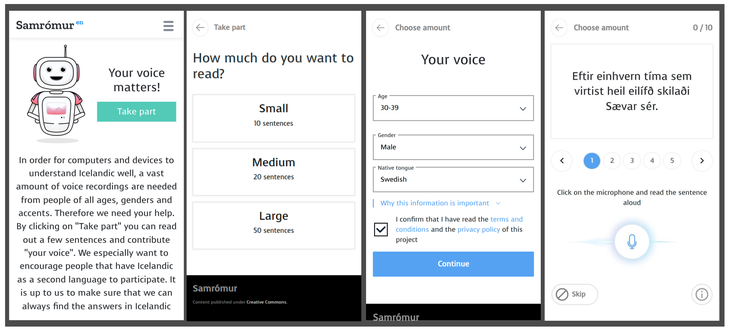
Staffan Hedström, David Erik Mollberg, Ragnheiður Þórhallsdóttir, Jón Guðnason
Reykjvík University

INTRODUCTION

Samrómur is platform for crowd-sourcing utterances for automatic speech recognition (ASR) data. The platform is inspired by mozilla common voice.

We have collected **2250 hours** of data from 20 thousand different speakers.

Three different ASR data sets has been released since the beginning of the project 2019.



METHOD

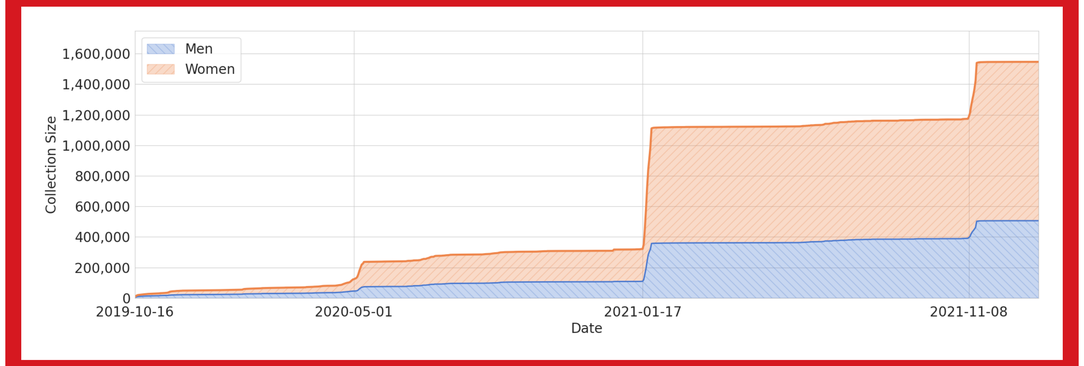
To engage the community, we have had an active social media presence organized three competitions. Our focus during the marketing of these competitions was to emphasize the importance of being able to use the Icelandic language in our day to day lives with our ever-evolving technology.

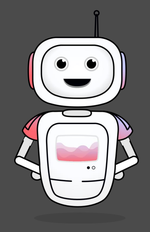
Two competitions were aimed towards primary schools. We had direct contact with teachers of the larger schools and the **president of Iceland** also helped with public announcements. The third competition was a competition between companies and institutions in Iceland.

Automatic verification: Two aligner tools, Marosijo and Montreal Forced Aligner, were used to process a subset of the data. The results produced a score for each utterance which was used as a basis to verify each utterance as invalid or potentially valid.

Engaging the community is the key to crowd-sourcing resources for low resource languages!

Through competitions we collected over **790 thousand** utterances in a week.





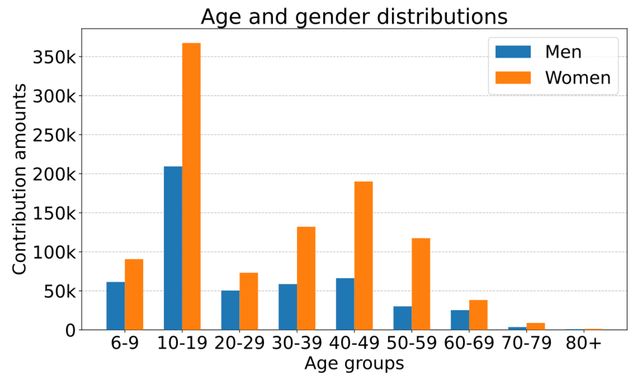
Platform: samromur.is
Email: samromur@ru.is
Paper: samromur.is/lrec2022paper

RESULTS

The competitions encouraged users to contribute many utterances. Over 10% donated over 100 utterances.

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

Women donates more than men, 66% vs 33%.



Manual verification: 12 students spent 400 man-hours verifying a total of 192,819 utterances. Yielding in 128,827 valid utterances and 63,992 invalid utterances.

Automatic verification: Out of 759,000 utterances processed, 60,354 utterances were determined to be invalid and 435,550 to be potentially valid utterances.

DISCUSSION

Competitions can be a great tool to engaging the community in crowd-sourced efforts. Competitions have downsides such as cheating, overly-excited contributions and so forth which lead to a high amounts of unusable utterances.

Manual efforts to verify the data is expensive and time consuming. The automatic verification process used here is good for removing invalid utterances and identifying potentially valid utterances.