

**Samrómur Milljón:
An ASR Corpus of One Million
Verified Read Prompts in Icelandic**

By Carlos Daniel Hernández Mena

samromur.is

The Samrómur Platform

Þín rödd skiptir máli!

Taka þátt

Til þess að tölvur og tæki skilji íslensku svo vel sé þá þarf mikinn fjölda upptaka af íslensku tali frá allskonar fólki. Þess vegna þurfum við þína aðstoð, með því að smella á „Taka þátt“ þá getur þú lesið upp nokkrar setningar og lagt „þína rödd“ af mörkum. Við viljum sérstaklega hvetja fólk sem hefur íslensku sem annað mál að taka þátt. Það er á okkar valdi að alltaf megi finna svar á íslensku.

Samrómur hófst í október 2019 og hingað til hafa um **28** þúsund manns lesið rúmlega **4156** klukkustundir eða **2.855.450** setningar. Hægt er að lesa meira um verkefnið hér. [Lesu meira hér.](#)

Lesnar setningar síðastliðinn mánuð

Dagur	Lesnar setningar
18. september	2853 þús.
19. september	2854 þús.
20. september	2854 þús.
21. september	2855 þús.
22. september	2855 þús.
23. september	2855 þús.
24. september	2856 þús.
25. september	2856 þús.
26. september	2856 þús.
27. september	2856 þús.
28. september	2856 þús.
29. september	2856 þús.
30. september	2856 þús.
1. október	2856 þús.

Language technology programme for Icelandic 2019-2023

L2 Speakers with Samrómur



**Computer-assisted pronunciation training
in Icelandic (CAPTinI): developing a method
for quantifying mispronunciation in L2 speech**

Catlin Richter¹, Branislav Bédi²,
Ragnar Pálsson³, and Jón Guðnason⁴

The CAPTinI Project

Samrómur Milljón

Samrómur Unverified 22.07

Samromur Unverified 22.07



“ Please use the following text to cite this item or export to a predefined format:

BIBTEX

CMDI

Hedström, Staffan; et al., 2022, *Samromur Unverified 22.07*, CLARIN-IS, <http://hdl.handle.net/20.500.12537/265>.



Clarín IS Repository

✎ Authors

[Hedström, Staffan](#) ; et al.

▼ show everyone

[Hedström, Staffan](#) ; [Fong, Judy Y.](#) ; [Þórhallsdóttir, Ragnheiður](#) ; [Mollberg, David Erik](#) ; [Guðmundsson, Smári Freyr](#) ; [Jónsson, Ólafur Helgi](#) ; [Þorsteinsdóttir, Sunneva](#) ; [Magnúsdóttir, Eydís Huld](#) ; [Gudnason, Jon](#)

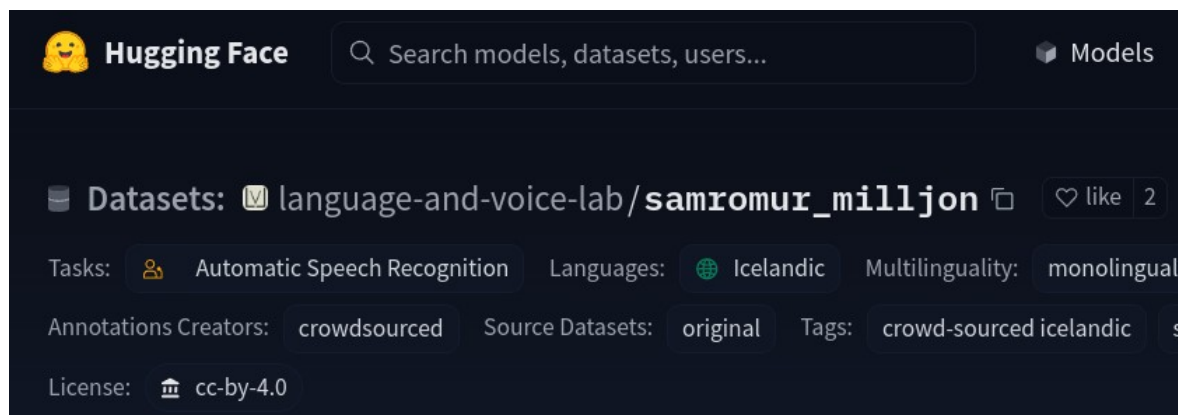
➔ Item identifier

<http://hdl.handle.net/20.500.12537/265>



2,159,314 (2,233 hours) speech recordings in Icelandic that are essentially unverified!

Corpus Characteristics

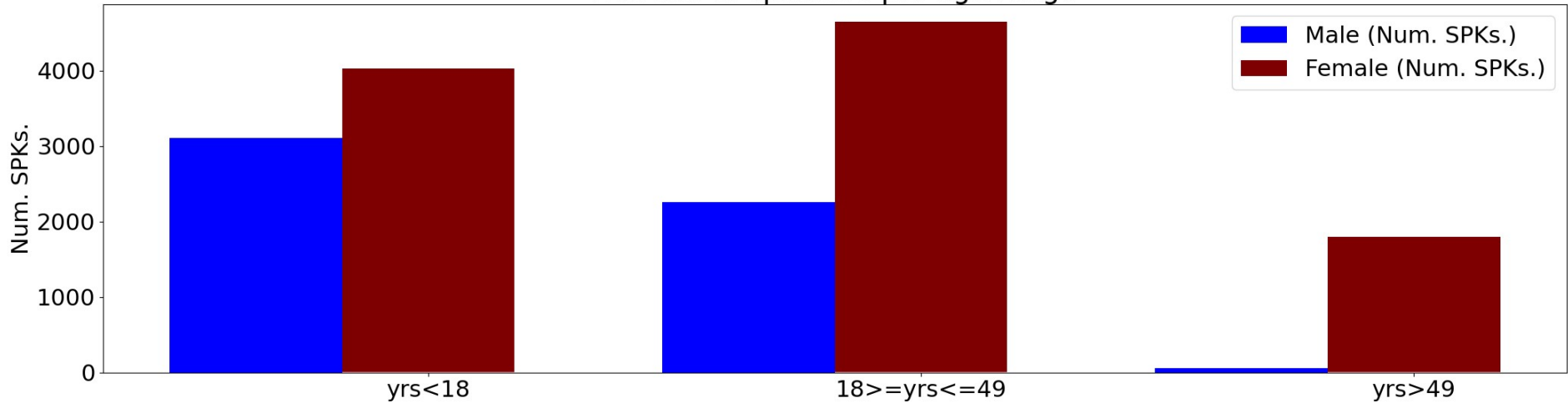


- 1,002,157 speech recordings
- 967 hours
- 16,604 unique speakers
- Range of age 4 to 79 years old
- Verified with NeMo, Wav2Vec2, Whisper and Faster-Whisper.

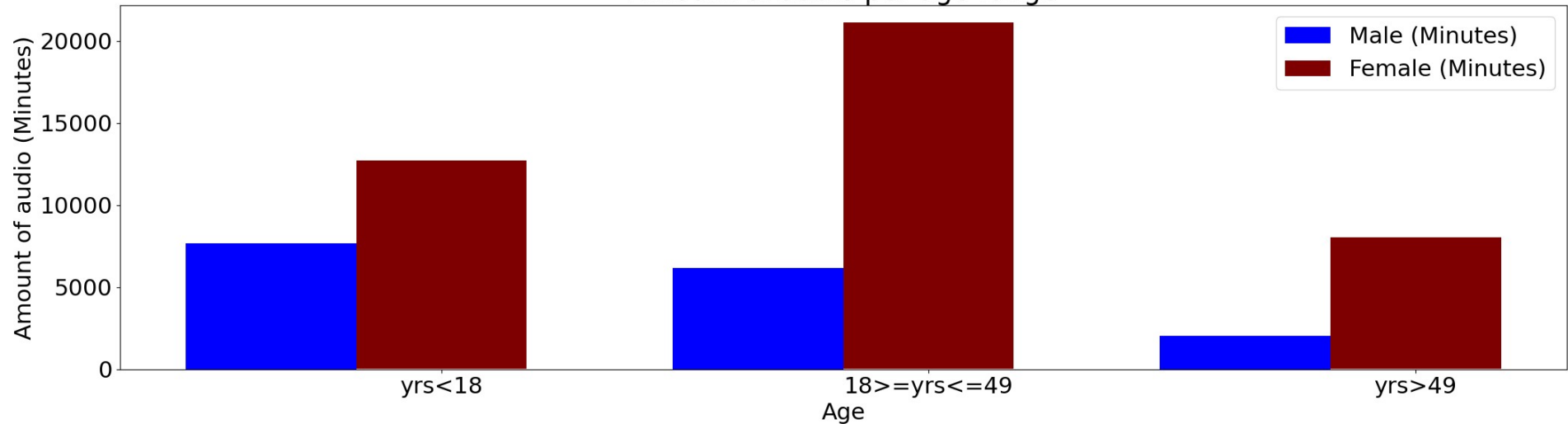
Gender	Female	Male	UNK
Duration	697h22m	264h28m	5h16m
Utterances	714,564	282,499	5,094
Speakers	10,447	5,948	209

Ranges of Age

Number of speakers per age range



Amount of audio per age range



The Verification

Data for the verification models

language-and-voice-lab/althingi_asr

Viewer • Updated Feb 24

514 hours

NeMo

language-and-voice-lab/malromur_asr

Viewer • Updated Feb 24

119 hours

Wav2Vec2

language-and-voice-lab/samromur_asr

Viewer • Updated Feb 24 • ↓ 6

114 hours

Whisper

language-and-voice-lab/samromur_children

Viewer • Updated 1 day ago • ↓ 85 • ♥ 1

127 hours



Faster-Whisper

• L2-Speakers Data (125h55m) **Unpublished material**

125 hours

875 + 125 hours of training data in a GPU Tesla A100 !

Verification with NeMo



 carlosdanielhernandezmena / [stt_is_quartznet15x5_ft_ep56_875h](#) 

- [Samrómur 21.05 \(114h34m\)](#).
- [Samrómur Children \(127h25m\)](#).
- [Malrómur \(119hh03m\)](#).
- [Althingi Parliamentary Speech \(514h29m\)](#).

875 hours

- Trained with 875 hours
- 2 million recordings in 7 hours
- No language model used
- **348,295 Perfect Matches**

Verification with Wav2Vec2

 carlosdanielhernandezmena/[wav2vec2-large-xlsr-53-icelandic-ep10-1000h](#) 

- [Samrómur 21.05 \(114h34m\)](#)
- [Samrómur Children \(127h25m\)](#)
- [Malrómur \(119hh03m\)](#)
- [Althingi Parliamentary Speech \(514h29m\)](#)
- L2-Speakers Data (125h55m) **Unpublished material**

1,000 hours

- Trained with 1,000 hours
- 2 million recordings in 7 days
- No language model used
- **1,002,218 Perfect Matches**

Verification with Whisper

📄 language-and-voice-lab/whisper-large-icelandic-30k-steps-1000h 📄

- Samrómur 21.05 (114h34m)
- Samrómur Children (127h25m)
- Malrómur (119hh03m)
- Althingi Parliamentary Speech (514h29m)
- L2-Speakers Data (125h55m) Unpublished material

1,000 hours

- Trained with 1,000 hours
- 2 million recordings in 36 days
- **Corrupted copy of the model!**
- 138,992 transcriptions in 2.5 days.
- 18,839 Perfect Matches!

This issue motivated the creation of the Faster-Whisper Model

Verification with Faster-Whisper

```
language-and-voice-lab/whisper-large-icelandic-30k-steps-1000h-ct2
```

- Samrómur 21.05 (114h34m)
- Samrómur Children (127h25m)
- Malrómur (119hh03m)
- Althingi Parliamentary Speech (514h29m)
- L2-Speakers Data (125h55m) **Unpublished material**

1,000 hours

This model was created from the Whisper model shown in the previous slide.

863,220 Perfect Matches!

This model was created with 2 lines of code!

Verification Results

<code>audio</code> audio	<code>speaker_id</code> string	<code>gender</code> string	<code>age</code> string	<code>duration</code> float32	<code>verified_with</code> string	<code>normalized_text</code> string
-----------------------------	-----------------------------------	-------------------------------	----------------------------	----------------------------------	--------------------------------------	--

Sys.	Matches	Percent.
V+N+F	325,713	32.50%
V+N+W	4,449	0.44%
V+F	537,453	53.62%
V+N	18,072	1.80%
V+W	14,390	1.43%
V	102,080	10.18%

- V = Wav2Vec2
- N = NeMo
- W = Whisper
- F = Whisper-Fast

More than 80% of the data was verified by at least 2 systems!

**Models trained
with Samrómur
Milljón**

The Wav2Vec2 Model trained with Samrómur Milljón

[language-and-voice-lab/wav2vec2-large-xlsr-53-icelandic-ep30-967h](#)

Dataset	V. M.	S. M.
Samrómur (Test)	9.847%	7.698%
Samrómur (Dev)	8.736%	6.786%
Samrómur Children (Test)	9.391%	6.467%
Samrómur Children (Dev)	6.055%	4.234%
Malrómur (Test)	5.643%	6.631%
Malrómur (Dev)	6.156%	5.836%
Althingi (Test)	11.437%	17.904%
Althingi (Dev)	11.093%	17.931%

V.M. = Verification Model | S.M. = Samrómur Milljón Model

The Whisper Models trained with Samrómur Milljón

[language-and-voice-lab/whisper-large-icelandic-62640-steps-967h](#)

Dataset	V. M.	S. M.
Samrómur (Test)	8.479%	7.762%
Samrómur (Dev)	7.299%	7.035%
Samrómur Children (Test)	7.743%	7.047%
Samrómur Children (Dev)	4.591%	4.425%
Malrómur (Test)	5.110%	11.511%
Malrómur (Dev)	5.286%	11.000%
Althingi (Test)	8.250%	16.189%
Althingi (Dev)	7.998%	16.007%

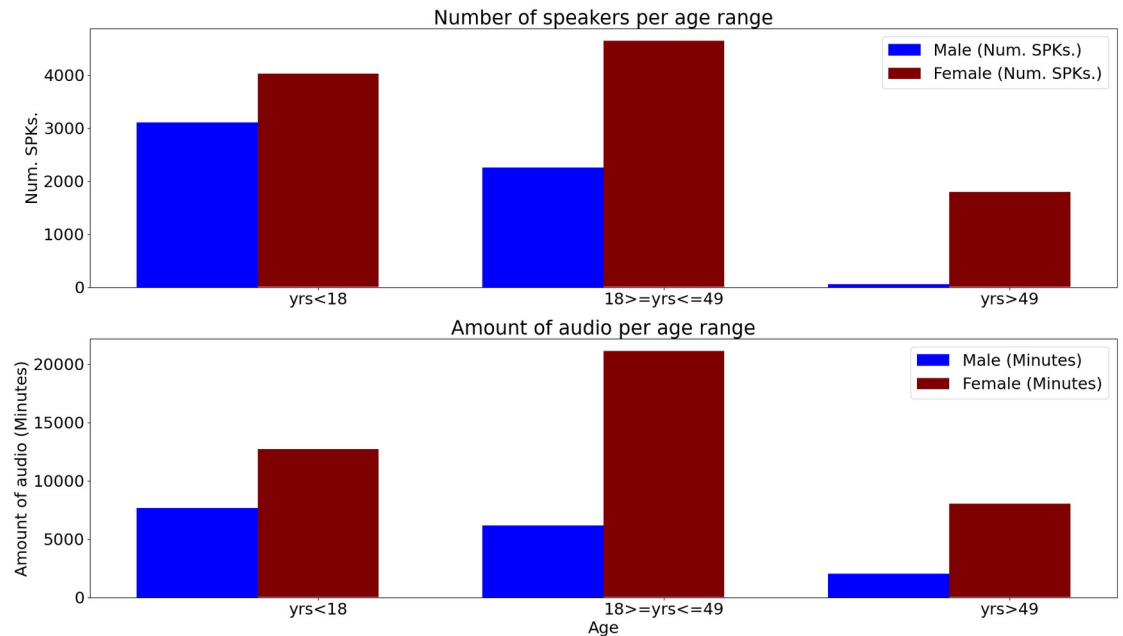
V.M. = Verification Model | S.M. = Samrómur Milljón Model

[language-and-voice-lab/whisper-large-icelandic-62640-steps-967h-ct2](#)

Conclusiones & Further Work

Further Work

- Verify the rest of the data of “Samrómur Unverified 22.07”.
- Organize collection campaigns to balance the data that needs to be balanced.
- Verify other information of the corpus apart from the transcriptions.



Conclusions

However, despite the pending tasks, we strongly think that the work presented in this paper is a relevant contribution to the speech technologies in Icelandic, a language that will not be under represented any more.

Thank you!