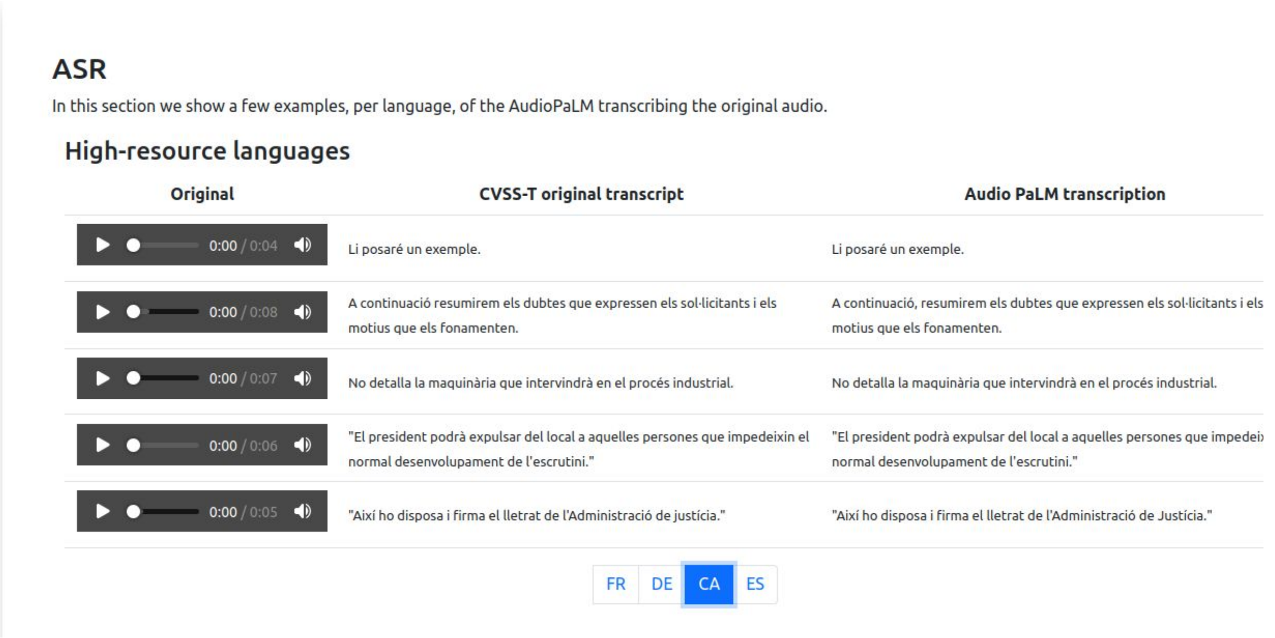
Becoming a High-Resource Language in Speech

The Catalan Case in the Common Voice Corpus

Carme Armentano-Oller, Montserrat Marimon & Marta Villegas

Catalan: a high resourced language?

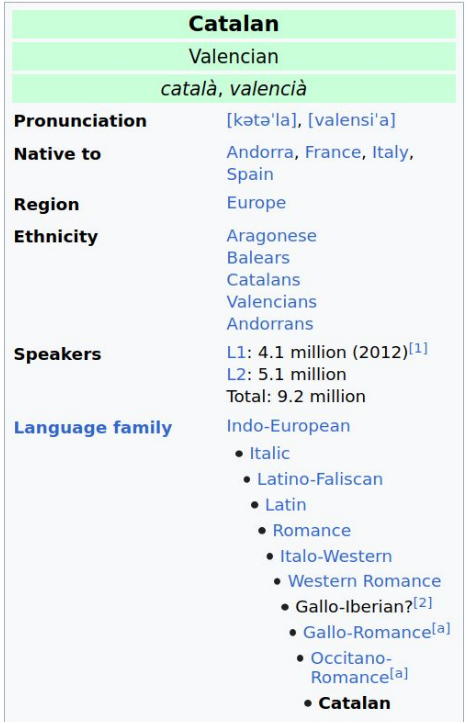


ASR

In this section we show a few examples, per language, of the AudioPaLM transcribing the original audio.

High-resource languages

Original	CVSS-T original transcript	Audio PaLM transcription
▶ 0:00 / 0:04	Li posaré un exemple.	Li posaré un exemple.
▶ 0:00 / 0:08	A continuació resumirem els dubtes que expressen els sol·licitants i els motius que els fonamenten.	A continuació, resumirem els dubtes que expressen els sol·licitants i els motius que els fonamenten.
▶ 0:00 / 0:07	No detalla la maquinària que intervindrà en el procés industrial.	No detalla la maquinària que intervindrà en el procés industrial.
▶ 0:00 / 0:06	"El president podrà expulsar del local a aquelles persones que impedeixin el normal desenvolupament de l'escrutini."	"El president podrà expulsar del local a aquelles persones que impedei normal desenvolupament de l'escrutini."
▶ 0:00 / 0:05	"Així ho disposa i firma el lletrat de l'Administració de justícia."	"Així ho disposa i firma el lletrat de l'Administració de Justícia."

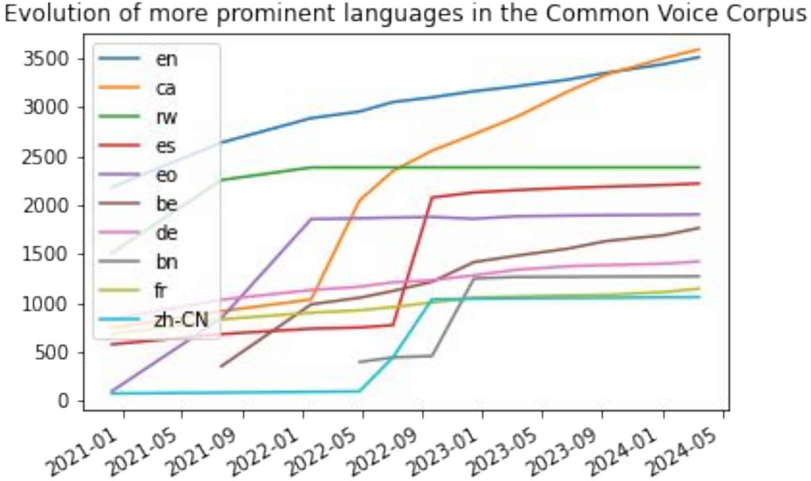


Catalan	
Valencian	
català, valencià	
Pronunciation	[kətə'la], [valensi'a]
Native to	Andorra, France, Italy, Spain
Region	Europe
Ethnicity	Aragonese Balears Catalans Valencians Andorrans
Speakers	L1: 4.1 million (2012)[1] L2: 5.1 million Total: 9.2 million
Language family	Indo-European • Italic • Latino-Faliscan • Latin • Romance • Italo-Western • Western Romance • Gallo-Iberian?[2] • Gallo-Romance[a] • Occitano-Romance[a] • **Catalan**

LA NOSTRA LLENGUA ÉS LA TEVA VEU

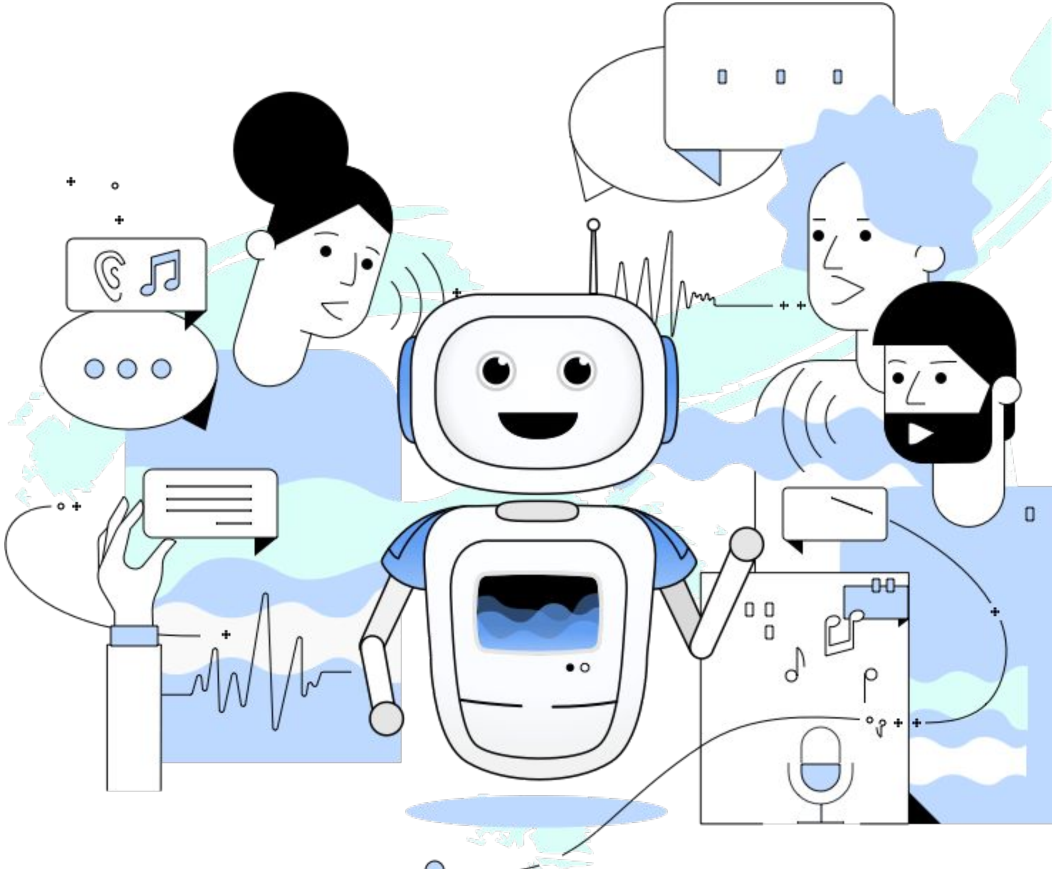
Fem que la tecnologia parli català.

Entra i dona la teva veu.



Evolution of more prominent languages in the Common Voice Corpus

Acquiring voice resources

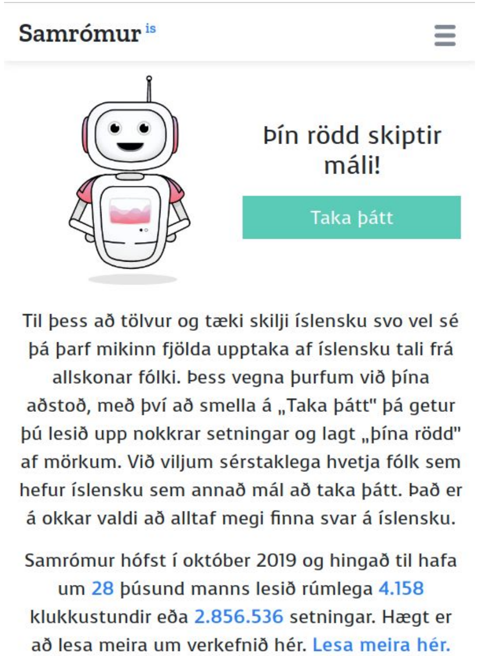


- Legal issues
- Technical difficulties
- Significant investments in equipment and personnel
- Need for diversity in accents and speakers

Crowd-sourced experiences

Icelandic

- Samrómur
- 1.5 million utterances (~ 2,250h)
- 2 years
- 20,000 speakers



Samrómur is

Þín rödd skiptir máli!

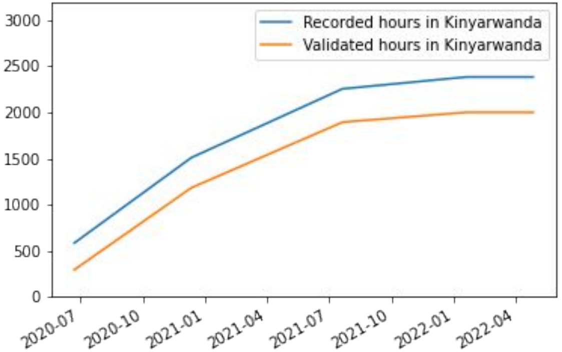
Taka þátt

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 4.158 klukkustundir eða 2.856.536 setningar. Hægt er að lesa meira um verkefnið hér. Lesa meira hér.

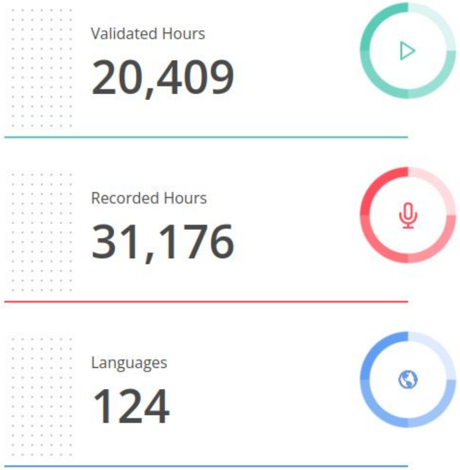
Languages of Rwanda

- project launched in 2019
- 2,388 h in Kinyarwanda
- 1,077 h in Kiswahili
- 583 h in Luganda



Mozilla Common Voice Corpus

The Mozilla Common Voice is a crowd-sourcing platform that allows obtaining an extensive, high-quality, publicly available voice dataset for the development of speech technologies



Validated Hours
20,409

Recorded Hours
31,176

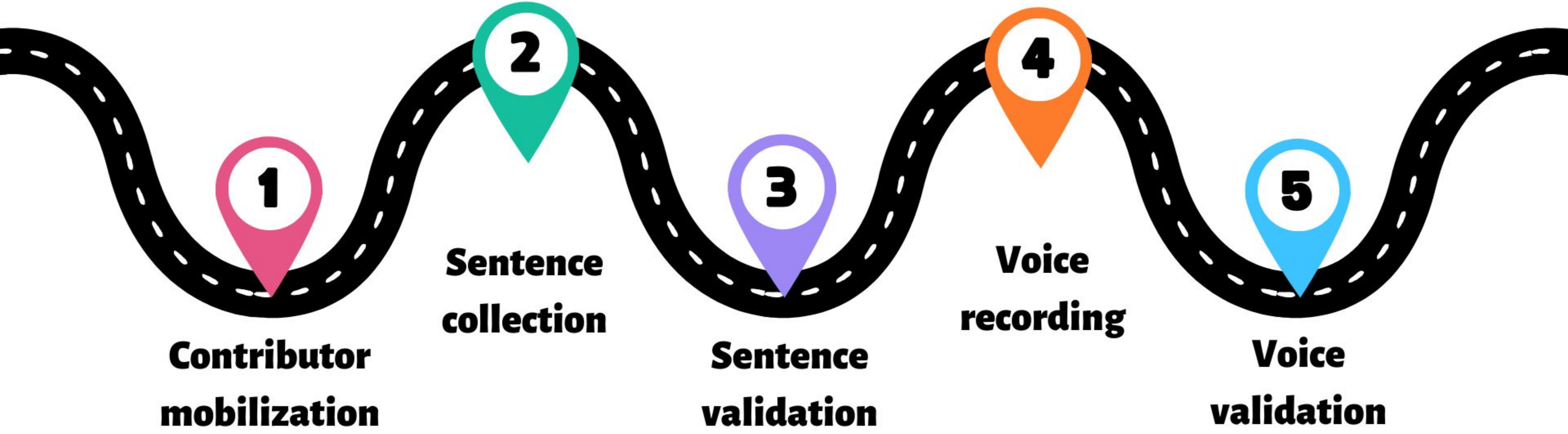
Languages
124

- Globally recognized as a reference dataset
- Multilingual nature
- ~ 1,000 hours of recorded Catalan data in early 2022

Objectives

- Reach 2,000 hours of voice data
- Address gender disparities
- Incorporating a wider range of accents and age groups
- Raise the international visibility of the Catalan language

Step by step: Challenges and Caveats



1 Contributor mobilization

2 Sentence collection

3 Sentence validation

4 Voice recording

5 Voice validation

1 Contributor mobilization

Institutional campaign endorsed by the Catalan Government.

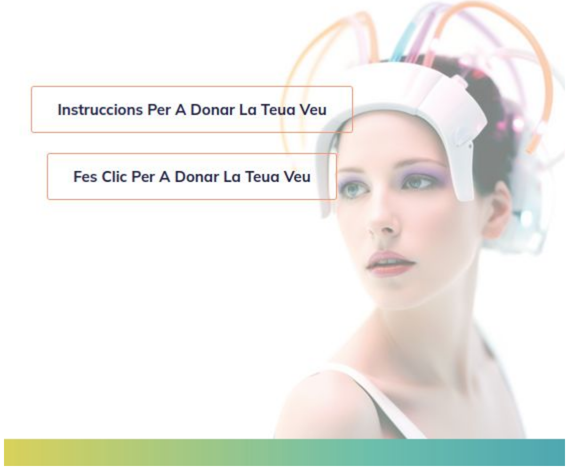
- Starting: February 2022
- Duration: six months

- Website
- TV and radio advertisements
- Social media
- Advertising posters
- Customized van
- Two stationary booths

Contributor mobilization II

Collaboration with:

- Language advocacy organizations
- Balearic Government
- Vives project


Instruccions Per A Donar La Teua Veu
Fes Clic Per A Donar La Teua Veu

Afegeix a sa llista de sa compra blat dindi.

First objective reached in three months!

2 Sentence collection

To obtain one hour of recordings in Catalan we need ~670 sentences.

Two options:

a) single sentence submission
b) bulk sentence submission

Several potential sources contacted

Doubts about ownership

Licence CC0

Automatic generation of sentences

3 Sentence validation

- Readable in 10 - 15 seconds
- Correct grammar and orthography
- Avoid:
  - certain characters ($, &, and emojis)
  - numeric formats
  - abbreviations
  - acronyms
  - offensive terms
  - personal names

Sentence filter: removal of 89% of the collected sentences

Manual quality validation

4 Voice recording

Sentences are recorded by volunteers

Registration option:

- gender
- accent
- age

Critical for mitigating biases!

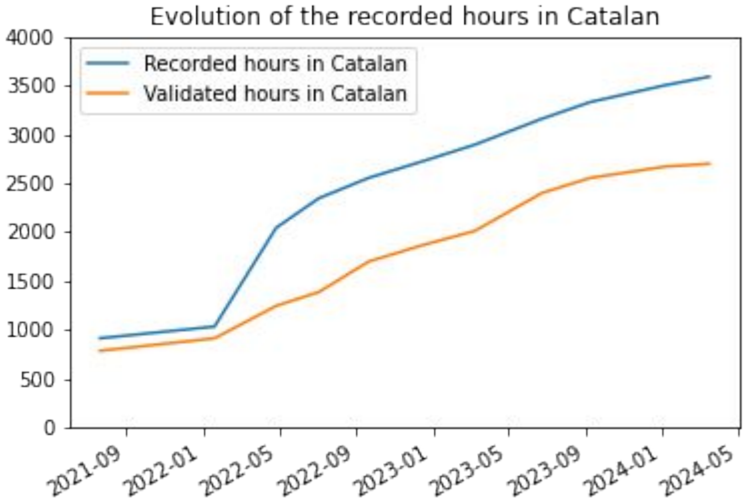
5 Voice validation

Each recorded sentence must be validated by at least two people

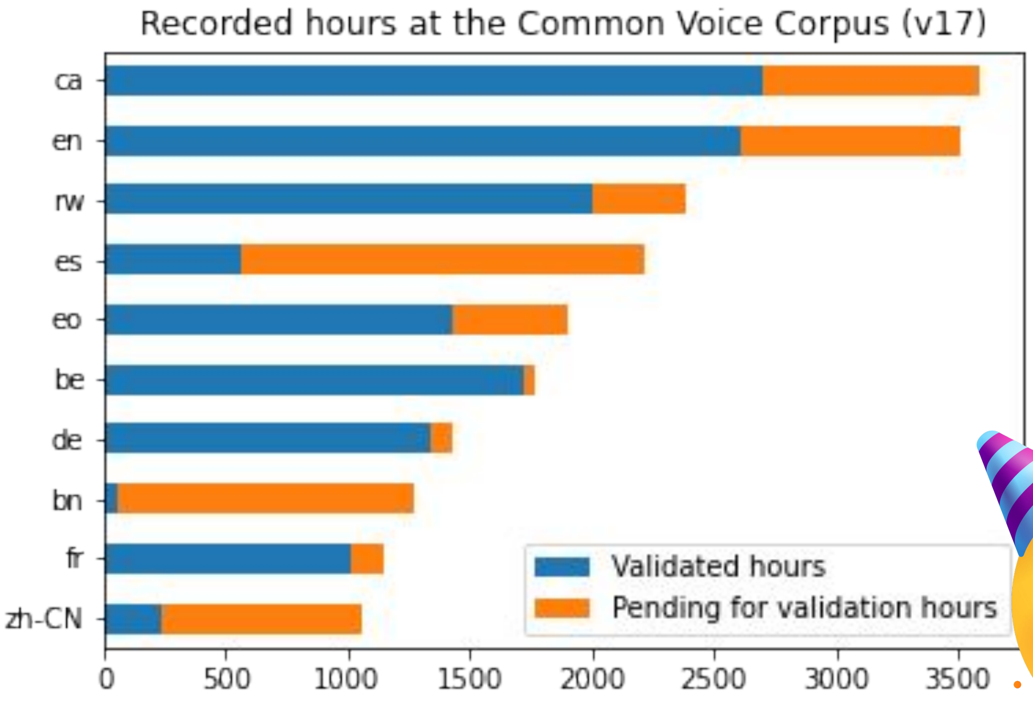
Well-defined guidelines

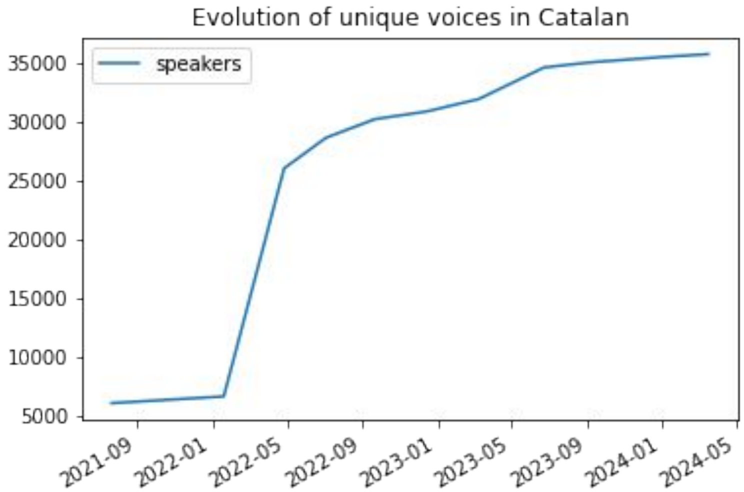
- coherence
- respect accent variations

Hired team of validators

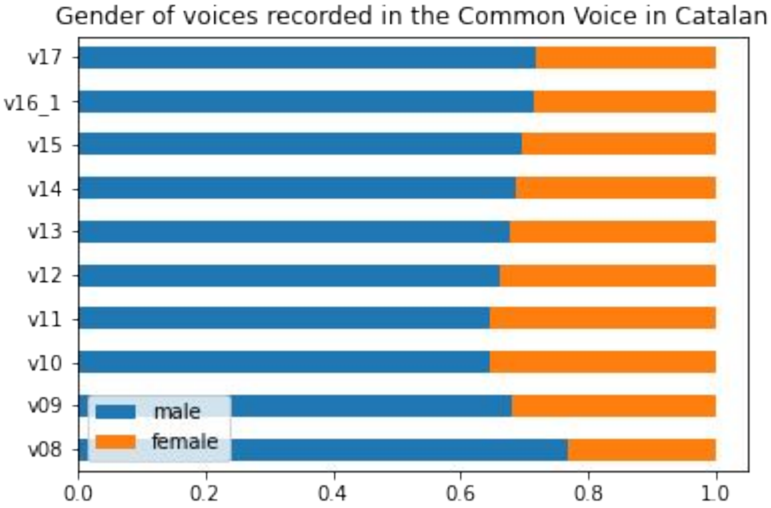


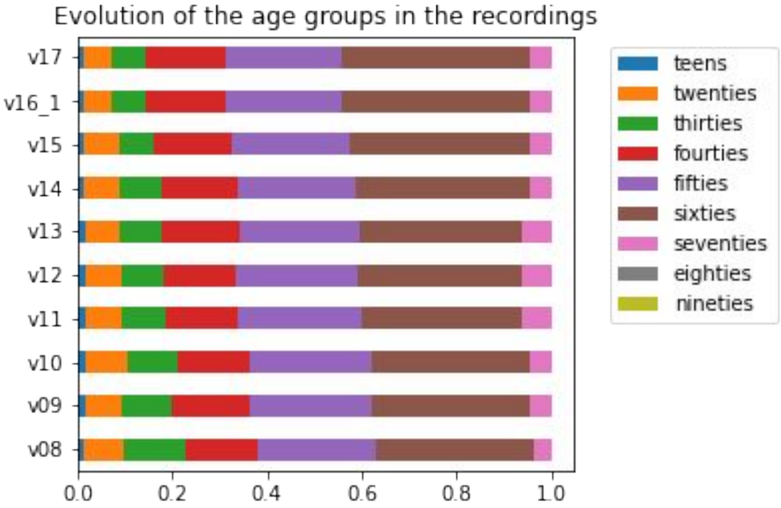
Results


Recorded hours at the Common Voice Corpus (v17)
ca
en
rw
es
eo
be
de
bn
fr
zh-CN
0
500
1000
1500
2000
2500
3000
3500
Validated hours
Pending for validation hours

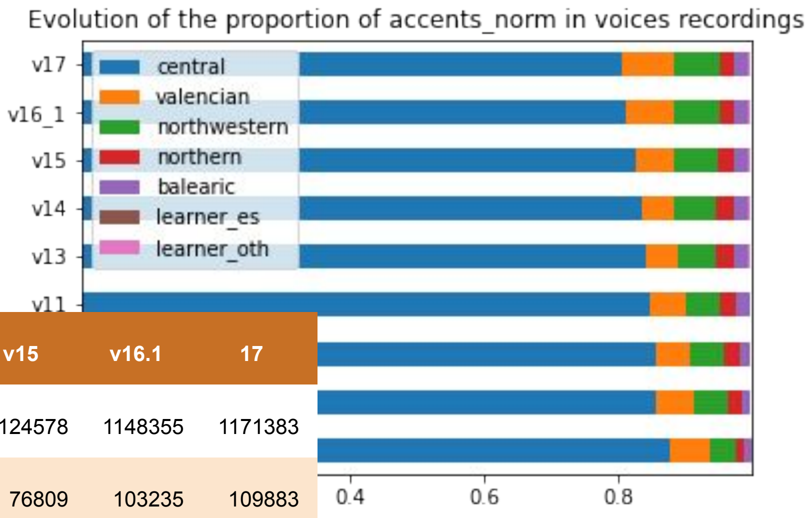

Evolution of unique voices in Catalan
speakers
35000
30000
25000
20000
15000
10000
5000
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
2024-05

Diversity in the corpus





Accents



	v7	v8	v9	v10	v11	v13	v14	v15	v16.1	17
central	421896	483723	709401	805843	860277	988170	1067042	1124578	1148355	1171383
valencian	29662	33248	46935	47806	53450	55048	59942	76809	103235	109883
northwestern	17463	21411	42291	49106	52081	67304	82279	91071	95145	99892
northern	6680	6984	16820	22211	23060	30746	31883	32011	32082	32087
balearic	4796	5963	9684	12781	21184	24151	27093	27681	27827	27960
other	661	ND	1108	1128	1202	1547	1579	1644	1754	1884
learner_es	6	146	279	304	324	413	538	553	1146	1146
learner_oth	ND	5	780	1481	2227	4034	5069	5794	5904	5904

Characteristics to consider

CC0 license:

- visibility
- long-term acknowledgment and usage
- difficult to find text

Community-driven project:

- dedicated community
- sustainability of the project
- collaborative efforts
- long decision-making processes
- not always align with the campaign's timeline or objectives

Conclusions

- significant corpus of voices
- voices diversity
- international profile