Becoming a High-Resource Language in Speech: The Catalan Case in the Common Voice Corpus

Carme Armentano-Oller, Montserrat Marimon, Marta Villegas
Barcelona Supercomputing Center - Centro Nacional de Supercomputación
Plaça d'Eusebi Güell, 1-3, Barcelona
{carme.armentano,montserrat.marimon,marta.villegas}@bsc.es

Abstract

Collecting voice resources for speech recognition systems is a multifaceted challenge, involving legal, technical, and diversity considerations. However, it is crucial to ensure fair access to voice-driven technology across diverse linguistic backgrounds. We describe an ongoing effort to create an extensive, high-quality, publicly available voice dataset for future development of speech technologies in Catalan through the Mozilla Common Voice crowd-sourcing platform. We detail the specific approaches used to address the challenges faced in recruiting contributors and managing the collection, validation, and recording of sentences. This detailed overview can serve as a source of guidance for similar initiatives across other projects and linguistic contexts. The success of this project is evident in the latest corpus release, version 16.1, where Catalan ranks as the most prominent language in the corpus, both in terms of recorded hours and when considering validated hours. This establishes Catalan as a language with significant speech resources for language technology development and significantly raises its international visibility.

Keywords: Catalan, Speech Resources, Common Voice

1. Introduction

Speech technologies form the cornerstone of cutting-edge AI applications that integrate voice interfaces for human-machine communication, such as virtual assistants, conversational agents, or driving assistants. Additionally, in virtual environments, such as videoconferencing platforms, the integration of multilingual subtitling tools is advancing rapidly. These tools allow diverse participants to express themselves in their native languages, while effectively conveying their messages to others who speak different languages. This feature, exclusive to technologically advanced languages, can prove highly advantageous for speakers of languages with a more limited presence, such as Catalan.

Acquiring voice resources needed for training high-quality speech recognition and synthesis models is a complex process that includes legal issues, technical difficulties, resource constraints, and the need for diversity in accents and speakers. It involves obtaining informed consent, handling privacy regulations, etcetera. The task also requires significant investments in equipment and personnel and includes the labor-intensive annotation of voice data. Finally, potential biases in data further complicate the process. However, overcoming these challenges is essential for developing speech recognition systems that accurately represent diverse linguistic and cultural backgrounds, ensuring fair access to voice-driven technologies.

In this paper we describe an ongoing effort to create an extensive, high-quality, publicly available voice dataset for future development of speech technologies in Catalan employing the Mozilla Common Voice crowd-sourcing platform. This work is part of the AINA project,[1] a five-year project launched in 2022 and financially supported by the Catalan Government to ensure the presence of Catalan in the digital era.

2. Similar projects

Crowd-sourcing for voice collection has proven its effectiveness in engaging individuals across various age groups, genders, and accents. For instance, Mollberg et al. (2020) describes the acquisition of a substantial and diverse corpus for speech recognition through crowd-sourced contributions. This data collection initiative used the web application Samrómur,[2] constructed upon Mozilla Foundation's open-source voice collection platform, Common Voice.[3] The primary objective of the project was to build a comprehensive speech corpus for automatic speech recognition in Icelandic. The data collection phase spanned just over two years, and in (Hedström et al., 2022), they reported having obtained a total of 1.5 million utterances (approximately 2,250 hours) from approximately 20 thousand distinct speakers. Additionally, Krewer (2023) outlines the strategies employed to create publicly accessible voice datasets

[1] https://projecteaina.cat/
[2] https://www.samromur.is/
[3] https://commonvoice.mozilla.org/en

LREC-COLING 2024, pages 2142–2148
20-25 May, 2024.

in Kinyarwanda, Kiswahili, and Luganda, using the Mozilla Common Voice platform. This project was launched in 2019, and to date these communities have successfully generated open voice datasets, including nearly 2,400 hours in Kinyarwanda (among the largest on Mozilla Common Voice), more than 1000 hours in Kiswahili, and nearly 600 hours in Luganda. These datasets have subsequently facilitated the development of machine learning models available for use by local innovators, such as a voice recognition model in Kinyarwanda.

3. Objectives

To construct a dataset suitable for training inclusive speech models, two crucial factors come into play: quantity and diversity. That is, it is crucial to acquire a substantial volume of data that incorporates a significant range of speaker diversity, including accent, age, and gender representation groups. Crowdsourcing has proven to be highly effective in achieving these goals.

Our primary goal was to reach 2,000 hours of voice data in the Mozilla Common Voice (MCV) corpus, elevating Catalan to a language with substantial speech resources. As we will see, we significantly surpassed this objective. Furthermore, we aimed to improve the dataset diversity by addressing gender disparities and incorporating a wider range of accents and age groups into the existing dataset.

On the other hand, we aimed to raise the visibility of the Catalan language on an international scale, showcasing it as a language with available resources and demonstrating the Catalan-speaking population's interest in accessing technology in their native language.

4. Corpus building

As mentioned before, for constructing the Catalan voice dataset, we rely on MCV (Ardila et al., 2020), a well-known platform and initiative that relies on crowdsourcing to collect open-source, multilingual speech data. The MCV initiative began in July 2017, initially focusing on English and later expanding to include support for any language in June 2018. In its latest update (version 16.1), the project has a significant community of volunteers, it supports 120 languages, and includes a substantial audio data collection of 30,329 hours.

We opted for the MCV platform based on three significant factors. Firstly, the MCV corpus is globally recognized as a reference dataset. Secondly, its multilingual nature opens the door to innovative techniques such as Speech-To-Speech (STS) translation or multilingual Speech-To-Text (STT) models, as demonstrated by the CoVoST project (Wang et al., 2020), for instance.[4] And thirdly, when our project started in early 2022, the platform already sheltered a considerable Catalan dataset of approximately 1,000 hours of recorded data, thanks to the efforts made by a group of volunteers, led by the non-profit organization Soft-Català.[5]

The necessary actions for building a large-scale voice dataset using the MCV platform are the following:

- **Contributor mobilization**. A crucial task in gathering data through crowdsourcing is to engage a substantial number of volunteers in the process of recording and validating sentences.
- **Sentence collection**. The initial step involves compiling a significant repository of sentences to be used for recording voice clips thereafter. These sentences must be diverse and free from copyright constraints, as MCV data are released under a CC0 License.[6]
- **Sentence validation**. Before sentences are uploaded, they need to be verified to ensure they meet the specific formal criteria required by MCV, while adhering to the project's licensing terms.
- **Voice recording**. Then, sentences submitted to the platform are recorded.
- **Voice validation**. Finally, recorded sentences must go through a validation process before they can be definitively incorporated into the corpus.

5. Challenges and Caveats

In the forthcoming sections, we explore the challenges we faced while executing the actions mentioned earlier to accomplish our project goals.

5.1. Contributor mobilization

The first big challenge when using crowdsourcing to construct a comprehensive dataset is effectively

[4] This flexibility is favored by the fact that the datset is published under the CC0 license, with the only restriction being the avoidance of speaker identification attempts. However, in light of the growing technological advancements in the field, there is an ongoing discussion about potentially imposing explicit limitations on its use for Text-to-Speech (TTS) applications (see https://discourse.mozilla.org/t/explicitly-forbidding-limiting-tts-usage/115072).

[5] www.softcatala.org

[6] https://creativecommons.org/public-domain/cc0/

mobilizing a significant number of volunteers, engaging them actively in the processes of recording and validating sentences. In our project, this objective was successfully achieved through the instrumental support of an institutional campaign formally endorsed by the Catalan Government.
Starting in February 2022 and extending for a duration of six months, the campaign was coordinated by a specialist firm and encompassed several key components. It began with the development of a website to serve as the primary gateway to the MCV initiative. Additionally, a compelling TV and radio advertisement[7] ran for six months across various platforms, including social media. A customized van, bearing the *AINA* slogan, housed two voice collection booths and traveled to events throughout the territory. Two stationary booths were also established at prominent gatherings. Finally, the campaign collaborated closely with SoftCatalà and other language advocacy organizations, emphasizing its commitment to promoting the Catalan language and linguistic diversity. Additionally, the reach of the campaign was extended to the Balearic Islands, in a Campaign coordinated with the Balearic Government, with a focus on capturing the linguistic diversity of the island's accents.[8]
As illustrated in Figure 1, displaying the counts of recorded and validated hours from July 2021 to January 2024, the launch of the institutional campaign was an extraordinary success, and more than 1,000 hours were recorded in just three months, marking a significant rise from 1,036.87 hours recorded in January 2022 (version 8) to 2045.27 in March of the same year (version 9). In just six months, more than one million sentences were recorded, equivalent to an increase of 1,500 hours.

5.2. Sentence collection

The process of creating a voice dataset begins with the collection of an extensive corpus of sentences, which will be used for recording corresponding voice clips. It's important to emphasize the substantial number of sentences required. To obtain one hour of recordings in Catalan, we estimate that approximately 670 sentences are necessary.[9]

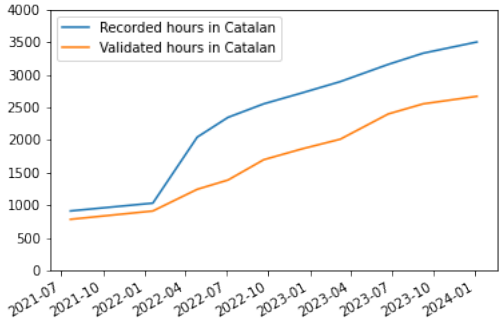

Figure 1: Recorded and validated hours in Catalan (July 2021-January 2024)

To collect sentences, the MCV platform offers a sentence collector, where volunteers can write sentences to be added to the corpus[10]. An alternative to this process is to make a massive contribution of sentences through a bulk submission.[11] Since the process of obtaining sentences through the platform is slow, and we anticipated a massive participation in the recordings, in our project we opted directly for the bulk submission of sentences.
To create a diverse sentence collection that includes various language variations and topics, we contacted several potential sources, including authors, publishers, media entities, and public administrations.
Texts published under a CC0 license can be added to the corpus directly, as allowed by the license. However, this license is uncommon and it is difficult to find texts that have it. For pre-existing works without public licenses, contributors must complete the MCV Contribution Agreement, waiving all associated copyrights.[12] It's worth emphasizing that the MCV project maintains a rigorous approach to this matter.
Regarding publishers, authors, and media outlets, initial responses were generally positive. However, uncertainties about rights ownership often restricted the text transfer process.
Faced with the difficulty in obtaining sentences with a CC0 license, we chose to also generate them automatically, from templates (which spoke of activities performed by various fictitious people, on various dates, in different places) or by modify-

[7] https://www.youtube.com/watch?v=paUhy6qlUXk
[8] https://www.youtube.com/watch?v=cRwMLub3hiY&list=PLv3GNuXpKzbO6alq-02I2eK4ZIJ7cfkN6
[9] In version 16.1 of the corpus, the average duration of each sentence in Catalan is 5.342 seconds. This information is available for all versions and languages at the following link: https://github.com/common-voice/cv-dataset/
[10] https://commonvoice.mozilla.org/en/write
[11] https://github.com/common-voice/common-voice/blob/main/docs/submitting-bulk-sentences.md
[12] https://github.com/common-voice/community-playbook/blob/draft/sub_pages/Common%20Voice%20ContributionAgmt%20-%20%5B1722573%5D%20-%20Template.pdf

ing previous phrases using language models.

5.3. Sentence validation

Sentences included in the corpus must adhere to specific guidelines concerning their length, punctuation, utilization of abbreviations, acronyms, and numeric elements to ensure they can be read in less than 15 seconds. To select sentences that meet the requirements, we implemented a filtering mechanism.[13] This filter eliminates sentences with fewer than 5 words or exceeding a predefined word limit, as well as those containing certain characters (such as $, &, and emojis) or some numeric formats (which might indicate time expressions or too long numbers). Additionally, sentences with one or more words entirely in capital letters (potentially acronyms) or those not beginning with a capital letter and not found in the Hunspell dictionary[14] are excluded. However, we accepted Out of Vocabulary (OOV) capitalized words, assuming they are proper nouns that can be pronounced in a Catalan accent even though they are from a foreign language (e.g., "George," "Facebook," etc.). Nevertheless, at some point, we decided to exclude sentences with more than 1/3 proper nouns to avoid sentences with an excessive number of potential foreign words.

When filtering the literary and journalistic texts, we excluded sentences that contained potentially inappropriate words, based on a list of offensive or potentially discomforting terms. In administrative and journalistic texts, we additionally eliminate sentences with personal names.

At the same time, to reduce the number of discarded sentences, the same script makes some automatic modifications to the sentences, to make them comply with the necessary quality criteria, including replacing sequences of more than one exclamation or question mark character for a single one, normalizing and fixing the quote marks, removing certain characters at the beginning of the line (*, §, –, numbers, etc.), transcribing some numbers, and developing some usual acronyms and abbreviations. In the case of literary texts, to preserve as many sentences as possible, some excluded sentences are modified manually to make them fit the established criteria.

The script has evolved over time and has undergone numerous adjustments based on experience. For example, we raised the maximum word limit from 14 to 18.

The execution of the script resulted in the removal of 89% of the collected sentences. This significant reduction highlights the need to accumulate an extensive amount of textual data.

Additionally, to uphold quality standards for the published sentences, we conducted a quality evaluation of a representative sample from each corpus, as outlined in the project documentation.[15] Initially, we performed the assessment using our in-house resources, but subsequently, we enlisted the assistance of two external validators. This quality control process proved crucial, as it revealed that some of the corpora we had initially prepared did not meet the required standards, leading to their exclusion. In other instances, it facilitated improvements to the sentence selection filter.

5.4. Voice recording

Sentences submitted to the platform are recorded via a microphone, using a computer, a tablet, or a mobile phone. The MCV platform provides volunteers with the option to register, although it is not mandatory. Registration plays a key role in collecting essential speaker demographic data (gender, age, and accent variety), which is critical for mitigating biases that could unfairly affect specific population groups during the training and evaluation of the models.

In terms of accent variety, the MCV strategy has undergone several adjustments. Initially, the platform provided a dropdown menu with accents traditionally associated with Catalan: Central, North-western, Northern, Balearic, and Valencian. Given the diverse demographics of Catalan speakers, we proposed the inclusion of options for individuals who speak Catalan as a second language and may have noticeable accents from their native language. As a result, we introduced the "learner_es" tag for those with Spanish as their first language and "learner_oth" for those with another first language. However, in February 2022, there was a shift from a closed list of accent options to an open field, allowing speakers to freely define their accents. While this change promoted inclusivity, it presented challenges in tracking campaign demographics.[16] Subsequently, in May 2023, a new approach was introduced, which offered both a closed list for selecting dialect variants and an open field for self-definition. This modification was in line with a platform promotion campaign by the Valencian government,[17] and it provided an oppor-

[13] The filter is available at https://github.com/projecte-aina/catalan_common_voice_filter

[14] https://github.com/hunspell/hunspell

[15] https://github.com/common-voice/common-voice/blob/841eaa518cbac644f6e66c5a7754aa09ecb0d025/docs/SENTENCES.md

[16] The progression of each accent's representation, as displayed in Figure 5, has been determined through a data normalization process with respect to the initial labels: Central, North-western, Northern, Balearic, Valencian, learner_es, and learner_oth.

[17] https://vives.gplsi.es/

tunity to request the inclusion of Valencian subvarieties in the list of options.

Regarding the platform's use by children, it's important to note that MCV's legal terms stipulate that individuals under the age of 19 require parental consent and can only participate under supervision.[18] This issue has been a topic of discussion within the platform's forums. However, given the lack of clear guidelines on how consent and supervision should be established, we opted not to proceed with any school-based campaigns.

As mentioned before, the campaign launch was an extraordinary success, generating over 233,000 recordings (approximately 400 hours) within a single week. This massive response led to a temporary server crash within the MCV infrastructure. Faced with this challenge, we explored the possibility of creating an alternative platform, similar to what was done for Icelandic (Hedström et al., 2022), which could have resolved the server issue but would have required splitting the Catalan oral corpus into two parts: Common Voice and the alternative platform. This was a scenario we preferred to avoid. Fortunately, the server problem was resolved within a reasonable time frame, allowing us to proceed with the campaign as initially planned.

5.5. Voice validation

In the final step, each recorded sentence must be validated by at least two volunteers before it can be included in the corpus.

Volunteers follow well-defined guidelines, drawn up by the community. Their primary task is to ensure the coherence between spoken words and the accompanying written text, taking into account potential accent variations and contributions from individuals with foreign accents. Any inconsistencies, including alterations in verb tenses, omitted or added words, inaudible sentence beginnings or endings, disruptive background noise, unclear audio volumes, or reader interruptions such as stumbling or repetition, result in the rejection of the respective recordings.

Figure 1 shows that the participants preferred the task of donating speech to that of validating, therefore, while we obtained a substantial amount of recorded sentences through crowdsourcing, validating this data did not raise the same level of active engagement.

To bridge this gap, we brought validators on board. In an initial phase, we hired a team of seven validators who reviewed and validated nearly 500,000 clips in approximately two months. In a second phase, we extended our validation efforts by bringing on board 12 additional validators for an eight-month period.

[18] https://commonvoice.mozilla.org/en/terms

During this time, we faced a new challenge: despite having more recorded sentences than validated ones, the platform didn't permit additional validations. We eventually realized that this limitation was because the same sentence could be recorded multiple times but was limited to a single validation. Consequently, only 75% of the corpus could be validated. After identifying this issue, we requested the removal of this limitation, and following community discussions, it was successfully resolved. Our team has validated approximately 790,000 sentences.

6. Results

6.1. Collected data

In the latest release of the corpus, version 16.1 (January 2024), Catalan stands as the most prominent language within the corpus, having 3,500 recorded hours, followed by English, with 3,438 recorded hours. In terms of validated hours, Catalan also leads with 2,670 hours, closely followed by English with 2,586 hours. In this regard, we have not only met our primary objective of accumulating 2,000 hours of voice data, but exceeded it, successfully establishing Catalan as one of the languages with substantial resources for language technology development. Figure 2 provides a snapshot of the current counts for recorded hours (validated and pending for validation) of the top 10 languages with the highest representation in the latest corpus release.

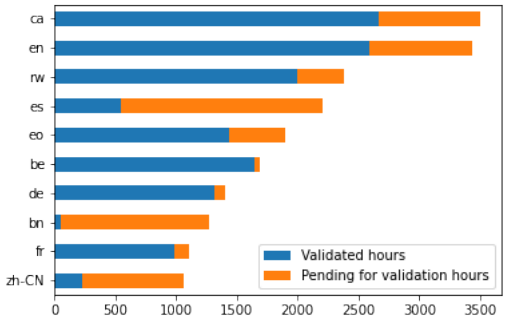

Figure 2: Recorded hours (validated and pending for validation) of the 10 languages with the highest representation in the corpus (v16.1)

6.2. Participant demographic

Despite the campaign's success, there was a decline in the percentage of sentences accompanied by demographic information, dropping from 82% to 55%. This decrease suggests that many new contributors chose not to provide this information. However, this trend began to reverse with version

12, and currently, in version 16.1, we observe that 70% of sentences include demographic data, indicating a positive shift. Figure 3 illustrates the progression of the percentage of sentences with linked demographic data, spanning from version 7 to version 16.1.

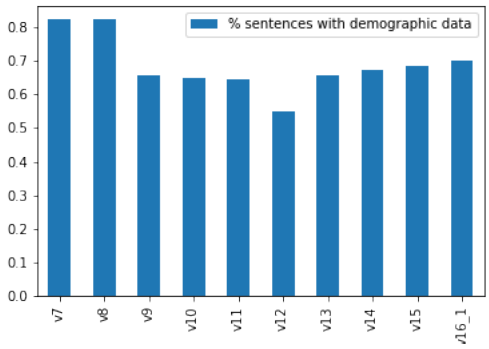

Figure 3: Progression of the percentage of sentences with associated demographic data

In terms of addressing the **gender gap**, we observed a notable rise in female participation, with the representation of female voices increasing from 22.9% (version 8) to 35.21% (version 11) of recordings with associated demographic data. Unfortunately, in the latest versions, this trend appears to be regressing, as shown in Figure 4.

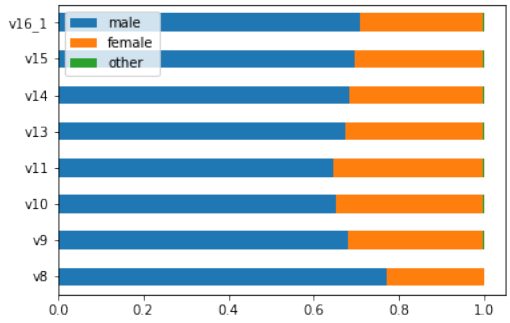

Figure 4: Gender evolution (in sentences with associated demographic information)

Regarding **accent variety**, the central accent remains dominant, mainly due to demographic factors, as it aligns with the area where the majority of the population resides. However, there has been a slight shift, with the central accent decreasing from 88% in version 8 to 84% in version 14.
The evolution of accent proportions can be observed in Figure 5.[19]

[19]We calculated this data by concatenating the data in the `validated.tsv`, `invalidated.tsv`, `other.tsv` splits published for each version in the official webpage https://commonvoice.mozilla.org/en/datasets. The version 12 files don't have the accent information.

Note that the Balearic accent, although currently representing only 2% of the corpus, has seen significant growth, going from 5,963 sentences recorded in version 8 to 27,827 in version 16.1.

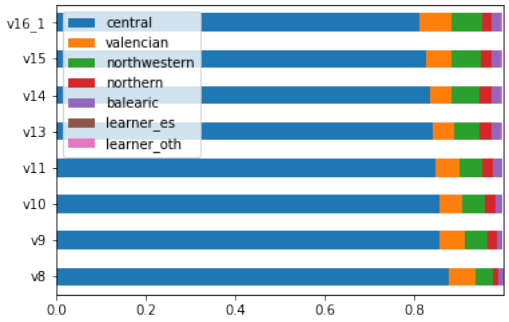

Figure 5: Accent evolution (in sentences with associated demographic information)

When looking at the evolution of recordings by **age group**, we observed an increase in contributions from individuals in their sixties, who were already the largest group at the start of the campaign. This highlights the need for future campaigns to target different age groups (see Figure 6).

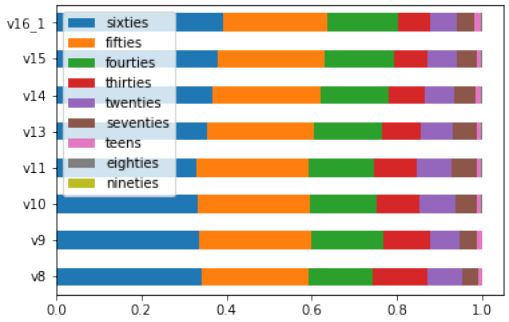

Figure 6: Age evolution (in sentences with associated demographic information)

6.3. Impact

The campaign had a clear goal of raising the international profile of the Catalan language. In this regard, we are extremely pleased with the achieved outcomes. Catalan's substantial representation in the MCV corpus has led to its inclusion in projects such as Meta's Seamless m4t[20] and Google's Au-

[20]https://ai.meta.com/blog/seamless-m4t/

dioPaLM,[21] where it is acknowledged as one of the four "high-resources" languages, together with French, German, and Spanish.

7. Lessons learned

Upon reviewing the presented results, it's evident that organizing a crowdsourcing campaign can lead to positive results when citizen engagement strategies are well-defined. In this context, the MCV platform stands out as a valuable resource. Nevertheless, the project displays distinctive characteristics that require careful consideration, as they offer both benefits and drawbacks:

- **Distributed under a CC0 license**. This grants the corpus substantial visibility, ensuring its long-term acknowledgment and usage. However, this approach does come with the disadvantage of making it notably difficult to identify texts that fall under this particular license.
- **Community-driven project**. With the support of a dedicated volunteer community, this model not only contributes to the sustainability of the corpus but also encourages collaborative efforts. However, it does come with the drawback of decision-making being a collective process, which may not always align with the campaign's desired timeline or objectives. In our case, this became particularly evident during the server outages at the beginning of the campaign, although in the end it was solved satisfactorily and on time.
- **A single set of texts for all language variants**. Currently, all language variants use a common set of sentences. We believe that it would be beneficial to provide the option of linking sentences with local expressions to their respective accents, allowing them to be read in a manner that aligns with the specific regional accent. Unfortunately, this feature is presently unavailable, resulting in some sentences being read in an unnatural manner.

In conclusion, despite some challenges, the campaign has enabled us to gather a significant corpus of voices and improve its diversity. Most importantly, it has elevated the international profile of the Catalan language.

8. Bibliographical References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Staffan Hedström, David Erik Mollberg, Ragnheiður Þórhallsdóttir, and Jón Guðnason. 2022. Samrómur: Crowd-sourcing large amounts of data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2311–2316, Marseille, France. European Language Resources Association.

Jan Krewer. 2023. Creating community-driven datasets: Insights from mozilla common voice activities in east africa.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jon Gudnason. 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. European Language Resources Association.

Changhan Wang, Juan Miguel Pino, Anne Wu, and Jiatao Gu. 2020. Covost: A diverse multilingual speech-to-text translation corpus. *CoRR*, abs/2002.01320.

9. Acknowledgements

An immense thanks to the entire AINA team, particularly Paul Andrei Petrea and Baybars Külebi, for their consistent participation and support, and Hannah Rose Galbraith for refactoring the sentence filter. We express deep gratitude to the entire MCV community, with special recognition to Francis Tyers, for their unwavering dedication in overcoming challenges during the campaign. Our sincere appreciation extends to the collective Soft-Català, especially Joan Montané, and other supporting organizations like Platforma per la Llengua, Òmnium Cultural, and many more. Special acknowledgment is given to the writers and editors who generously contributed, including Grup Enciclopèdia Catalana, VilaWeb, Racó Català, El Cèrvol, Secretaria de Política Lingüística, Màrius Serra, Carles Cortés, and Joan Pujolar. Lastly, immense thanks to the over 35,000 individuals who lent their voices to this project; your contributions have been invaluable to our success.
This work has been promoted and financed by the Generalitat de Catalunya through the AINA project.

[21] https://google-research.github.io/seanet/audiopalm/examples/