Building Collaborative Speech Corpora for Low-Resource Languages: The Galician Dataset in Mozilla Common Voice

Adina Ioana Vladu, Elisa Fernández Rei, María Pérez Lago
Instituto da Lingua Galega
Universidade de Santiago de Compostela
{adina.vladu, elisa.fernandez, mariaperez.lago}@usc.gal

Abstract

This paper presents the methodology and outcomes of building collaborative speech corpora in Mozilla Common Voice (MCV), focusing on the Galician case within *Proxecto Nós*. We describe the organization of voice collection campaigns –on-site events, student participation, *Validatón* marathons, and corporate collaboration– and analyze the results in MCV v22.0. While the dataset has achieved a modest scale, major gaps remain in metadata completeness and dialectal tagging, with implications for ASR performance. Drawing on our experience, we highlight effective strategies for engagement, such as transparent communication, cultural identification, and user-friendly tools. We conclude with lessons learnt for improving data representativeness, participant retention, and ethical governance. The observations are specific to the Galician case study but may inform similar efforts in other lesser-resourced languages.

Keywords: Mozilla Common Voice, Galician, crowdsourcing, speech corpus, low-resource languages, metadata, FAIR data

1. Introduction

Speech technologies have become integral to human–computer interaction, yet their development remains highly uneven across the world's languages. For low-resource and regional languages, the lack of large, publicly available corpora of transcribed speech continues to be a major barrier to inclusion in data-driven speech technologies such as automatic speech recognition (ASR) and text-to-speech (TTS). Building such resources requires not only technical infrastructure but also a social and ethical framework capable of mobilizing speakers and guaranteeing fair, transparent use of their contributions.

The Mozilla Common Voice (MCV)[1] project provides one of the most significant examples of open, community-driven corpus building. Its participatory model (collecting, validating, and releasing voice data under open licenses) has enabled many low-resource languages to gain initial representation in the digital sphere (Krewer, 2023). However, the translation of this general framework into specific linguistic and sociocultural contexts requires careful methodological adaptation.

This paper examines the processes, challenges, and methodological implications of corpus construction in Common Voice through the Galician case. Within the framework of *Proxecto Nós* (de Dios-Flores et al., 2022; Vladu et al., 2022), the initiative combines institutional leadership, community engagement, and open-data principles to build a publicly accessible speech corpus for a lesser-resourced language. The case study illustrates how participatory design and FAIR-data management can support large-scale voice collection in a feasible and ethically robust manner, while also highlighting the structural limitations encountered in the process.

The paper is organised as follows: Section 2 offers an overview of the state of language technology in Galician, Section 3 presents two success stories in crowdsourcing data for lesser-resourced languages and what can be learnt from their example, Section 4 outlines the MCV framework, Section 5 highlights the building of the Galician presence in MCV, Section 6 reflects on the process of building the corpus and its limitations, and Section 7 presents the conclusions to the work.

2. Language Technology in Galician

The development of language technology (LT) is essential in the current digital society, culture and economy. Such technology, widely supported in languages in high demand worldwide, is also necessary for smaller and less economically powerful languages.

Galician is considered as a low-to-medium resource language: while the availability of textual data and resources is increasingly favorable for most language processing technologies, a significant gap remains in the area of multimedia data. Speech technologies, though advances have been made in recent years, are comparatively less developed, largely due to the limited quantity and diversity of available corpora. In addition, only a small portion of the existing resources are freely available as open-access materials, which further

[1] https://commonvoice.mozilla.org

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1710–1720
11–16 May 2026.

restricts their widespread use (Ramírez Sánchez and García-Mateo, 2022).

For TTS resources, notable open-access contributions include Cotovía (Banga et al., 2012) (a TTS system for Galician and Spanish), and several datasets: the CRPIH_UVigo-GL-Voices dataset[2], a multi-speaker resource with a total of approximately 21 hours of speech recordings with the corresponding text, and two large-scale high-quality single-speaker TTS datasets, Nos_Celtia-GL (Vázquez Abuín et al., 2023)[3] and Nos_Brais-GL (Vladu et al., 2025)[4].

In ASR, Galician benefits from several publicly accessible speech datasets. Nos_ParlaSpeech-GL (Magariños et al., 2023)[5] contains approximately 1,600 hours of audio and transcripts of Galician parliamentary sessions. Nos_TranscriSpeech-GL (Vladu et al., 2023)[6] is a 53-hour manually transcribed and speech-to-text aligned dataset covering multiple domains (conferences, debates, speeches, and interviews). With 250 hours of speech, the FalAI dataset[7] is the largest publicly available dataset for spoken language understanding. Additional resources include Google's Open SLR77 (approx. 10 hours)[8] and the Galician subset of FLEURS[9]. Although some of these datasets contain many hours of speech, they do not necessarily constitute a representative sample of the Galician language in terms of linguistic diversity (e.g., accents and speakers). Furthermore, some of them are domain-specific (Nos_ParlaSpeech-GL, FalAI), which imposes important limitations.

3. Crowdsourcing Data for Lesser-Resourced Languages: Success Stories

The previous section outlined the resource landscape for Galician LT, highlighting a relative gap in large, open, and diverse speech data. Crowdsourcing offers an effective way to address this gap while keeping costs manageable and governance open. Two recent initiatives–one within Galician (FalAI) and one in a comparable context (the Catalan language in MCV)–illustrate complementary strategies from which lessons learnt can be applied to our approach to Galician MCV.

The FalAI dataset (Pineiro-Martin et al., 2024) is a recent illustrative example of a crowdsourced approach to speech data creation. Developed by Balidea and the AtlanTTic research center at the University of Vigo, the project involved the design of a set of about 3.500 multi-domain sentences, paired with a large-scale public speech data collection campaign in which volunteers were invited to record themselves reading 30 sentences using a web-based tool specifically developed for this purpose.

The campaign achieved remarkable success thanks to its effective dissemination strategy and strong ludic and participatory component. It gained visibility through social networks, online platforms, and the press, and was supported by well-known Internet personalities who encouraged public participation. The recording collection process was conceived as a competitive challenge, including prizes and a contest between municipalities, which helped mobilize participants. As a result, more than 11,000 people contributed over 250 hours of speech, far exceeding the initial 100-hour target. As explained in Pineiro-Martin et al. (2024), a few key takeaways from this experience are that transparent communication on data purpose and privacy, culturally significant prompts that strengthen identification with the project, engaging advertising, and an accessible, multi-device recording interface foster participation and trust, while strategic collaboration accelerates execution and impact.

Within MCV, the Catalan AINA initiative demonstrates how government-backed, community-driven mobilisation can transform a lesser-resourced language into a high-resource one (Armentano-Oller et al., 2024). The project combined bulk CC0 sentence acquisition and curation, coordinated contributor mobilisation in multiple media and collaboration with volunteer groups to achieve large-scale recording, and paid validation waves to address the imbalance between recorded and validated data.

The Catalan effort achieved over 3,500 recorded hours in MCV, raising the international visibility of the language and enabling its inclusion in ASR and TTS projects. The case underscores transferable lessons: pair mass collection campaigns with robust sentence governance and validation capacity; communicate licensing/consent clearly; plan for metadata and dialect coverage; and leverage public–private–community alliances to sustain scale and diversity, as well as to foster and maintain long-term user retention.

Together, the FalAI and AINA initiatives offer complementary insights for the ongoing development of the Galician corpus in Common Voice, such as: (i) building trust through transparent communica-

[2] https://zenodo.org/records/8027725
[3] https://zenodo.org/records/7716958
[4] https://zenodo.org/records/14265241
[5] https://huggingface.co/datasets/proxectonos/Nos_Parlaspeech-GL
[6] https://huggingface.co/datasets/proxectonos/Nos_Transcrispeech-GL
[7] https://huggingface.co/datasets/GTM-UVigo/FalAI
[8] https://www.openslr.org/77/
[9] https://huggingface.co/datasets/google/fleurs

tion on data use and privacy; (ii) maintaining cultural proximity in prompts to enhance motivation; (iii) pairing large-scale campaigns with sustained at-home participation through registered profiles; (iv) reinforcing validation capacity through community events and trained reviewers; and (v) establishing long-term alliances with public institutions, academia, and civic organisations.

4. The Common Voice Framework

4.1. Collaborative and Open Methodology

Common Voice was launched by Mozilla to democratize the creation of voice data and reduce the dependency of speech technology on proprietary datasets (Ardila et al., 2020). Its design rests on inclusivity, transparency, and openness: anyone can participate by submitting short textual prompts, validating others' contributions or recording or validating readings of said utterances; all resulting data are released under the CC0 license. This openness ensures full reusability, aligning with the FAIR principles (Findable, Accessible, Interoperable, and Reusable)(Wilkinson et al., 2016).

Each language community coordinates data gathering, communication, and validation tasks. The uniform technical framework provided by Mozilla guarantees that datasets are comparable across languages while allowing cultural and linguistic customization. For many languages, Common Voice has become the first large-scale, publicly available speech corpus, serving both as a training resource and a tool for digital language activism (Ardila et al., 2020). However, even for languages with existing technological support, Common Voice represents a strategic contribution by expanding and diversifying the open data ecosystem. The diversity of speakers contributing to the platform, spanning different ages, accents and recording conditions, provides data that can enhance the robustness and representativeness of speech models, while the continuous and community-driven nature of data collection allows corpora to grow and be updated over time at a relatively low cost. In this sense, even for higher-resourced languages, Common Voice contributes to speech technologies by broadening the diversity of the training data and reducing reliance on proprietary resources.

4.2. Quality Assurance and Workflow

The Common Voice pipeline integrates human validation at multiple points to balance scale and quality. Textual prompts are curated by the local linguistic community to ensure orthographic and syntactic correctness and to avoid copyright infringement, or are added by volunteers and validated following dedicated guidelines[10]. Recordings are carried out by volunteers, generally using their own devices. Each audio clip must then be validated by two independent listeners, following provided guidelines[11]. This process, illustrated in Figure 1, combines openness with minimal consensus-based quality control, but can also introduce incoherences that stem from crowdsourced human validation and other challenges typical of open participation: uneven recording and text quality, varying degrees of representation of dialects and registers, and the risk of data imbalance (gender, age, etc.). These issues must be addressed locally through supplementary curation and data collection strategies.

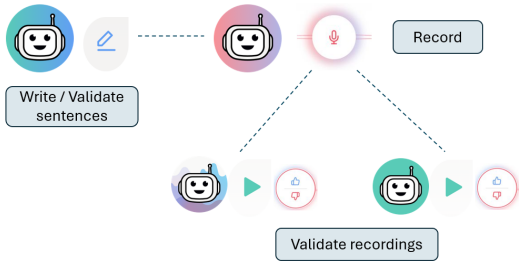


Figure 1: *Workflow of collaborative corpus creation in MCV.*

5. The Galician Common Voice Initiative

5.1. Motivation and Institutional Ecosystem

The integration of Galician into Common Voice is part of the broader *Proxecto Nós*[12], an initiative led by Instituto da Lingua Galega (ILG) and Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) from the University of Santiago de Compostela (USC) with the support of the Xunta de Galicia, currently within the framework of state-wide TL projects such as ILENIA (Külebi et al., 2024)[13] and ALIA[14]. Its objective is to situate the Galician language within the digital economy and ensure its presence in the technological landscape through open, reusable resources.

MCV was chosen as the main platform for speech collection due to its open-data policy, existing infras-

[10] https://commonvoice.mozilla.org/gl/guidelines?tab=scripted-speech#public-domain

[11] https://commonvoice.mozilla.org/gl/guidelines

[12] https://nos.gal

[13] https://proyectoilenia.es

[14] https://alia.gob.es

tructure, and visibility in the international research community. It required no proprietary software, offered immediate hosting and data distribution, included support from a dedicated international community, and allowed the Galician corpus to remain interoperable with other multilingual datasets such as the CoVoST corpora (Wang et al., 2020a,b). Methodologically, this decision positioned Galician not only as a language beneficiary but also as a contributor to the global ecosystem of open speech data.

5.2. Preparing the Oral Corpus

In the process of creating spoken datasets, both scripted and spontaneous speech are necessary, as each modality has its own characteristics and poses different challenges. In the case of scripted speech, the act of reading conditions pronunciation and intonation, generally reducing the appearance of dialectal or colloquial features, since reading usually brings speakers closer to a standard or more refined variety. While MCV has mainly focused so far on read-aloud speech, its recent expansion to include spontaneous speech aims to solve some of the existing limitations by including more naturally-occurring different accents, lexical and morphological variants, code switching, or other typically oral phenomena. Our observations in this paper focus solely on the read-aloud speech corpus.

5.2.1. Linguistic Variation

In the development of a read-sentences Galician speech corpus on MCV, a few linguistic particularities must be considered. First, a relevant aspect, common to the majority of languages present on the Mozilla platform, is dialectal variation. Galician is generally described as comprising three broad dialectal areas (western, central and eastern), which exhibit noticeable phonetic, morphological and lexical differences. Phonetic variation includes phenomena such as *gheada* –the articulation of /g/ as [h], [ɦ], or [x]– and *seseo* –the absence of the /θ/ sound, as opposed to the standard variety. These variations have a direct impact on the quality of speech recognition systems, which must be trained with samples from different areas to ensure balanced performance.

The close contact between Galician and Spanish frequently manifests in oral interaction through code-switching, hybrid forms, and phonetic realizations influenced by Spanish. A representative corpus should therefore reflect this linguistic reality without excessive normalization.

Phonetic and phonological traits, as well as reduction, elision, and assimilation of sounds are also frequent processes that can further complicate the task of ASR.

This linguistic variation causes certain difficulties both in the creation of text corpora for MCV and in the collaborative process of validating utterances by users.

5.2.2. Building the Read Text Corpus

In order to be representative and balanced, a dataset should reflect current language use. In the case of voluntary-lead, sentence-based read datasets, a few more criteria must be met: sentences must be easily readable, content appropriate, attractive, as well as grammatically correct. More necessary conditions for datasets designed for ASR are: sentences must be short enough to correspond to audios under 30 seconds, and the text should not allow ambiguous readings, meaning that numbers, abbreviations, acronyms, or foreign words must have a single clear pronunciation to avoid problems with ASR training. Additionally, the MCV open data policy requires that sentences submitted to the platform be copyright-free (CC0).

Following these requirements, the Galician MCV text corpus[15] was created starting from public texts or texts that were CC0-released by public or private entities: the Galician Parliament, local broadcasting companies, digital and printed newspapers, and public cultural and linguistic entities.

Based on these materials, trained linguists implemented a structured processing pipeline[16], combining automatic scripts and manual supervision. Texts were segmented into sentences, filtered to a maximum of 14 words, cleaned of unwanted characters and formatting inconsistencies, and normalized to comply with MCV criteria. The workflow included the normalization of numbers, abbreviations and acronyms, removal of foreign words (except adapted forms and common toponyms/anthroponyms), elimination of duplicates, and content filtering.

Given the absence of highly reliable deep grammatical correction tools for Galician, orthographic and grammatical revision required a semi-automatic approach: automatic spell-checkers (Hunspell[17] and later LanguageTool[18]) were combined with curated regular-expression lists and supervised mass corrections. Random samples of each subcorpus were manually reviewed to refine correction rules and ensure linguistic quality, readability, and semantic coherence. This initial curation phase entailed a substantial methodological

[15] https://github.com/proxectonos/nos_gl_CC0

[16] https://github.com/proxectonos/nos_gl_CC0/tree/main/Scripts

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

[18] https://github.com/languagetool-org/languagetool

and human effort before integration into the platform.

Even though the coexistence and close contact of Galician and Spanish is a reality that spoken corpora must take into account, we decided not to include code-switching into our dataset, as it is a phenomenon that occurs mostly in spoken Galician but is generally absent in scripted speech, except from frequently used words.

The Common Voice Analyzer tool for version 22.0 of the Galician MCV corpus[19] (MCV-22-gl) shows that the text dataset comprises 696,448 sentences with a total of 6,030,332 words, most added by bulk submission and thus already validated by linguists. The proportion of unique words (types) to all words (tokens) in the corpus, i.e., the type/token ratio, is 49.67% (comparable to Catalan 58.85%, English 57.29%), indicates moderate lexical diversity. For the phonetic analysis, the dataset was automatically transcribed and syllabified using Cotovía[20], which allowed us to compute syllable counts and diphone inventories in a systematic manner. Phonetically, the dataset contains a total of 12,915,511 syllables, with 5,861 different units, and 29,989,344 total diphones, with 1,101 distinct units. In the latter calculation, we treated the pause at the beginning and end of each utterance as an element within the diphone inventory, and we distinguished between stressed and unstressed vowels. It should also be noted that the corpus includes some common foreign words (mainly anthroponyms and toponyms), which slightly increase phonetic variability.

Altogether, as summarized in Table 1, the main quantitative traits of the text corpus indicate a broad lexical and phonetic coverage of the Galician vocabulary, sound inventory and coarticulatory patterns.

Statistic	Value
Sentences	696,448
Validated sentences	696,265
Total words	6,030,332
Average words / sentence	8.66
Total tokens (types)	121,385
Average token occurrence	49.67
Total syllable count	12,915,511
Syllables (individual units)	5,861
Total diphone count	29,989,344
Diphones (individual units)	1,101

Table 1: Summary statistics of the MCV-22-gl text corpus.

[19] https://analyzer.cv-toolbox.web.tr/examine/gl/22.0

[20] https://gtm.uvigo.es/en/transfer/software/cotovia

5.3. Gathering Spoken Data: Voice Donation Campaigns

Once the Galician sentence database was consolidated, a multi-channel and long-term strategy was designed to coordinate and sustain the collection of spoken data. Starting from an initial value of 18 recorded hours and 161 speakers, the campaign sought to: (i) introduce Mozilla Common Voice to the general public; (ii) foster user loyalty and repeated participation; (iii) test different modalities of voice collection (online, in-person, institutional, and corporate); (iv) observe user behavior and identify technical or motivational barriers; and (v) ensure the sustained growth of the Galician dataset within Common Voice.

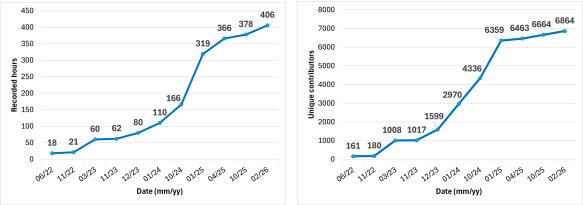
The strategy unfolded in successive phases between 2022 and 2025, combining in-person and digital actions to reach diverse demographic and linguistic profiles. An initial pilot campaign in December 2022 tested in-person collection through mobile recording stands at university sites. Despite logistical limitations, such as exam periods, holiday closures, technical issues, and slow user registration, the campaign yielded valuable insights into user behavior, motivational factors, and technical barriers. It produced 19 hours of recordings from nearly 400 speakers, representing an initial baseline.

The lessons from this pilot informed the design of subsequent campaigns. The 2023–2024 strategy combined targeted digital communication with onsite collection. A multimedia advertising campaign redirected users to a centralized landing page[21] which explained the goals of the speech collection and modeled participation. The ads achieved over one million views and generated 25,000 visits to the landing page, of which 20% resulted in voice donations. Subsequently, fieldwork was carried out across 17 Galician municipalities, focusing on demographic diversity and regional dialects. By the end of 2024, 151 new hours of recorded speech had been collected, more than doubling the existing corpus. An important observation was that participation rose sharply during and following dissemination campaigns and slowly outside of active campaigns. Figures 2a and 2b summarize these milestones.

5.4. Academic Volunteer Networks

A distinctive feature of the Galician model is its integration with educational programs. As of late 2024, a university credit scheme enables Galician-speaking students to earn academic recognition towards their undergraduate degree by contributing voice recording and validation on the platform,

[21] https://doagalego.nos.gal



(a) Recorded hours (b) Contributors

Figure 2: *Evolution of the Galician presence in MCV.* Sharp rises in recorded hours and unique contributors correspond to dissemination and voice collection campaigns.

framing data donation as both an introduction to language technology, and civic participation in digital language preservation. Monitoring through user dashboards allowed transparent tracking of individual effort and performance. From the launch of the program, 40 students contributed more than 60 hours of voice recording and validated audio.

Additionally, to encourage and systematize quality control, the ILG launched collective half-day validation marathons known as *Validatóns*, open to the students of Galician language at the USC. These sessions combined awareness-raising, data gathering and practical training, ensuring that participants applied consistent linguistic and technical validation criteria. The 2024 edition validated more than 23,000 fragments in a single day, or 40 hours (three times the typical daily target). The 2025 event added another 10,000 validated audio fragments, together with 6 recorded hours; also, participants created 1,366 new sentences to be read aloud.

5.5. Corporate and Public-Sector Engagement

In 2025, the *Faino x Nós* campaign extended participation to the corporate sphere. 38 employees of the data-centered company SDG Group donated more than 10 hours of recordings within their workplace, linking voice contribution to corporate social responsibility and data ethics. This model demonstrated how partnerships with industry can generate both data and visibility, reinforcing the notion of shared linguistic responsibility in technological ecosystems.

Beyond data quality, these events fostered a sense of linguistic community, a key factor in sustaining participation.

6. Methodological Reflections

6.1. Representativeness and Metadata

A crucial methodological challenge lies in achieving representativeness across sociolinguistic variables. Systematic metadata collection remains essential to guarantee transparency and reproducibility. Common Voice stores basic demographic attributes voluntarily provided by users; in the Galician context, these were complemented internally by orientative field notes on the gender, age and accent of participants in on-site voice collection sessions. However, such contextual records cannot be integrated into the platform metadata and therefore remain external to the publicly available dataset.

In the case of MCV-22-gl, the dataset contains a total of 262,400 clips corresponding to approximately 364.54 hours of recorded speech. Among these, 136.74 hours are validated by at least two independent users, representing 37% of the total available recordings.

Gender distribution. According to the metadata release of MCV-22-gl, the dataset includes 6,610 unique contributors, of whom 682 (10.3%) identify as female, 366 (5.5%) as male, 6 (0.1%) as other, and 5,556 (84.0%) do not have gender metadata associated. However, metadata at the level of individual voice recordings shows that contributors do not participate equally in terms of the amount of audio they provide. As a result, the gender distribution of contributors does not directly reflect the gender distribution of the recordings included in the final dataset: 34,348 clips (13.09%) are tagged as *male*, 56,595 (24.57%) as *female*, and 1,229 (0.47%) as *other* or *nonbinary*, while 170,228 clips (64.87%) lack any gender information. Figure 3 illustrates this distribution.

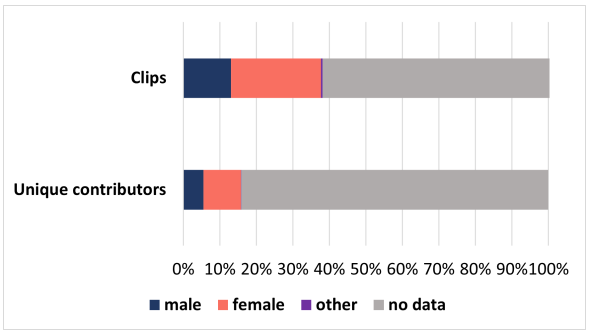


Figure 3: *Gender distribution of recorded clips and contributors.*

The imbalance between specified and unspecified metadata reflects a structural limitation of voluntary participation: contributors tend to participate without registering a profile, or otherwise may skip

demographic questions, resulting in significant underreporting. From a model-training perspective, this opacity can lead to hidden biases, as speaker distributions by gender cannot be reliably weighted.

A large part of these results can be linked to the conditions under which the public dissemination and voice donation campaigns were carried out. Because many participants recorded their contributions on shared or public devices, concerns about privacy and data security often discouraged user registration. As a result, the majority of contributors chose to donate their voices anonymously, which significantly reduced the amount of demographic metadata associated with individual recordings. This underscores not only the need for clear communication about privacy safeguards, but also the structural limitations of large-scale in-person collection campaigns. Within the MCV framework, such campaigns are highly effective for initial dissemination and awareness-raising; however, long-term implication and consistent metadata collection requires sustained engagement, encouraging participants to continue donating their voices from home, on private devices, through registered user profiles that allow the inclusion of demographic metadata.

Age distribution. Age statistics in MCV-22-gl show that 24.67% of recordings come from contributors aged between 0–39, 19.97% from those aged 40–69, and only 0.05% from contributors over 70. Again, more than half of the total data collected lacks explicit age metadata. Figure 4 illustrates the age distribution of voice clips and unique contributors throughout the dataset.

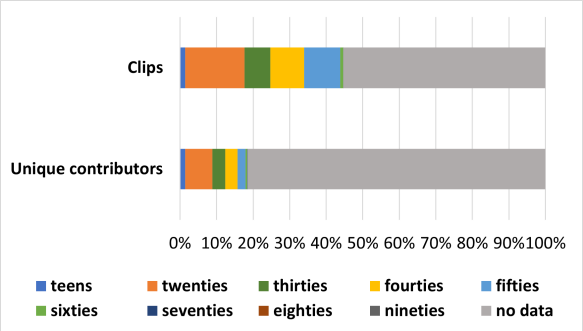


Figure 4: *Age distribution of recorded clips and contributors.*

The age imbalance in the registered metadata is typical of digital participation patterns, with younger speakers overrepresented in online and mobile-based initiatives. Several factors contribute to this phenomenon: older adults are often less exposed to university-based or social media campaigns, may feel less comfortable with digital interfaces, and in some cases lack regular access to the Internet or recording devices. Additionally, privacy concerns and unfamiliarity with the registration process can further discourage participation. For ASR development, this skew affects model robustness in recognizing the voices of elderly speakers, who may exhibit distinct acoustic, articulatory, and prosodic characteristics.

Dialectal variation. The Galician corpus currently lacks explicit annotation of dialectal or regional variants or integration of lexical dialectal traits, a limitation that reduces its representativity. While MCV introduced variant tagging for its written corpora in 2022, implementing such annotation for Galician speech data poses challenges. The dialectal continuum between western and eastern varieties does not correspond to discrete, easily classifiable boundaries, and contributors are not always able to self-identify with a specific regional label. MCV does not capture additional metadata such as speakers' place of origin.

Moreover, phonetic traits that characterize Galician varieties, such as *gheada* and *seseo*, are often attenuated or neutralized in read speech, even when they remain salient in spontaneous oral interaction. These factors make dialect tagging in Common Voice both technically and sociolinguistically complex. Nonetheless, on-site voice collection campaign records indicate a considerable geographical spread among contributors, including coastal, inland, and urban speakers, suggesting that a degree of dialectal diversity is already reflected in the data, although lacking the identifying metadata. Specific dialectally marked text subcorpora and adopting MCV variant tagging, despite described difficulties, might help address this issue in the future. In addition, variant-identification ASR models might prove helpful in annotating already recorded audios.

MCV-22-gl and other European languages. In MCV-22, Galician ranked among the eight languages with the highest number of recorded hours (364.54 hrs), positioning it above many European languages and within the upper tier of the platform in absolute terms. However, its validated portion (136.74 hrs, 37%) was comparatively modest, well below the validation rates of most languages within this group. Figure 5 offers a visual comparison.

As figures 6 and 7 illustrate, in terms of gender and age metadata of recorded clips, Galician presents a markedly different demographic profile compared to the remaining 14 languages with the highest volume of recorded data. Galician stands out for its high proportion of recordings without declared gender (65%) or age (55%). These values are substantially higher than in most other top-tier languages (typically between 24% and 47%).

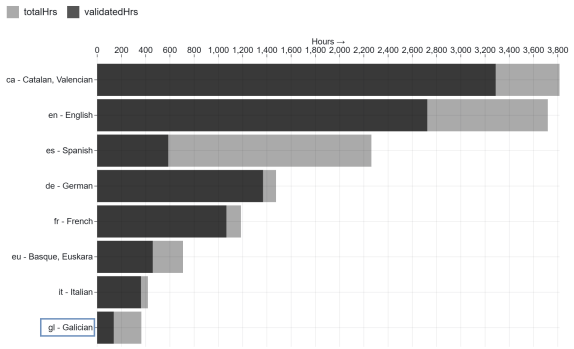


Figure 5: *Recorded and validated hours for the 8 European languages with most data in MCV-22.*

Among recordings with specified gender, female recordings surpass male recordings, contrasting with the male-dominated distributions observed in most other high-volume languages. In terms of age, participation concentrates mainly in the twenties and fifties, with very limited representation of younger teenagers or older speakers. Overall, the distinctiveness of the Galician corpus lies less in clear gender and age imbalance than in pronounced metadata sparsity.

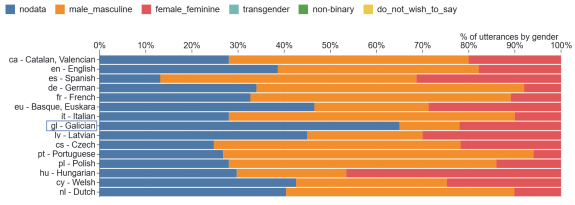


Figure 6: *Gender distribution of recorded clips for the 15 European languages with most data in MCV-22.*

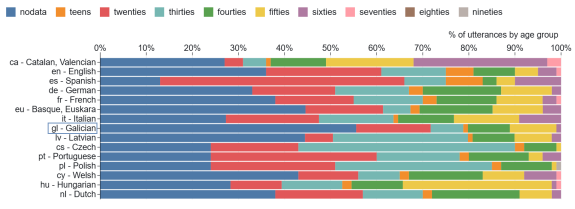


Figure 7: *Age distribution of recorded clips for the 15 European languages with most data in MCV-22.*

Current State of the Common Voice-GL dataset. In the latest, yet unreleased, version of MCV, the Galician language presents a significant advance in the total number of sentences available for recording (1,001,996 sentences). This places it close to Spanish (1,078,181 sentences) and Catalan (1,162,365 sentences), two of the most prominent languages in the Common Voice corpus and with the greatest number of resources for the development of linguistic technologies in the Iberian peninsula.

The Galician corpus currently comprises approximately 409 hours of recordings contributed by 6,866 speakers. Since the publication of MCV-22-gl, four professional linguists have performed a massive validation task, bringing the total number of validated hours to 282[22]. The current validation rate stands at around 69%, so encouraging collaboration not only in the recording of sentences but also in their subsequent validation continues essential.

As commented briefly in Section 5, a new functionality was recently added to the Common Voice collaborative platform: the collection of spontaneous speech. By answering a set of prompts, volunteers can record short and spontaneous answers that will register their dialectal variety. Similarly to the read speech functionality, community collaboration is key in all aspects, as transcribing the collected speech fragments and validating said transcriptions is also done collaboratively on the platform. Currently, the Galician spontaneous speech dataset contains 2,184 clips representing 5 hours of recorded speech (1.88 hours transcribed) from 109 speakers[23].

Discussion. From a methodological standpoint, the Galician dataset exemplifies the delicate balance to be achieved between scale and representativeness in open voice corpora. On the one hand, the use of a standardized, user-friendly global infrastructure ensures usability and comparability, and allows integration into multilingual models. On the other: read-speech limits the diversity of the corpus by reducing or eliminating oral phenomena; reliance on self-reported metadata, voluntary participation, and the manner in which contributors are recruited reduces the control over corpus diversity; and continued effort is needed to maintain and increase interest and public contribution.

6.2. Data Quality, Validation, and Ethics

The dual validation system ensures minimal consensus but does not replace expert review. Human variability in judgments, driven by natural language diversity and variability of recording conditions, can affect the reliability of acceptance ratios. To mitigate this, our team of Galician linguists developed internal validation guidelines based on linguistic correctness, intelligibility, and technical adequacy, establishing a reproducible framework for volunteer training. In addition, as already mentioned, a large part of the available voice recordings were

[22] Data retrieved Mar 2, 2026
[23] Data retrieved Feb 20, 2026

validated by paid contributors with a background in Galician linguistics in order to ensure the quality of the corpus.

Ethical transparency was also central: the data used to create the text datasets was freely donated to the project by the respective rights owners, and all contributors were informed that recordings were openly licensed and reusable for research and public-benefit technologies, preventing misconceptions about data ownership.

7. Discussion and Conclusions

The Galician case demonstrates both the potential and the structural constraints of collaborative speech corpus building within MCV. Through large-scale dissemination campaigns, academic volunteer networks, validation marathons, and other types of collaboration, the initiative significantly increased the volume of recorded data. However, analysis of MCV-22-gl shows that the corpus remains modest in scale and limited in demographic and dialectal traceability, with implications for downstream ASR robustness and bias control.

A first structural tension concerns the volume of recorded versus validated audio. In MCV-22, Galician ranked among the eight languages with the highest number of recorded hours but had a lower validation rate than most languages in the same tier. This imbalance reflects a recurrent bottleneck in open pipelines: while recording can be rapidly boosted through high-impact campaigns, validation requires sustained capacity and consistent engagement. Professional linguist validation can have a significant effect, although community validation remains essential to sustain growth.

A second axis involves representativeness and metadata completeness. The Galician corpus exhibits particularly high levels of missing demographic metadata compared to other high-volume languages in MCV-22. This limits the capacity to assess speaker balance and mitigate potential bias in model training. The issue is closely linked to collection conditions: on-site campaigns relying on shared devices and privacy-sensitive participants reduce registration rates and thus demographic traceability. Sustained at-home participation through registered profiles appears essential for improving metadata quality.

A third constraint derives from the reliance on read speech. Although methodologically controlled and well suited for ASR-oriented corpora, read-aloud data attenuates dialectal traits and does not naturally capture phenomena such as code-switching. The absence of explicit dialectal annotation further restricts sociolinguistic analysis. While the introduction of spontaneous speech in MCV opens a complementary path, the current Galician spontaneous dataset remains limited in scale.

Finally, the initiative highlights the importance of institutional governance and ethical transparency. CC0 licensing and FAIR alignment ensure reusability, but require careful sentence sourcing and extensive linguist-led curation, demonstrating that open participation must be complemented by structured oversight.

Despite the identified limitations, in the Galician case, Common Voice has proven to offer a robust infrastructure for gathering data that can lead to the development of equitable multilingual AI. The Galician corpus is now integrated into multilingual training pipelines and contributes to speech technologies such as Mozilla's own fine-tuned Whisper ASR models[24]. Ultimately, the Galician experience in MCV shows that, although challenges remain, sustained institutional collaboration and community engagement can make Common Voice a stable and practical framework for developing open, representative speech resources for lesser-resourced languages.

8. Acknowledgements

This research was carried out within *Proxecto Nós*, funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the projects ILENIA (ref. 2022/TL22/00215337) and Desarrollo Modelos ALIA.

We gratefully acknowledge the commitment of Xunta de Galicia and its Secretaría Xeral da Lingua, as well as the departments of Normalización Lingüística of the participating municipalities for their essential role in the on-site voice collection campaign. We also sincerely thank the Instituto da Lingua Galega (ILG) for its continued institutional support throughout the different phases of this initiative, and its members, as well as the members of Proxecto Nós, for their collaboration. Our gratitude likewise extends to the University of Santiago de Compostela (USC) for its broader institutional commitment to advancing the Galician language in the digital sphere.

We wish to express our special gratitude to Francis Tyers and the Mozilla Foundation team for their technical assistance and continued guidance throughout the integration of Galician into the Common Voice platform. Their expertise was instrumental in ensuring the growth and visibility of the

[24] https://huggingface.co/mozilla-ai/whisper-large-v3-gl, https://huggingface.co/mozilla-ai/whisper-small-gl, https://huggingface.co/mozilla-ai/whisper-large-v3-turbo-gl

Galician dataset within the global Common Voice ecosystem.

We are also deeply indebted to Noelia Díaz García, Daniel Fernández López, and Patricia Ramos Maceiras for their meticulous work and long hours devoted to the curation and correction of the Galician text corpus used in Common Voice. Our thanks also extend to all the public and private institutions that generously donated textual materials, and to the members of Proxecto Trasno, whose early efforts enabled the initial stages of Galician participation in Common Voice.

We acknowledge the invaluable contribution of all those who participated in the dissemination and voice collection events across Galicia, which played a crucial role in mobilizing contributors and increasing public engagement. Special thanks are due to Rubén Cela for his ongoing dedication, coordination, and commitment to the long-term success and continuity of this project.

Finally, we express our deepest gratitude to the entire community of volunteers and collaborators whose recordings and validations made this corpus possible. Their collective contribution embodies the spirit of open, community-driven language technology that underpins this project.

9. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus.

Carme Armentano-Oller, Montserrat Marimon, and Marta Villegas. 2024. Becoming a High-Resource Language in Speech: The Catalan Case in the Common Voice Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2142–2148, Torino, Italia. ELRA and ICCL.

Eduardo Rodríguez Banga, Carmen García-Mateo, Francisco Méndez-Pazó, Manuel González-González, and Carmen Magariños. 2012. Cotovía: an open source TTS for Galician and Spanish. In *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH*, pages 308–315.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The Nós Project: Opening routes for the Galician language in the field of language technologies. In *Proc. of the Workshop Towards Digit. Lang. Equality within the 13th Lang. Resour. and Eval. Conf.*, pages 52–61, Marseille, France. ELRA.

Jan Krewer. 2023. Creating Community-Driven Datasets: Insights from Mozilla Common Voice Activities in East Africa. Technical report, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn, Germany.

Baybars Külebi, Inma Hernáez, Elisa Fernández Rei, Andres Montoyo, Sarah Solito, Carme Armentano-Oller, Javier Hernando, Eva Navas, Carmen Magariños, Adina Vladu, Ibon Saratxaga, Jon Sánchez, Victor García Romillo, Asier Herranz, Christoforos Souganidis, Noelia García, Antonio Moscoso Sánchez, Xose Luis Regueira, Francisc Dubert, and Yoan Gutiérrez. 2024. Speech Technologies in the ILENIA Project: Generating Resources to Develop Voice Applications in the Official Languages of Spain. In *IberSPEECH 2024*, pages 289–292.

Andres Pineiro-Martin, Carmen Garcia-Mateo, Laura Docio-Fernandez, Maria del Carmen Lopez-Perez, and Jose Gandarela-Rodriguez. 2024. FalAI: A dataset for end-to-end spoken language understanding in a low-resource scenario. In *Proc. of the 2024 Joint Int. Conf. on Comput. Ling., Lang. Resour. and Eval. (LREC-COLING 2024)*, pages 7107–7116, Torino, Italy. ELRA and ICCL.

José Manuel Ramírez Sánchez and Carmen García-Mateo. 2022. Report on the Galician Language. Report D1.15, ELE.

Adina Ioana Vladu, Iria de Dios-Flores, Carmen Magariños, John E. Ortega, José Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. Proxecto Nós: Artificial intelligence at the service of the Galician language. In *CEUR Workshop Proc.*, volume 3224, pages 26–30. CEUR-WS.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020b. Covost 2: A massively multilin-

gual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Mark D Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1).

10. Language Resource References

Magariños, Carmen and Adrián Vidal Miguéns and Adina Ioana Vladu and Noelia García Díaz and Marta Vázquez Abuín and Ainhoa Vivel Couso and Daniel Bardanca and Elisa Fernández Rei. 2023. *Nos_ParlaSpeech-GL: Galician ASR corpus*. Zenodo. PID https://doi.org/10.5281/zenodo.7913218. Dataset. Available under CC BY 4.0 license.

Vladu, Adina Ioana and Marta Vázquez Abuín and Elisa Fernández Rei and Noelia García Díaz and Adrián Vidal Miguéns and Carmen Magariños. 2023. *Nos_TranscriSpeech-GL: Galician ASR corpus*. Zenodo. PID https://doi.org/10.5281/zenodo.7717140. Dataset. Available under CC BY 4.0 license.

Vladu, Adina Ioana and García Díaz, Noelia and Regueira Fernández, Xosé Luís and Magariños, Carmen and Moscoso Sánchez, Antonio and Fernández López, Daniel and Fernández Rei, Elisa and Dubert-García, Francisco. 2025. *Nos_Brais-GL: Galician TTS corpus*. Zenodo. PID https://doi.org/10.5281/zenodo.14265241. Dataset. Available under CC BY 4.0 license.

Vázquez Abuín, Marta and García Díaz, Noelia and Vladu, Adina Ioana and Magariños, Carmen and Vidal Miguéns, Adrián and Fernández Rei, Elisa. 2023. *Nos_Celtia-GL: Galician TTS corpus (1.0.0.)*. Zenodo. PID https://doi.org/10.5281/zenodo.7716958. Dataset. Available under CC BY 4.0 license.