Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning

Anonymous ACL submission

Abstract

Our quality audit for three widely used public multilingual speech datasets—Mozilla Common Voice 17.0, FLEURS, and VoxPopuli—shows that in some languages, these datasets suffer from significant quality issues. We believe addressing these issues will make these datasets more useful as evaluation sets, and improve downstream models. We divide these quality issues into two categories: micro-level and macro-level. We find that macro-level issues are more prevalent in less institutionalized, often under-resourced languages. We provide a case analysis of Taiwanese Southern Min (nan_tw) that highlights the need for proactive language planning (e.g. orthography prescriptions, dialect boundary definition) and enhanced data quality control in the process of Automatic Speech Recognition (ASR) dataset creation. We conclude by proposing guidelines and recommendations to mitigate these issues in future dataset development, emphasizing the importance of sociolinguistic awareness in creating robust and reliable speech data resources.

1 Introduction

The emergence of massively multilingual speech datasets has significantly advanced the performance of various speech technologies in recent years, particularly for low-resource languages. These datasets are crucial for training and evaluating state-of-the-art ASR models like Whisper (Radford et al., 2023), Google USM (Zhang et al., 2023), SeamlessM4T (Barrault et al., 2023), MMS (Pratap et al., 2024), and Gemini (Gemini Team et al., 2024), and also enable advances in cross-lingual speech representation learning (Babu et al., 2022; Conneau et al., 2021) and downstream applications like multilingual speech generation and understanding (Le et al., 2023; Rubenstein et al., 2023). However, despite their growing importance, the quality of these datasets remains surprisingly under-researched.

Prior work on data collection and curation (Penedo et al., 2024; Goyal et al., 2022; Kreutzer et al., 2022) has acknowledged the generally lower quality of web-scraped data, but these efforts primarily focused on text. Similarly, research on ASR data augmentation for low-resource languages (Casanova et al., 2023; Bartelds et al., 2023; Tsoukala et al., 2023) has not addressed data quality issues adequately. The community-driven Mozilla Common Voice project (Ardila et al., 2020), for example, lacks well-documented quality control processes for its diverse text sources (Wikipedia and volunteer contributions) and subsequent audio recordings, which makes the quality and reliability largely unknown.

Inspired by Kreutzer et al. (2022)'s audit methodology for text datasets, we conduct a thorough quality assessment of three widely used multilingual speech datasets: Mozilla Common Voice 17.0 (MCV17, Ardila et al. 2020), FLEURS (Conneau et al., 2023), and VoxPopuli (Wang et al., 2021). Our investigation employs both quantitative and qualitative methods. We calculate metrics including Signal-to-Noise Ratio (SNR), Voice Activity Detection (VAD), median utterance duration, and median word count for each language subset. Additionally, we asked native speaker volunteers, covering around 40 languages, detailed in Table 16 in Appendix, to review 100 randomly sampled sentences (text and audio) for coherence, audio-text alignment, dialect, topic domain, and language ID from each language subsets of the datasets.

Our analysis reveals serious data quality issues, particularly in under-resourced, less-institutionalized languages (Sections 3 and 4). As one example, the nan_tw (Taiwanese Southern Min) subset in MCV17 showcases a multitude of these issues, rendering it nearly unusable without significant cleaning or restructuring (Section 5).

We find a strong positive correlation between a language's institutionalization status and its dataset quality, a crucial factor often overlooked in ASR research. We discuss the impact of these issues on downstream research and applications (Section 6), and propose mitigation guidelines (Section 6.2). We draw conclusions and suggest future work in Section 7.

2 Datasets

In this work, we study three multilingual datasets: Mozilla Common Voice 17.0 (MCV17), FLEURS, and VoxPopuli.

- Mozilla Common Voice (Ardila et al., 2020) is a community-driven project proposed by Mozilla. Sentence sourcing, recording and reviewing are all contributed by volunteers. Version 17 supports 124 locales.

- FLEURS (Conneau et al., 2023) uses the text from the machine translation corpus FLoRes-101 (Goyal et al., 2022). Sentences are extracted from the English Wikipedia and translated to 101 languages by professional translators. Each sentence is recorded by 3 native speakers and invalid recordings are discarded.

- VoxPopuli (Wang et al., 2021) uses speech and transcriptions from European Parliament event recordings. The paired data contains 16 European languages.

3 Micro-Level Issues

Micro-level issues in ASR datasets typically stem from inadequate quality control, supervision, or management during data collection and curation. They are often detectable via automatic metrics and can be mitigated programmatically. Micro-level issues are language-agnostic and may exist in any language subsets.

3.1 Extremely Short Duration

In VoxPopuli the median utterance duration ranges from around 6 to 13 seconds and FLEURS ranges from 9 to 24 seconds. However, MCV17 displays a concerning trend of extremely short utterances. Detailed plots are shown in Appendix C. In MCV17, 35 languages have median utterance duration under 4 seconds. And for most languages, over 99% of the utterances are under 10 seconds. We also find some extreme cases such as `nan_tw` (Southern Min, Taiwan), `sr` (Serbian), and `br` (Breton), whose median utterance durations are below 3 seconds, and 99% of the utterances are below 7 seconds. We manually inspect those language subsets and discover that the utterances of these languages are mostly short phrases or isolated words, as shown in Figure 5. This reveals a lack of quality control of text-prompts in MCV17. Why does this matter? Today's ASR and TTS models sometimes fail to generalize to inputs of lengths not seen in training (Narayanan et al., 2019; Varis and Bojar, 2021). Without this insight, a model trained on MCV17 `nan_tw` might fail on longform tasks like video subtitling. Mitigation is possible on the level of model architecture (e.g. more flexible attention variants).

3.2 Low Proportion of Speech

To assess the proportion of actual speech content within the audio data, we employed a neural model to classify speech and non-speech segments in all three datasets. While VoxPopuli generally demonstrates high speech content, with all 14 languages having at least 89% speech, we identified several languages in other datasets with significantly lower proportions, specifically, Basaa (`bas`), Zaza (`zza`), Serbian (`sr`) in MCV17, and Danish (`da_dk`) in FLEURS exhibiting less than 50% speech content, detailed in Appendix D. We manually inspected these languages and found two primary causes:

- The speaker's voice is too distant from the microphone, resulting in poor audio capture (e.g. the `da_dk` training set in FLEURS).

- As pointed out in Section 3.1, the short text prompts led to a disproportionate amount of silence between the start and stop of the recording process, as contributors interacted with their recording devices.

Similar issues have also been reported by the public[1] which further validates our findings. This issue combined with extremely short utterance durations severely limits the amount of usable speech data. For instance, while `nan_tw` in MCV17 contains 21 hours of audio, only 48.3% constitutes actual speech, resulting in merely 10 hours of usable data. While in general this micro-level issue affects TTS more than ASR, using the raw audio duration e.g. for re-balancing the amounts of training data per

[1] A large number of incorrect audio samples on FLEURS reported in April 2023 https://huggingface.co/datasets/google/fleurs/discussions/16

language can lead to unexpected results. This can be mitigated by using the number of transcribed words (or tokens) as the most fundamental unit for quantifying speech data.

3.3 Imbalanced Topic Domains

Assessing topic domain balance is challenging to quantify programmatically. Therefore, we adopted a qualitative approach, sampling text prompts and corresponding audio data from each dataset. We asked native speakers to evaluate whether these sentences were representative of typical, everyday conversations—a common target domain for general-purpose ASR systems. However, we acknowledge that other domains (e.g., news broadcasts, academic lectures) might be relevant for specific applications. This investigation revealed significant topic domain imbalances in several languages within both MCV17 and FLEURS, particularly concerning the assumption of everyday conversation as the target. In FLEURS, the reliance on text prompts from the FLoRes-101 dataset (Goyal et al., 2022), ultimately sourced from Wikipedia, leads to a dominance of formal, literary, and encyclopedic sentences.

We also identified an issue in MCV17, where a number of sentences across multiple languages exhibit high repetitiveness, often lacking meaningful content and suggesting machine generation using fixed templates (see Table 11 in Appendix). It's important to acknowledge that these datasets are built on the valuable contributions of volunteers. Given the community-driven nature of text prompt creation in Common Voice, it is possible that these repetitive sentences were introduced due to varying interpretations of the guidelines or a lack of awareness about their potential impact on model training. This highlights the need for clearer guidelines and more robust review processes within community-driven projects.

The problem of domain mismatch has been discussed in the context of monolingual datasets in (Likhomanenko et al., 2021). In multilingual datasets, a poor coordination in topic selection across languages might lead to incorrect interpretation of ASR and TTS model quality. While there is no easy mitigation, at the very least this information should be exposed in machine-readable metadata.

3.4 Lack of Speaker Diversity

The final micro-level issue we identified is a lack of speaker diversity. We calculated the average recording time per speaker for each language in MCV17, as detailed in Tables 13 and 14 in the Appendix. Notably, Macedonian (`mk`) exhibits an average of 1.20 hours of audio per speaker, with only 19 unique speakers in total, indicating a high concentration of data from a limited number of contributors. Data of some languages such as Zulu (`zu`), Northern Sotho (`nso`), Haitian Creole (`ht`) consist of recordings from only a single speaker.

This introduces a high risk of overfitting to the speakers' age, gender and dialects, as well as a risk of an overlap between training and test data. Whenever the speaker statistics are available as metadata, dataset users can focus on the subset of languages with diverse data, which produce reliable results.

4 Macro-Level Issues

Macro-level issues often arise from overlooking a language's sociolinguistic context during dataset design. These issues are typically not detectable through automatic metrics, requiring manual inspection and linguistic expertise for diagnosis. They are particularly prevalent in less-institutionalized languages, often characterized by complex phenomena like digraphia or diglossia. Our analysis of VoxPopuli, which primarily includes more-institutionalized languages, further supports this claim, as it revealed no such macro-level issues.

4.1 Unspecified Writing System in Digraphic Languages

Digraphia is a sociolinguistic phenomenon where a single language is written using multiple writing systems (Dale, 1980). Contemporary examples include Serbian (Latin and Cyrillic scripts), Malay (Latin and Jawi scripts), and Punjabi (Gurmukhi and Shahmukhi scripts). Jung and Kim (2023) further categorize digraphia into two types: complementary and exclusive. Japanese is an well-known example of complementary digraphia, where four scripts, Kanji, Hiragana, Katakana, and Latin letters are combined in common Japanese texts. In ASR, exclusive digraphia introduces multiple parallel written forms that can create ambiguities in audio transcription at the script level, while complementary digraphia can cause ambiguities at the orthography / spelling level.

For ASR, the definition of "writing system" extends beyond script to include orthography and spelling rules. Minor variations such as "color" vs. "colour" in American and British English, can impact evaluations based on Word Error Rate (WER), where a spelling difference counts as a word error. While these variations have a limited effect on English ASR performance due to their infrequency and can be addressed with rule-based normalization, languages with multiple coexisting spelling standards for substantial portions of their vocabulary pose a significantly greater challenge, as discussed below.

4.1.1 Norwegian Bokmål / Nynorsk

Norwegian exemplifies a digraphic language with two synchronic written standards: Bokmål (nb_no) and Nynorsk (nn_no). Table 1 illustrates the differences with an example sentence. Notably, 4 out of 8 words differ in spelling, including common words like *jeg* (Bokmål) and *eg* (Nynorsk), both meaning "I" in English. Both MCV17 and FLEURS include Norwegian, but Common Voice uses nn_no (purportedly Nynorsk) while FLEURS uses nb_no (purportedly Bokmål). To verify the actual proportions of Bokmål and Nynorsk in these subsets, we employed a classification script detailed in Algorithm 1.

English	Have I covered myself with song and playing the harp.
Bokmål	Har **eg dekt** meg med **song** og **harpespel**.
Nynorsk	Har **jeg dekket** meg med **sang** og **harpespill**.

Table 1: Spelling differences between Bokmål and Nynorsk. In this example, a mismatching orthography of the same sentence will lead to a 50% WER even the transcription is completely "correct".

Table 2 presents the classification results. Neither MCV17 nor FLEURS contains purely Nynorsk or Bokmål as their language codes suggest. MCV17's nn_no subset contains 8.1% Bokmål sentences, while FLEURS' nb_no subset contains 8.8% Nynorsk. This finding suggests a possible lack of control over the sourcing of text sentences.

We hypothesize that neglecting the Bokmål / Nynorsk distinction in an ASR dataset significantly affects model evaluation using WER. To test this, we evaluated a Conformer Hybrid Autoregressive Transducer (HAT) model (Variani et al., 2020) of 120M parameters trained with Norwegian Bokmål data on both datasets. Table 3 shows the results. While the model exhibits similar deletion and insertion error rates across both datasets, the substitution error rate is nearly 25% higher on MCV17's nn_no subset. Manual inspection of the substitution errors revealed that most are indeed orthographic variants from Bokmål. This supports our hypothesis that mixing different writing systems in an ASR dataset significantly impacts downstream evaluations.

	MCV17 nn_no		FLEURS nb_no	
Nynorsk (nn_no)	764	(65.1%)	323	(**8.8%**)
Bokmål (nb_no)	96	(**8.1%**)	2682	(72.7%)
Mixed	161	(13.7%)	344	(9.3%)
Unmarked	153	(13.0%)	338	(9.1%)
Total sentences	1174	validated	3687	

Table 2: Classification of Norwegian text prompts in MCV17 and FLEURS.

WER [%]	MCV17 nn_no	FLEURS nb_no
Total	49.1	23.8
Del/Ins/Sub	11.8 / 1.6 / **35.0**	11.1 / 2.2 / **10.0**

Table 3: WER of a Norwegian Bokmål Conformer ASR model on MCV17 and FLEURS test splits.

4.1.2 Risks of Unverified Script Assumptions in Digraphic Languages

Digraphia is not a static structure; it evolves, sometimes rapidly. The "default" script for a given language can shift, as is currently occurring in several post-Soviet countries (Jung and Kim, 2023). For instance, Kazakhstan is transitioning from Cyrillic to Latin script by 2025, and Mongolia is restoring the Mongolian (Bichig) script by 2025 according to the government's plan. Neither MCV17 nor FLEURS include script codes in their locale codes, leading to implicit, unverified assumptions about the script used in the text data. These assumptions, listed in Table 15 in Appendix, introduce significant risks for downstream applications, potentially compromising dataset usability in the near future.

4.2 Ambiguous Register or Variety in Diglossic Languages

Diglossia describes a community's use of two distinct language varieties in a compartmentalized manner: a "High" (H) variety for typically formal contexts and a "Low" (L) variety for everyday conversation, with the community perceiving these varieties as a single language (Ferguson, 1996). This duality may lead to ambiguous language code interpretations in ASR dataset construction. For example, "I speak and write Chinese" could imply "I speak and write Mandarin", or "I speak Cantonese and write Mandarin". Such ambiguity can result in datasets pairing audio and text from mutually unintelligible varieties. Contemporary examples of diglossic languages include Standard Arabic (Fusha) and its vernacular dialects (Brustad, 2017; Ferguson, 1996), Hong Kong Chinese / Cantonese (Snow, 2010), Standard German / Swiss German, Classical Tibetan / vernacular Tibetan (Roche, 2017), Persian (Mahmoodi-Bakhtiari, 2018) and Bengali (Dil, 1986). Our analysis of MCV17 and FLEURS revealed serious issues with register and variant confusion, particularly in Arabic and Hong Kong Chinese / Cantonese. Since VoxPopuli does not include any diglossic languages, we find no such issues in VoxPopuli.

4.2.1 Arabic

Modern Standard Arabic (MSA, or Fusha) is a classic example of diglossia (Ferguson, 1959), employed in formal contexts such as religion and education, while regional dialects (e.g. Āmmiyya / Dārija) prevail in everyday conversation. These dialects can differ significantly from MSA (Høigilt and Mejdell, 2017). We analyzed the Arabic text prompts in MCV17 (`ar`) and FLEURS (`ar_eg`) using a classification tool (Algorithm 2). Table 4 shows that both datasets mainly contain MSA, with some dialectal Arabic or mixed forms in MCV17's `ar` subset. Interestingly, FLEURS, labeled `ar_eg` (presumably Egyptian Arabic), contains almost exclusively MSA. This could be due to a combination of factors: the intended dataset composition, the selection of source material, and the transcribers' interpretation of the term "Arabic", which can be ambiguous in a diglossic context. This highlights the complex interplay between dataset composition, content selection, and transcribers' preference and understanding of the task in diglossic situations.

	MCV17 `ar`		FLEURS `ar_eg`	
Fusha	**5963**	**(76.3%)**	**2787**	**(98.6%)**
Dialect	991	(1.2%)	0	
Mixed	1648	(2.1%)	25	(0.88%)
Unmarked	15881	(20.3%)	15	(0.53%)
#sentences	78157	validated	2827	

Table 4: Classification of text prompts in MCV17 `ar` and FLEURS `ar_eg`. Nearly all data of FLEURS `ar_eg` is in Fusha.

We speculate that the prevalence of MSA in FLEURS, despite the `ar_eg` label, might be partly a consequence of adopting a strict language code formatting standard that does not adequately represent the nuances of diglossic languages. While regional codes exist within the ISO 639-3 standard, there is no widely accepted code for MSA within the ISO 3166-1 alpha-2 country code framework, which is often used in conjunction with language codes to form locale identifiers. The dataset creators might have chosen `eg` (Egypt) as the most readily available option to satisfy the IETF BCP 47 language tag specifications, despite it not accurately reflecting the linguistic reality of MSA as a supra-regional standard. This exemplifies a broader challenge: the limitations of current language classification systems (particularly within ISO 639-3) in representing regional and supra-regional varieties, especially those that do not align neatly with national boundaries. For instance, there is no distinct code for specific varieties of British English (e.g., Scottish English) or for African American Vernacular English (AAVE) within the current framework, hindering the development of ASR or TTS systems tailored to these communities. The Arabic case in FLEURS, therefore, serves as a microcosm of the larger issue: the need for more flexible and nuanced language classification systems to capture the full spectrum of linguistic diversity, particularly for diglossic languages and regional or supra-regional varieties.

4.2.2 Hong Kong Chinese / Cantonese

Hong Kong presents another case of diglossia (Snow, 2010), where Standard Written Chinese (SWC), a written register largely based on Modern Standard Mandarin, is used in formal writing and spoken Cantonese is used in everyday conversation. According to Unicode CLDR version 44.0,[2] `zh` is interpreted as "Chinese", referring to SWC, while Cantonese has its own ISO 639-3 code, `yue`. However, the ambiguous nature of the code `zh` often leads to inconsistent interpretations in the industry.

We used the `canto-filter` package[3] developed by Lau et al. (2024) to classify text prompts into four categories. The results, shown in Table 5, highlight significant inconsistencies across datasets. Common Voice's `yue` subset aligns relatively well with spoken Cantonese, with most prompts being in written Cantonese. However, Common Voice's `zh_hk` subset contains a mixture of SWC and Cantonese, while FLEURS' `yue_hk` subset consists almost entirely of SWC. This suggests that the `yue_hk` label in FLEURS is a misnomer and should likely be `zh_hk`.

[2] Territory-Language Information
[3] pypi.org/project/canto-filter/

	MCV17 zh_hk		MCV17 yue		FLEURS yue_hk	
SWC	**8851**	**(9.6%)**	15	(0.1%)	**2803**	**(89.8%)**
Cantonese	37357	(40.3%)	16466	(75.7%)	**0**	
Mixed	299	(0.3%)	40	(0.2%)	0	
Unmarked	46113	(49.8%)	5238	(24.1%)	317	(10.2%)
#sentences	92620		21759		3120	

Table 5: Classification of text prompts in MCV17 yue, zh_hk, and FLEURS yue_hk. None of the FLEURS yue_hk is Cantonese.

4.3 Ambiguous Scoping of Target Dialect Continuum

Both MCV17 and FLEURS utilize ISO 639-3 codes to identify languages. However, neither dataset provides detailed metadata describing the specific dialects represented within each language subset, leaving the interpretation of these codes open to contributors. Our investigation revealed that the ff_sn (Fula, Senegal) subset in FLEURS only includes the Peul dialect spoken across Senegal. This is significant because the Fula variety of the most population is typically the Guinean variant, not the one found in the dataset. Another case we identify is the kea_cv (Cape Verdean Creole, a.k.a Kabuverdianu) in FLEURS, which consists entirely of the Sotavento (Southern Islands) variant. The omission of the Barlavento (Northern Islands) variant could limit the dataset's usefulness for developing ASR and TTS systems that are robust across the entire dialect continuum of Cape Verdean Creole.

The lack of specificity in dialect scoping can lead to significant ambiguity regarding the actual dialectal scope of the data, potentially skewing the representativeness of the dataset for the target language as a whole. Researchers and developers might assume that a dataset labeled with a particular code encompasses the full range of variation within that language, when in reality it might only represent a specific dialect or a subset of dialects. This can result in inflated performance metrics if ASR models are evaluated only on the dialects present in the training data, while performance on other dialects remains unknown and potentially lower. In the context of TTS, having no control over the exact dialect might strike some sensitive nerve among native speakers, if the dialect boundaries align with socio-economic disparities or historical and geopolitical injustices.

Mitigation is possible to a certain degree if the metadata contains the dialect annotations for every utterance.

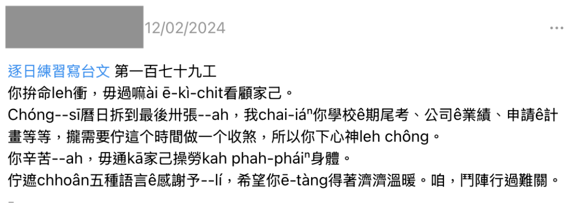

Figure 1: A Threads post written in Taiwanese Southern Min. Sinographs and Latin letters are used interchangeably without a codified convention.

5 An Extreme Case Analysis: MCV17 nan_tw (Taiwanese Southern Min)

5.1 Social and Historical Context

Taiwanese Southern Min (TSM) has historically been a predominantly unwritten language. Various writing systems emerged sporadically, including full Sinograph (a.k.a Chinese characters or Han script), Tâi-lô romanization, Church romanization (Peˈh-ōe-jī), and a mixture of Sinographs and romanization (Ota, 2005; Alivin, 1999). While the Taiwan Ministry of Education introduced the *Taiwanese Southern Min Recommended Characters*[4] between 2007 and 2009, and a *Dictionary of Frequently-Used Taiwan Minnan*[5] to promote a standard orthography using full Sinographs, as of early 2025, the language community has not yet reached a consensus. As examples, a substantial proportion of pages of Southern Min Wikipedia use Latin letters, while everyday usage on social media platforms like Threads (Figure 1) displays a varied mix of Sinographs and romanization. Unlike the more established complementary digraphia in Japanese, TSM lacks widely adopted rules for script choice, resulting in a more undetermined and evolving orthographic landscape.

5.2 Data Quality Issues

Table 12 in Appendix presents a snapshot of the nan_tw subset in MCV17, revealing critical issues that severely compromise its usability:

1. **Dictionary Structure:** The dataset resembles a dictionary dump, with nearly every entry consisting of single words or short phrases, rather than complete sentences.

2. **Duplicate Text Prompts:** Each prompt is written in both full Sinograph and full Latin (Tâi-lô), resulting in redundant entries.

[4] 臺灣閩南語推薦用字700字表
[5] 教育部臺灣台語常用詞辭典

3. **Text-Audio Misalignment:** Voice contributors apparently read each prompt (single words or phrases) only once, leading to misalignment between the text (containing both writing systems) and the corresponding audio.

Furthermore, many validated sentences (text prompts not yet paired with audio) are in `zh_cn` (Mandarin Chinese in simplified characters), posing a significant risk of language contamination and potentially introducing the "mixed-language" issues discussed in Section 4.2.2.

5.3 Root Causes of the `nan_tw` Data Issues

To understand the rationale behind the dictionary-style structure and dual-script representation in MCV17 `nan_tw`, we contacted the voice and text contributors. They explained that most TSM speakers are not proficient in reading and writing TSM. Moreover, among those who can read and write the language, preferences and proficiency levels for different writing standards vary considerably. Some are only comfortable with Sinographs, others with Romanization, while some prefer a mixture of both. To maximize participation and facilitate data collection, the contributors opted to include both Sinographs and Romanization, aiming to ensure that all potential contributors could read the prompts.

This situation highlights a common challenge faced by many under-resource or less-institutionalized languages when developing ASR datasets. ASR fundamentally involves transcribing audio, the spoken form, into text, the written form. But what if the language has no "written form"? The motivations of building ASR for such languages may differ fundamentally from those with well-established written traditions, necessitating distinct approaches to dataset construction. We will explore their implications in Section 6.2.

6 Discussions

6.1 Impacts on Downstream Research and Applications

The quality issues discussed previously have significant ramifications for downstream research and applications. Shorter utterance lengths correlate with higher WER (Li et al., 2023, 2022), and excessive silence negatively impacts emotion recognition (Perez et al., 2022). Lack of speaker and topic diversity can introduce biases related to gender, age, and regional accents (Feng et al., 2024; Garnerin et al., 2021). While micro-level issues are often detectable programmatically, macro-level issues, stemming from sociolinguistic factors, can have more insidious and far-reaching impacts.

A compelling example is presented in Costa-jussà et al. (2022), where a language identification (LangID) system, evaluated on the FLoRes-200 dataset (the text source for FLEURS), failed to distinguish between `zh_hk` (Hong Kong Chinese) and `yue` (Cantonese), see the third confusion matrix in Figure 9 in the original paper. This stems exactly from the "wrong language" issue in FLEURS' `yue_hk` subset, which predominantly contains Standard Written Chinese rather than Cantonese, as discussed in Section 4.2.2 and publicly noted.[6] Consequently, downstream models like Whisper-v3 struggle with these languages, exhibiting inconsistent outputs and unpredictable "auto-translations" between varieties.[7] Furthermore, model distillation can exacerbate these issues, as seen in the WER degradation from 10.8% to 46.1% on Common Voice 15.0 `yue`.[8]

6.2 Addressing Macro-Level Issues: The Role of Language Planning in ASR Dataset Creation

The macro-level issues identified in this study, unlike micro-level issues that can be mitigated programmatically (e.g. Rai et al. (2024)), necessitate a more fundamental approach to dataset creation, particularly for less-institutionalized languages. We argue that incorporating principles of language planning into the ASR dataset construction process is crucial for these languages.

Many widely spoken languages such as English, Spanish, Mandarin, are institutionalized and standardized due to a process called language planning (Cooper, 1989). This typically involves language planning agencies (LPAs), e.g. a country's education ministry, establishing a normative orthography, grammar, and lexicon to guide a diverse speech community. Consequently, speakers of these languages generally have a clear understanding of how to read and write their language "correctly". However, with over 7,000 spoken languages and only around 200 countries in the world, most languages have not enjoyed such privilege, leaving speakers without standardized written forms.

The decentralized, community-driven nature of

[6] github.com/facebookresearch/flores/issues/61
[7] github.com/openai/whisper/discussions/366
[8] github.com/openai/whisper/discussions/2363

Common Voice, while valuable for participation and diversity (Ardila et al., 2020), can exacerbate this issue. It can lead to implicit language planning decisions being made by communities without adequate expertise or consensus, as exemplified by the deicsion of merging Norwegian Nynorsk and Bokmål in Common Voice.[9] As Common Voice expands to more languages with complex sociolinguistic backgrounds, such as Konkani,[10] we anticipate a rise in these macro-level issues without proactive intervention.

To mitigate these challenges and guide the creation of future massively multilingual ASR datasets, particularly for under-resourced and less-institutionalized languages, we propose the following guidelines:

1. **Sociolinguistic Assessment and Intervention**: Before dataset creation, conduct a thorough sociolinguistic survey of the target language, including demographics, literacy rates, writing systems, diglossia/digraphia, and other relevant factors. If literacy is low or a standardized writing system is lacking, develop an intervention plan involving contributor training and detailed transcription guidelines to ensure dataset consistency.

2. **Collaborative Scope Definition**: Establish a framework involving language experts, linguists, and community members to define the precise scope of the dataset, including script, orthography, dialect, and register. Document these decisions transparently, ensuring they are informed by both expert knowledge and community needs.

3. **Multi-Level Quality Assurance**: Implement rigorous quality assurance throughout the dataset building process. This includes micro-level checks (e.g., filtering silent/noisy recordings, removing repetitive prompts) and macro-level checks (e.g., verifying adherence to the defined scope by rejecting recordings from non-target dialects or text in non-target orthographies).

4. **Comprehensive and Transparent Metadata**: Release each dataset with detailed metadata for each language subset. Building on the data statement practice by Bender and Friedman (2018), include macro-level information (e.g., chosen dialect, script, orthography) and micro-level statistics (e.g., average utterance duration, speaker demographics) to ensure transparency and inform downstream users.

When a language lacks a standardized or widely adoped written form, the motivation for building ASR systems may shift from building primarily a tech product to a tool for corpus planning and educational initiatives. As demonstrated by Williams et al. (2000) and Kumar et al. (2012), ASR can be employed to reduce illiteracy and promote the use of a standard written form within a community.

7 Conclusions

This study investigated the quality of prominent public ASR datasets, highlighting the critical need for sociolinguistic awareness, especially for less-institutionalized languages, and emphasizing the importance of incorporating language planning principles into dataset creation. We proposed guidelines for future ASR dataset creation, focusing on sociolinguistic assessment, informed scope definition, rigorous quality assurance, and comprehensive metadata release (Section 6.2).

For downstream users of existing datasets, we strongly recommend manual inspection, language verification by native speakers or linguists, and data cleaning/filtering before use. When training multilingual models, precise scoping of language/dialect variants is crucial to avoid mixed-language issues. Employing more flexible evaluation metrics than WER and CER, such as those proposed by Nigmatulina et al. (2020) and Karita et al. (2023), are also recommended.

Future work should address the limitations of current language classification systems, particularly the ISO 639-3 standard, in representing the full spectrum of linguistic diversity. More flexible and granular classification systems that capture the dynamic sociolinguistic nuances is crucial for creating ASR datasets that accurately reflect real-world language use. Furthermore, future research should focus on developing practical tools and frameworks for implementing our proposed guidelines, exploring the effectiveness of different language planning strategies in the context of ASR, and investigating how ASR can be leveraged to support language revitalization and educational initiatives, especially for under-resourced languages.

[9] discourse.mozilla.org/t/merging-norwegian-nynorsk-and-norwegian-bokmal/130474

[10] github.com/common-voice/common-voice/issues/4454

8 Limitations

While this study covered over 40 languages, a significant number remain uninspected in the evaluated datasets. Future work should extend our investigation to these languages and focus on developing practical tools and frameworks for implementing the proposed guidelines. Further research could also explore the effectiveness of various language planning strategies in the context of ASR and investigate how ASR can be leveraged to support language revitalization and educational initiatives, particularly for under-resourced and less-institutionalized languages.

9 Ethical Considerations

There are no known ethical concerns or risks associated with this work.

References

Lin Alivin. 1999. *Writing Taiwanese: The development of modern written Taiwanese*. Department of Asian and Middle Eastern Studies, University of Pennsylvania.

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. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, Incheon, Korea.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 715–729, Toronto, Canada.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Kristen Brustad. 2017. Diglossia as ideology. In *The politics of written language in the Arab world*, pages 41–67. Brill.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Interspeech*, pages 1244–1248, Dublin, Ireland.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*, Brno, Czechia.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Robert L. Cooper. 1989. *Language planning and social change*. Cambridge UniversityPress.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ian R.H. Dale. 1980. Digraphia. *International Journal of the Sociology of Language*, 1980(26):5–14.

Afia Dil. 1986. Diglossia in Bangla: A study of shifts in the verbal repertoire of the educated classes in Dhaka, Bangladesh. *The Fergusonian Impact*, 2:451–65.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Charles A. Ferguson. 1996. Epilogue: diglossia revisited. *Understanding Arabic: Essays in contemporary Arabic linguistics in honor of El-Said Badawi*, pages 49–67.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the impact of gender representation in ASR training data: A case study on Librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jacob Høigilt and Gunvor Mejdell. 2017. *The Politics of Written Language in the Arab World: Writing Change*. Brill, Leiden, The Netherlands.

Youngjoo Jung and Bora Kim. 2023. Coexistence of multiple writing systems: Classifying digraphia in post-socialist countries. *Journal of Eurasian Studies*, 14(2):139–150.

Shigeki Karita, Richard Sproat, and Haruko Ishikawa. 2023. Lenient evaluation of Japanese speech recognition: Modeling naturally occurring spelling inconsistency. In *Proceedings of the Workshop on Computation and Written Language (CAWL)*, pages 61–70, Toronto, Canada. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. 2012. Improving literacy in developing countries using speech recognition-supported games on mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing systems*, pages 1149–1158, Austin, Texas, USA.

Chaak-ming Lau, Mingfei Lau, and Ann Wai Huen To. 2024. The extraction and fine-grained classification of written Cantonese materials through linguistic feature detection. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 24–29, Torino, Italia. ELRA and ICCL.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022. Fusing ASR outputs in joint training for speech emotion recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7362–7366, Singapore.

Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. 2023. ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. In *Interspeech*, pages 1244–1248, Dublin, Ireland.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2021. Rethinking evaluation in ASR: Are our models robust enough? In *Interspeech*, pages 311–315, Brno, Czechia.

Behrooz Mahmoodi-Bakhtiari. 2018. *Spoken vs. written Persian: Is Persian diglossic?*, pages 183–212. De Gruyter Mouton, Berlin, Boston.

Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N. Sainath, and Trevor Strohman. 2019. Recognizing long-form speech using streaming end-to-end models. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 920–927, Singapore.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain. International Committee on Computational Linguistics (ICCL).

Katsuhiro J. Ota. 2005. An investigation of written Taiwanese. Master's thesis, University of Hawaii, August.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, pages 79155–79172, Red Hook, NY, USA. Curran Associates Inc.

Matthew Perez, Mimansa Jaiswal, Minxue Niu, Cristina Gorrostieta, Matthew Roddy, Kye Taylor, Reza Lotfian, John Kane, and Emily Mower Provost. 2022. Mind the gap: On the value of silence representations to lexical-based speech emotion recognition. In *Interspeech*, pages 156–160, Incheon, Korea.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research (JMLR)*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518, Honolulu, Hawaii, USA.

Anand Kumar Rai, Siddharth D Jaiswal, Shubham Prakash, Bendi Pragnya Sree, and Animesh Mukherjee. 2024. DENOASR: Debiasing ASRs through Selective Denoising. In *IEEE International Conference on Knowledge Graphs (ICKG)*, Abu Dhabi, UAE.

Gerald Roche. 2017. Introduction: The transformation of Tibet's language ecology in the twenty-first century. *International Journal of the Sociology of Language*, 2017(245):1–35.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. AudioPaLM: A large language model that can speak and listen. Technical report.

Don Snow. 2010. Hong Kong and modern diglossia. *International Journal of the Sociology of Language*, 2010(206):155–179.

Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. ASR pipeline for low-resourced languages: A case study on Pomak. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45, Dubrovnik, Croatia. Association for Computational Linguistics.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (HAT). In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6139–6143.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8246–8257.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.

Susan M. Williams, Don Nix, and Peter Fairweather. 2000. Using speech recognition technology to enhance literacy instruction for emerging readers. In *International Conference of the Learning Sciences*, pages 115–120.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. Technical report.

A License or Terms for Use of Artifacts

In this paper we studied three datasets VoxPopuli, FLEURS and MCV17. VoxPopuli is under CC-BY-NC 4.0 license[11], FLEURS is under CC-BY 4.0 license[12] and MCV17 is under CC0 1.0 license[13]. In addition, we used canto-filter package in 4.2.2. The package is under MIT license[14]. The usage of each dataset and package in our work is consistent with its intended use according to the license.

B Meta-Information of VoxPopuli, FLEURS and MCV17

The VoxPopuli corpus contains 1.8k hours of transcribed speech in 16 European languages, detailed in Table 6.

code	name	hours	code	name	hours
cs	Czech	62	hu	Hungarian	63
de	German	282	it	Italian	91
en	English	543	lt	Lithuanian	2
es	Spanish	166	nl	Dutch	53
et	Estonian	3	pl	Polish	111
fi	Finnish	27	ro	Romanian	89
fr	French	211	sk	Slovak	35
hr	Croatian	43	sl	Slovene	10

Table 6: List of languages and hours in VoxPopuli

FLEURS contains English utterances their translations and readings into 101 languages. Around 2009 English sentences are extracted from FLoRes101 corpus. Each sentence is recorded by 3 native speakers and the invalid recordings are discarded, which makes a total of 1.4k hours of speech and aroudn 12 hours in each language. The full list of languages is shown in Table 7 and 8.

code	name	code	name
af_za	Afrikaans, South Africa	el_gr	Greek, Greece
am_et	Amharic, Ethiopia	en_us	English, United States
ar_eg	Arabic, Egypt	es_419	Spanish, Latin America
as_in	Assamese, India	et_ee	Estonian, Estonia
ast_es	Asturian, Spain	fa_ir	Persian, Iran
az_az	Azerbaijani, Azerbaijan	ff_sn	Fulah, Senegal
be_by	Belarusian, Belarus	fi_fi	Finnish, Finland
bg_bg	Bulgarian, Bulgaria	fil_ph	Filipino, Philippines
bn_in	Bengali, India	fr_fr	French, France
bs_ba	Bosnian, Bosnia	ga_ie	Irish, Ireland
ca_es	Catalan, Spain	gl_es	Galician, Spain
ceb_ph	Cebuano, Philippines	gu_in	Gujarati, India
ckb_iq	Central Kurdish, Iraq	ha_ng	Hausa, Nigeria
cmn_hans_cn	Mandarin, China	he_il	Hebrew, Israel
cs_cz	Czech, Czech Republic	hi_in	Hindi, India
cy_gb	Welsh, United Kingdom	hr_hr	Croatian, Croatia
da_dk	Danish, Denmark	hu_hu	Hungarian, Hungary
de_de	German, Germany	hy_am	Armenian, Armenia

Table 7: List of languages in FLEURS, Part 1

[11] github.com/facebookresearch/voxpopuli/blob/main/LICENSE
[12] huggingface.co/datasets/google/fleurs
[13] huggingface.co/datasets/mozilla-foundation/common_voice_17_0
[14] pypi.org/project/canto-filter/

code	name	code	name
id_id	Indonesian, Indonesia	ny_mw	Chichewa, Malawi
ig_ng	Igbo, Nigeria	oc_fr	Occitan, France
is_is	Icelandic, Iceland	om_et	Oromo, Ethiopia
it_it	Italian, Italy	or_in	Odia, India
ja_jp	Japanese, Japan	pa_in	Punjabi, India
jv_id	Javanese, Indonesia	pl_pl	Polish, Poland
ka_ge	Georgian, Georgia	ps_af	Pashto, Afghanistan
kam_ke	Kamba, Kenya	pt_br	Portuguese, Brazil
kea_cv	Kabuverdianu, Cape Verde	ro_ro	Romanian, Romania
kk_kz	Kazakh, Kazakhstan	ru_ru	Russian, Russia
km_kh	Khmer, Cambodia	sd_in	Sindhi, India
kn_in	Kannada, India	sk_sk	Slovak, Slovakia
ko_kr	Korean, South Korea	sl_si	Slovenian, Slovenia
ky_kg	Kyrgyz, Kyrgyzstan	sn_zw	Shona, Zimbabwe
lb_lu	Luxembourgish, Luxembourg	so_so	Somali, Somalia
lg_ug	Ganda, Uganda	sr_rs	Serbian, Serbia
ln_cd	Lingala, DRC	sv_se	Swedish, Sweden
lo_la	Lao, Laos	sw_ke	Swahili, Kenya
lt_lt	Lithuanian, Lithuania	ta_in	Tamil, India
luo_ke	Luo, Kenya	te_in	Telugu, India
lv_lv	Latvian, Latvia	tg_tj	Tajik, Tajikistan
mi_nz	Māori, New Zealand	th_th	Thai, Thailand
mk_mk	Macedonian, North Macedonia	tr_tr	Turkish, Turkey
ml_in	Malayalam, India	uk_ua	Ukrainian, Ukraine
mn_mn	Mongolian, Mongolia	umb_ao	Umbundu, Angola
mr_in	Marathi, India	ur_pk	Urdu, Pakistan
ms_my	Malay, Malaysia	uz_uz	Uzbek, Uzbekistan
mt_mt	Maltese, Malta	vi_vn	Vietnamese, Vietnam
my_mm	Burmese, Myanmar	wo_sn	Wolof, Senegal
nb_no	Norwegian Bokmål, Norway	xh_za	Xhosa, South Africa
ne_np	Nepali, Nepal	yo_ng	Yoruba, Nigeria
nl_nl	Dutch, Netherlands	yue_hant_hk	Cantonese, Hong Kong
nso_za	Northern Sotho, South Africa	zu_za	Zulu, South Africa

Table 8: List of languages in FLEURS, Part 2

The statistics of MCV17 meta-information can be found in Table 9 and 10. The full dataset contains 124 locales. Each locale contains 3 categories: validated, invalidated and other. Validated means the data clip has received more than 2 validations and the upvotes > downvotes; Invalidated means the data clip has received more than 2 validations but the upvotes $\leq$ downvotes; Other means the data clip has not received 2 or more validations.

When analyzing the speaker diversity in Section 3.4, we used all the 3 categories. In other analyses, we only used the validated data and skipped locales with 0 or 1 utterances.

code	name	total hours	validated hours	code	name	total hours	validated hours
ab	Abkhaz	84.41	59.86	es	Spanish	2219.39	561.82
af	Afrikaans	0.56	0.27	et	Estonian	59.97	45.85
am	Amharic	2.63	1.55	eu	Basque	672.41	273.48
ar	Arabic	155.81	90.27	fa	Persian	415.36	363.42
as	Assamese	3.28	2.72	fi	Finnish	21.65	13.35
ast	Asturian	1.91	0.81	fr	French	1147.41	1013.03
az	Azerbaijani	0.45	0.19	fy_nl	West Frisian	212.09	68.73
ba	Bashkir	268.17	257.77	ga_ie	Irish	10.47	5.48
bas	Basaa	2.82	2.16	gl	Galician	131.47	65.57
be	Belarusian	1765.25	1717.14	gn	Guarani	27.56	3.66
bg	Bulgarian	20.77	16.46	ha	Hausa	12.17	3.94
bn	Bengali	1272.95	53.51	he	Hebrew	6.06	2.22
br	Breton	26.87	18.55	hi	Hindi	20.7	14.11
ca	Catalan	3586.54	2697.77	hsb	Upper Sorbian	3.02	2.43
ckb	Central Kurdish	172.16	130.21	ht	Haitian Creole	0.01	0
cnh	Chin Haka	6.04	2.4	hu	Hungarian	172.63	92.64
cs	Czech	262.67	76.09	hy_am	Armenian	47.36	22.27
cv	Chuvash	27.51	24.36	ia	Interlingua	17.04	13.73
cy	Welsh	156.58	123.08	id	Indonesian	64.53	28.93
da	Danish	12.6	11.68	ig	Igbo	8.77	0.02
de	German	1423.9	1333.93	is	Icelandic	0.08	0.02
dv	Dhivehi	64.21	38.87	it	Italian	395.57	354.96
dyu	Dyula	0.49	0.32	ja	Japanese	476.98	124.31
el	Greek	31.5	18.64	ka	Georgian	214.21	138.8
en	English	3507.22	2614.76	kab	Kabyle	689.77	566.48
eo	Esperanto	1904.97	1433.1	kk	Kazakh	3.46	2.13
es	Spanish	2219.39	561.82	kmr	Kurmanji Kurdish	99.79	67.57
et	Estonian	59.97	45.85	ko	Korean	5.57	1.72

Table 9: List of locales and hours in Common Voice 17.0, Part 1

code	name	total hours	validated hours	code	name	total hours	validated hours
ky	Kyrgyz	47.6	38.42	sah	Sakha	12.65	8.31
lg	Ganda	559.22	436.7	sat	Santali	1.02	0.57
lij	Ligurian	3.29	2.8	sc	Sardinian	1.95	1.5
lo	Lao	0.37	0.2	sk	Slovak	26.89	22.1
lt	Lithuanian	25.21	23.71	skr	Saraiki	6.59	4.2
ltg	Latgalian	23.3	20.67	sl	Slovenian	14.99	11.38
lv	Latvian	276.61	222.45	sq	Albanian	1.97	1.94
mdf	Moksha	0.5	0.49	sr	Serbian	6.76	5.01
mhr	Meadow Mari	301.25	280.45	sv_se	Swedish	54.5	45.38
mk	Macedonian	22.89	7.82	sw	Swahili	1084.72	399.48
ml	Malayalam	10.11	3.46	ta	Tamil	403.84	232.59
mn	Mongolian	23.08	13.17	te	Telugu	2.3	0.26
mr	Marathi	27.48	18.75	th	Thai	422.96	171.29
mrj	Western Mari	36.77	33.63	ti	Tigrinya	0.11	0.03
mt	Maltese	17.21	8.48	tig	Tigre	2.69	1.08
myv	Erzya	3.2	3.15	tk	Turkmen	6.52	2.73
nan_tw	Southern Min (Taiwan)	20.21	5.73	tok	Toki Pona	18.86	13.59
ne_np	Nepali	1.54	0.81	tr	Turkish	122.06	117.28
nhi	Nahuatl	0.03	0.02	tt	Tatar	31.24	30.59
nl	Dutch	119.6	109.48	tw	Twi	0.27	0.16
nn_no	Norwegian Nynorsk	1.67	1.42	ug	Uyghur	236.63	199.46
nso	Northern Sotho	0.03	0	uk	Ukrainian	111.71	97.44
oc	Occitan	12.83	2.25	ur	Urdu	231.8	63.52
or	Odia	12.51	4.4	uz	Uzbek	263.24	99.63
os	Ossetic	0.31	0.28	vi	Vietnamese	18.76	5.65
pa_in	Punjabi	3.99	2.01	vot	Votic	0.29	0.06
pl	Polish	175.85	166.71	yi	Yiddish	0.05	0.04
ps	Pashto	2.11	1.69	yo	Yoruba	7.29	5.07
pt	Portuguese	210.97	174.17	yue	Cantonese	177.76	23.4
quy	Quechua (Ayacucho)	0.01	0	zgh	Tamazight	1.47	0.51
rm_sursilv	Romansh Sursilvan	10.91	6.53	zh_cn	Chinese (China)	1061.32	233.77
rm_vallader	Romansh Vallader	4.26	2.46	zh_hk	Chinese (Hong Kong)	138.23	107.42
ro	Romanian	46.8	19.85	zh_tw	Chinese (Taiwan)	126.04	77.1
ru	Russian	273.89	234.46	zu	Zulu	0.05	0
rw	Kinyarwanda	2384	2001.34	zza	Zaza	0.45	0.33

Table 10: List of locales and hours in Common Voice 17.0, Part 2

C Median Utterance Duration of VoxPopuli, FLEURS and MCV17

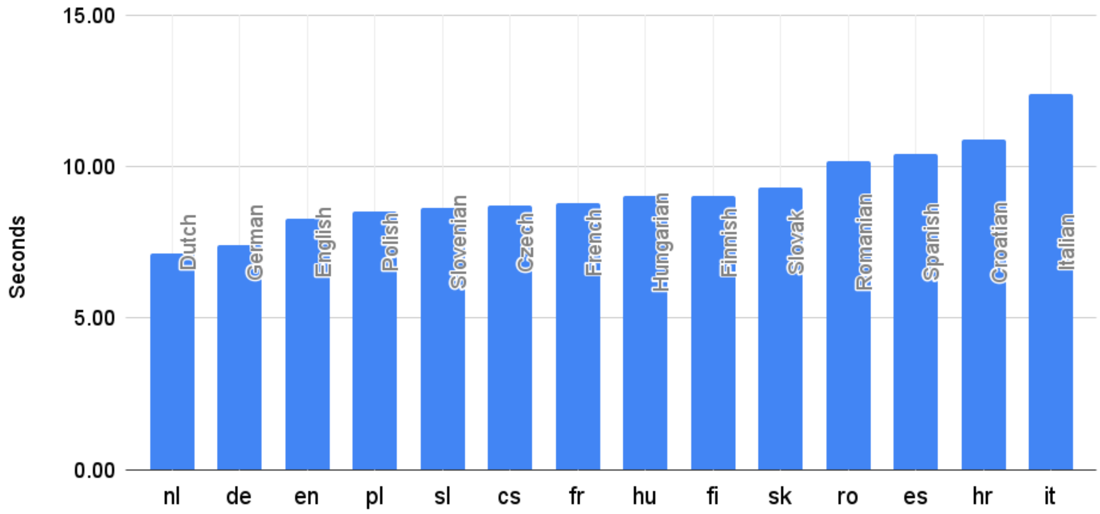

Figure 2: Median utterance duration of the 14 languages in VoxPopuli. All languages have a median utterance duration of at least 7 seconds.

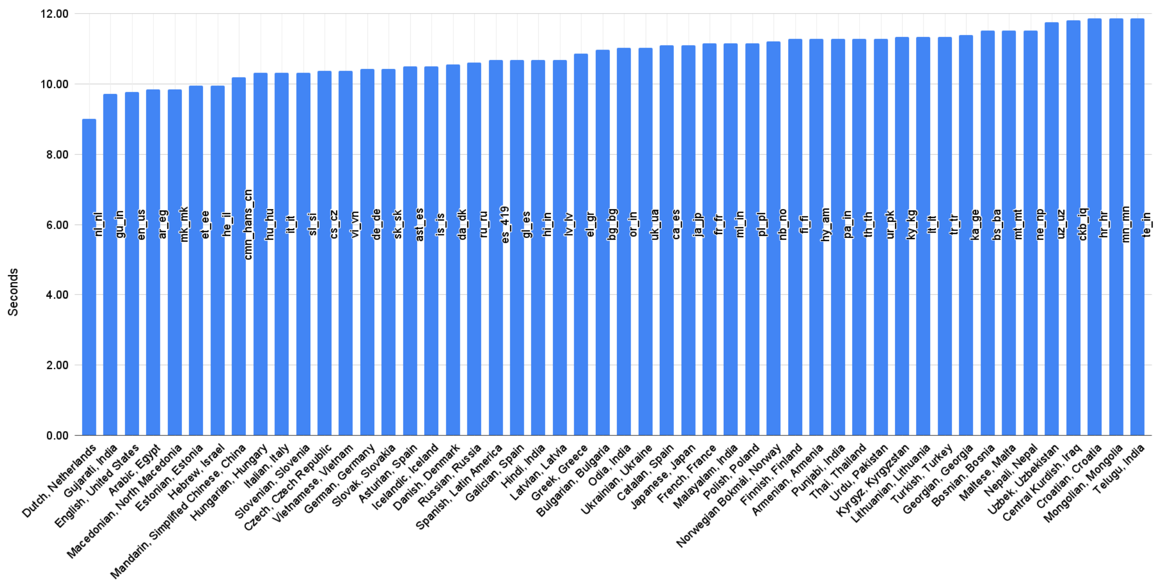

Figure 3: 50 languages with the shortest median utterance duration in FLEURS. All languages have a median utterance duration of at least 5 seconds.

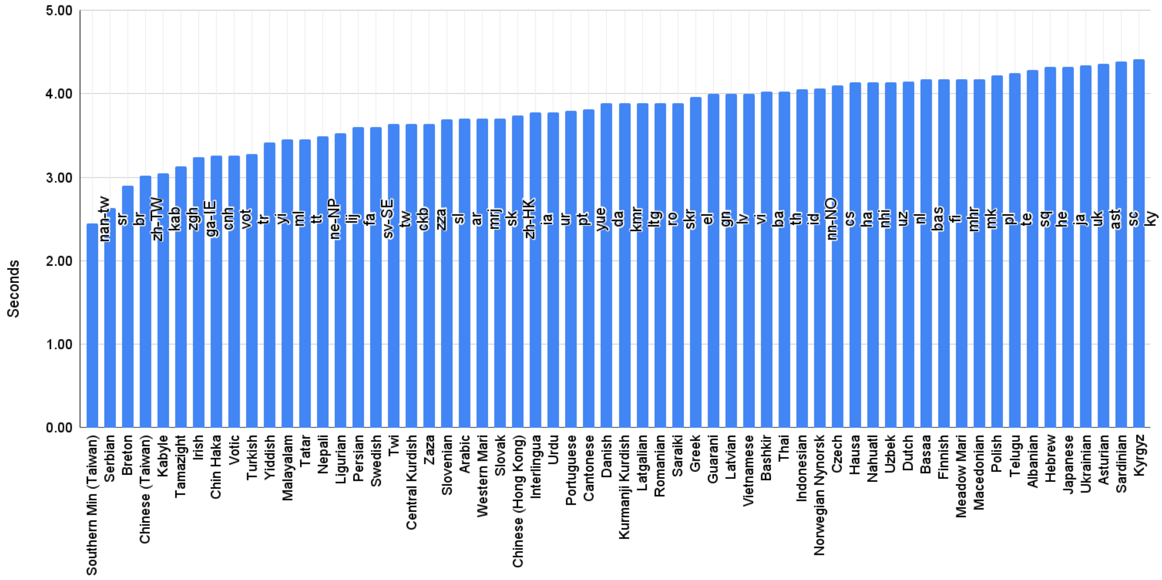

Figure 4: 60 languages with the shortest median utterance duration in Mozilla Common Voice 17.0. Taiwanese Southern Min (nan_tw) has a median duration of only 2.45 seconds.

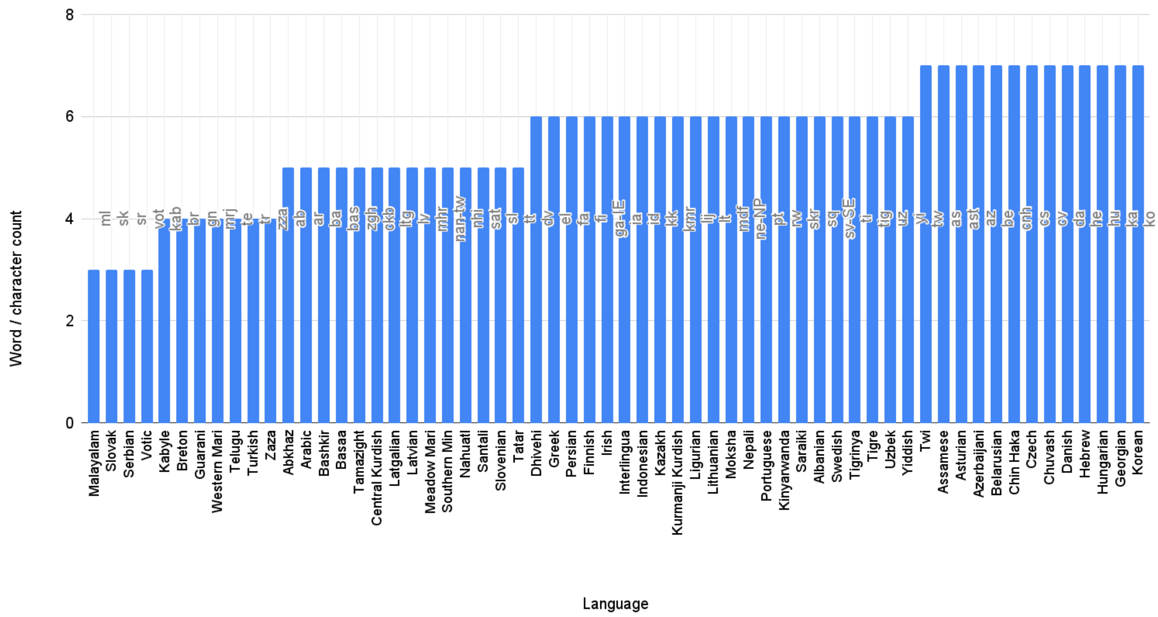

Figure 5: 60 languages with the shortest median text prompt length (measured by word or character count) in Mozilla Common Voice 17.0. Note that nan_tw doesn't have the shortest length because the words are duplicated in two writing systems.

D Speech and Silence Percentage of VoxPopuli, FLEURS and MCV17

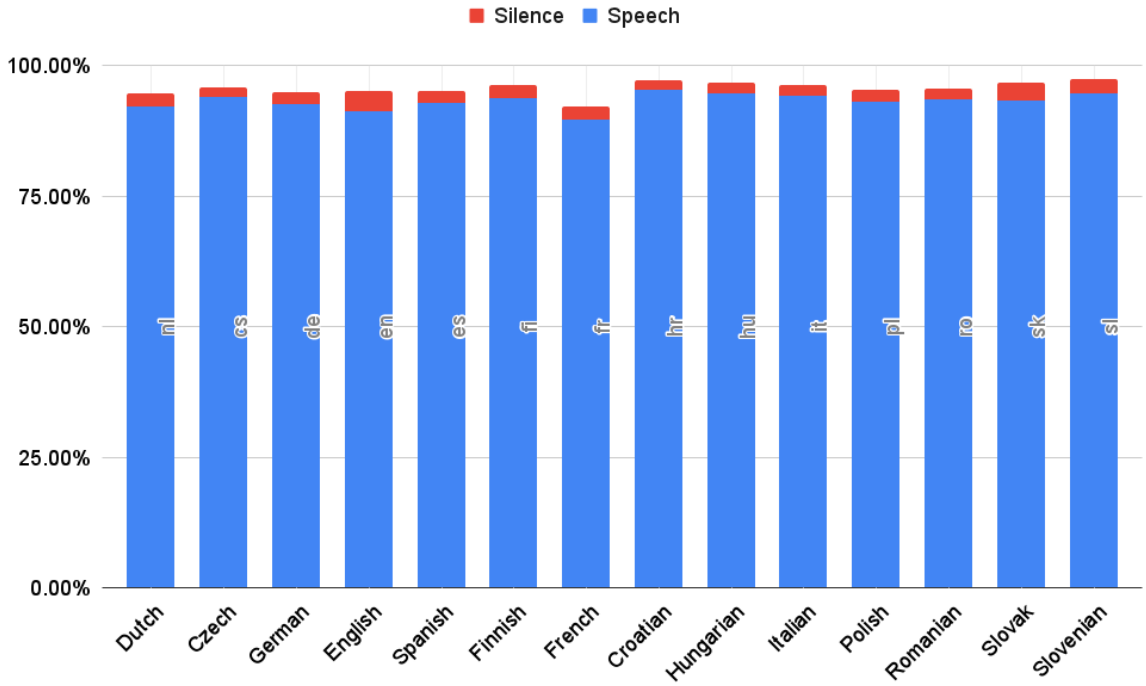

Figure 6: Percentages of speech and silence in VoxPopuli. All languages have more than 90% speech in the utterances. Note that the percentages of silence and speech don't add up to 100% because we omit other types of sound such as music, noise, laughter, etc.

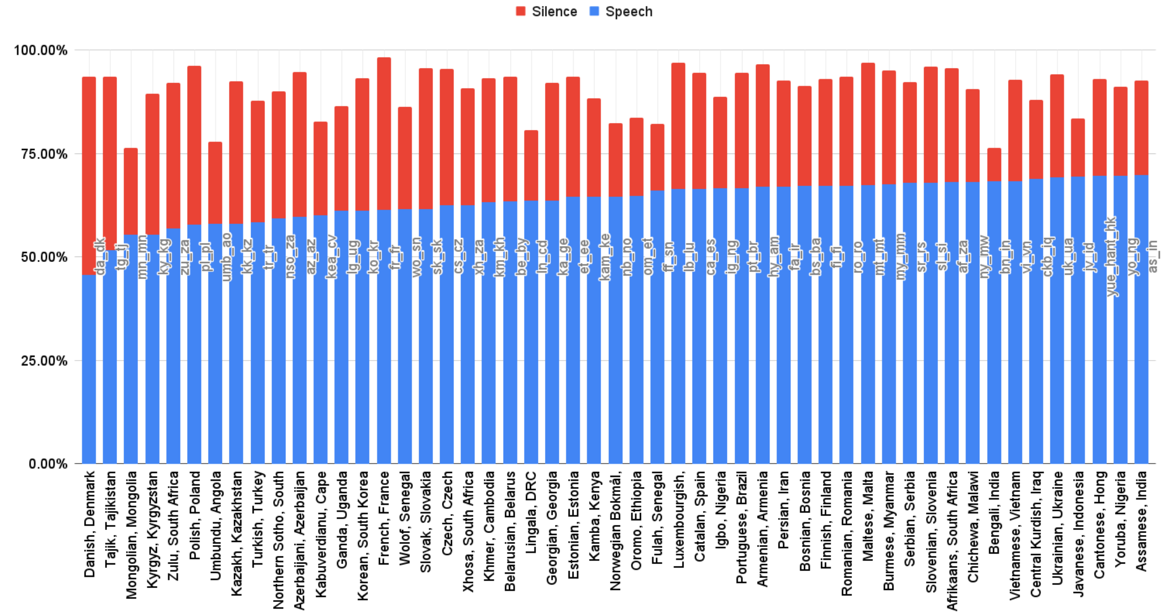

Figure 7: 50 languages with the lowest speech proportion in FLEURS. Danish (da_dk) has less than 50% speech in the audio recordings. Note that the percentages of silence and speech don't add up to 100% because we omit other types of sound such as music, noise, laughter, etc.

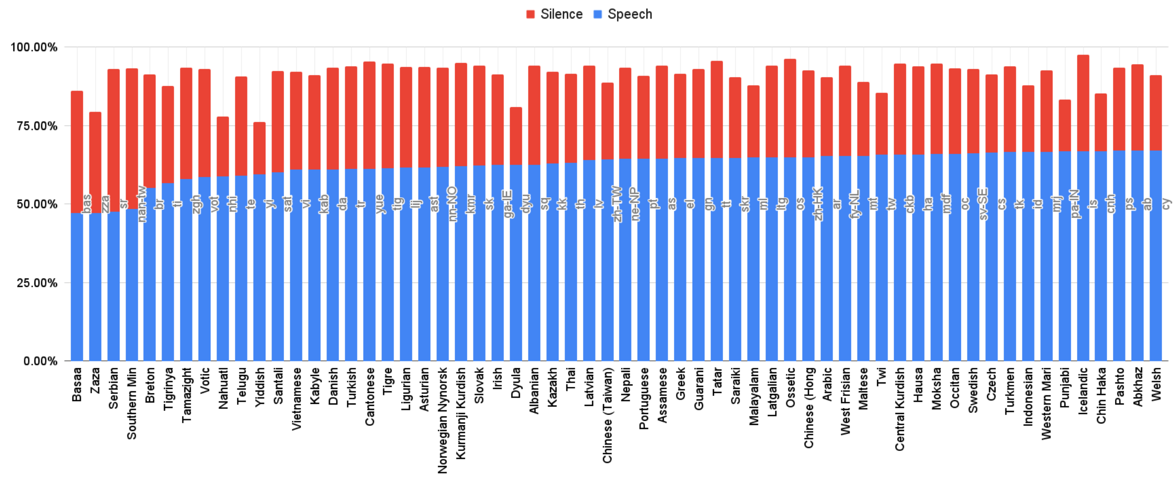

Figure 8: 60 languages with the lowest speech proportion in Mozilla Common Voice 17.0. Basaa, Zaza, Serbian and Southern Min Taiwan have less than 50% speech in the audio recordings. Note that the percentages of silence and speech don't add up to 100% because we omit other types of sound such as music, noise, laughter, etc.

E Example Sentences of MCV17

zh_cn (Chinese, China)	ca (Catalan)
殿试登进士第三甲第一百一十一名。	No he anat mai a Agost.
殿试登进士第三甲第一百七十一名。	No he anat mai a Aigües.
殿试登进士第三甲第一百三十一名。	No he anat mai a Aiora.
殿试登进士第三甲第一百三十九名。	No he anat mai a Aiòder.
殿试登进士第三甲第一百三十五名。	No he anat mai a Alaior.
殿试登进士第三甲第一百九十名。	No he anat mai a Alaró.
殿试登进士第三甲第一百五十名。	No he anat mai a Alaior.
殿试登进士第三甲第一百八十七名。	No he anat mai a Albaida.
殿试登进士第三甲第一百八十九名。	No he anat mai a Albatera.
殿试登进士第三甲第一百六十七名。	No he anat mai a Alberic.
...	...

Table 11: Examples of highly repetitive, seemingly machine-generated sentences in Common Voice 17.0. We find at least hundreds of such sentences in zh_cn, zh_tw, zh_hk and ca. We suspect that similar issues may exist in other languages of MCV17 that we haven't analyzed.

nan_tw sentences in MCV17
竹仔籃（Tik-á-nâ）
竹南鎮（Tek-lâm-tìn）
竹南鎮（Tik-lâm-tìn）
竹坑口（Tik-khinn-kháu | Tek-khi-kháu）
竹子腳（Tek-á-kha）
竹崎鄉（Tik-kiā-hiong）
竹東鎮（Tik-tang-tìn）
竹東（Tik-tang）
竹田鄉（Tik-tshân-hiong）
竹田（Tik-tshân）
...

Table 12: A snapshot of MCV17 nan_tw's text prompts. Each row is a single word or phrase with its romanization. For homographs, multiple pronunciations are appended to the Sinographs.

F Average Hours per Speaker of MCV17

code	name	unique speakers	total hours	average hours per speaker
rw	Kinyarwanda	1131	2,384.0	2.11
mk	Macedonian	19	22.9	1.20
eo	Esperanto	1739	1,905.0	1.10
lg	Ganda	657	559.2	0.85
sw	Swahili	1452	1,084.7	0.75
ur	Urdu	349	231.8	0.66
mrj	Western Mari	60	36.8	0.61
mhr	Meadow Mari	496	301.3	0.61
kab	Kabyle	1547	689.8	0.45
ta	Tamil	906	403.8	0.45
mr	Marathi	90	27.5	0.31
ha	Hausa	40	12.2	0.30
ba	Bashkir	917	268.2	0.29
he	Hebrew	21	6.1	0.29
cs	Czech	983	262.7	0.27
ia	Interlingua	67	17.0	0.25
myv	Erzya	13	3.2	0.25
cv	Chuvash	112	27.5	0.25
ps	Pashto	9	2.1	0.23
be	Belarusian	8291	1,765.25	0.21
ab	Abkhaz	403	84.41	0.21
yue	Cantonese	913	177.76	0.19
ug	Uyghur	1258	236.63	0.19
dv	Dhivehi	357	64.21	0.18
kmr	Kurmanji Kurdish	561	99.79	0.18
lij	Ligurian	19	3.29	0.17
ky	Kyrgyz	283	47.60	0.17
gn	Guarani	164	27.56	0.17
bg	Bulgarian	134	20.77	0.15
zh_cn	Chinese (China)	7005	1,061.32	0.15
hsb	Upper Sorbian	21	3.02	0.14
sc	Sardinian	14	1.95	0.14
br	Breton	207	26.87	0.13
ka	Georgian	1679	214.21	0.13
tok	Toki Pona	149	18.86	0.13
hy_an	Armenian	390	47.36	0.12
uz	Uzbek	2170	263.24	0.12
rm_sursilv	Romansh Sursilvan	90	10.91	0.12
tt	Tatar	258	31.24	0.12

Table 13: Top 40 languages with the highest average total hours per speaker in Common Voice 17.0. As a reference, English has 0.04 hours per speaker.

code	name	unique speakers	total hours	average hours per speaker	code	name	unique speakers	total hours	average hours per speaker
zu	Zulu	1	0.05	0.05	bas	Basaa	36	2.82	0.08
nso	Northern Sotho	1	0.03	0.03	nn_no	Norwegian Nynorsk	38	1.67	0.04
ht	Haitian Creole	1	0.01	0.01	te	Telugu	39	2.30	0.06
nhi	Nahuatl	2	0.03	0.02	ha	Hausa	40	12.17	0.30
quy	Quechua (Ayacucho)	2	0.01	0.01	as	Assamese	46	3.28	0.07
yi	Yiddish	3	0.05	0.02	rm_vallader	Romansh Vallader	53	4.26	0.08
zza	Zaza	4	0.45	0.11	sq	Albanian	55	1.97	0.04
is	Icelandic	4	0.08	0.02	skr	Saraiki	57	6.59	0.12
vot	Votic	6	0.29	0.05	mrj	Western Mari	60	36.77	0.61
tw	Twi	6	0.27	0.05	ia	Interlingua	67	17.04	0.25
ti	Tigrinya	6	0.11	0.02	pa_in	Punjabi	68	3.99	0.06
os	Ossetic	8	0.31	0.04	mr	Marathi	90	27.48	0.31
ps	Pashto	9	2.11	0.23	rm_sursilv	Romansh Sursilvan	90	10.91	0.12
mdf	Moksha	11	0.50	0.05	ko	Korean	90	5.57	0.06
myv	Erzya	13	3.20	0.25	yo	Yoruba	108	7.29	0.07
sat	Santali	13	1.02	0.08	sah	Sakha	111	12.65	0.11
lo	Lao	13	0.37	0.03	cv	Chuvash	112	27.51	0.25
sc	Sardinian	14	1.95	0.14	tk	Turkmen	112	6.52	0.06
zgh	Tamazight	17	1.47	0.09	ig	Igbo	114	8.77	0.08
mk	Macedonian	19	22.89	1.20	or	Odia	125	12.51	0.10
lij	Ligurian	19	3.29	0.17	bg	Bulgarian	134	20.77	0.15
he	Hebrew	21	6.06	0.29	ml	Malayalam	134	10.11	0.08
hsb	Upper Sorbian	21	3.02	0.14	oc	Occitan	145	12.83	0.09
af	Afrikaans	23	0.56	0.02	tok	Toki Pona	149	18.86	0.13
tig	Tigre	24	2.69	0.11	sr	Serbian	153	6.76	0.04
az	Azerbaijani	26	0.45	0.02	sl	Slovenian	154	14.99	0.10
ast	Asturian	29	1.91	0.07	gn	Guarani	164	27.56	0.17
am	Amharic	30	2.63	0.09	kk	Kazakh	166	3.46	0.02
ne_np	Nepali	32	1.54	0.05	ga_ie	Irish	192	10.47	0.05
dyu	Dyula	33	0.49	0.01	br	Breton	207	26.87	0.13

Table 14: Bottom 60 languages with the least unique voice contributors in Common Voice 17.0. As a reference, English has 92325 speakers, with each contributing 0.04 hours on average

G Script Choices of Digraphic Languages in MCV17 and FLEURS

Language name	Scripts that can be written in	MCV17	FLEURS
Central Kurdish (Sorani)	Cyrillic, Hawar (Latin), Sorani (Arabic)	`ckb`: Arabic	`ckb_iq`: Sorani (Arabic)
Dyula	Latin, N'Ko	`dyu`: Latin	N/A
Fula	Adlam, Ajami (Arabic), Latin	N/A	`ff_sn`: Latin
Malay	Jawi (Arabic), Latin	`ms`: Latin	`ms_my`: Latin
Mongolian	Cyrillic, Mongolian (Bichig)	`mn`: Cyrillic	`mn_mn`: Cyrillic
Northern Kurdish (Kurmanji)	Cyrillic, Hawar (Latin), Sorani (Arabic)	`kmr`: Hawar (Latin)	N/A
Serbian	Cyrillic, Latin	`sr`: Cyrillic	`sr_rs`: Cyrillic and Latin
Punjabi	Gurmukhi, Shahmukhi (Arabic)	`pa_in`: Gurmukhi	`pa_in`: Gurmukhi
Tamazight (Berber a.k.a Amazigh)	Arabic, Latin, Tifinagh	`zgh`: Tifinagh	N/A
Uzbek	Arabic, Cyrillic, Latin	`uz`: Latin	`uz_uz`: Latin
Votic	Cyrillic, Latin	`vot`: Latin	N/A

Table 15: Script choices and assumptions of digraphic languages in Common Voice 17.0 and FLEURS.

H Text Classification Logic

Algorithm 1 Norwegian Orthography Classification Logic

procedure CLASSIFYBOKMALNYNORSK($sentence$)
  $nynorskMarkers \leftarrow$ "ikkje", "eg", "eit", "eitt", "me", "ho", "hjå", "kva", "kven", "noko", "nokre", "sjå", "skule", "kor", "fyrst", "mykje", "òg", "medan"
  $bokmaalMarkers \leftarrow$ "ikke", "jeg", "et", "en", "vi", "hun", "hos", "hva", "hvem", "noe", "noen", "se", "skole", "hvor", "først", "mye", "også", "mens"
  $sentence \leftarrow$ Convert $sentence$ to lowercase
  $nynorskCount \leftarrow 0$
  $bokmaalCount \leftarrow 0$
  for $marker \in nynorskMarkers$ **do**
    if $marker$ exists as whole word in $sentence$ **then**
      $nynorskCount \leftarrow nynorskCount + 1$
    end if
  end for
  for $marker \in bokmaalMarkers$ **do**
    if $marker$ exists as whole word in $sentence$ **then**
      $bokmaalCount \leftarrow bokmaalCount + 1$
    end if
  end for
  $nynorskCount \leftarrow nynorskCount +$ Count of words ending in 'a'
  $bokmaalCount \leftarrow bokmaalCount +$ Count of words ending in 'en'
  if $nynorskCount > bokmaalCount$ **then**
    return "Nynorsk"
  else if $bokmaalCount > nynorskCount$ **then**
    return "Bokmål"
  else if $nynorskCount = bokmaalCount$ and $nynorskCount > 0$ **then**
    return "Mixed"
  else
    return "Unmarked / Neutral"
  end if
end procedure

Algorithm 2 Arabic Fusha / Dialect Classification Logic

```
procedure IsFusha(sentence)
    fushaMarkers ← List of Fusha (Modern Standard Arabic) markers
    dialectMarkers ← List of dialectal Arabic markers
    fushaScore ← 0
    dialectScore ← 0
    for each marker in fushaMarkers do
        if marker found in sentence then
            fushaScore ← fushaScore + 1
        end if
    end for
    for each marker in dialectMarkers do
        if marker found in sentence then
            dialectScore ← dialectScore + 1
        end if
    end for
    if fushaScore = 0 and dialectScore = 0 then
        return "unmarked"
    else if fushaScore > dialectScore then
        return "fusha"
    else if fushaScore < dialectScore then
        return "dialect"
    else
        return "mixed"
    end if
end procedure
```

I Languages Inspected by Native Speaker Volunteers in this Study

We asked native speaker volunteers from our co-workers to review 100 randomly sampled sentences (text and audio) for coherence, audio-text alignment, dialect, topic domain, and language ID from each language subsets of the datasets. The language list is shown in 16.

Name	MCV17 code	FLEURS code	Name	MCV17 code	FLEURS code
Amharic	am	am_et	Lingala		ln_cd
Arabic (Egypt)	ar	ar_eg	Mandarin Chinese (China)	zh_cn	cmn_hans_cn
Basaa	bas		Mandarin Chinese (Taiwan)	zh_tw	
Bashkir	ba		Meadow Mari	mhr	
Cantonese	yue	yue_hk	Moksha	mdf	
Cape Verde Creole		kea_cv	Norwegian (Nynorsk / Bokmål)	nn_no	nb_no
Catalan	ca	ca_es	Oromo		om_et
Chichewa		ny_mw	Persian	fa	fa_ir
Chinese (Hong Kong)	zh_hk		Russian	ru	ru_ru
Chuvash	cv		Shona		sn_zw
Danish, Denmark	da	da_dk	Southern Min	nan_tw	
Dyula	dyu		Spanish	es	
English (United States)	en	en_us	Swahili	sw	sw_ke
Fulah (Senegal)		ff_sn	Twi	tw	
Ganda	lg	lg_ug	Ukrainian	uk	
German	de		Votic	vot	
Greek	el	el_gr	Western Mari	mrj	
Hausa		ha_ng	Wolof		wo_sn
Igbo		ig_ng	Yoruba		yo_ng
Japanese	ja	ja_jp	Zulu	zu	zu_za
Korean	ko	ko_kr			

Table 16: Languages with native speaker volunteers available for qualitative inspections in this study.

J Acknowledgments

The authors gratefully acknowledge the helpful comments from [anonymous collaborators]