Voices from the Margins: Modeling Linguistic Diversity in Spontaneous Speech for Low-Resource Languages

Vitthal Bhandari **Tiya Kumar** **Kate Mulhern**
Department of Linguistics
University of Washington
{vitthal1, tiyakr, mulhernk}@uw.edu

Abstract

We conduct Automatic speech recognition (ASR) experiments on the Common Voice Spontaneous Speech dataset by Mozilla Data Collective, consisting of 21 low-resource languages across four continents of the world. We fine-tune popular multilingual speech models on all languages of this dataset, and observe that while a single-best-model solution doesn't exist, the Massively Multilingual Speech model and Whisper achieve superior performance on certain languages. Through n-gram language modeling decoding experiments, we observe a significant improvement in error rate over greedy decoding by up to 27.3%. We follow our experiments with a close linguistic error analysis of the best performing models on Scots (sco) and Nubi (kcn) - two of the languages in our dataset, with very little prior audio and text modeling research. We highlight the morphosyntactic errors induced during speech recognition and perform a holistic analysis of these languages. We finally advocate for the importance of building efficient and accurate ASR tools for modeling speech in endangered languages with scarce resources, and their applications to language revitalization, language learning assistance, and accessibility. The code can be found at https://github.com/vitthal-bhandari/low-resource-asr/

1 Introduction

Progress in the field of automatic speech recognition (ASR) for high-resource languages has led to several large multilingual speech models with human-level performance on popular benchmarks (Omnilingual et al., 2025; Pratap et al., 2024; Radford et al., 2023; Babu et al., 2022; Yadav and Sitaram, 2022). Of equal importance is the need to model the linguistic diversity in the majority of the 7,000+ languages in the world that are either endangered, on the brink of extinction, low-resource, or have few native speakers left.

Building speech tools (such as ASR models) for indigenous languages has applications in language documentation (Jimerson et al.; Shi et al., 2021; Jimerson and Prud'hommeaux, 2018), language revitalization (Mainzinger, 2024; Zhang et al., 2022), community language learning (van Doremalen et al., 2016; Dolinska et al., 2024; sec, 2003; Sun, 2023; Xiao and Park, 2021), and user accessibility (Wald and Bain, 2008; Butler et al., 2019; Morales et al., 2013; Guo et al., 2020). Not only can ASR tools help generate valuable synthetic data to help augment languages with scarce resources, leading to higher resources (Venkateswaran and Liu, 2024; Tjandra et al., 2020), they can also be used to assist linguists and fieldworkers in improving transcription error rates and time-to-transcribe when building gold standard corpora (Prud'hommeaux et al., 2021) for indigenous languages.

These are not standalone issues. The use of NLP tools in language preservation ensures that seminal information about cultural artefacts can be passed down to future generations (Koc, 2025; Gedeon et al., 2024; Dueck, 2024; Murshed et al., 2025).

In this work, we fine tune and evaluate popular multilingual speech models on 21 low-resource languages and employ n-gram modeling to enhance decoding accuracy. Our empirical analysis sheds light on the efficacy of using these models for building ASR tools and highlights significant gaps in model performance, thereby justifying the need to create more resources and build corpora for underserved languages worldwide.

Our paper has three contributions. First, we provide a detailed background of two languages in our dataset - Scots (sco) and Nubi (kcn) in order to highlight their history, syntax, morphology, lack of labeled speech corpora, and difficulty in modeling speech tools (§3, §4). Second, we provide a comprehensive review of error rates after fine-tuning three popular multilingual models and compare their performance. We augment beam search with

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 37–54

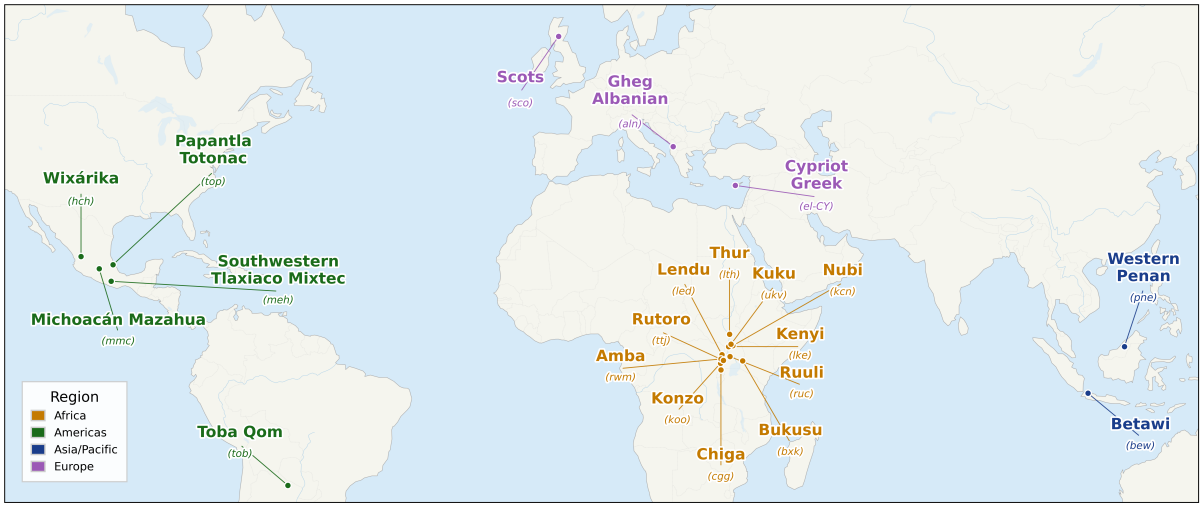

Figure 1: Geographic Distribution of Languages in the Mozilla Common Voice Spontaneous Speech Dataset. Note that color groupings are geographically (not linguistically) informed and that national borders don't necessarily reflect the distribution of the languages.

an external n-gram language model estimated from training text, and combine acoustic and LM scores via shallow fusion(§5). Finally, we perform an extensive linguistic-error analysis of the transcripts generated for the test set by our best-performing models on Scots and Nubi and highlight key challenges and problems that occur with our fine-tuned models (§7).

Our choice of Scots and Nubi for linguistic error analysis is motivated by two reasons: (1) both share a "shadow" relationship with a high-resource language (English for Scots, Arabic for Nubi), making them a natural minimal pair for asking whether genetic/lexical proximity to a well-modeled language helps ASR, and (2) all authors are native English speakers, which enabled meaningful analysis of Scots transcripts at the lexical and phonological level.

In the next section (§2), we provide a brief overview of our dataset and some initial preprocessing steps taken to clean and split it.

2 Dataset Overview

2.1 Shared Task Information

For this study, we utilized data from the Mozilla Common Voice Spontaneous Speech ASR Shared Task. This shared task is based on the recently released spontaneous speech datasets from Mozilla Common Voice. In these datasets, participants freely respond to prompts, and the responses are transcribed and validated, providing a diverse set of examples suitable for training and evaluating our models (Mozilla Data Collective, 2026). For all 21 languages in the dataset, we list the countries of origin, vitality, and number of speakers in Appendix A. Hours spent on training, development, and test sets for each language are also included in Appendix A. In Appendix B, we list complete statistics of utterance durations for all languages. Of particular importance is the number **P97.5**, which indicates the 97.5th percentile of utterance duration (in seconds). We use this number for each language to drop superfluous utterances with a duration greater than P97.5 to avoid CUDA errors during training.

2.2 Train/Dev & Test Splits

The data was originally divided into train/development and test sets. We used the provided training set for model learning and the validation set to tune hyperparameters and monitor performance. Our access to the gold labels for the test data set was limited, preventing us from being able to evaluate the model on the actual test data. To adapt to this, we held out 45 minutes of data from the validation set, serving as a replacement for the gold labels, and got an estimate of the model's performance. By enacting this approach, we ensured the distinct use of a train/dev and test split while gathering meaningful results.

For reproducibility, we release the exact utterances used across all language splits (for training, validation, and testing) here[1].

[1] https://github.com/vitthal-bhandari/low-resource-asr/tree/master/results/splits

3 Background

In this section, we provide a detailed background of two languages from our dataset - Scots (sco) and Nubi (kcn) - which form the basis of our linguistic error analysis later on. Through the representation of these two languages, we aim to highlight the difficulties in modeling speech from similar low-resource, typologically diverse languages.

3.1 Scots

Scots (ISO 639: sco) is a language native to the United Kingdom and the Republic of Ireland, with over 1,508,540 speakers and 2,444,659 reporting being able to speak, read, or write the language[2]. It is a minority language in Europe and is considered vulnerable according to UNESCO; its Agglomerated Endangerment Status according to Glottolog is threatened (Moseley, 2010; Hammarström et al., 2026). It belongs to the Indo-European language family and has roots in Old English. Similar to English, Scots follows the SVO (Subject-verb-object) word order. For example, "*I eat lettuce*" becomes "*Ah eat lettuce*" in Scots. There are 5 mainly recognized dialects: Central, Southern, Northern, Insular, and Ulster. Many scholars argue whether Scots is its own language or a dialect of English (Kortmann et al., 2008; Trudgill, 1984). However, the Scottish government recognizes Scots as its own distinct language. Despite its over 1 million speakers worldwide, Scots is considered a low-resource language due to limited parallel corpora and gold standard annotations (Lameris and Stymne, 2021). Recently, Scots has been added to the Scottish curriculum in schools in an attempt to revitalize and maintain use of the language.

Modeling Scots for speech-based LLMs is challenging due to its unique position on a linguistic continuum with Scottish Standard English (SSE). This often results in code-mixing, making it difficult for models to delineate language boundaries. Additionally, Scots features high regional variation in its sound system, and the vowel length is determined by the phonetic environment, a feature not present in standard English. In Scots, plural verbs take an -s suffix (e.g., "the men is lachin") unless the subject is an immediately adjacent pronoun (Millar, 2018, 2023; Purves and Society., 1997). These factors, combined with the "Scottish Cringe" - a socio-cultural internalized feeling of linguistic inferiority that can lead to code-switching in formal recording environments — make it difficult to obtain truly representative data (MoChridhe, 2020).

[2] https://www.gov.scot/policies/languages/scots

3.2 Nubi

Nubi (ISO 639: kcn) is an Arabic-based Creole language spoken by approximately 50,000 people, primarily in Uganda and Kenya (Gussenhoven, 2006; Owens, 1991; Avram, 2020). It's Agglomerated Endangerment Status according to Glottolog is shifting, and the Catalogue of Endangered Languages lists it as threatened (Hammarström et al., 2026; Campbell et al., 2022). It evolved in the late 19th century within military camps in Sudan and Upper Egypt, separating from its lexifier, Sudanic Arabic, around 1885 (Wellens, 2005; Kihm, 2011; Owens, 2014). Approximately 90% of its basic vocabulary is derived from Arabic (Owens, 1985). Similar to English, Nubi has a balanced 5 vowel system and has two primary dialects: Ugandan (Bombo) and Kenyan Nubi (Owens, 2006).

Modeling Nubi presents several unique challenges for speech LLMs. Phonologically, Nubi exhibits significant variation between lento (slow) and allegro (fast) speech. In allegro forms, vowels are frequently elided through processes of syncope (internal deletion) and apocope (final deletion), which often obscures the underlying CV syllable structure and complicates phonetic boundary detection for speech models (Wellens, 2003; Owens, 1985). Furthermore, while Nubi has lost the pharyngealized and geminate consonants of its Arabic lexifier, it has "imported" phonemes such as /p/, /v/, and /ɲ/ from African substrate and adstrate languages like Swahili and Luganda.

Morphosyntactically, Nubi lacks grammatical gender and the complex person-number verbal inflections found in Arabic. Plural marking is optional and can be indicated through suffixation (e.g., *-á*, *-ín*) or stress shifts. This grammatical optionality and the use of suprasegmental features—where final stress and high pitch are used to distinguish passive from active verb forms—make Nubi particularly difficult to model accurately through audio alone (Wellens, 2003).

Syntactically, Nubi follows a strict Subject-Verb-Object (SVO) word order (Wellens, 2003; Amer and Iryna, 2023). Minor category lexical items, such as the definite article (de) and cardinal numerals, are postnominal (following the noun), diverging from the prenominal structures typical of Arabic. These combined factors make it difficult to use multilingual LLMs for Nubi ASR.

4 Related Works

4.1 Automatic Speech Recognition

ASR for low-resource and endangered languages is a well-established research topic with a long history, resulting in a vast body of prior literature (Besacier et al., 2014). Methodologically, the field has progressed from early acoustic modeling using graphemes, Hidden Markov Models (HMMs), and Dynamic Time Warping to modern Connectionist Temporal Classification (CTC) and wav2vec-based speech Large Language Models (Le and Besacier, 2009; Ranathunga et al., 2023; Pratap et al., 2024; Babu et al., 2022; Radford et al., 2023). Despite these advancements, the majority of the 7000+ languages in the world are still low-resource, with many of them being endangered. Organizing fieldwork to obtain realistic recordings from native speakers and gold labels from capable translators is an extremely complex task. For instance, Eischens and Hedding (2024) perform extensive fieldwork for San Martín Peras Mixtec, which is a variety of Mixtec (in this paper, we analyse `WER` on Southwestern Tlaxiaco Mixtec, which is distant from this dialect but has substantial overlap).

Recent research at the University of Washington has leveraged large multilingual models to mitigate these issues. Liang and Levow (2025) benchmarked `MMS` and `XLS-R` on Cicipu, Mocho', Toratán, Ulwa, and Upper Napo Kichwa, finding that fine-tuned multilingual ASR models can substantially reduce the transcription burden for low-resource languages. Additionally, (Mainzinger and Levow, 2024) investigated ASR for the American indigenous language Mvskoke, and their analysis supported the above findings.

Future work on expanding the scope of speech technologies to endangered languages should be carried out in tandem with HCI researchers (Reitmaier et al., 2022), local communities, and both direct and indirect stakeholders (Imam et al., 2025; Alabi et al., 2025).

4.2 Resources for Scots

The Scots language represents a linguistic continuum between Scottish Standard English and "Broad Scots", featuring high regional variation (Douglas, 2003). Despite having over 1.5 million speakers, it remains low-resource in natural language processing (NLP), with a critical scarcity of annotated audio data compared to higher-resource languages (Blaschke et al., 2023).

4.2.1 Speech Corpora and Accessibility

The primary resource for the language is the Scottish Corpus of Texts & Speech (SCOTS), which offers over 800,000 words of oral history, interviews, and casual conversation (Anderson et al., 2007). This dataset is publicly available and free of charge, providing synchronized orthographic transcriptions and metadata.

Other significant, publicly accessible Scots speech resources include:

- The Scots Syntax Atlas (SCOSYA): Comprises 275 hours of conversational audio from 530 speakers across 146 locations, accompanied by acceptability judgments (Smith et al., 2019; Adger et al., 2023).
- Google's Multi-speaker British Isles Accents: An open-source dataset containing high-quality audio with roughly 10 hours specifically dedicated to Scottish accents (Demirsahin et al., 2020).
- Freiburg English Dialect Corpus (FRED): Includes approximately 300 hours of speech from the UK, with specific subsets for the Hebrides and Scottish Highlands (Anderwald and Wagner, 2007).
- Mozilla Common Voice: Provides spontaneous speech data for Scots, used in recent low-resource ASR challenges (Ardila et al., 2020).

While the total volume of documented Scots audio exceeds 600 hours, the percentage of gold-standard labeled data suitable for training neural models is much smaller. Most corpora are labeled with orthographic transcriptions, though some, like the Google accent dataset, provide high phoneme coverage for phonetic analysis. Works such as MoChridhe (2020) examine SCOTS and the Wee Windaes[3] project within the framework of digital humanities and critical engagement.

4.2.2 NLP Research on Scots Speech/Text

NLP work on Scots speech was historically scarce, but recent efforts have shifted toward ASR using transformer-based architectures. Babu et al. (2022) utilized mere 2 hours of Scots data to pre-train the XLS-R model, highlighting the extreme data constraints researchers face. Rafkin et al. (2026)

[3] `https://wee-windaes.nls.uk`

explored task arithmetic by fine-tuning Whisper-tiny and Whisper-large-v3. They leveraged the genetic relationship between Scots and English to improve performance on spontaneous speech sets. Lameris and Stymne (2021) developed Part-of-Speech (POS) tagging models specifically for Scots using annotated sentences derived from the SCOTS corpus.They manually tagged a small set of data, examined zero-shot and transfer learning methods to English, and fine-tuned a model to determine parts of speech. Sonderegger et al. (2022) built the Integrated Speech Corpus Analysis (ISCAN) system to perform large-scale automated acoustic analysis across Scots corpora, such as SoTC and SCOTS (The SPADE Project).

4.3 Resources for Nubi

Speech modeling for Nubi faces the typical hurdles of low-resource and endangered languages. Nubi currently lacks a robust digital presence, with existing written materials often unstandardized and insufficient for complex model training. Recent efforts have shifted toward systematic documentation; notably, Otieno (2024) developed a framework to build linguistic corpora for Kisii Town Heritage Nubian, which involves collecting recorded audio, video, and manual transcripts. Additionally, the Spontaneous Speech Dataset by Mozilla Common Voice serves as an important, albeit low-resource, external resource for naturalistic audio (Ardila et al., 2020).

To the best of our knowledge, there is currently no documented prior work in the sources regarding speech or text modeling for the Nubi language.

5 Experimental Setup

We fine-tune three multilingual ASR models on all 21 languages: MMS (`facebook/mms-1b-all`) (Pratap et al., 2024), XLS-R (`facebook/wav2vec2-xls-r-1b`) (Babu et al., 2022), and Whisper Large-v3 (`openai/whisper-large-v3`) (Radford et al., 2023). All three are trained on multilingual data and have comparable parameter counts (1B for MMS and XLS-R, 1.5B for Whisper), providing a reasonable basis for practitioner-oriented comparison. Of the 21 languages, MMS has prior exposure to Toba Qom and Greek (of which Cypriot Greek is a dialect); XLS-R and Whisper to Greek only.

Audio is resampled to 16 kHz mono, and transcripts are normalized with a language-agnostic cleaning function. Utterances above the 97.5th-percentile duration per language are excluded during training to avoid memory errors.

For MMS and XLS-R, we adopt parameter-efficient fine-tuning via bottleneck adapters (Houlsby et al., 2019), freezing the pre-trained backbone and updating only the adapter layers and the CTC head ($\sim$0.25% of parameters). Whisper is fine-tuned end-to-end using `Seq2SeqTrainer` with the pre-trained tokenizer intact. MMS and XLS-R are trained for 15 epochs (lr $= 1\times10^{-3}$, batch size 16); Whisper for 8 epochs (lr $= 1\times10^{-5}$, same batch schedule). The best checkpoints are selected by validation `WER`.

We note that these models differ in architecture (encoder-decoder vs. self-supervised CTC) and fine-tuning regime (full vs. adapter). Our goal is not to isolate architectural effects but to benchmark a suitably wide variety of practitioner-accessible options under realistic low-resource constraints, thus providing a substrate for future researchers to build upon.

We report `WER` as the primary metric and `CER` as secondary, computed using the Hugging Face `evaluate` library. For MMS and XLS-R, we additionally evaluate beam search decoding with a unigram vocabulary derived from training transcripts using `pyctcdecode` (Heafield, 2011), and further with 4-gram ARPA language models. All experiments were run on a single NVIDIA L40/L40S GPUs on the University of Washington Hyak cluster, requiring 1–3 GPU hours per language.

We provide further details about the experimental setup in Appendix C.

6 Results

We present the results of our fine-tuning experiments in Table 1. The results are separated by continent of origin. For MMS and XLS-R, we collected `WER` and `CER` with and without unigram Language Model (LM) decoding. The bolded and underlined numbers are the best `WER` and `CER` for a given language, respectively. Interestingly, MMS achieves the lowest `WER` amongst all three models for 9 out of the 21 languages. Whisper achieves the lowest `WER` for 8 of 21 languages, whereas XLS-R only achieves this for 4 languages. Similarly, MMS achieves the lowest `CER` for 13 of the 21 languages. 6 languages achieve their lowest `CER` with XLS-R, whereas only 2 do so with Whisper.

Languages		MMS-1B-All				XLS-R-1B				Whisper Large-v3	
		w/o LM		w/ LM		w/o LM		w/ LM		Full fine-tuning	
Language	ISO 639 Code	W	C	W	C	W	C	W	C	W	C
Africa											
Bukusu	bxk	0.520	0.148	**0.512**	0.147	0.688	0.191	0.684	0.198	0.532	0.173
Chiga	cgg	**0.473**	0.118	0.475	0.122	0.759	0.230	0.749	0.222	0.524	0.165
Nubi	kcn	0.625	0.292	0.622	0.288	0.588	0.285	**0.570**	0.325	0.627	0.385
Konzo	koo	0.682	0.175	0.686	0.185	0.844	0.237	0.840	0.302	**0.643**	0.199
Lendu	led	0.358	0.130	0.348	0.127	0.322	0.120	0.318	0.120	**0.308**	0.126
Kenyi	lke	**0.553**	0.140	0.556	0.139	0.815	0.256	0.783	0.254	0.581	0.172
Thur	lth	1.001	0.771	0.999	0.780	0.364	0.156	0.362	0.159	**0.325**	0.167
Ruuli	ruc	0.589	0.137	**0.583**	0.136	0.697	0.187	0.689	0.218	0.634	0.203
Amba	rwm	0.603	0.197	0.594	0.195	0.561	0.195	0.557	0.205	**0.531**	0.201
Rutoro	ttj	0.242	0.043	**0.241**	0.043	0.349	0.063	0.347	0.064	0.283	0.088
Kuku	ukv	0.422	0.133	0.415	0.131	0.394	0.127	**0.381**	0.124	0.406	0.141
Americas											
Wixárika	hch	0.677	0.161	0.673	0.160	0.560	0.130	0.551	0.130	**0.509**	0.149
Southwestern Tlaxiaco Mixtec	meh	0.423	0.170	**0.420**	0.169	0.445	0.183	0.436	0.178	0.634	0.466
Michoacán Mazahua	mmc	0.764	0.332	0.765	1.121	0.715	0.297	**0.715**	0.946	0.842	0.622
Toba Qom	tob	0.595	0.194	**0.595**	0.413	0.657	0.203	0.653	0.338	0.634	0.280
Papantla Totonac	top	0.639	0.157	**0.638**	0.155	1.000	0.969	1.623	0.998	1.023	0.595
Asia & Pacific											
Betawi	bew	0.499	0.171	**0.494**	0.170	0.778	0.305	0.763	0.297	0.548	0.349
Western Penan	pne	0.348	0.127	0.342	0.125	0.380	0.151	0.366	0.146	**0.264**	0.113
Europe											
Gheg Albanian	aln	0.616	0.275	0.607	0.273	0.813	0.371	0.762	0.343	**0.472**	0.272
Cypriot Greek	el-CY	0.464	0.144	0.458	0.141	0.922	0.392	0.924	0.480	**0.456**	0.260
Scots	sco	0.306	0.113	0.300	0.111	0.302	0.115	**0.297**	0.113	0.556	0.418
Average		**0.543**	**0.197**	**0.539**	**0.244**	**0.617**	**0.246**	**0.637**	**0.293**	**0.540**	**0.264**

Table 1: Word Error Rate (W) and Character Error Rate (C) of **MMS-1B-All**, **XLS-R-1B**, and **Whisper-large-v3** fine-tuned on 21 low-resource languages, rounded to three decimal places. Results are reported without (**w/o**) and with (**w/**) unigram decoding for **MMS-1B-All** and **XLS-R-1B**. Lowest WER is **bolded** and lowest CER is underlined for each language. Row tints indicate geographic region: Africa, Americas, Asia/Pacific, Europe.

LM decoding consistently improves model performance. When fine-tuning MMS, using LM decoding reduces the WER of 17 languages, whereas with XLS-R, LM decoding helps reduce the WER of 19 languages. For languages where the use of LM decoding worsened the WER, the increase is not more than 1% absolute WER. These include Chiga, Konzo, Kenyi, Michoacán Mazahua, and Cypriot Greek. An exception to this trend is Papantla Totonac.

Of all languages, MMS achieves the best performance on Rutoro (WER = 0.241 and CER = 0.043), XLS-R achieves the best performance on Scots (WER = 0.297 and CER = 0.113), and Whisper performs best on Western Penan (WER = 0.264 and CER = 0.113).

6.1 How Much Data is Enough Data?

We attempt to ablate the number of training hours to capture model performance deterioration. In Table 2 we evaluate the models on two training splits of Scots and Nubi - a 1 hr training data split and a 50% split (3 hrs for Nubi and 5 hrs for Scots). We observe that Whisper outperforms other models in extremely low-resource scenarios by significant margins. Whisper's low WER on just 1 hr of Scots

Model \ Split	Nubi (kcn)		Scots (sco)	
	1 hr	3 hr	1 hr	5 hr
MMS-1B-all	1.197	0.673	0.996	**0.338**
XLS-R-1b	1.040	0.752	0.989	0.482
Whisper large-v3	**0.883**	**0.641**	**0.353**	0.476

Table 2: Word Error Rate (%) of **MMS-1B-All**, **XLS-R-1B**, and **Whisper Large-v3** on **Nubi** (kcn) and **Scots** (sco) across two training splits: **One** (1 hr) and **Mid** (50%). **Bold** indicates the lowest WER per language.

Language	Greedy	Unigram	4-gram	%Δ_A
Nubi (kcn)	0.556	0.543	0.482	-13.1%
Lendu (led)	0.322	0.318	0.278	-14.0%
Thur (lth)	0.368	0.367	0.311	-15.5%
Kuku (ukv)	0.394	0.381	0.352	-10.4%
Wixárika (hch)	0.560	0.556	0.506	-9.6%
Betawi (bew)	0.778	0.761	0.617	-20.7%
Cypriot Greek (el-CY)	0.923	0.924	0.671	-27.3%
Scots (sco)	0.302	0.297	0.244	-19.2%

Table 3: XLS-R-1B decoding results for 8 selected languages. **4-gram**: Set A (α=0.5, β=1.0, beam= 100). %Δ_A: percentage WER change relative to greedy decoding. Intensity of Green is proportional to $|\Delta_A|$.

is evidence that model performance in a given language is directly correlated with the amount of training data in that language (Whisper is trained on $438k+$ hrs of English audio alone).

6.2 Effect of n-gram Language Decoding

To improve CTC decoding beyond unigram baselines, we augment CTC beam search with an external n-gram language model estimated from training text, combining acoustic and LM scores via shallow fusion (weighted by α and β). We sweep $n \in \{3,4\}$, α, β, and beam width across all 21 languages for the models. Table 3 provides WER for select languages for the best-performing sweep parameters, and Table 6 in Appendix D gives full sweep results across all models and settings.

We observe that 4-gram LM decoding consistently outperforms greedy and unigram decoding for most languages, with improvements up to 27.3% in WER. Improvements are geographically broad, while only a few languages show mild degradation under specific settings overall.

6.3 What Drives Cross-Model Variance?

A natural question is whether the observed performance gaps reflect architectural differences, language properties, or our fine-tuning and decoding setup. While we cannot fully disentangle these factors, three patterns in our results emerge. First, Whisper's advantage is concentrated in languages where either the target language or a closely related high-resource language is well represented in its pretraining data: it achieves the lowest WER on Scots in the 1-hour setting (Table 2), Western Penan, Gheg Albanian, and Cypriot Greek — all languages with substantial English, Indonesian/Malay, or Greek pretraining exposure to draw on. Second, MMS performs best on languages with simpler orthographies and character inventories well-covered by its 1162-language pretraining (e.g., Rutoro, Chiga, Bukusu), where its language-specific adapters appear to give it a head start over architectures without per-language conditioning. Third, the largest gains from n-gram LM decoding (Table 6) accrue to XLS-R rather than MMS, suggesting that XLS-R's CTC outputs are more under-constrained at the lexical level and benefit disproportionately from external lexical priors. Architectural and pretraining-distribution effects, therefore, appear intertwined; isolating them would require controlled pretraining ablations beyond the scope of this work.

7 Linguistic Analysis

In addition to our WER and CER results, we were interested in analyzing the errors of two languages in particular to determine if there were linguistic features that proved to be especially difficult for the model, or if there were noticeable differences in the kinds of errors made by the model for the two languages. For this closer analysis we chose Scots and Nubi because both are closely associated with a high-resource language and might therefore be expected to perform well, but had different outcomes with the ASR models. Through our analysis we hoped to determine a likely cause for these contrasting results. To perform this error analysis, we used the XLS-R outputs of Scots and Nubi, taking approximately 10% of the data outputs and performing an error analysis of substitutions, deletions, and insertions. Because Scots shares so much of its lexicon and syntax with English, we were also able to make some phonological inferences for Scots.

7.1 Scots

Of the 75 generated transcripts in Scots, eight were examined. Many of the errors in the Scots data appear to be misalignments with Scots phonology and

English orthography. This is particularly apparent in the vowel substitution errors, which appear to align well with Scots vowels. For example, the transcription of "walking" as "walken" is potentially the result of the realized vowel being lower than the "standard" English vowel that is represented by the letter "i." This word is also an example of a complication that arises when analyzing transcription errors in an orthography that uses digraphs. The use of "ng" to represent the phoneme /ŋ/ turns one phoneme into two characters, and when the model transcribes "walken" with only an "n" instead of an "ng," it is analyzed as a deletion error, when it is perhaps more linguistically accurate to analyze it as a substitution error between the phonemes /n/ and /ŋ/.

There were several consonant substitutions, particularly in English words, that do not align with a potential phonological and orthographic disagreement between Scots and English. For example, the word "gymnastics" was transcribed as "Jimnastic" and "physique" was transcribed as "fhasic." This does not appear to represent some underlying phonological difference in the consonants, but rather a failure of the model to select the correct characters.

It was common for word-final characters to be deleted, both consonants and vowels, as evidenced by words like "hole" becoming "hol" and "chill" becoming "chil." These characters do not directly represent phones in the spoken data, and as such we would not expect these deletions to be the result of some difference between the Scots phonology and the English orthography.

As previously noted, much of the Scots lexicon is shared with English. In this dataset, the Scots words that are not common with English tend to be short, high-frequency words. While there were some errors, these uniquely Scots words like *oot* "out" and *maist* "most" tended to be transcribed faithfully. This may be due to their frequency in the data, but it may also be a result of their relatively phonetic spellings, in contrast to the English words.

It's also important to note the difference in punctuation between the languages' transcripts. The Nubi data has no punctuation of any kind and relatively little capitalization. In contrast, the Scots transcripts have significant punctuation, including hyphens denoting interrupted speech and apostrophes in contractions. This punctuation might have increased the error rates in Scots, as punctuation increases the number of possible characters for the model to choose from, while not corresponding to phonetic information in the data, and still contributing to error counts.

7.2 Nubi

Our analysis of Nubi was limited by our relative unfamiliarity with the language and orthography of the dataset, meaning phonological and morphosyntactic analysis was not possible. However, we were able to identify some potential trends in the errors. Of the 100 transcripts in the test, ten were analyzed for common substitution, deletion, and insertion errors in comparison with the provided gold transcriptions.

Certain substitutions between vowels appeared throughout the entries, especially between the pairs "i" and "e" and "o" and "u". Assuming that the orthography used in this dataset is typical, this is likely representative of the relative similarities of the pairs of front unrounded vowels and back rounded vowels. Similarly, the nasals "m" and "n" were commonly substituted, which is unsurprising given their similarity and the generally lower perceptual differences between nasals as compared to vowels or non-nasal consonants (Hura et al., 1992). There was also a substitution that occurred in a specific environment, which may indicate an error happening with the LM. In all instances of the word "ab", the model transcribed "al", but the "b" > "l" substitution did not occur outside of that specific environment.

Vowel deletions were very common throughout the data, for each of the five vowels, words initially, medially, and finally. There does not appear to be a significant pattern to the single vowel deletions. There was a pattern, however, with the deletion of repeated vowels, which may indicate a problem with the model. Every time there was a repeated vowel in the gold transcript, the ASR model would transcribe only the first of the repeating vowels and delete the rest. Consonant deletions were less common than vowel deletions, generally, and the majority of the deleted consonants were nasals. In addition to the nasals, there was a trend of "w" being deleted intervocalically. Without knowledge of the orthography, it is difficult to confidently infer a reason for this, but Nubi does have the phoneme /w/, which may be represented by the letter "w" (Wellens, 2003). If so, this could indicate a failure of the model to recognize the approximant in this environment, but it may also be a reflection of a phonological phenomenon.

In two of the transcripts, there were English loanwords in the transcript that proved to be difficult for the model to transcribe. The loanwords included "mindset", "school", "government", "busy", "creative", and "typhoid". Some of these words resulted in a series of substitution errors, for example, "mindset" became "ma endist" and "busy" became "bse," whereas "school" was deleted in its entirety, and "creative" resulted in a mix of substitution and deletion with "kwet." This indicates that the model struggles with multilingual ASR.

The most surprising and egregious errors made by the model were insertions that significantly outnumbered the actual word count of the gold transcript. For example, one entry was eight words long, but the model returned a transcript that was 120 words long; another was five words long and returned a transcript that was 72 words long. This would significantly affect WER and may be the cause of Nubi's relatively poor results in the tests.

7.3 Scots vs. Nubi: A Comparative View

Both languages stand in a "shadow" relationship to a high-resource language — English for Scots and Arabic for Nubi - with varied error patterns. Scots errors cluster at the interface between Scots phonology and English orthography: vowel substitutions reflecting genuine Scots realizations rendered in English-like spellings, non-phonetic word-final deletions, and punctuation-driven noise. The model carries a strong English prior, producing Scots-adjacent transcripts mis-anchored to English conventions. Nubi errors, by contrast, lack any comparable anchor: segmental confusions are broader and less systematic, and the most consequential failures are catastrophic insertion blowups (e.g., 8-word references producing 120-word hypotheses), consistent with a decoder that fails to terminate reliably. Proximity to a high-resource language thus appears double-edged — it supplies useful priors but may also impose a wrong frame, whereas its absence can leave the decoder structurally unstable.

8 Conclusion

In this work, we benchmarked MMS, XLS-R, and Whisper on 21 low-resource languages from Mozilla Common Voice Spontaneous Speech and analyzed both model-level trends and language-specific errors. MMS delivered the strongest overall performance, while XLS-R achieved the largest relative gains from n-gram LM decoding, with improvements up to 27.3% WER over greedy decoding. Across models, 3-gram and 4-gram decoding consistently outperformed unigram decoding, confirming that explicit n-gram LM integration is crucial for stronger CTC ASR. Our linguistic analysis of Scots and Nubi further showed recurring substitution, deletion, and insertion patterns tied to orthography, phonology, and punctuation. In Nubi, severe insertion-heavy outputs suggest transcription instability under low-resource conditions. Our findings reinforce that careful decoding and language-aware analysis are essential for robust ASR in endangered language settings, and for practical revitalization and accessibility tools development.

9 Limitations

Despite strong results from fine-tuned models on some languages, we witnessed poor performance on certain other languages, such as Konzo (koo) and Papantla Totonac (top). This shows that mere fine-tuning is sometimes not enough to obtain reasonable ASR transcription accuracy. Other techniques, such as data augmentation and transfer learning, should be taken into consideration.

Another limitation of our work is the dataset itself, which has between 4 and 14 hours of training data across all 21 languages. This is certainly not enough data to sufficiently train billion-parameter models to human-level accuracy.

A major limitation of our work is the lack of a comprehensive linguistic error analysis shaping a narrative across all 21 languages. We hoped to perform further analysis across the morphological typologies, but were limited by time.

We would also like to highlight Meta's Omnilingual ASR model, which has been trained on 1600+ languages, including 20 of the 21 languages from our research (Omnilingual et al., 2025). We have not included ASR results from this model, as it was recently released, and we urge future researchers to work with and support such projects.

A further limitation concerns the comparability of the three models we evaluate. Whisper is an encoder-decoder model trained with weak supervision and fine-tuned end-to-end, while MMS and XLS-R are self-supervised CTC models fine-tuned via lightweight adapters. Although we chose these models because they are the most widely used multilingual ASR systems available to practitioners, the differences in architecture, pretraining objec-

tive, and fine-tuning regime mean that observed performance gaps cannot be cleanly attributed to any single factor. Our discussion in §6.3 surfaces likely contributors, but disentangling architecture from pretraining-distribution effects would require controlled ablations in the future.

Relatedly, the amount of training data per language varies substantially (Table 4) — from roughly 4 hours for Toba Qom to over 14 hours for Ruuli — as do the number of contributing speakers and utterance length distributions (Table 5). This makes cross-linguistic comparisons of WER difficult to interpret as comparisons of language difficulty per se: a language with more training hours, more speaker diversity, or shorter utterances has structural advantages independent of its linguistic properties. Our 1-hour and 50% ablation in Table 2 partially addresses this for Scots and Nubi, but a fuller study would normalize training conditions across all 21 languages.

Finally, our investigation of n-gram language model decoding is limited to shallow fusion with unigram and 4-gram models estimated from training transcripts, and our linguistic analysis of punctuation effects in Scots is qualitative. A more systematic study of how lexical resources, LM order, and punctuation handling interact with model architecture across typologically diverse languages remains an important direction for future work.

10 Ethical Considerations

Our work uses the Mozilla Common Voice Spontaneous Speech dataset, which is released under a CC0 license and collected under Mozilla's own consent and contributor framework (Ardila et al., 2020; Mozilla Data Collective, 2026). We did not collect new data, conduct fieldwork, or interact directly with speakers of any of the 21 languages studied. This shapes both what our work can claim and where its ethical risks lie.

First, we are not members of the Scots or Nubi speech communities, and our linguistic analysis in §7 is therefore an external reading constrained by the orthographies and conventions chosen by the dataset's contributors and validators. We have tried to be explicit about this limitation, particularly for Nubi, where our unfamiliarity with the orthography prevented deeper morphosyntactic analysis. Conclusions about either language should be taken as hypotheses to be verified by community linguists rather than settled findings.

Second, ASR systems for low-resource and endangered languages can cause real harm if deployed without community oversight. Transcription errors of the kinds we document — particularly the catastrophic insertion failures observed in Nubi — could distort downstream documentation, language-learning tools, or accessibility applications, and could misrepresent how a language sounds or behaves. We therefore caution against treating our fine-tuned models as deployment-ready artifacts; they are benchmarks, not products. Any downstream use should involve review by speakers and community stakeholders, with attention to the specific failure modes documented in §7.

Finally, we recognize that benchmarking work like ours can itself shape research priorities for under-resourced languages by privileging those with existing labeled data. We have tried to mitigate this by performing per-language analysis rather than reporting only averages, and by foregrounding linguistic detail for two specific languages, but we acknowledge that the choice of dataset constrains which communities receive research attention.

11 Acknowledgements

We would like to thank the Student Technology Fund at the University of Washington for providing access to its Hyak GPU clusters for our model fine-tuning and evaluation. We also acknowledge the efforts of the Hyak team in helping us navigate the distributed GPU cluster as first-time users. Our work was possible in large parts due to the consistent feedback and ideas given by Prof. Gina-Anne Levow. We acknowledge her support in helping us shape this manuscript in its current form and structure.

References

2003. Automatic speech recognition for second language learning: How and why it actually works.

D Adger, E Jamieson, J Smith, G Thoms, and C Heycock. 2023. 'when intuitions (don't) fail': combining syntax and sociolinguistics in the analysis of scots. *English Language & Linguistics*.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Ahmed Amer and Lenchuk Iryna. 2023. Relexification and dialect levelling in the genesis of creoles: the case of the arabic-based creole, nubi. *Research Result. Theoretical and Applied Linguistics*, 9(2):49–72.

Jean Anderson, Dave Beavan, and Christian Kay. 2007. *SCOTS: Scottish Corpus of Texts and Speech*, pages 17–34. Palgrave Macmillan UK, London.

Lieselotte Anderwald and Susanne Wagner. 2007. *FRED — The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data*, pages 35–53. Palgrave Macmillan UK, London.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Andrei A Avram. 2020. Substrate and adstrate influence on (ki) nubi: Evidence from early records. *Academic Journal of Modern Philology*, (10):7–21.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Janine Butler, Brian Trager, and Byron Behm. 2019. Exploration of automatic speech recognition for deaf and hard of hearing students in higher education classes. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 32–42, New York, NY, USA. Association for Computing Machinery.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, and Kaori Ueki. 2022. The catalogue of endangered languages (elcat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source multi-speaker corpora of the English accents in the British isles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6532–6541, Marseille, France. European Language Resources Association.

Joanna Dolinska, Shekhar Nayak, and Sumittra Suraratdecha. 2024. Akha, dara-ang, karen, khamu, Mlabri and urak lawoi' language minorities' subjective perception of their languages and the outlook for development of digital tools. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 94–99, St. Julians, Malta. Association for Computational Linguistics.

Fiona M. Douglas. 2003. The scottish corpus of texts and speech: Problems of corpus design. *Literary and Linguistic Computing*, 18(1):23–37.

Gerhard W Dueck. 2024. Using ai to help preserve indigenous oral histories. In *2024 IEEE International Humanitarian Technologies Conference (IHTC)*, pages 1–5. IEEE.

Ben Eischens and Andrew A. Hedding. 2024. San martín peras mixtec. *Journal of the International Phonetic Association*, 54(2):811–852.

Sanchit Gandhi. 2022. Fine-tune whisper for multilingual asr with transformerss. `https://huggingface.co/blog/fine-tune-whisper`. [Blog post; Accessed on 11 March 2026].

Mugisho Matabaro Gedeon, Swati Samantaray, and Kwigomba Bulonza René. 2024. *Changing the Trajectory: Preserving the Linguistic Diversity of Shi Language Using AI and NLP*, pages 57–69. Springer Nature Singapore, Singapore.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in ai for people with disabilities sbg@a research roadmap. *SIGACCESS Access. Comput.*, (125).

Carlos Gussenhoven. 2006. Between stress and tone in nubi word prosody. *Phonology*, 23(2):192–223.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. Glottolog 5.3. Available online at http://glottolog.org, Accessed on 2026-03-18.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Susan L Hura, Björn Lindblom, and Randy L Diehl. 1992. On the role of perception in shaping phonological assimilation rules. *Language and speech*, 35(1-2):59–72.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud'hommeaux. Improving asr output for endangered language documentation. *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Alain Kihm. 2011. Plural formation in nubi and arabic: A comparative study and a word-based approach. *Brill's Journal of Afroasiatic Languages and Linguistics*, 3(1):1 – 21.

Vincent Koc. 2025. Generative ai and large language models in language preservation: Opportunities and challenges. *ArXiv*, abs/2501.11496.

Bernd Kortmann, Clive Upton, Edgar W. Schneider, Kate. Burridge, and Rajend. Mesthrie. 2008. Varieties of english.

Harm Lameris and Sara Stymne. 2021. Whit's the richt pairt o speech: PoS tagging for Scots. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48, Kiyv, Ukraine. Association for Computational Linguistics.

Viet-Bac Le and Laurent Besacier. 2009. Automatic speech recognition for under-resourced languages: Application to vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1471–1482.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.

Julia Mainzinger. 2024. Technology and language revitalization: A roadmap for the mvskoke language. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 7–12, St. Julians, Malta. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Robert McColl Millar. 2018. *Modern Scots: An Analytical Survey*. Edinburgh University Press.

Robert McColl Millar. 2023. *A History of the Scots Language*. Oxford University Press.

Race MoChridhe. 2020. *Digital humanities and critical engagement: The case of the Scottish Corpus of Texts Speech and Wee Windaes*, 1st edition edition, page 32–51. Routledge.

Santiago Omar Caballero Morales, Gladys Bonilla Enríquez, and Felipe Trujillo Romero. 2013. Speech-based human and service robot interaction: An application for mexican dysarthric people. *International Journal of Advanced Robotic Systems*, 10(1):11.

Christopher Moseley. 2010. *Atlas of the world's languages in danger*, 3 edition. UNESCO Publishing, Paris.

Mozilla Data Collective. 2026. Dataset: Mozilla data collective. Accessed: 2026-03-23.

Aref A. Murshed, Ali alrahamneh, Al-Hareth Alhalalmeh, and Mohammed Al-Badawi. 2025. *The Role of Technology in Preserving Indigenous Cultures and Languages*, pages 2399–2409. Springer Nature Switzerland, Cham.

Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Peter Nyansera Otieno. 2024. Framework for building linguistic corpora for a large language model project for the heritage nubian language of kenya. *Journal of Languages, Linguistics and Literary Studies*, 4(3):139–144.

Jonathan Owens. 1985. The origins of east african nubi. *Anthropological Linguistics*, 27(3):229–271.

Jonathan Owens. 1991. Nubi, genetic linguistics, and language classification. *Anthropological Linguistics*, 33(1):1–30.

Jonathan Owens. 2006. Creole arabic. *Encyclopedia of Arabic Language and Linguistics, Leiden–Boston, Brill*, pages 518–527.

Jonathan Owens. 2014. The morphologization of an arabic creole. *Journal of Pidgin and Creole Languages*, 29(2):232–298.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation Conservation*, 15:491–513.

David. Purves and Saltire Society. 1997. A scots grammar : Scots grammar and usage : Scots that haes–.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Emma Rafkin, Dan DeGenaro, and Xiulin Yang. 2026. Task arithmetic with support languages for low-resource asr. *Preprint*, arXiv:2601.07038.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

SIL International. 2024. Ethnologue: Languages of the world. https://www.ethnologue.com. Accessed: 2026-03-23.

Jennifer Smith, David Adger, Brian Aitken, Caroline Heycock, E Jamieson, and Gary Thoms. 2019. The scots syntax atlas. https://scotssyntaxatlas.ac.uk. [Accessed on 16 March 2026].

Morgan Sonderegger, Jane Stuart-Smith, Michael McAuliffe, Rachel Macdonald, and Tyler Kendall. 2022. Managing data for integrated speech corpus analysis in speech across dialects of english (spade). In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Weina Sun. 2023. The impact of automatic speech recognition technology on second language pronunciation and speaking skills of efl learners: a mixed methods investigation. *Frontiers in Psychology*, Volume 14 - 2023.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.

P. Trudgill. 1984. *Language in the British Isles*. Cambridge University Press.

Joost van Doremalen, Lou Boves, Jozef Colpaert, Catia Cucchiarini, and Helmer Strik. 2016. Evaluating automatic speech recognition-based language learning systems: a case study. *Computer Assisted Language Learning*, 29(4):833–851.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for Hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66, St. Julians, Malta. Association for Computational Linguistics.

Patrick von Platen. 2021. Fine-tuning xls-r for multi-lingual asr with transformers. https://huggingface.co/blog/fine-tune-xlsr-wav2vec2. [Blog post; Accessed on 11 March 2026].

Patrick von Platen. 2023. Fine-tuning mms adapter models for multi-lingual asr. https://huggingface.co/blog/mms_adapters. [Blog post; Accessed on 11 March 2026].

Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6(4):435–447.

Inneke Hilda Werner Wellens. 2003. *An Arabic creole in Africa: the Nubi language of Uganda*. Ph.D. thesis, [Sl: sn].

Inneke Hilda Werner Wellens. 2005. *The Nubi language of Uganda: an Arabic creole in Africa*, volume 45. Brill.

Wenqi Xiao and Moonyoung Park. 2021. Using automatic speech recognition to facilitate english pronunciation assessment and learning in an efl context: Pronunciation error diagnosis and pedagogical implications. *Int. J. Comput.-Assist. Lang. Learn. Teach.*, 11(3):74–91.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

A Dataset Overview and Corpus Statistics

ISO 639	Language	Country	Vitality	Speakers	Train (h)	Dev (h)	Test (h)
aln	Gheg Albanian	Albania, Kosovo, N. Macedonia	Institutional	4,705,860	6h 40m	1h 29m	0.75
bew	Betawi	Indonesia	Endangered	6,810,000	5h 47m	2h 19m	0.75
bxk	Bukusu	Kenya, Uganda	Institutional	1,225,900	10h 14m	1h 58m	0.75
cgg	Chiga	Uganda	Institutional	2,950,000	8h 2m	1h 3m	0.75
el-CY	Cypriot Greek	Cyprus	Institutional	1,189,200	6h 47m	1h 17m	0.75
hch	Wixárika	Mexico	Stable	66,700	6h 17m	58m	0.75
kcn	Nubi	Kenya, Uganda	Stable	51,100	10h 29m	1h 2m	0.75
koo	Konzo	DRC	Institutional	1,140,000	11h 26m	1h 6m	0.75
led	Lendu	DRC, Uganda	Stable	1,760,000	11h 56m	50m	0.75
lke	Kenyi	Uganda, DRC	Stable	101,000	8h 30m	1h 22m	0.75
lth	Thur	South Sudan	Stable	113,000	12h	51m	0.75
meh	Sw. Tlaxiaco Mixtec	Mexico	Stable	527,000	6h 43m	1h	0.75
mmc	Michoacán Mazahua	Mexico	Stable	154,000	7h 11m	1h 31m	0.75
pne	Western Penan	Malaysia	Endangered	3,400	8h 20m	1h 17m	0.75
ruc	Ruuli	Uganda	Stable	255,000	14h 13m	1h 5m	0.75
rwm	Amba	Uganda, DRC	Endangered	64,700	10h	1h 8m	0.75
sco	Scots	United Kingdom	Stable	1,599,200	6h 50m	59m	0.75
tob	Toba Qom	Argentina	Endangered	32,140	4h 27m	1h 58m	0.75
top	Papantla Totonac	Mexico	Stable	256,000	5h 37m	1h 36m	0.75
ttj	Rutoro	Uganda	Institutional	1,050,000	12h 25m	1h 27m	0.75
ukv	Kuku	South Sudan, Uganda	Endangered	102,000	8h 29m	50m	0.75

Table 4: Details of all 21 languages in the dataset w.r.t country of origin, current vitality status, number of speakers (all three as reported by Ethnologue (SIL International, 2024)), as well as sizes of the training/development/test sets.

B Utterance Statistics for Each Language

ISO 639	Language Name	N	Median (s)	Mean (s)	P95 (s)	P97.5 (s)	Max (s)
aln	Gheg Albanian	1367	23.40	23.47	36.21	38.86	73.37
bew	Betawi	1175	25.02	28.27	61.14	70.47	157.00
bxk	Bukusu	2542	17.86	17.31	27.11	29.27	72.18
cgg	Chiga	2625	12.28	13.84	28.46	34.11	86.98
el-CY	Cypriot Greek	1150	25.69	29.39	65.73	79.43	271.08
hch	Wixárika	1317	16.74	21.89	55.09	70.46	114.30
kcn	Nubi	2369	17.28	18.68	41.06	50.22	104.44
koo	Konzo	2889	16.67	16.64	27.40	31.86	112.97
led	Lendu	2576	19.01	19.09	25.74	26.64	54.04
lke	Kenyi	2408	17.10	16.07	25.24	27.53	47.74
lth	Thur	2880	17.17	17.42	24.41	28.30	1324.08
meh	Southwestern Tlaxiaco Mixtec	819	29.52	37.24	90.02	118.16	234.86
mmc	Michoacán Mazahua	682	39.60	50.13	129.44	144.61	332.82
pne	Western Penan	2267	16.42	16.73	21.74	24.19	36.32
ruc	Ruuli	2597	20.70	22.29	44.70	54.37	130.54
rwm	Amba	2103	20.27	20.43	31.78	36.59	94.43
sco	Scots	569	48.53	54.26	116.50	129.64	298.66
tob	Toba Qom	1255	15.70	21.30	51.59	59.55	191.92
top	Papantla Totonac	289	90.22	104.92	212.39	253.01	499.00
ttj	Rutoro	2741	18.65	19.21	28.80	32.31	82.08
ukv	Kuku	2109	16.96	17.26	28.82	33.85	114.77

Table 5: For each language in the dataset, columns indicate the number of utterances (**N**), median, mean, & max duration, along with 95th (P95) & 97.5th (P97.5) percentiles of utterance duration (Mozilla Data Collective, 2026).

C Detailed Experimental Setup

We experiment with a number of settings to evaluate the performance of popular ASR systems on languages with varying linguistic features and available labeled speech data.

C.1 Models

For fine-tuning on all 21 languages in our dataset, we consider three models - Massively Multilingual Speech (MMS) (`facebook/mms-1b-all`), XLS-R (`facebook/wav2vec2-xls-r-1b`), and Whisper (`openai/whisper-large-v3`). The choice of models for this experiment is driven by two factors: (1) all three models are trained on multilingual datasets, and (2) the models have a comparable number of trainable parameters - 1B for MMS and XLS-R and 1.5B for Whisper, making the performance comparison fair.

The Massively Multilingual Project (Pratap et al., 2024) is based on the Wav2Vec 2.0 architecture (Baevski et al., 2020) and trained on publicly available labeled audio recordings of people reading the New Testament in 1162 languages (more than $45k$ hours). Of the 21 languages in our dataset, this checkpoint is trained[4] on Toba Qom (`tob`) and Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

XLS-R (Babu et al., 2022) is also based on the Wav2Vec 2.0 architecture and trained on $436k$ hours of publicly available unlabeled speech recordings. The training covers 128 languages. Of the 21 languages in our dataset, this model is trained on 2 hours of Scots (`sco`) and $17k$ hours of Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

Whisper is trained on $680k$ hours of weakly labeled data across 99 different languages, collected from the internet (Radford et al., 2023) using a sequence-to-sequence transformer architecture. Of the 21 languages in our dataset, this model is trained only on Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

C.2 Pre-processing

All models use a common audio and text preprocessing pipeline. Audio files are loaded from disk with soundfile and converted to 16 kHz mono `float32` waveforms. Entries with non-positive duration or missing/empty transcriptions are removed before further processing. Transcripts are normalized using a light-weight but language-agnostic text cleaning function. After cleaning, any remaining empty or whitespace-only transcripts are discarded. To control pathological outliers in utterance length, we compute per-language duration statistics using a separate corpus analysis script. During training, we drop any train/validation utterances with raw duration above the 97.5th-percentile duration (`p97_5_sec`) per language to avoid running into CUDA errors during training.

For CTC-based MMS and XLS-R models, we construct a character-level vocabulary per language by aggregating all unique characters from the cleaned train and validation transcripts. Space is replaced by a word-delimiter symbol (|), and special `[UNK]` and `[PAD]` tokens are appended. For Whisper, we reuse the pre-trained tokenizer without modification and rely on its internal normalization and special tokens.

C.3 Fine-tuning Strategies

For MMS and XLS-R we adopt a parameter-efficient fine-tuning regime based on adapter layers. Houlsby et al. (2019) proposed adapter modules as a means to introduce trainable layers in existing architectures to allow parameter-efficient fine-tuning. In MMS, we follow the adapter design described in the HuggingFace blog by von Platen (2023): the base encoder is loaded, lightweight bottleneck adapters are initialized in each encoder block, and the underlying pre-trained parameters are frozen. Only the adapters and the language-specific CTC head are updated during fine-tuning, resulting in a small fraction (0.25%) of trainable parameters relative to the 1B-parameter backbone while preserving its cross-lingual representations.

For XLS-R, we mimic this approach by enabling attention adapters and follow the settings given by von Platen (2021).

Whisper Large-v3 is fine-tuned using full-model sequence-to-sequence training as explained by Gandhi (2022). In all Whisper experiments, we keep the pre-trained tokenizer and text normalization behavior intact, following the original Whisper training and evaluation protocol (Radford et al., 2023).

C.4 Training Details

Training is implemented using the Hugging Face `Trainer` (for MMS and XLS-R) and `Seq2SeqTrainer` (for Whisper) APIs with

[4] https://dl.fbaipublicfiles.com/mms/asr/mms1b_all_langs.html

language-specific adapters and tokenizers. For MMS and XLS-R, we train for 15 epochs by default, with a per-device batch size of 2 and gradient accumulation over 8 steps (effective batch size of 16 utterances). We use AdamW with a learning rate of 1×10^{-3}, 100 warmup steps, and gradient checkpointing enabled to reduce memory footprint.

For Whisper, we fine-tune `WhisperForConditionalGeneration` using 8 epochs, the same nominal batch size and gradient accumulation schedule, and a learning rate of 1×10^{-5}. We started our experiments with an initial set of hyperparameters inspired by Hugging Face blogs (von Platen, 2023, 2021; Gandhi, 2022) and then iteratively tuned them based on the initial results and training logs.

For all models, we save checkpoints every 100 steps, evaluate every 100 steps, retain at most 4 checkpoints per run, and load the best checkpoint according to validation Word Error Rate (`WER`).

C.5 Language Model Decoding

For the CTC-based MMS and XLS-R models, we optionally augment greedy decoding with external n-gram language models using `pyctcdecode`[5], but without an explicit n-gram language model. After training a model for a given language, we collect the cleaned training transcripts and derive a unigram word list, which provides a lexicon and approximate word frequencies. We then construct the CTC label set from the tokenizer vocabulary in index order, mapping the pad token to the CTC blank (empty string) and the word-delimiter token to a space, following standard practice for CTC decoding. In this configuration, pyctcdecode performs beam search using the CTC scores and the unigram vocabulary, but does not incorporate any KenLM n-gram probabilities (Heafield, 2011). That is, decoding corresponds to "beam search with unigrams (no ARPA)" in our logs. We apply this unigram-only beam search decoder only at evaluation time on the validation and test splits; all training and model selection are based on greedy CTC decoding. For Whisper, we do not use any external language model.

C.6 Evaluation Metrics

We report word error rate (`WER`) as the primary evaluation metric and character error rate (`CER`) as a secondary metric, computed using the `WER` and `cer` implementations from the Hugging Face `evaluate` library. For CTC-based MMS and XLS-R models, we convert logits to token sequences via argmax (greedy decoding) or beam search (with or without LM), then replace any -100 labels with the tokenizer's pad token id before decoding. For Whisper, the predictions are generated token sequences; we similarly replace masked label positions with the pad token before decoding.

C.7 Compute

All experiments were run on the University of Washington Hyak cluster using single-GPU jobs. All adapter runs used an NVIDIA `L40` or `L40S` GPU with `bf16` or `fp16` training enabled, requiring approximately 1–2 GPU hours per language and split. Whisper fine-tuning, which trains the full encoder–decoder in `float32`, required 2-3 GPU hours per language.

[5] https://github.com/kensho-technologies/pyctcdecode

D Results for n-gram LM Decoding Sweep

		MMS-1B-All						XLS-R-1B					
Language	**ISO 639**	**Greedy**	**Unigram**	**Set A**	**Δ_A**	**Set B**	**Δ_B**	**Greedy**	**Unigram**	**Set A**	**Δ_A**	**Set B**	**Δ_B**
Africa													
Bukusu	bxk	0.518	0.512	0.529	+0.010	0.570	+0.051	0.685	0.682	0.603	−0.082	0.644	−0.041
Chiga	cgg	0.479	0.484	0.479	−0.001	0.516	+0.037	0.760	0.751	0.636	−0.125	0.722	−0.038
Nubi	kcn	0.605	0.601	0.479	−0.126	0.523	−0.083	0.556	0.543	0.482	−0.073	0.494	−0.061
Konzo	koo	0.680	0.725	0.739	+0.059	0.920	+0.240	0.848	0.843	0.775	−0.073	0.916	+0.068
Lendu	led	0.360	0.349	0.277	−0.082	0.290	−0.070	0.322	0.318	0.278	−0.045	0.286	−0.037
Kenyi	lke	0.553	0.554	0.545	−0.008	0.577	+0.024	0.818	0.772	0.680	−0.138	0.690	−0.128
Thur	lth	1.001	0.999	0.970	−0.031	0.957	−0.044	0.368	0.367	0.311	−0.057	0.320	−0.049
Ruuli	ruc	0.590	0.583	0.572	−0.018	0.612	+0.022	0.699	0.693	0.637	−0.062	0.686	−0.013
Amba	rwm	0.607	0.597	0.538	−0.069	0.579	−0.028	0.568	0.563	0.495	−0.073	0.518	−0.050
Rutoro	ttj	0.243	0.240	0.229	−0.014	0.237	−0.005	0.349	0.350	0.290	−0.059	0.318	−0.031
Kuku	ukv	0.422	0.414	0.361	−0.060	0.377	−0.045	0.394	0.381	0.352	−0.041	0.351	−0.043
Americas													
Wixárika	hch	0.678	0.671	0.555	−0.123	0.608	−0.070	0.560	0.556	0.506	−0.054	0.521	−0.040
SW Tlaxiaco Mixtec	meh	0.423	0.418	0.359	−0.063	0.388	−0.035	0.446	0.437	0.371	−0.075	0.391	−0.056
Mich. Mazahua	mmc	0.763	0.766	0.705	−0.058	0.738	−0.026	0.714	0.714	0.658	−0.056	0.682	−0.033
Toba Qom	tob	0.595	0.592	0.603	+0.007	0.660	+0.065	0.657	0.653	0.582	−0.075	0.641	−0.016
Papantla Totonac	top	0.639	0.639	0.671	+0.032	0.741	+0.102	1.000	1.621	1.558	+0.558	2.421	+1.421
Asia & Pacific													
Betawi	bew	0.501	0.494	0.441	−0.060	0.473	−0.028	0.778	0.761	0.617	−0.161	0.643	−0.135
Western Penan	pne	0.348	0.341	0.264	−0.084	0.279	−0.069	0.381	0.365	0.284	−0.097	0.293	−0.088
Europe													
Gheg Albanian	aln	0.615	0.609	0.503	−0.112	0.545	−0.070	0.813	0.763	0.596	−0.217	0.622	−0.192
Cypriot Greek	el-CY	0.464	0.458	0.358	−0.106	0.375	−0.089	0.923	0.924	0.671	−0.252	0.808	−0.115
Scots	sco	0.323	0.321	0.271	−0.053	0.282	−0.042	0.302	0.297	0.244	−0.058	0.250	−0.052
Avg (all 21)		**0.543**	**0.541**	**0.498**	**-0.046**	**0.536**	**-0.008**	**0.616**	**0.636**	**0.554**	**-0.063**	**0.629**	**0.013**

Table 6: Full n-gram LM decoding ablation for all 21 languages (MMS-1B-All and XLS-R-1B, **All** training split). **Set A**: 4-gram LM, α=0.5, β=1.0, beam= 100. **Set B**: 3-gram LM, α=0.2, β=1.0, beam= 50. Δ_A and Δ_B are the absolute WER change vs. greedy decoding; green = improvement, red = degradation. Colour intensity is proportional to magnitude of $|\Delta|$ (capped at Δ=0.30). Row tints on section headers indicate geographic region.