Toward Equitable Speech Recognition: Linguistic Structure, Representation, and Fairness in Automatic Speech Recognition

Siyu Liang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Gina-Anne Levow, Chair

Richard Wright, Chair

Alicia Beckford Wassink

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Toward Equitable Speech Recognition:
Linguistic Structure, Representation, and Fairness in Automatic Speech Recognition

Siyu Liang

Co-Chairs of the Supervisory Committee:
Professor Gina-Anne Levow
Department of Linguistics

Professor Richard Wright
Department of Linguistics

Automatic Speech Recognition (ASR) systems have achieved near-human transcription accuracy on high-resource languages, yet their benefits remain unevenly distributed across the world's languages and speaker communities. This dissertation investigates three interlocking sources of ASR inequity: data scarcity for endangered and low-resource languages, internal representational biases in multilingual ASR architectures, and the socially-situated harms experienced by speakers of underrepresented language varieties.

At the data and adaptation level, the dissertation benchmarks fine-tuned multilingual ASR models, MMS-1B (Pratap et al., 2024) and XLS-R-300m (Babu et al., 2021), on five typologically diverse endangered languages drawn from linguistic fieldwork archives with systematic control of training data duration. Results show that parameter-efficient adapter fine-tuning could outperform full fine-tuning under one hour of data, that both models reach comparable performance at approximately one hour, and that both struggle persistently with complex phonological inventories—such as tone, nasality, and consonant length—in ways that aggregate error metrics obscure.

At the model-internal level, a set of experiments probe the decoding mechanisms of Whisper, a large multilingual ASR model, across languages in Latin and diverse scripts. Sub-token probing reveals that higher-resource languages have higher decoder confidence,

lower predictive entropy, better-ranked correct tokens, and more diverse alternative candidates. Sub-token usage patterns cluster typologically, not merely by resource tier, showing that linguistic structure shapes internal representations in addition to training scale. A scaled follow-up study across languages in diverse scripts finds that sub-token vocabulary activation is largely independent of pre-training hours, converging instead after approximately two hours of audio, and that rank-frequency distributions can be explained by shared orthography in addition to resource differences.

At the social and user level, a mixed-methods survey of speakers across four U.S. dialect communities documents the lived experiences of ASR bias: invisible labor (code-switching, hyper-articulation, emotional management), internalized self-blame despite critical awareness of systemic exclusion, and varying rates of technology abandonment. These findings demonstrate that accuracy-centric fairness metrics miss critical dimensions of harm borne by affected users.

Together, the studies in this dissertation argue that equitable speech recognition requires linguistically-grounded design and evaluation at every level of the pipeline—from input data and phonological structure, through model-internal representations, to the social consequences of system implementation.

ACKNOWLEDGMENTS

First and foremost, I would like to thank the members of my committee: Gina-Anne Levow, Richard Wright, and Alicia Beckford Wassink. Gina gave me the freedom to explore what sparked my interest and supported me whenever I needed guidance and advice. Richard has always been a pleasant and resourceful mentor. Alicia introduced me to the field of sociolinguistics and opened my perspectives in ways I didn't think I could. I truly could not have asked for a better committee.

I'm also thankful to Nicolas Ballier for our meaningful collaboration and for being a wonderful host in Paris. At UW, I have been fortunate to learn from great teachers and researchers in the department: Shane Steinert-Threkeld, Fei Xia, Sharon Hargus, Myriam Lapierre, Basia Citko, Toshiyuki Ogihara, and Emily Bender. I'd like to thank Laura McGarrity for teaching me the art of teaching throughout the years. I'm also grateful to the faculty across campus who indulged my intellectual curiosity: Zev Handel, whose class on historical Chinese phonology I will remember dearly; Talant Mawkanuli, for his vast knowledge of Turkic linguistics; S'yama Than Than Win, for introducing me to Burmese; and Yunxiao Xiao, whose class on Chinese paleography I enjoyed immensely. I would also like to thank the wonderful staff in the department for all their help: Karoliina Kuisma, Chris Shaeffer, Joyce Parvi, Anna Schnell, and Brandon Graves.

I have been fortunate to be surrounded by a wonderful community at UW and in Seattle. Connor Boyle was the first friend I made during student orientation and still remains and will always be my best friend. I thank him for all the conversations and adventures in and around Seattle and frequent pickups at SeaTac, which made the city feel a bit more like home and harder to say goodbye to. Emily Ahn was a kind and supportive graduate student mentor, and a great friend. I thank Ray Gagne for surviving the cohort with me and making the graduate student office a place I looked forward to being in a bit more. I will always

miss the meals, drinks, and jokes I shared with Haotian Zhu and Chenxi Li. I'm grateful to Yadi Peng, with whom I share some of my happiest memories over meals, on walks, and in front of the silver screen. Trent Ukasick has always been a pleasant friend to talk to. I'm grateful for all the wonderful time I spent with Katie Lindekugel, Vipasha Bansal, Cassie Maz, Zhaxi Zerong, Christopher Haberland, and Gita Dhungana, and for all the help from Sara Ng and Agatha Downey. Thank you to Jingnong Qu for the wonderful time and conversations we had together. Thank you, Naomi Shapiro, for the time in the pro-seminar and beyond and for your training in the importance of work-life balance. I thank Yuan Chai for always being supportive and helpful. I'd also like to thank Grainger Lanneau, Gloria Lee, and Linjia Weng in the Department of Asian Languages and Literature for their friendship.

I also thank my advisors and mentors from my time at Georgetown: Amir Zeldes, Hannah Sande, and Elizabeth Zsiga, who helped lay the foundations for me to pursue computational linguistics, fieldwork, and phonetics. I'm grateful to my Georgetown friends for their continued support throughout this journey: Janet Yang, Logan Peng, Yilun Zhu, Zhuosi Luo, and Jordan MacKenzie. Thank you to Candice Luo for our time together trying our hand in the world of start-ups. And to all my friends whose names I cannot enumerate here: you have made this journey as painless as it could be.

I am also grateful for the financial support from the UW Global Innovation Fund and the Imre Boba Summer Research Fellowship.

Importantly, I'm grateful for the unwavering support from my family, some of whom could not get to read this but would have been overjoyed to witness this accomplishment.

Last but not least, I should probably also thank my 21-year-old self, who was curious and reckless enough to choose this path and make it happen.

DEDICATION

To 英姿, 斌姿, 現姿

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAE: African American English

ASR: Automatic Speech Recognition

AST: Acoustic Saturation Time

BPE: Byte Pair Encoding

CER: Character Error Rate

ELAR: Endangered Languages Archive

MMS: Massively Multilingual Speech

NLP: Natural Language Processing

PCA: Principal Component Analysis

SSL: Self-Supervised Learning

T-SNE: t-distributed Stochastic Neighbor Embedding

WER: Word Error Rate

XLS-R: Cross-Lingual Speech Representations

書不盡言，言不盡意。

Writing does not fully capture speech;

speech does not fully capture meaning.

——《周易·繫辭上》

Xici I, Zhouyi

Chapter 1

INTRODUCTION

1.1 The Uneven Promise of Automatic Speech Recognition

Automatic Speech Recognition (ASR) has become increasingly present in daily lives. It is embedded in smartphones and speakers, in customer service systems and medical transcription tools, in accessibility technologies and court reporting software. For hundreds of millions of users each day, the expectation is simple: speak, and the machine will understand. To much of the global population, that expectation is increasingly met. Large multilingual models—such as Whisper (Radford et al., 2023), wav2vec 2.0 (Baevski et al., 2020), MMS (Pratap et al., 2024), XLS-R (Babu et al., 2021)—now process dozens or hundreds of language varieties in a single architecture, achieving word error rates that rival skilled human transcribers on benchmark datasets. These systems represent a genuine scientific achievement: the result of years of work in self-supervised learning, multilingual transfer, and at-scale data collection.

Yet the communities that benefit most from these advances are usually communities that were already well-served by earlier and smaller ASR systems. High-resource languages with large volumes of transcribed audio, standard orthographies, and culturally dominant status continue to receive the most accurate recognition. Endangered and low-resource languages—whose speakers are most in need of accessible transcription technology to support language documentation, revitalization, and archive access—remain largely outside the practical reach of state-of-the-art ASR. Speakers of underrepresented dialect varieties, such as African American English, face word error rates that are systematically and substantially higher than those experienced by speakers of prestige varieties (Koenecke et al., 2020; Wassink et al., 2022). And the costs of those failures are borne unequally and invisibly:

in the code-switching and hyper-articulation that speakers perform to make failing systems work, in the self-blame that follows repeated technological rejection, in the decisions to abandon technologies that have become increasingly unavoidable (Mengesha et al., 2021).

The central claim of this dissertation is that these patterns of inequity are not incidental to the development of modern ASR—they are structural. They emerge from decisions made at every level of the pipeline: what data is collected and from whom, how models represent and process linguistic structure internally, and whose standard of adequate performance defines evaluation success. A complete account of ASR inequity therefore requires investigation at multiple levels simultaneously, and it requires methods that extend beyond the aggregate error rate metrics that currently dominate the field.

1.2 The Analytical Framework

This dissertation organizes its investigation around three levels of analysis that are distinct but inter-connected.

At the **data and adaptation level**, the question is whether—and under what conditions—existing multilingual models can be brought to bear on languages for which they were not designed. Endangered languages and unrecorded-until-recently fieldwork languages have none of the infrastructure that standard ASR development assumes: no large transcribed corpora, no standardized orthographies, no benchmark datasets against which to measure progress. The transcription of a single hour of fieldwork audio can require fifty or more hours of expert labor (Shi et al., 2021), creating a bottleneck that traps linguistic knowledge in archives and limits the use of analytical tools that depend on large aligned text-speech collections. If multilingual pre-training has genuinely learned transferable representations of cross-linguistic phonological structure, fine-tuning on a handful of recorded utterances should yield usable transcription; if the promise of transfer is illusory for languages outside a model's training distribution, the picture is considerably bleaker.

At the **model-internal level**, the question is whether the *representations* that large multilingual ASR models learn are linguistically equitable—whether they encode languages with different levels of pre-training exposure in qualitatively similar ways, or whether lower-resourced languages are processed through a different and less capable internal mechanism.

Aggregate error metrics can mask this distinction: a model may produce highly accurate output for high-resource languages because it has developed rich, language-specific representations, while producing passable but internally confused output for low-resource languages because it has relied on more generic approximation. To understand the nature of this internal inequity requires moving beyond commonly used aggregate metrics to the model's decoding behavior, probing the hypotheses it entertains at each generation step and the confidence it associates with them.

At the **social and user level**, the question is whether accuracy-centric evaluation captures the full scope of harm that ASR inequity causes. Error rates measure what the system gets wrong; they do not measure what the experience of repeated system failure does to the people who live with it. The labor of making a failing system work—adjusting pronunciation, switching register, repeating utterances, abandoning the technology entirely—is real, ongoing, and unequally distributed. The psychological toll of feeling inadequate in one's native language variety, of internalizing as personal failure what is structurally a design exclusion, is real and undocumented by any benchmark. Fairness assessments that ignore these experiential dimensions are not merely incomplete; they systematically understate harm in the populations where it is concentrated.

These three levels of analysis are analytically distinct but practically interdependent. The data and adaptation problems at the pipeline level produce the internal representational disparities that probing reveals; those internal disparities produce the output failures that users experience; and the accumulated experience of failure shapes communities' relationships with technology in ways that further entrench exclusion. An investigation that addresses only one level leaves the others unexplained. This dissertation examines all three in sequence.

1.3 The Role of Linguistic Analysis

A methodological commitment runs through the studies in this dissertation: linguistic analysis is not supplementary to technical ASR research but constitutive of it. Understanding why a model fails for a fieldwork language requires knowing what phonological features and processes are present, not merely measuring the error rates. Understanding why low-resource

languages are represented by multilingual models in similar ways requires knowing what typological and orthographic properties can be used to characterize them. Understanding why speakers of underrepresented and stigmatized dialects talk differently when interacting with ASR systems requires understanding sociolinguistic research on style-shifting, covert prestige, and the implications of language standardization.

The gap between the landscape of ASR research and linguistics is not merely disciplinary. It has practical consequences: models evaluated only on metrics such as WER miss phonological error patterns that matter for language documentation; models benchmarked only on accuracy miss representational inequities that manifest in under-the-hood decoding uncertainty; and fairness frameworks built only on performance metrics miss the social reality of harm in communities already subject to linguistic discrimination. This dissertation is an argument, instantiated in four empirical studies, that linguistic grounding is not just complementary to technical research, but indispensable for making technical research meaningful to the population it should benefit.

1.4 Research Questions

This dissertation is organized around three research questions, each addressed by a chapter.

1. **RQ1 (Adaptation):** How do parameter-efficient and full fine-tuning strategies compare when adapting large multilingual ASR models to extremely low-resource fieldwork languages, and what role does phonological inventory complexity play in residual error patterns that persist even after fine-tuning?

2. **RQ2 (Sub-token probing):** Do sub-token decoding metrics—how highly the correct token is ranked, how confident the decoder is, how uncertain its predictions are, how diverse its alternate candidates are, and how token-usage patterns cluster across languages—reveal systematic differences in how a large multilingual ASR model processes languages across training-data tiers? If so, do those differences align with typological structure, resource level, or both, and are they reducible to training data volume?

3. **RQ3 (User experience):** How do speakers of non-mainstream U.S. dialect varieties describe their experiences with ASR systems, and what forms of invisible labor, emotional harm, and internalized self-blame does accuracy-centric fairness evaluation fail to capture?

These three questions build on each other. RQ1 establishes that phonological complexity produces persistent errors even after adaptation—errors that aggregate metrics underreport. RQ2 shows that those aggregate gaps have internal correlates: resource-disadvantaged languages are processed with lower decoder confidence, higher uncertainty, and less typologically organized internal representations—and that this representational disadvantage is not simply a function of data volume, since sub-token utilization saturates quickly and is governed primarily by orthographic structure. Simply scaling data cannot straightforwardly fix these inequities. And RQ3 reveals that the ultimate recipients of all these technical failures—the people whose speech is misrecognized—bear real and measurable costs that none of the current metrics capture. Taken together, the three research questions argue that ASR inequity is deeper than data scarcity, that it is structural and multi-level, and that addressing it requires methods and evidence from linguistics, interpretability research, and social science simultaneously.

1.5 Dissertation Overview

Chapter 2 addresses RQ1 through a controlled fine-tuning study spanning five endangered languages drawn from the Endangered Languages Archive (ELAR)[1] fieldwork archives—Cicipu, Mocho', Toratán, Ulwa, and Upper Napo Kichwa—with training data scaled from ten minutes to two hours. The study benchmarks MMS-1B against XLS-R-300m under adapter-only and full fine-tuning conditions and conducts a phonologically informed error analysis that categorizes failures by tonal, nasal, and segmental length features.

Chapter 3 addresses RQ2 by probing Whisper-large-v2's sub-token decoding behavior across 20 Latin-script languages and then 29 additional languages spanning diverse scripts

[1] `https://www.elararchive.org`

and resource levels. Five sub-token probing metrics—correct-token rank, decoder confidence, predictive entropy, alternate-candidate diversity, and cross-language usage clustering—reveal that lower-resource languages are decoded through a generic regime with lower confidence, higher entropy, and typologically undifferentiated token usage, a pattern not visible in WER. Scaling the analysis to 49 languages shows that sub-token vocabulary activation converges after roughly two hours of audio regardless of resource tier, and that rank-frequency distributions track orthographic structure rather than pre-training scale, demonstrating that the representational disadvantage is not reducible to data volume.

Chapter 4 addresses RQ3 through a mixed-methods survey of 215 speakers across four U.S. dialect communities (Atlanta, Gulf Coast, Miami Beach, and Tucson). Quantitative analysis of structured questionnaire items documents the rates of reported exclusion, coping behavior, and technology abandonment across sites. Qualitative thematic analysis of open-ended narratives reveals the internalized-blame paradox, the experience of linguistic invisibility, and the sophisticated critical awareness that coexists with ongoing self-blame.

Chapter 5 synthesizes the findings across levels of analysis, articulates the dissertation's contributions to linguistics, NLP and speech technology research, and AI fairness scholarship, examines the limitations of each study, and identifies the most important directions for future work.

Chapter 2

ASR ADAPTATION FOR ENDANGERED FIELDWORK LANGUAGES

Overview

This chapter addresses **RQ1**: how do parameter-efficient and full fine-tuning strategies compare when adapting large multilingual ASR models to extremely low-resource fieldwork languages, and what role does phonological inventory complexity play in residual error patterns? It is based on Liang and Levow (2025), presented at the Fourth Workshop on NLP Applications to Field Linguistics (Field Matters, ACL 2025). The study benchmarks MMS-1B and XLS-R-300m on five endangered languages from the ELAR archive—Cicipu, Mocho', Toratán, Ulwa, and Upper Napo Kichwa—finding that adapter fine-tuning outperforms full fine-tuning under one hour of data and that phonological inventory complexity predicts residual error patterns that aggregate metrics obscure.

2.1 ASR and the Fieldwork Gap

Automatic Speech Recognition (ASR) has achieved significant breakthroughs in recent years, with deep learning-based models reported to reach near-human word error rates for high-resource languages (Radford et al., 2023; Baevski et al., 2020). However, these advancements have largely been driven by massive transcribed datasets (e.g. Chang et al. (2022); Panayotov et al. (2015); Godfrey et al. (1992)), leaving a substantial performance gap for low-resource languages, particularly those encountered in linguistic fieldwork (Guillaume et al., 2022a). Fieldwork speech data presents distinct challenges, including spontaneous speech, varied recording setups, and typologically diverse linguistic features, all of which could degrade the performance of ASR models trained on standardized speech corpora.

Linguistic fieldwork plays a critical role in preserving endangered languages and documenting linguistic diversity. These recordings capture not only the linguistic structures of

a language, but also oral traditions, discourse patterns, and sociolinguistic variations (Himmelmann, 1998; Austin and Sallabank, 2011). However, while some well-researched low-resource languages have substantial datasets (Guillaume et al., 2022b; Przezdziak, 2024), there is usually limited data to bootstrap an ASR model for most field linguists. Evaluations of the ASR approaches usually tend to focus on one language (Jones et al., 2024; Rijal et al., 2024; Guillaume et al., 2022b; Mainzinger and Levow, 2024) or are inconsistent regarding data size, genre, etc. in the sample (Jimerson et al., 2023). Evaluations of models for low-resource languages also tend to favor clean, good quality, read speech (Rijal et al., 2024; Mainzinger and Levow, 2024; Jimerson et al., 2023), compared with the noisier and more spontaneous speech of fieldwork recordings.

2.1.1 The transcription bottleneck

Linguistic fieldwork plays a crucial role in documenting endangered and under-researched languages, yet the process of manually transcribing recordings remains a significant barrier: transcribing a single hour of audio in a newly documented language can require up to 50 hours of work (Shi et al., 2021). Moreover, many of the fieldwork languages also lack standardized orthographies, requiring a handful of trained linguists to make discerning decisions during transcription. As a result, the volume of untranscribed linguistic data continues to grow, creating a severe bottleneck in language documentation, analysis, and distribution (Anastasopoulos and Chiang, 2018; Bird, 2020; Thieberger, 2012).

The dependence on large transcribed datasets for training ASR models exacerbates this issue, as most endangered and low-resource languages lack sufficient annotated speech data to support model development (Levow et al., 2021). Without adequate transcriptions, traditional supervised ASR methods remain ineffective, requiring alternative approaches that can leverage limited data more efficiently (Dunbar et al., 2019; Baevski et al., 2020, 2021).

2.1.2 ASR for Low-resource Data

The application of ASR in linguistic fieldwork closely parallels the development of low-resource ASR research. Since the 2010s, much of this work has focused on languages from the IARPA Babel project, which served as a cornerstone for ASR development in low-resource settings (Miao et al., 2013; Cui et al., 2014; Grézl et al., 2014). Research leveraging Babel datasets introduced key techniques such as transfer learning, multilingual adaptation, and data augmentation, which have since become fundamental to ASR advancements in under-documented languages (Zhang et al., 2014; Khare et al., 2021; Vanderreydt et al., 2022; Guillaume et al., 2022a).

The widespread adoption of the Kaldi toolkit (Povey et al., 2011) further propelled ASR research in these domains, enabling the development of reproducible pipelines and fostering the open distribution of Kaldi-compatible datasets (Yadava and Jayanna, 2017; Milde and Köhn, 2018; Adams et al., 2021; Zhang et al., 2022). Concurrently, researchers have explored approaches such as transfer learning and fine-tuning from multilingual pre-trained models (Guillaume et al., 2022a; **?**) or adapting English-centric models to new linguistic domains (Kim et al., 2021; Thai et al., 2020). Additionally, self-supervised and semi-supervised learning approaches have gained traction as viable solutions for overcoming transcription scarcity, further bridging the gap between ASR and field linguistics (Babu et al., 2021; Baevski et al., 2021).

2.1.3 Fine-tuning Pre-trained ASR Models

Fine-tuning pre-trained ASR models has emerged as a key approach for improving recognition accuracy in low-resource settings, particularly for linguistic fieldwork recordings (Guillaume et al., 2022a; Pillai et al., 2024; Nowakowski et al., 2023). Self-supervised learning models, such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), MMS (Massive Multilingual Speech) Model (Pratap et al., 2024), and XLS-R (Babu et al., 2021), have demonstrated the ability to learn generalized speech representations from large-scale multilingual datasets, significantly reducing the need for extensive transcriptions in under-documented languages. Studies on specific low-resource languages, such as Bribri (Coto-

Solano, 2021), Japhug (Guillaume et al., 2022b), Mvskoke (Mainzinger and Levow, 2024), and the Čakavian dialect of Croatian (Jones et al., 2024), underscore the benefits of adapting large multilingual models and report considerable reductions in error rates even with very limited data.

Nevertheless, recent work suggests that no single architecture or end-to-end approach consistently outperforms others under extremely low-resource conditions (Jimerson et al., 2023). Some studies advocate experimenting with multiple toolkits and hyperparameter configurations to identify solutions best suited to the language at hand. Indeed, while fully fine-tuning massive models can be effective, it often requires large amounts of computational resources, can risk overfitting with very small datasets, and demands updating millions of parameters.

Instead of modifying all model parameters, adapters introduce small trainable layers while keeping the base pre-trained model frozen, thereby reducing memory requirements and improving efficiency (Houlsby et al., 2019). This makes them particularly useful for linguistic fieldwork applications, where data is scarce and computational resources are limited. The MMS model developed by Meta (Pratap et al., 2024) integrates adapter layers specifically designed for ASR, enabling efficient adaptation to new languages with minimal training data. Studies in low-resource settings (Bai et al., 2024; Mainzinger and Levow, 2024) have shown that adapter-based fine-tuning can achieve performance comparable to full fine-tuning while requiring significantly fewer trainable parameters. By avoiding overfitting on small datasets and focusing on the most relevant parameters for language adaptation, adapter-based methods offer an attractive balance between accuracy and efficiency, an approach increasingly vital to sustaining language documentation efforts in the face of extremely scarse resources.

In contrast to earlier studies that often focus on a single language or on clean, scripted corpora, our work systematically evaluates both MMS and XLS-R in truly low-resource fieldwork conditions spanning multiple typologically diverse languages. By examining noise-heavy, spontaneous recordings rather than controlled speech, we test model adaptability in settings that more accurately reflect real-world linguistic documentation. Further, our fine-grained error analysis explores how each model handles the nuanced phonological features

that typify endangered and under-documented languages—details that have often been overlooked in prior research. Together, these innovations provide a clearer roadmap for linguists seeking practical ASR solutions under extreme data scarcity and diverse orthographic conventions.

2.2 Endangered Fieldwork Languages Data

We test the performance of fine-tuned ASR models on five typologically varied low-resource languages: Cicipu (ISO639-3: awc, McGill (2012)), Mocho' (ISO639-3: mhc, Pérez González (2018)), Toratán (ISO639-3: rth, (Jukes, 2010)), Ulwa (ISO639-3: yla, (Barlow, 2018a)), and Upper Napo Kichwa (ISO639-3: quw, (Grzech, 2020)). These languages span multiple language families and exhibit distinct phonetic, phonological, and morphological features. The data is drawn from the Endangered Languages Archive (ELAR)[1], where gold-standard transcriptions can be derived from the recordings and the corresponding time aligned transcriptions in the ELAN (Brugman and Russel, 2004) format. The dataset encompasses a variety of genres, such as greetings, narratives, ritual discourse, interviews, elicitation sessions, folktales, and cultural practices, the details of which are given in Table A.1 of Appendix A.2. Table 2.1 provides an overview of key linguistic features, including vowel and consonant inventories as well as tonal systems.

2.2.1 Recording Conditions and Data Availability

Given that the recordings were made in naturalistic fieldwork environments, they exhibit acoustic idiosyncrasies that could pose significant challenges for ASR. There's background noise from outdoor settings, such as wind, animals, and community sounds, in many of the recordings. We also observe code-switching, usually in the regional dominant languages. For example, Toratán speakers are also speakers of Manado Malay (with various degrees of fluency), and loans from Malay are generally not adapted to Toratán phonology (Himmelmann and Wolff, 1999). We also observe significant Spanish code-mixing in Mocho', as well as some English content in Cicipu.

[1] `https://www.elararchive.org/`

Language	**Family**	**Region**	**#V**	**#C**	**#T**	**Features**
Cicipu	Niger-Congo	Nigeria	28	27	4	$\tilde{\text{V}}$, V, C
Mocho'	Mayan	Mexico	10	27	2	V
Toratán	Austronesian	Indonesia	5	21	0	
Ulwa	Keram	Papua New Guinea	8	13	0	[-voice, +son]
Upper Napo Kichwa	Quechuan	Ecuador	8	20	0	V

Table 2.1: Linguistic data used for the study, showing language family, region spoken, phoneme inventory size, tones, and phonological features such as nasality, vowel length, consonant gemination, etc.

The dataset sizes vary, with archived speech ranging from approximately 2 to 22 hours per language. However, not all archived data have been transcribed. This variation reflects real-world constraints in linguistic fieldwork, where some languages have more extensive documentation than others.

2.2.2 From Fieldwork Recordings to Training Sets

All recordings were resampled to 16 kHz (the training sampling rate for both MMS and XLS-R), converted to mono-channel WAV format, and aligned with their corresponding transcriptions. During transcription pre-processing, we referenced the phonological description of each language to ensure that punctuation marks or special characters used to denote phonological features were retained (see Table A.1 of Appendix A.2). Audio segments explicitly transcribed as non-linguistic sounds, such as laughter, were excluded from the dataset. We also removed utterances that contain only filler words, such as 'mhm', 'aaa', etc. A breakdown of the audio lengths before and after pre-processing is given in Table A.2 of Appendix A.2.

We created four total train+dev duration configurations for each language—10, 30, 60, and 120 minutes—before splitting the data into training (90%) and development (10%) sets. In addition, we set aside a fixed 10-minute test set for final evaluation. To maintain

Language	#Spk	Avg. Leng.	#Char
Cicipu	33	2.1	93
Mocho'	6	2.0	29
Toratán	13	2.35	27
Ulwa	6	3.65	25
U.N. Kichwa	16	3.79	33

Table 2.2: Dataset statistics for different languages, including the number of speakers, average utterance length in seconds, and character inventory.

consistency and facilitate interpretation, larger dataset splits were structured as supersets of smaller ones. We did not designate a held-out speaker, as field linguists typically work with a limited number of consultants and would prioritize consistent model performance across familiar speakers (Liu et al., 2023). Details of the cleaned dataset are shown in Table 2.2. The last column of the table lists the number of unique characters used in the language. Due to different transcription conventions, features such as nasalization, vowel length and voicing could be indicated with diacritics or extra letters. Therefore, although transcriptions are meant to be phonemic, the number of unique characters might not match the number of contrastive vowels and consonants. In the case of Cicipu, its unusually large inventory of 93 characters is due to the number of all possible combinations of nasality, tone, and vowel quality marking.

2.3 Model Adaptation Strategies

This study investigates the effectiveness of fine-tuning multilingual ASR models to address the unique challenges posed by low-resource linguistic fieldwork recordings. By evaluating the performance of two state-of-the-art models, we aim to determine how fine-tuning can enhance recognition accuracy on typologically diverse, low-resource languages. In addition, we discuss the impact of key factors, including training data size, model choice, and pre-trained model features, to provide practical insights for ASR adaptation in fieldwork contexts.

2.3.1 Two Multilingual Architectures

Our goal is to fine-tune state-of-the-art multilingual ASR models that have been pre-trained on large-scale speech corpora. Specifically, we evaluate models from Meta's Massively Multilingual Speech (MMS) project (Pratap et al., 2024) alongside XLS-R (Babu et al., 2021), a widely used multilingual ASR model.

MMS is based on the wav2vec 2.0 framework (Baevski et al., 2020), which employs self-supervised learning to extract generalized speech representations from vast amounts of unlabeled audio. The model has been trained on 1,406 languages, making it one of the most comprehensive ASR models for multilingual speech recognition. Specifically, we choose MMS-1B-l1107, a 1-billion parameter model fine-tuned specifically for ASR with an additional 2-million parameter adapter, which supports ASR in 1107 languages out of the box (Houlsby et al., 2019). The adapter facilitates efficient language-specific fine-tuning while preserving the generalized multilingual knowledge encoded in the base model.

XLS-R (Babu et al., 2021) is a multilingual ASR model pre-trained on 128 languages using the wav2vec 2.0 framework. It has been extensively used for low-resource ASR and cross-lingual transfer learning, making it a strong baseline for evaluating ASR performance in linguistic fieldwork settings. Unlike MMS, which is trained on over 1,000 languages, XLS-R has been optimized for a balanced selection of 128 languages, with a strong focus on phonetic diversity. This makes it particularly useful for comparison against MMS models to assess the effectiveness of scaling multilingual pre-training to extremely low-resource languages. The XLS-R-300m with 300 million parameters is chosen for the study.

None of the languages used in the study, with the exception of Upper Napo Kichwa, is included in the pre-training data of the two models. A discussion of the possible effects is included in Section 2.4.2.

2.3.2 Training and Evaluation Setup

Following the fine-tuning procedure outlined by von Platen[2] and using the Hugging Face Transformers library (Wolf et al., 2020), we implement model-specific strategies. For MMS-

[2] `https://huggingface.co/blog/mms_adapters`

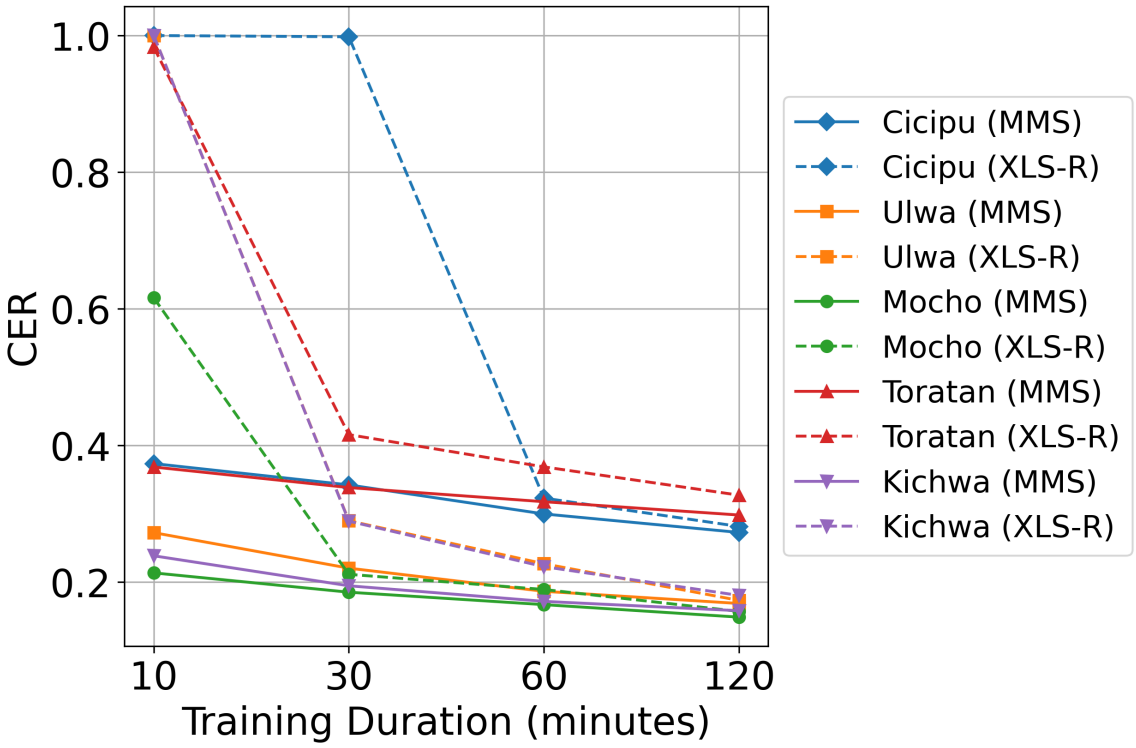

Figure 2.1: CER comparison for MMS-1b1107 and XLS-R-300m models across five languages. The MMS model performs markedly better under extremely low-resource settings (less than 1 hour), but XLS-R performs similarly well with 2 hours of data. Points are connected to aid trend reading and do not imply performance at intermediate durations.

1B-l1107, only the adapter layers are fine-tuned, with the base model frozen. For XLS-R, the entire model is fine-tuned. To assess the effect of data size, models are trained on subsets of 10, 30, 60, and 120 minutes of transcribed fieldwork data per language. Early stopping, based on the development set Character Error Rate (CER), mitigates overfitting. Hyperparameter details, including batch size, learning rate, optimization strategy, and training details are provided in Appendix A.1.

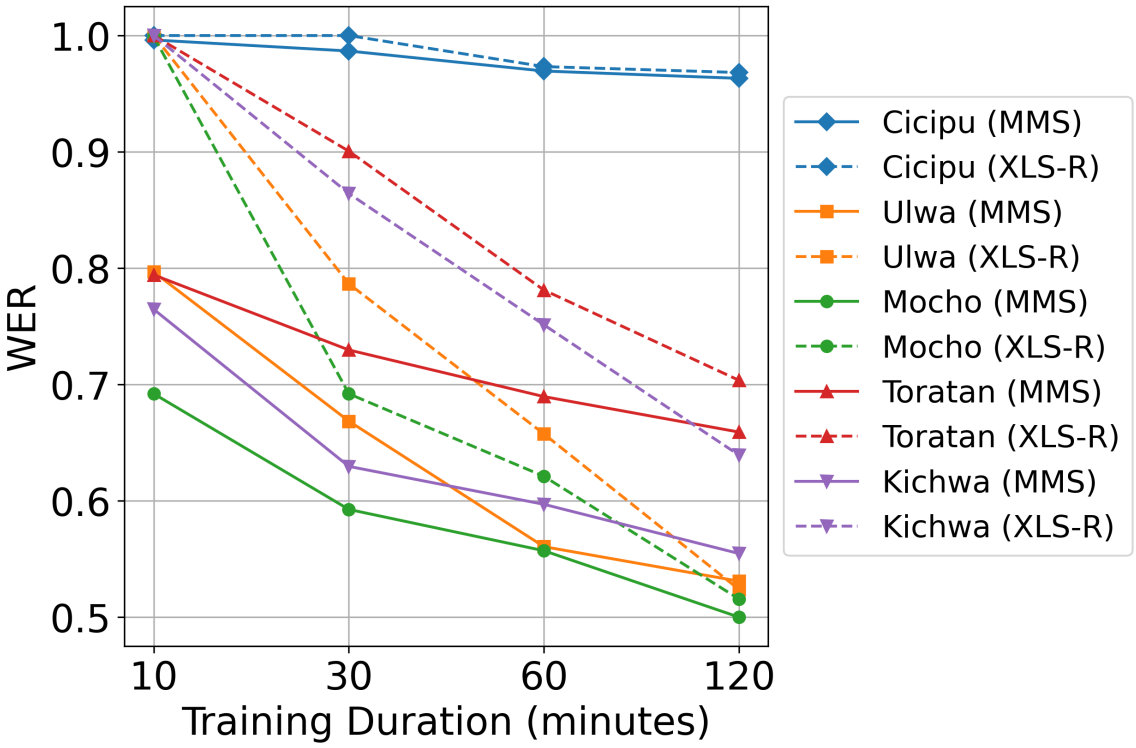

Figure 2.2: WER comparison for MMS-1b1107 and XLS-R-300m models across five languages.

2.4 Model Performance Across Languages and Data Conditions

Overall, the performance of the MMS and XLS-R models in automatic speech recognition (ASR) tasks on low-resource language data is comparable, though nuanced differences emerge depending on the availability of training data. In general, the MMS model outperforms XLS-R under extremely low-resource conditions (i.e., less than one hour of transcribed data). However, XLS-R demonstrates a marked improvement as the size of the training data scales beyond this threshold, ultimately becoming on par with MMS. Figure 2.1 and Figure 2.2 illustrate these trends in Character Error Rate (CER) and Word Error Rate (WER) metrics across the five languages in this study.

In our analysis, CER serves as the primary evaluation metric. Unlike WER, which

operates at the word level and often presupposes well-defined word boundaries and stable orthographic forms, CER better reflects the needs of field linguists, who often lack enough data to be able to use language models. In fieldwork contexts, the primary goal of transcription is to capture phonetic accuracy, particularly in languages without standardized orthographies. During model training, the best model metric is set to CER to ensure that model performance aligns with the core task of producing reliable phoneme-level transcriptions from spontaneous and noisy speech data.

2.4.1 How Much Data Is Enough?

The relationship between training data size and model performance is critical in understanding model suitability for field linguistics applications. For both MMS and XLS-R, performance improvements start to plateau when training data exceeds approximately one hour. This finding suggests steady although diminishing returns beyond this point, aligning with previous observations in low-resource ASR research (Guillaume et al., 2022a).

In scenarios where less than one hour of data is available, MMS consistently achieves lower error rates, likely due to its extensive multilingual pre-training on over 1,000 languages. For field linguists dealing with extremely limited resources, we thus recommend fine-tuning MMS to achieve acceptable ASR performance. However, when approximately one hour of data or more is obtainable, XLS-R becomes a more effective option due to its improving performance with increasing data volumes. This suggests that one hour of transcribed data serves as a practical threshold for developing a robust fine-tuned ASR system in fieldwork contexts.

It is worth noting that Cicipu exhibits particularly high error rates under extreme data scarcity, including a character error rate approaching 1.0 at 10 minutes of data for both models and at 30 minutes for XLS-R. Cicipu's unusually large orthographic inventory (93 unique characters reflecting combinations of nasality, tone, and vowel quality) requires more training examples to accurately learn the mapping from acoustics to graphemes. Consequently, with only a few minutes of labeled data, neither model can fully learn Cicipu's complex phonological and orthographical features.

2.4.2 Adapter Fine-tuning vs. Full Fine-tuning

The MMS model excels in settings with minimal data due to its multilingual pretraining on a vast corpus that includes low-resource languages. Moreover, unlike XLS-R—whose core version is primarily self-supervised and not initially fine-tuned for ASR—MMS has already undergone a large-scale ASR fine-tuning step. This means that MMS starts off with more task-specific parameters, making it more effective than XLS-R in extremely low-data regimes. However, MMS's reliance on primarily read speech data (e.g. Bible translations) may limit its adaptability to spontaneous speech environments, which are common in linguistic fieldwork recordings.

In contrast, XLS-R benefits from a more diverse training corpus that encompasses conversational and spontaneous speech, allowing it to generalize better once sufficient data becomes available. Indeed, Mainzinger and Levow (2024) reported superior performance of MMS over XLS-R when fine-tuning Mvskoke—likely due to both the advantage of MMS's ASR fine-tuning and the fact that much of the Mvskoke training material was similarly read or scripted speech.

Since several related dialects of Kichwa (Eberhard et al., 2024, 2025) were included in the MMS pre-training dataset, we investigated whether the performance gap between MMS and XLS-R would be *larger* for Kichwa than for the other languages in our study. Specifically, we fit a linear mixed-effects model (with random intercepts for each language) to our character error rate (CER) data, using *model* (MMS vs. XLS-R), *time*, and an indicator *similar* (1 = Kichwa, 0 = other languages) as fixed effects. If Kichwa had benefitted disproportionately from MMS's pre-training, we would have observed a significant positive interaction in the model. However, the interaction term (*model* $\times$ *similar*) was small and not statistically significant ($\beta = 0.021$, $p = 0.896$), indicating that while MMS outperforms XLS-R overall, the additional advantage for Upper Napo Kichwa is not discernibly greater than for the other languages.

Further research is needed to evaluate whether the performance trend continues with larger datasets, particularly for languages with similar phonological and morphological complexity as those in this study. Additionally, the effectiveness of adapter-based fine-tuning

for MMS suggests that optimizing model architecture for scalable adaptation could yield further improvements.

2.4.3 Where Models Still Fail

We performed a phonologically informed error analysis on two of the languages in our dataset, Cicipu and Mocho', both of which exhibit segmental contrasts that could be challenging for ASR models. Cicipu's orthography explicitly marks tone and nasality with diacritics and differentiates both vowel and consonant length with doubled letters (e.g., 'aa' for long vowels, 'tt' for geminate consonants), making it well suited for evaluating the system's performance on these phonological categories. Mocho' similarly features a vowel length distinction encoded with doubled vowel letters. Other languages in our dataset, such as Ulwa and Upper Napo Kichwa, contain too few instances of long vowels or voiceless sonorants for a robust category-based analysis.

Category-Level Error Rates

To quantify performance on these features, we leverage character-level alignments (details in Appendix A.3) and calculate phonological segment error rates (Table 2.3). For each category $C \in \{\text{Tone}, \text{Nasality}, \text{V\_length}, \text{C\_length}\}$, we sum all substitutions, deletions, and insertions that affect that category and normalize by the total number of reference tokens L_C exhibiting C.

As shown in Table 2.3, both MMS and XLS-R struggle with Cicipu's tone, nasality, and consonant length, each exhibiting error rates in the range of 30–38%. By contrast, vowel-length confusion is comparatively low (7–9%). For Mocho', the long–short vowel distinction remains problematic, with error rates around 35–38%. These findings suggest that neither model has a strong advantage for these particular phonological categories; even after fine-tuning, nuanced contrasts such as nasality and tone remain challenging.

Lang	Model	Tone	Nas.	V-Len	C-Len
Cicipu	MMS	0.199	0.337	0.239	0.174
	XLS-R	0.215	0.337	0.248	0.183
Mocho'	MMS	—	—	0.154	—
	XLS-R	—	—	0.160	—

Table 2.3: Phonological error rates (0–1 scale) for Cicipu and Mocho'. Cicipu shows higher confusion in tone, nasality, and consonant length, while Mocho' displays more issues with vowel length. Both models (MMS vs. XLS-R) yield broadly similar error patterns.

Lang	Category	S%	D%	I%
Cicipu	Tone	47.62	32.11	20.27
	Nasality	0.00	70.00	30.00
	V-Len	55.58	27.86	16.56
	C-Len	44.18	33.90	21.92
Mocho'	V-Len	58.29	26.86	14.86

Table 2.4: Percentage of substitutions (S%), deletions (D%), and insertions (I%) in XLS-R 120 min output, for each phonological category in Cicipu and Mocho'.

Substitution, Deletion, and Insertion Patterns

For further investigation, we performed a more detailed error analysis on the output from the 120-minute model of XLS-R for Cicipu and Mocho'. Table 2.4 breaks down each category by the percentage of substitutions (S), deletions (D), and insertions (I).

In Cicipu, nearly half of the tone errors ($\sim 48\%$) are substitutions, indicating confusion over which tone mark to apply, while around a third are deletions and the remainder insertions. Nasality errors, by contrast, skew heavily toward deletions ($\sim 70\%$), suggesting the model often fails to detect nasal vowel features. Substitutions are rare for nasality, indicating that the system either omits it entirely or adds it spuriously rather than confusing

it with another tone diacritic. Vowel-length errors ($\sim$ 44% for deletions and insertions) and consonant-length errors ($\sim$ 55% for deletions and insertions) reflect a high level of segment-level confusion, whereas confusion with a different vowel ($\sim$ 56% substitutions) or consonant ($\sim$ 44% substitutions) is also frequent. For Mocho', vowel-length confusion is likewise dominated by substitutions ($\sim$ 58%), a pattern similar to Cicipu which reveals that XLS-R often misidentifies one segment in the long vowel. This pattern is also consistent with CTC-style decoding behavior, where duration contrasts are encoded implicitly through repeated frame-level labels and can become unstable under low-resource training. The distributions point out that while the model does capture some acoustic correlates of nasality, tone, and length, it nevertheless struggles to map them consistently to diacritics and extended graphemes in low-resource scenarios.

The substitution counts in Table 2.4 combine two distinct events for the length categories: a change in vowel or consonant quality and a change in length. To isolate the length contrast itself, we restrict attention to substitutions that alter length while leaving segment quality unchanged, and classify each by direction. Table 2.5 reports, for the XLS-R 120-minute output, the proportion of reference long segments recovered with correct length, the proportion rendered short or dropped, and the proportion of reference short segments rendered long.

The pattern is sharply one-directional. Long vowels and geminate consonants are recovered correctly only about 60–70% of the time, and when they fail they are almost always shortened or dropped rather than over-lengthened; the reverse error is roughly an order of magnitude rarer, and for Cicipu consonants it is well under 1%. Consonant gemination is the most fragile contrast, with nearly a third of reference geminates degeminated or dropped. This asymmetry follows from the CTC-style decoding noted above: duration is encoded through repeated frame-level labels, and under low-resource training the model tends to collapse these repetitions, biasing it toward under- rather than over-production of length.

Overall, these patterns highlight the persistent challenge of representing languages with complex orthographies and rich phonological inventories. Even after multilingual pretraining and fine-tuning, contrasts such as tone or nasality may be overlooked when the

	Vowel	Consonant
Cicipu		
Long kept	0.72	0.59
Long collapsed	0.21	0.33
Short kept	0.81	0.87
Short lengthened	0.03	0.01
Mocho'		
Long kept	0.66	—
Long collapsed	0.31	—
Short kept	0.89	—
Short lengthened	0.02	—

Table 2.5: Directional length-error rates for the XLS-R 120-minute output, grouped by language. *Long kept* and *short kept* are the proportions of reference long and short segments recovered with correct length; *long collapsed* is the proportion of long segments shortened or dropped, and *short lengthened* the proportion of short segments rendered long. Rates are conditional on the reference length and do not sum to one, since segment deletions and quality-only substitutions are not shown.

amount of transcribed data is minimal. Addressing these gaps may require linguistically informed data augmentation, specialized adapter modules, or loss functions that explicitly emphasize distinct phonological categories. In extremely low-resource settings, such targeted methods could provide the additional examples and acoustic cues needed for more accurate transcription of endangered languages.

2.5 Practical Takeaways for Field Linguists

Our experiments show that fine-tuned multilingual ASR models can potentially reduce the transcription burden for endangered and low-resource languages. Across five typologically diverse languages, MMS proved more effective with extremely limited labeled data, whereas XLS-R caught up once approximately one hour of transcribed material was avail-

able. By using Character Error Rate (CER) rather than Word Error Rate (WER), we focus on phoneme-level accuracy—a more direct measure for languages without standardized orthographies. Despite improvements in overall accuracy, both models struggled with challenging phonological categories in Cicipu, such as tone and consonant length, and exhibited a high rate of vowel-length confusion in Mocho'. These findings confirm that current multilingual ASR systems are indeed helpful for language documentation but still require targeted adaptations to handle nuanced phonological contrasts in under-resourced settings.

2.6 Future Directions

Although our study confirms that fine-tuning multilingual ASR models can substantially reduce transcription overhead for low-resource languages, several research directions remain promising for further performance gains. One compelling approach is continued pre-training (CoPT) on unlabeled in-language audio. DeHaven and Billa (2022) show that CoPT on a wav2vec 2.0–based multilingual model can match or outperform pseudo-labeling techniques while being more computationally efficient. Similarly, Nowakowski et al. (2023) demonstrate that CoPT on about 234 hours of Sakhalin Ainu audio yields a considerable reduction in error, beyond what standard multilingual fine-tuning achieves. CoPT has also proven effective in domain adaptation, especially for noisy data or new speaker types (Attia et al., 2024). While the scarcity of fieldwork data could limit the scale of CoPT, even incremental benefits may substantially ease the manual transcription effort.

A second avenue is leveraging diverse augmentation methods to enlarge the effective training set. Self-training (pseudo-labeling) uses an initial ASR model to generate transcripts for unlabeled audio, which can then be added to the training pool. This method has consistently boosted low-resource ASR performance (Bartelds et al., 2023), particularly when coupled with filtering or iterative refinement. TTS-based augmentation offers another option: if a target-language text-to-speech system is available, synthesizing speech from text yields additional "perfectly labeled" data, potentially improving recognition robustness (ibid). Finally, common audio perturbations, speed/pitch changes, SpecAugment, and noise injection, remain valuable for avoiding overfitting and preparing the model for real-world variability.

A final challenge involves better capturing difficult phonological categories, such as tone, nasality, and consonant length. Adapters in MMS could be extended or reconfigured to emphasize language-specific features, while training regimes could incorporate acoustic or phonological priors explicitly. Future work might integrate fine-grained linguistic annotations (if available) or employ specialized masking strategies during CoPT to boost the model's sensitivity to subtle contrasts. Combining these techniques into user-friendly toolkits will be essential for widespread adoption by field linguists, who often have limited computational resources yet require high-accuracy, phoneme-level transcriptions for documenting and revitalizing endangered languages.

2.7 Limitations and Caveats

Despite promising results, several specific limitations affect the generalizability and applicability of our approach. The most critical limitation is related to the data size and representativeness of the linguistic diversity considered. Our study focused on a small number of typologically diverse languages, each with relatively limited datasets ranging from just a few minutes to two hours. As such, the models' performances may not generalize to other endangered or low-resource languages with distinct phonological or orthographic features.

Additionally, due to constraints inherent in linguistic fieldwork, the training and test datasets often contained data from the same speakers, potentially inflating model accuracy estimates. Future research should validate these findings with genuinely held-out speakers to better gauge model robustness to speaker variability.

Moreover, the orthographic inconsistencies and the absence of standardized orthographies in our datasets likely influenced model performance, especially for phonologically complex categories like tone, vowel length, and nasality. This issue highlights a broader limitation: ASR models trained under these conditions may struggle to generalize to spontaneous and noisy field recordings, especially when orthographic conventions vary within and across datasets.

Finally, computational resource limitations (training on a single NVIDIA T4 GPU with constrained runtimes) restrict our ability to fine-tune larger models or extensively optimize

hyperparameters, which may have further improved performance. Addressing these limitations would require additional computational resources and potentially more extensive data augmentation strategies tailored explicitly to low-resource linguistic contexts.

Postlude

This chapter addressed **RQ1**: how do parameter-efficient and full fine-tuning strategies compare when adapting large multilingual ASR models to endangered fieldwork languages, and what role does phonological inventory complexity play in residual error patterns? The central finding is that adapter fine-tuning is more data-efficient below one hour, while aggregate error metrics mask systematic failures on the phonological features, such as tone, nasality, consonant and vowel length.

Since this work was conducted, multilingual ASR has continued to expand in scale. For example, omnilingual ASR (Keren et al., 2025) supports over 1,600 languages and demonstrates that new languages can be supported with only a handful of in-context examples. Yet the questions this chapter raises remain open: there might always be languages that are left out, and broader language coverage does not automatically mean better coverage of the phonological features that distinguish the languages left out. The goal is not merely that every language appear on a supported-language list, but also that every language be supported for the diverse linguistic features that it possesses.

As model architectures and size contiinue to evolve, the implication that carries forward is not primarily about which architecture to prefer. It is about what aggregate metrics conceal: both models improved substantially across most error categories while barely learning the categories that matter most for fieldwork linguists. Whether this residual failure reflects something about how the model encodes languages internally — rather than simply how much data it has seen — is the question that motivates Chapter 3.

Chapter 3

SUB-TOKEN PROBING OF MULTILINGUAL ASR

Overview

This chapter addresses **RQ2**: do sub-token decoding metrics reveal systematic differences in how a large multilingual ASR model processes languages across training-data tiers, do those differences align with typological structure or resource level, and are they reducible to training data volume? The chapter is based on two publications: Liang et al. (2025), which introduces five decoder-level probing metrics applied to 20 Latin-script languages, and a follow-up study Liang et al. (2026), which scales the analysis to 49 languages across diverse scripts to examine the disparities in relation to training data volume.

The experiments indicate that lower-resource languages are decoded through a generic regime of lower confidence, higher entropy, and typologically undifferentiated token usage, while higher-resource languages form coherent typological clusters. Yet sub-token vocabulary activation converges after roughly two hours of audio regardless of resource tier, and the distributional shape of sub-token usage reflects orthographic and typological structure rather than training scale. Together, these findings show that the representational disadvantage of low-resource languages is not simply a function of insufficient data; it is shaped by tokenizer design and the structure of the shared decoding hypothesis space.

3.1 Why Move Beyond Aggregate Error Rates

Large multilingual automatic speech recognition (ASR) models like Whisper (Radford et al., 2023) demonstrate impressive capabilities on high-resource languages, yet their performance often degrades significantly for low-resource languages (Javed et al., 2022), while concerns about fairness across diverse linguistic groups persist (Zee et al., 2024). Aggregate metrics such as Word Error Rate (WER) can obscure the nuanced ways these models falter internally and may not capture critical issues like hallucination (Koenecke et al., 2024).

This necessitates deeper analysis of the decoding process itself, with some prior work also highlighting the value of evaluating models at the sub-unit level, for instance, in assessing calibration (Ballier et al., 2024b).

This chapter posits that a granular, sub-token level investigation of Whisper's decoder is crucial for a more comprehensive understanding of these performance variations. We use the term 'sub-token' to refer to the sub-word units (e.g., Byte Pair Encoding (BPE) units) that models like Whisper generate, often broadly called 'tokens' in the literature. Recognizing that tokenization strategies can themselves introduce biases and affect model behavior (Petrov et al., 2023; Ahia et al., 2024), we track five concrete properties of decoding across languages: how highly the correct token is ranked, how confident the decoder is in its chosen token, how uncertain the candidate distribution is (entropy), how diverse the alternatives are, and how token-usage patterns cluster across languages.

We first apply five probing metrics to Whisper-large-v2 across 20 Latin-script languages spanning three resource tiers, demonstrating that higher-resource languages consistently benefit from more robust decoding metrics and that sub-token usage patterns reveal typologically coherent clusters, while unrelated low-resource languages are decoded through a generic regime. We restrict this set of experiments to Latin-script languages because the per-token metrics are not meaningfully comparable across scripts whose sub-tokens encode different linguistic granularities. We then extend the analysis to 49 languages across diverse scripts and resource levels using script-agnostic measures of vocabulary activation, asking how much audio is required to observe the model's learned sub-token inventory and whether the representational disadvantage documented in the first set of experiments is reducible to data volume. The results show that sub-token vocabulary activation converges after roughly two hours of audio regardless of resource tier, and that rank–frequency distributions are shaped by orthographic structure rather than pre-training scale alone—indicating that simply scaling data does not close the equity gap.

3.2 Decoder, Tokenization, and Resource Disparity

3.2.1 Whisper and Tokenisation

Whisper is an influential foundation encoder-decoder Transformer model (Radford et al., 2023). It was trained using large-scale weak supervision on approximately 680,000 hours of multilingual audio data, covering a wide array of tasks, including speech transcription and translation. This extensive pre-training enables strong zero-shot performance.

A core component of Whisper's architecture, as well as that of many modern large language and speech models, is its tokenization strategy. As detailed by Radford et al. (2023), Whisper utilizes two separate Byte Pair Encoding (BPE) vocabularies: one derived from the GPT-2 tokenizer (Sennrich et al., 2016; Radford et al., 2019) for English-only models, and a distinct, refitted vocabulary of the same size for multilingual models. This refitting was intended to avoid excessive fragmentation on other languages since the BPE vocabulary is English only (Radford et al., 2023). During decoding, the model generates a sequence of these sub-tokens, typically guided by special tokens like a language ID. While sub-word units allow handling large vocabularies and morphological variations more effectively than word-level tokenization, and can facilitate cross-lingual transfer (Conneau et al., 2020), Radford et al. (2023) themselves acknowledge potential limitations, particularly for languages distant from the Indo-European family which forms the bulk of the training data. They note that performance outliers could be due to a lack of transfer across languages and that the BPE tokenizer could be a poor match for these languages or variations in data quality.

However, the nature of sub-word tokenization, especially in a multilingual context, is not without its challenges. The way texts are segmented into tokens can vary significantly across languages, potentially leading to disparities in processing efficiency, context window utilization, and even model performance (Petrov et al., 2023). For instance, some languages might be systematically broken into more tokens than others for equivalent semantic content, an issue explored in the context of text-based LLMs (Petrov et al., 2023; Ács, 2019). Such tokenization artifacts can contribute to unfairness, as models might inherently find it more complex to process or learn representations for languages that result in longer token sequences (Ahia et al., 2024). While recent research also explores discrete acoustic or se-

mantic tokens for ASR (Guo et al., 2026; Cui et al., 2024), the BPE approach as employed in Whisper remains a common paradigm, making the study of its sub-token characteristics critical. Typological studies further link BPE compression behavior to morphological structure (Gutierrez-Vasques et al., 2023), while scaling analyses demonstrate that tokenization strategy substantially impacts performance across 70+ languages (Tjandra et al., 2023).

3.2.2 Beam Search Decoding

For generating transcriptions, Whisper typically employs beam search decoding. Beam search maintains a set of k (the beam width) most probable partial hypotheses (sequences of tokens). At each step, it extends these hypotheses with possible next tokens and reranks them based on their cumulative probabilities, pruning the set back to k hypotheses. This exploration of multiple paths aims to find an overall sequence with a higher likelihood than what might be found through a purely greedy approach. Our analysis focuses on the characteristics observed along the single, final beam path chosen by the model, representing its ultimate transcription output. Probing model predictions, particularly the probability scores assigned by the decoder to chosen sub-tokens, provides insights into the model's decision-making at each step of this generation process. These output probabilities are commonly used as a direct measure of model confidence in the ASR literature (e.g., Jiang, 2005; Ballier et al., 2024b,a; Aggarwal et al., 2025). However, it is also well-established that the raw softmax probabilities from deep neural networks may not always be well-calibrated and can exhibit overconfidence (Guo et al., 2017).

3.2.3 Resource Disparity and Fairness in Multilingual ASR

A significant challenge in developing truly equitable multilingual ASR systems is the vast disparity in available training resources across languages. While Whisper's pre-training dataset is exceptionally large, the distribution of data per language varies by orders of magnitude (Radford et al., 2023). This imbalance directly impacts model performance, generally leading to superior results for higher resource languages (Javed et al., 2022). Scaling models and data (e.g., Tjandra et al. (2023) and Pratap et al. (2024)) can improve overall perfor-

mance but does not necessarily resolve fairness issues or guarantee equitable performance across all languages and speaker groups (Zee et al., 2024). In fact, Zee et al. (2024) found that larger models can sometimes exhibit greater worst-case performance disparities.

The challenges for low resource languages are multifaceted. Beyond raw data quantity, issues include the pre-training coverage of diverse scripts (Pfeiffer et al., 2021; Muller et al., 2021), the quality of sub-token vocabularies for these languages (Downey et al., 2023), and the potential for "capacity dilution" where a fixed-size model struggles to adequately represent many languages (Conneau et al., 2020). These factors can lead to higher error rates, lower model confidence, and increased susceptibility to issues like hallucination (Koenecke et al., 2024) for low resource languages. Prior work often evaluates these disparities at the word or utterance level (e.g., WER), with dedicated studies benchmarking performance on specific low-resource language sets like Pashto, Punjabi, and Urdu (Sehar et al., 2025). Recent work has also begun to probe representation and decoding mechanisms in large speech language models: Langedijk et al. (2024) proposed layerwise interpretability methods for encoder–decoder transformers, while related analyses highlight structural bias in multilingual decoders (Koenecke et al., 2024; Muller et al., 2021) and the importance of uncertainty calibration (Guo et al., 2017). Our research instead asks: how do decoder-level uncertainties and hypothesis characteristics manifest differently between resource tiers even before a full word or sentence is output? This sub-token perspective is crucial for understanding the foundational biases and uncertainties that may contribute to downstream performance gaps and for developing targeted interventions.

3.2.4 Corpus Design and Resource Evaluation

A recurring challenge in ASR corpus development is determining how much speech data is needed for robust evaluation and meaningful cross-lingual comparison. While large-scale efforts such as Common Voice provide the scale for such analyses (Ardila et al., 2020), most prior work on multilingual ASR evaluation has focused on *fairness* and *balance* across resource levels rather than on *sufficiency* itself—how much data is enough for a corpus to fully engage a model's capacity (Javed et al., 2022; Sehar et al., 2025; Swain et al., 2024;

Zee et al., 2024; Guo et al., 2026). These studies emphasize representativeness and equity across languages but do not explicitly quantify when additional audio ceases to provide new information for the model. We approach corpus design from this angle, introducing a behavioral criterion for corpus completeness: the *acoustic saturation time* (AST), defined as the point beyond which additional speech yields diminishing activation of new sub-tokens, reframing corpus evaluation as a model–data interaction problem.

3.3 Dataset

3.3.1 Decoder Probing: Latin-Script Languages

To investigate Whisper's sub-token decoding characteristics across different linguistic contexts and resource availability, we curated a diverse set of 20 languages. These languages were categorized into three tiers, High, Medium, and Low resource for better visualization, based on their pre-training exposure in Whisper's training data composition (Radford et al., 2023). It should be noted that the definition does not correspond to actual resource levels in the real world beyond Whisper's training dataset. The selected languages are detailed in Table 3.1. English, while having the highest pre-training exposure in Whisper's training data, was intentionally excluded from our analysis. Its training data volume is disproportionately larger (over 430,000 hours) compared to the other high resource languages analyzed (e.g., German with approximately 13,000 hours), which would skew the comparative analysis across resource tiers and make it less informative for studying graduated cross-linguistic differences.

To maintain consistency in sub-token analysis and simplify cross-linguistic comparisons at the sub-token level, our analysis primarily focuses on a subset the languages supported by Whisper and available on Common Voice. We restricted our scope to languages written in the Latin script for two reasons. First, pilot analyses of non-Latin scripts (e.g., Chinese, Japanese, Tibetan, Amharic) revealed systematic transcription failures such as outputs rendered in Latin characters rather than the intended script, which would have confounded our cross-linguistic comparisons and PCA analysis. Second, and more fundamentally, the five decoding metrics we define—rank, confidence, entropy, diversity, and sub-token usage—are

not directly comparable across scripts: a BPE sub-token in a Latin-script language and a logographic token in Chinese represent fundamentally different linguistic units, so cross-script comparisons of these per-token metrics would conflate tokenization granularity with decoding behavior. Restricting to a single script family holds this variable constant and isolates the effects of training resource level and typological structure. Additionally, we made an explicit decision to exclude languages for which our initial baseline ASR performance using Whisper-large-v2 was exceptionally poor (specifically, WER higher than 60%). Preliminary qualitative analyses on languages such as Uzbek, Swahili and higher resource Danish revealed that the model frequently misrecognized the target language entirely or produced outputs with non-canonical orthography, making a meaningful analysis impractical.

For each of the selected languages, we used approximately 10 minutes of speech data randomly sampled from the validated subset in the Common Voice 17.0 dataset (Ardila et al., 2020). To assess whether 10 minutes adequately represents the sub-token space, we computed a token coverage curve on held-out Estonian: cumulative unique decoder sub-tokens observed as duration increases ($1\times$10min $\ldots$ $6\times$10min = 60min). The first 10 minutes cover $\sim$73% of the unique sub-tokens that appear by 60 minutes; 30 minutes reach $\sim$89%, and 40 minutes $\sim$94%, after which gains are $\leq$6% per additional 10 minutes. This motivates our use of 10 minutes as a practical representative sample (frequent tokens dominate decoder behaviour), while we revisit duration choice in Section 3.7.5.

3.3.2 Saturation Probing: Languages Across Diverse Scripts

To test whether the patterns observed across twenty Latin-script languages generalize across writing systems and a broader resource spectrum, we conduct further experiments on a total of 49 languages from the Common Voice 17.0 dataset (Ardila et al., 2020), selected from 65 languages that are both in the dataset and supported by Whisper for inference, shown in Figure 3.8. This set spans a broad typological and orthographic range, including languages across various language families, in diverse scripts such as Latin, Cyrillic, Arabic, Devanagari, Chinese-Japanese (CJ), etc., as well as in disparate levels of resource during Whisper's pre-training. English is again excluded to avoid extreme leverage from its disproportionate

Resource Tier	Language	Training Hours
High	German	13,344
	Spanish	11,100
	French	9,752
	Portuguese	8,573
	Turkish	4,333
Medium	Italian	2,585
	Swedish	2,119
	Dutch	2,077
	Catalan	1,883
	Finnish	1,066
	Indonesian	1,014
	Hungarian	379
	Romanian	356
	Norwegian	266
Low	Welsh	73
	Lithuanian	67
	Latvian	65
	Azerbaijani	47
	Estonian	41
	Basque	21

Table 3.1: Whisper training hours by language, categorized by resource level.

training size. Language inclusion was based on a quality filter described in Section 3.6.1.

For the expanded analysis, audio is processed cumulatively at 10–120 minute checkpoints (10-minute steps), so each longer window strictly contains all shorter windows. This cumulative design enables the saturation analyses described below.

3.4 Methodology of Sub-token Probing

3.4.1 Sub-token Extraction

Our methodology involves two key stages: generating hypothesis transcriptions and capturing the decoder's state at each generation step. For each audio utterance, we first obtain a transcription using Whisper-large-v2 with beam search (beam size=5, temperature=0.2). We provide the correct language ID token in the initial prompt to ensure transcription in the target language. We use the official open-source implementation and enforce target-language decoding by prefixing the decoder with `<|sot|>`, the language ID token `<|`ℓ`|>`, and `<|transcribe|>`. Audio is resampled to 16 kHz. This process yields the hypothesis sequence of sub-tokens $C = (c_1, c_2, \ldots, c_{N_H})$.

Once this hypothesis is generated, we re-trace its generation path step-by-step. For each position s in sequence C, we:

1. Provide the decoder with the audio features (encoder output) and the prefix of already chosen sub-tokens $(c_1, c_2, \ldots, c_{s-1})$

2. Extract the decoder's full probability distribution over all possible sub-tokens for step s

3. Record the top-$K_{cand} = 50$ sub-token candidates $T_{s,K_{cand}} = (t_{s,1}, \ldots, t_{s,K_{cand}})$ along with their respective log-probabilities

This re-tracing procedure captures the decoder's internal state—its ranked candidates and their probabilities—at each decision point in the transcription process. These snapshots form the basis for calculating all our analytical metrics and provide insight into the model's decision-making across different languages. In practical terms, the five metrics below ask

whether the model is correct early, confident when correct, uncertain when struggling, broad or narrow in its alternatives, and typologically structured or generic in its token usage.

3.4.2 Five Decoding Metrics

We compute several metrics by analyzing the beam search path of hypotheses generated by Whisper. For each utterance, we denote $C = (c_1, c_2, \ldots, c_{N_H})$ as the sequence of N_H sub-tokens in the hypothesis. At each step s in the generation process:

- c_s is the sub-token selected by beam search
- $T_{s,K_{cand}} = (t_{s,1}, t_{s,2}, \ldots, t_{s,K_{cand}})$ represents the top-K_{cand} (50 in our experiments) sub-token candidates predicted by the decoder, with corresponding probabilities $(p_{s,1}, p_{s,2}, \ldots, p_{s,K_{cand}})$

For metrics requiring comparison against ground truth, we use the reference transcription $G = (g'_1, g'_2, \ldots, g'_{N_R})$, tokenized with Whisper's tokenizer. Metrics are aggregated over all relevant items for a given language L, with N_{H_L} denoting the total sub-tokens generated across all hypotheses and N_{R_L} the total sub-tokens in all reference transcriptions for language L.

Average Rank of Correct Sub-token

This metric assesses how highly the ground truth sub-tokens rank among the decoder's predictions. We use Levenshtein alignment to map each reference sub-token g'_k to a hypothesis sub-token c_s (or identify it as a deletion). For each reference sub-token g'_k:

- If g'_k is aligned with hypothesis sub-token c_s: The rank of g'_k is its 1-indexed position in $T_{s,K_{cand}}$. If g'_k is not found within the top-K_{cand} list, its rank is assigned a penalty value of $K_{cand} + 1$.
- If g'_k is deleted: Its rank is also assigned the penalty value $K_{cand} + 1$.

The average rank for a language L is the mean of these individual ranks across all reference sub-tokens:

$$\overline{\text{Rank}}(g')_L = \frac{1}{N_{R_L}} \sum_{k=1}^{N_{R_L}} R(g'_k)$$

A lower average rank indicates that the correct sub-tokens are more frequently found among the model's top predictions.

To illustrate this process, Table 3.2 shows the alignment between ground truth and model-generated sub-tokens for a Turkish phrase. The table demonstrates possible alignment operations: equal (where the model correctly identified the token), replace (where the model chose a different token), and delete (where a ground truth token has no corresponding model token).

Position	**GT**	**Output**	**Operation**	**Rank**
0	⟨—BOS—⟩	⟨—BOS—⟩	equal	1
1	Sil	S	replace	1
2	ah	el	replace	5
3	lar	am	replace	7
4	...	lar	replace	4
5	⟨—EOS—⟩	(deleted)	delete	$K+1$
6		A	insert	–
7		ley	insert	–
8		kum	insert	–
9		.	insert	–
10		⟨—EOS—⟩	equal	2

Table 3.2: Alignment between ground truth (GT) sub-tokens ("Silahlar..."/"Weapons...") and model output sub-tokens ("Selamun Aleykum."/"Peace be upon you.").

For this example, the average rank is calculated as: $\overline{\text{Rank}} = \frac{1+1+5+7+4+(K+1)+2}{7} = \frac{20+(K+1)}{7} = 10.14$ with $K = 50$.

Confidence

Confidence measures the average probability assigned to the sub-token c_s that was ultimately chosen at each step along the beam search path:

$$\overline{\text{Conf}}_L = \frac{1}{N_{H_L}} \sum_{s=1}^{N_{H_L}} p(c_s|\text{utterance}, c_{<s})$$

where $p(c_s|\text{utterance}, c_{<s})$ is the probability assigned to c_s given the utterance audio input and previous tokens.

Entropy

Token-level entropy quantifies the uncertainty in the model's predictions, calculated over the top-$K_H = 50$ predicted sub-tokens. After normalizing the probabilities of these candidates, the entropy (in bits) at step s is:

$$H_s = -\sum_{i=1}^{K_H} p'_{s,i} \log_2 p'_{s,i}$$

where $p'_{s,i}$ are the renormalized probabilities over the top K_H candidates. The average entropy $\bar{H}_L$ for a language is reported as the mean across all decoding steps.

Alternate-candidate Diversity

This metric evaluates the variety within the set of predicted candidates using the Type-Token Ratio (TTR). Specifically, we calculate the TTR of the non-top-1 candidates within the top-$K_D = 50$ predictions. For each language, we collect all sub-tokens that appear as candidates ranked from 2 to K_D across all decoding steps. The diversity is then computed as:

$$\text{Diversity}_L = \frac{|\text{unique non-top-1 tokens in L}|}{|\text{total non-top-1 tokens in L}|}$$

This approach explores the richness of the hypothesis space beyond the model's single best guess. Our underlying assumption is that lower-resourced languages might exhibit less diversity in these alternative candidates, potentially reflecting a more constrained or less nuanced hypothesis space learned by the model due to limited training data. A higher TTR indicates a broader range of unique alternatives being considered.

Sub-token Use Visualization

To visualise cross-lingual patterns in sub-token usage, we build a frequency vector for each language from the sub-tokens that appear among the top-$K_{\text{PCA}} = 10$ candidates at every decoding step. These frequency vectors represent the distribution of sub-token IDs from Whisper's multilingual vocabulary of 51,865 tokens (Radford et al., 2023). Using a wider window (e.g. $K_{\text{PCA}} = 50$) would fold in many low-frequency alternates and blur fine-grained distinctions, so we fix K at 10 to keep language differences salient.

After standardizing these vectors, we apply two dimensionality reduction techniques: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA projects the data onto the first two principal components to reveal linear relationships and global structure, while t-SNE emphasizes local neighborhoods and non-linear clustering patterns. The two approaches allow us to examine both broad cross-linguistic trends and fine-grained language-specific patterns in sub-token usage. For t-SNE, we used scikit-learn t-SNE with perplexity set to 20 (clamped automatically to $(n-2)/3$ if the number of languages was small).

3.4.3 Token Discovery and Saturation Modeling

For the expanded 49-language analysis, we additionally track how the model's sub-token vocabulary is activated over cumulative audio duration. For each language L and elapsed audio duration t (in minutes), let $\mathcal{V}_L(t)$ denote the set of sub-tokens that have appeared among the model's Top–K decoder candidates by time t, and let $N_L(t) = |\mathcal{V}_L(t)|$ be the running discovery count. $N_L(t)$ measures how much of the model's sub-token space is engaged during the task. We examine this process along four dimensions: (1) the total number of discovered sub-tokens across languages; (2) the rate of sub-token discovery over time; (3) the rank–frequency distribution of sub-token usage; and (4) the average sub-token length as an indicator of segmentation granularity.

Empirically, $N_L(t)$ follows an asymptotic growth curve with diminishing marginal discovery. We fit, per language, an exponential saturation model

$$N_L(t) \;=\; A_L\big(1 - e^{-k_L t}\big) + B_L,$$

where A_L is the growth amplitude, k_L the saturation rate, and B_L a small offset capturing early activation. We summarize sufficiency via the *acoustic saturation time* (AST), defined as the time to reach 90% of the fitted asymptote:

$$t_{90,L} \;=\; -\ln(0.1)/k_L.$$

$t_{90,L}$ provides a model-based estimate of when additional audio yields minimal discovery of new sub-tokens.

3.4.4 Rank–Frequency Analysis

To characterize how activations are organized at saturation, we compute rank–frequency distributions at the 120-minute checkpoint and fit Zipf-type laws. Specifically, we compare a plain Zipf model and the Zipf–Mandelbrot variant,

$$f(r) \;=\; C\,(r+\beta)^{-\alpha},$$

where r is rank, α controls tail steepness, and β captures heavy-head curvature. Fits are obtained in log–log space via least squares, and model choice is guided by Akaike Information Criterion (AIC). The parameters (α, β) index the concentration of sub-token usage and head–tail balance.

3.4.5 Segmentation Granularity

We quantify sub-token segmentation using the frequency-weighted mean sub-token length

$$\bar{\ell}_L = \frac{\sum_i f_i \cdot |t_i|}{\sum_i f_i},$$

where f_i denotes the empirical frequency of sub-token t_i in the discovered set at the given analysis horizon. We relate $\bar{\ell}_L$ to $\log_{10}$ pre-training hours and, where appropriate, analyze within-script subsets (e.g., Latin) to control for orthographic effects.

3.5 Decoding Differences Across Latin-Script Languages

Our analysis of Whisper's sub-token decoder reveals systematic variations in its behavior that correlate strongly with language resource levels, quantified by training hours. As a baseline, Figure 3.1 reports the Word Error Rate (WER) of the languages studied.

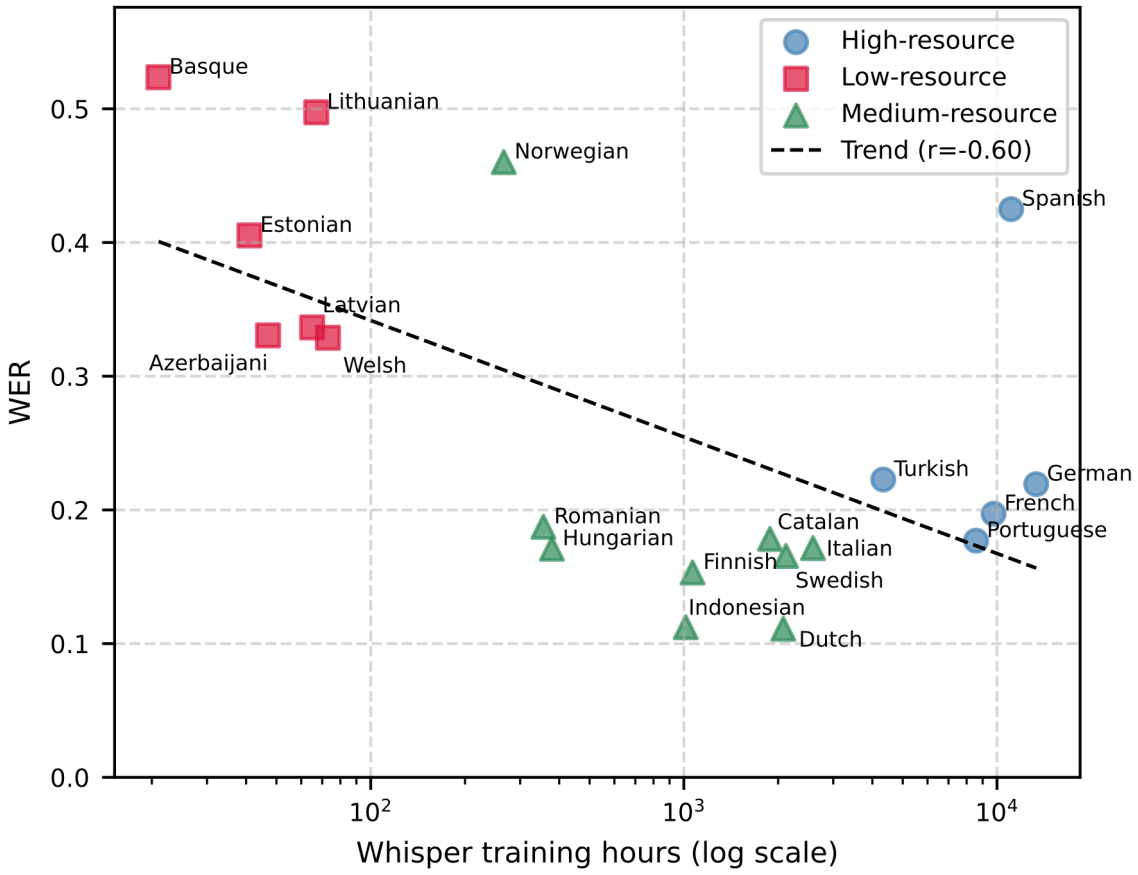

Figure 3.1: WER of Whisper on the Common Voice dataset versus training hours. Higher resource languages tend to have lower WER (p-value < 0.001).

Consistent with previous findings we observe that WER is generally lower for languages with more training hours (Radford et al., 2023). Spanish stands with WER higher than its training data volume would predict; as a pluricentric language spanning many national standards, its Common Voice data is likely more heterogeneous than that of most other high-resource languages, which would penalize a single model fit. Our subsequent sub-token level analyses aim to delve deeper into the internal decoding characteristics that might contribute to these performance differences.

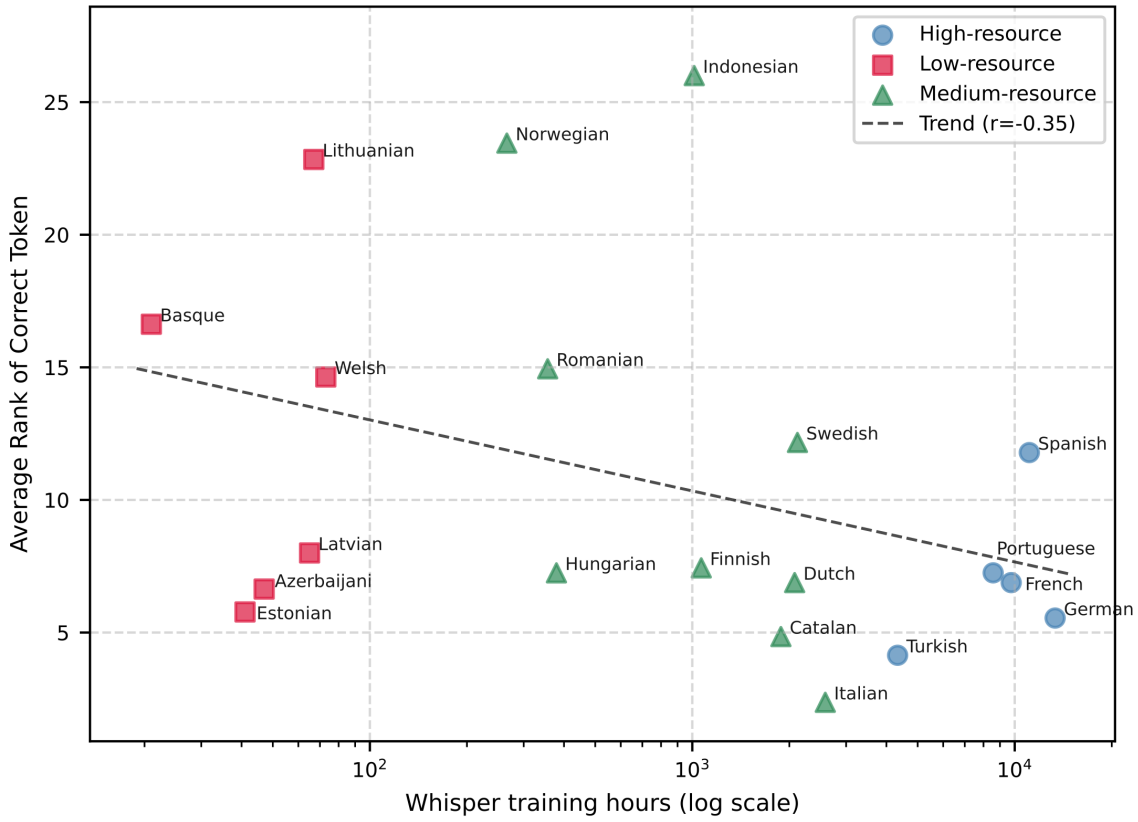

Figure 3.2: Average rank of the correct sub-token versus Whisper training hours. Higher resource languages tend to have correct tokens ranked higher (closer to 1, p-value < 0.001).

3.5.1 Rank of Correct Sub-token

The average rank of the correct sub-token correlates with language resource levels, as shown in Figure 3.2. Higher resource languages have their correct sub-tokens ranked higher. This indicates that for higher resource languages, the beam search predominantly follows the locally highest-probability path. Despite the general trend, we do observe some deviations in individual languages.

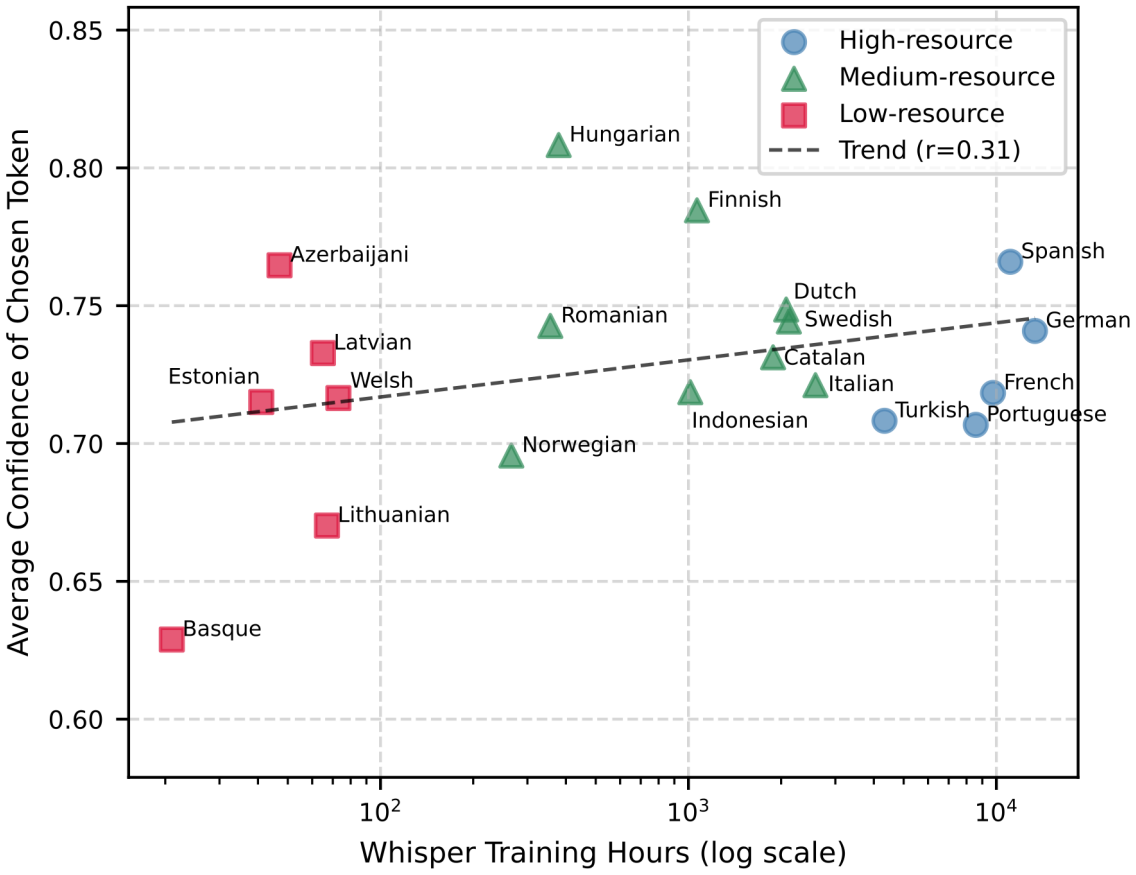

Figure 3.3: Average model confidence (probability of chosen sub-token) versus Whisper training hours. Higher-resource languages generally exhibit higher confidence (p-value < 0.001).

3.5.2 Decoder Confidence and Predictive Entropy

Decoder confidence, measured as the average probability of the chosen sub-token, shows a strong positive correlation with language training hours, as shown in Figure 3.3. High-resource languages tend to exhibit higher average confidence values.

Conversely, predictive entropy, calculated over the top-5 predicted sub-tokens, demonstrates a negative correlation with training hours, as shown in Figure 3.4. High-resource languages show lower average entropy, indicating more peaked and certain predictive dis-

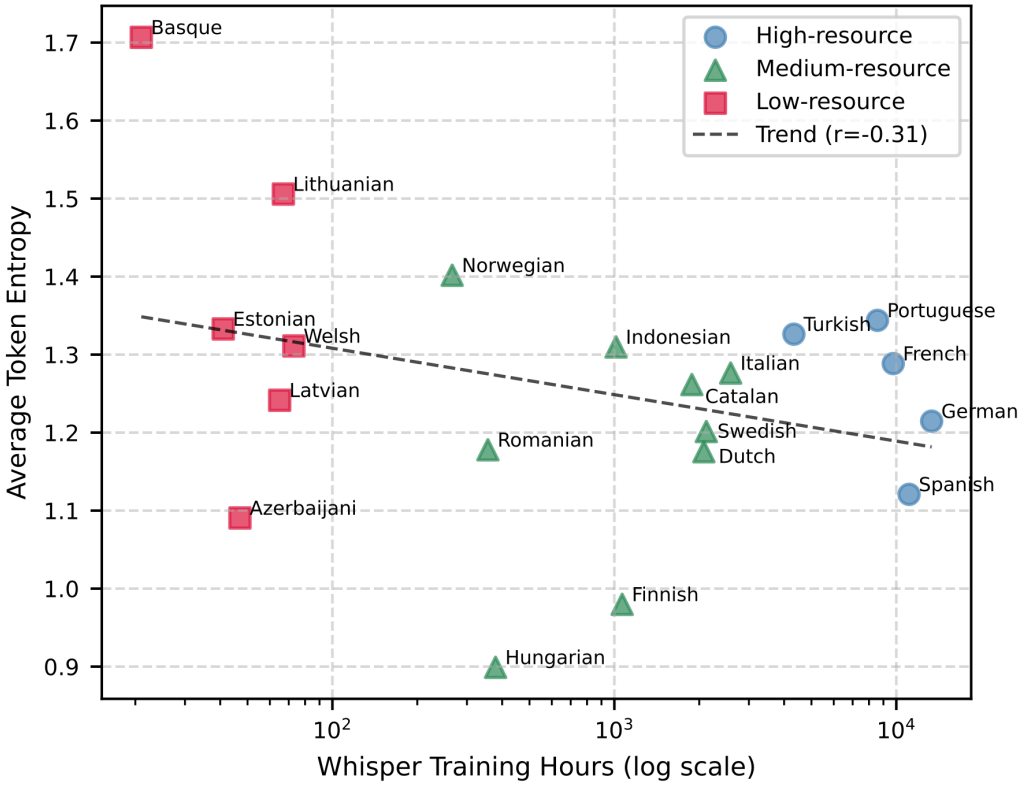

Figure 3.4: Average predictive token entropy of the top-50 candidates versus Whisper training hours. Higher resource languages generally exhibit lower entropy (p-value < 0.001).

tributions. This inverse relationship between confidence and entropy is expected: when the model is more confident in its chosen token, its distribution over alternatives is sharper (less entropic).

3.5.3 Alternate-Candidate Diversity

Alternate-candidate diversity, measured as the TTR of non-top-1 candidates within the top-5 predictions, exhibits a generally positive correlation with language resource levels, as shown in Figure 3.5. Higher resource languages tend to populate the upper range of diversity scores. This suggests that for higher resource languages, the model often considers a richer set of unique alternatives beyond its top choice. The overall trend indicates that

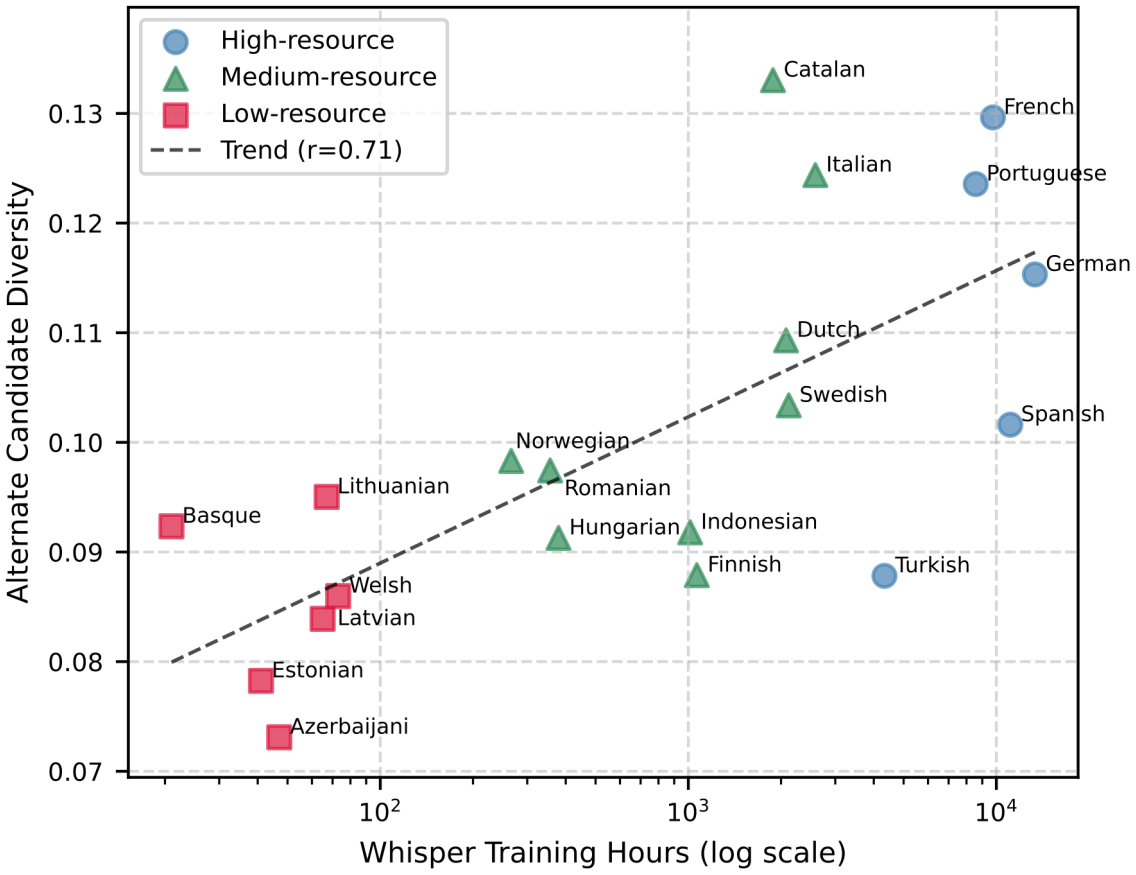

Figure 3.5: Alternate-candidate diversity (TTR of non-top-1 candidates in top-50 candidates) versus Whisper training hours. Higher resource languages tend to have higher alternate-candidate diversity (p-value < 0.001).

increased training data may lead to a more varied set of hypotheses being considered.

3.5.4 PCA and t-SNE Clustering of Sub-token Usage

The PCA visualization of language-specific sub-token frequency vectors (from top-10 candidates) reveals distinct clustering patterns that reflect both typological relationships and resource levels, as presented in Figure 3.6. A prominent observation is the separation of language families. Romance languages (Spanish, French, Portuguese, Italian, Catalan,

and Romanian) form a relatively cohesive cluster. Similarly, Turkic languages (Turkish and Azerbaijani), North Germanic languages (Swedish and Norwegian), Uralic languages (Finnish and Estonian, and to a lesser extent Hungarian), tend to stay clustered closer together and distinctly separated from other groups along the principal components. The clustering is also interesting as it groups languages of varying resource levels together, suggesting that shared linguistic structures (e.g., common morpho-phonological features captured in common sub-tokens) could influence sub-token usage patterns.

Notably, typologically unrelated lower resource languages, such as Welsh, Lithuanian, Latvian, and Basque, cluster together despite their significant linguistic differences. This unexpected grouping suggests these languages are handled similarly by Whisper's decoder not because of linguistic similarity, but due to their shared status as low-resource languages in the training data. Unlike high-resource languages that form distinct family-based clusters, these unrelated low-resource languages appear to share common sub-tokenization characteristics that transcend actual linguistic relationships, indicating deficiencies in the model's sub-token representations where it may be falling back to more generic decoding patterns due to insufficient exposure during training.

The t-SNE analysis (Figure 3.7) provides additional insights into local clustering patterns while confirming several key observations from the PCA. Several neighborhoods remain stable across both dimensionality reduction techniques, particularly Lithuanian–Latvian, Finnish–Estonian, and Norwegian–Swedish pairs, suggesting robust family- or script-level commonalities in sub-token usage. However, t-SNE reveals that family structure is weaker when emphasizing local neighborhoods: Romance languages do not form a single tight cluster as they do in PCA, with Spanish/Portuguese/French appearing more separated. This suggests that while broad cross-lingual trends are largely linear and captured effectively by PCA, fine-grained, language-specific token frequencies can create local distinctions that pull related languages apart in the t-SNE embedding.

The resource-linked patterns observed in PCA are reinforced in t-SNE, where medium-resource languages occupy several tight neighborhoods, while unrelated low-resource languages (Welsh, Lithuanian, Latvian, Azerbaijani) continue to cluster together. This consistency across both methods strengthens the interpretation of a "generic decoding" regime

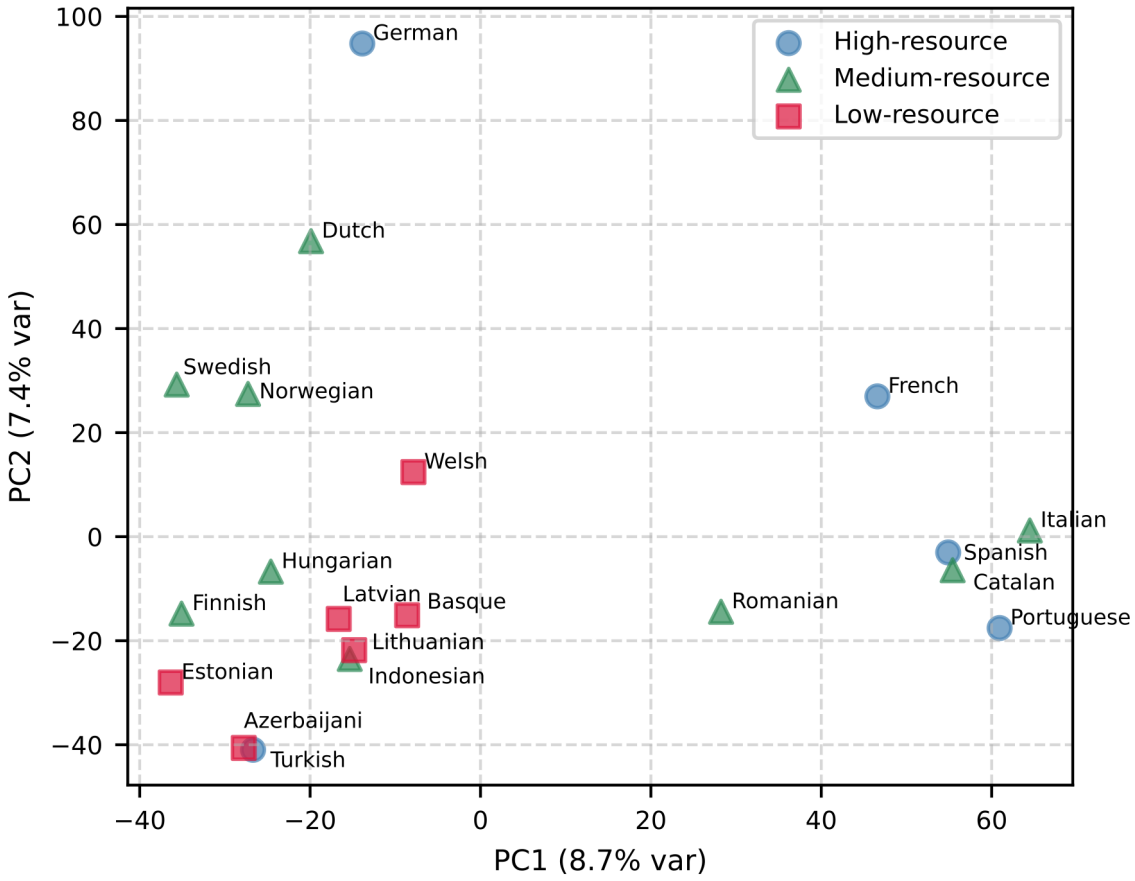

Figure 3.6: PCA of sub-token usage frequency vectors (top-10 candidates).

under data scarcity. Notably, Estonian and Basque emerge as outliers in both visualizations, suggesting these languages exhibit particularly idiosyncratic sub-tokenization patterns that warrant further investigation.

3.6 Sub-token Vocabulary Activation Across Diverse-Script Languages

The results above establish that lower-resource languages are decoded through a generic regime with lower confidence, higher entropy, and typologically undifferentiated sub-token usage. A follow-up question is whether this representational disadvantage is simply proportional to training data volume, and therefore whether more data is the path to equity. To

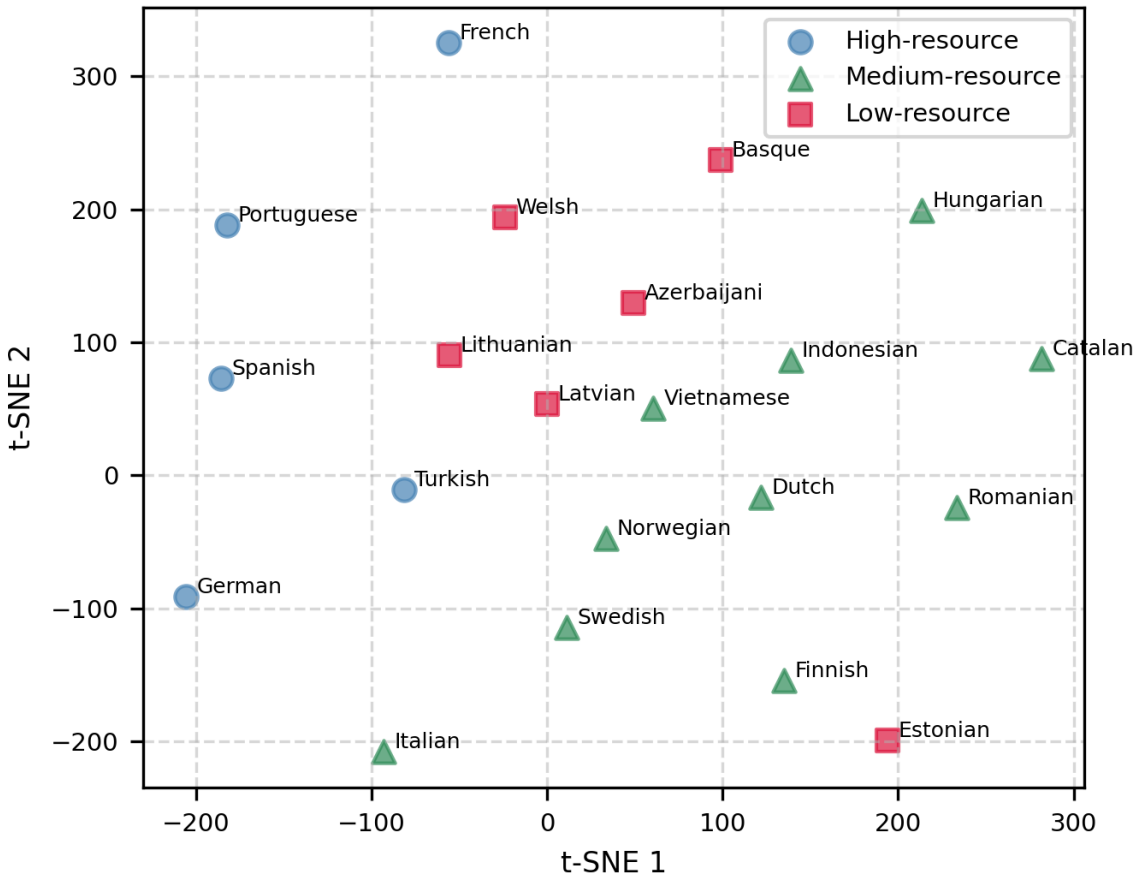

Figure 3.7: t-SNE of sub-token usage frequency vectors (top-10 candidates).

test this, we scale the analysis to 49 languages across diverse scripts. Because the per-token decoding metrics from the first phase are not comparable across scripts (a Latin BPE fragment and a logographic Chinese token encode different amounts of linguistic information), we shift to script-agnostic measures: sub-token vocabulary breadth, discovery rate over time, and distributional shape, none of which require cross-script per-token comparison.

3.6.1 Language Selection and Filtering

We performed Whisper inference across 65 languages covered by both Whisper and Common Voice 17.0. To ensure interpretable cross-linguistic comparison, we confined our primary

analysis to languages with adequate transcription quality, defined as having a Character Error Rate (CER) below 30% on the 10-minute subset. This threshold yielded 49 languages for inclusion. Figure 3.8 summarizes CER values by language, with the red dashed line marking the inclusion threshold.

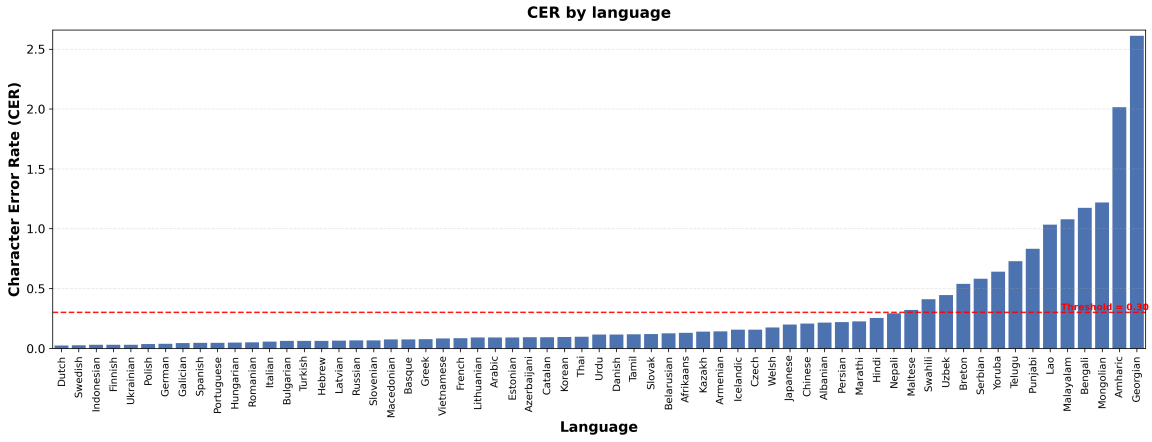

Figure 3.8: Character Error Rate (CER) at 10 minutes for all 65 languages, sorted by CER. The red dashed line denotes the 30% inclusion threshold; only languages below the line (n=49) are included.

Among the 16 excluded high-CER languages, errors primarily reflected either script mismatch or language ID confusion. Script mismatches occurred when Whisper transcribed in an unintended orthography, for example, producing Latin-script output for Amharic rather than Ge'ez. Language confusion arose when the model disregarded the provided language token, such as transcribing Lao speech as Thai.

3.6.2 Sub-token Breadth and Training Scale

Figure 3.9 shows the relationship between the model's training data size and the number of unique sub-tokens discovered during decoding. Each point represents one language, with color intensity indicating training resource level on a continuous logarithmic scale from red (lower resource) to blue (higher resource).

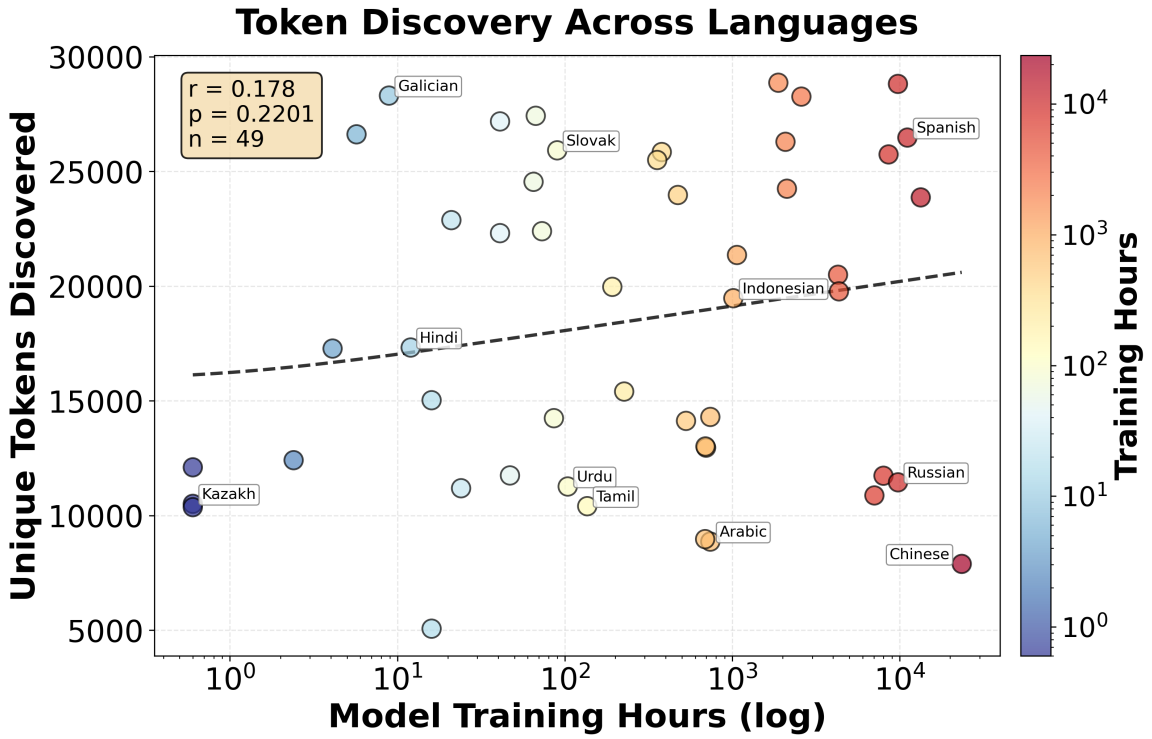

Figure 3.9: Sub-token discovery after 120 minutes across 49 languages. Each point represents one language, with color indicating training data hours and dashed line showing the logarithmic trend. Training data hours show weak, non-significant correlation with sub-token count.

Across all languages, a mean of 18,472 unique sub-tokens (median = 19,474; range = 5,069–28,868) were utilized at the end of inference for 120 minutes. Training hours of the model exhibited only a weak, non-significant correlation with sub-token discovery ($r = 0.178$, $p = 0.22$), indicating that larger pre-training data volumes do not systematically increase the model's comparative sub-token utilization.

Although the languages represent a diverse range of orthography, even when restricting analysis to Latin-script languages only, training hours remained non-significant ($r = 0.256$, $p = 0.18$, $n = 29$). However, a linear regression analysis revealed that writing system accounts for substantial variance in sub-token discovery ($R^2 = 0.634$, $F = 5.81$, $p < 0.001$).

Latin-script languages discovered significantly more tokens (mean = 22,885) compared to non-Latin scripts with a +12,017 token effect ($p < 0.001$). In contrast, languages using Chinese-Japanese (CJ) characters (mean = 9,385), Hebrew (mean = 8,973), and Arabic script (mean = 10,441) consistently discovered fewer tokens. Cyrillic-script languages occupied the intermediate position (mean = 12,768).

3.6.3 Convergence Pattern

To assess how pre-training scale influences the rate at which languages reach sub-token coverage, we fit asymptotic growth curves to sub-token discovery trajectories over time. Three languages (Afrikaans, Azerbaijani, Icelandic) were excluded due to stagnant growth ($< 10\%$ increase from baseline or coefficient of variation < 0.05), leaving 46 languages for analysis.

Figure 3.10 shows fitted saturation curves, with individual trajectories (thin lines) and script-median trends (thick lines). The median time to 90% coverage was $t_{90} = 115.7$ minutes (IQR: 105.4–128.8), and languages reached on average 94.3% of their fitted asymptote at 120 minutes. The exponential model has strong fits across all languages (mean $R^2 = 0.996$, minimum $R^2 = 0.987$).

Across languages, t_{90} exhibited a weak but positive correlation with log training hours (r=0.209, p=0.164, n=46), indicating that languages with larger pre-training data tended to reach full sub-token coverage more gradually. Within the Latin subset, this relationship strengthened and reached statistical significance (r=0.401, p=0.042, n=26). A multiple regression including both log training hours and script accounted for 51.3% of the variance in t_{90} (R^2=0.513, adjusted R^2=0.356; F=3.26, p=0.0039).

Besides the scale of data in pre-training, writing system exerted a systematic effect. Latin-script languages reached saturation fastest (median $t_{90} = 109.3$ min; n=26), followed by Hangul (97.7 min; n=1) and Devanagari (120.6 min; n=3). Chinese and Japanese showed the slowest saturation (median $t_{90} = 155.0$ min; n=2), with Thai (156.0 min; n=1) and Arabic (142.5 min; n=3) displaying similarly delayed convergence. These cross-script differences parallel the sub-token inventory asymmetries reported in Section 3.6.2, suggesting that or-

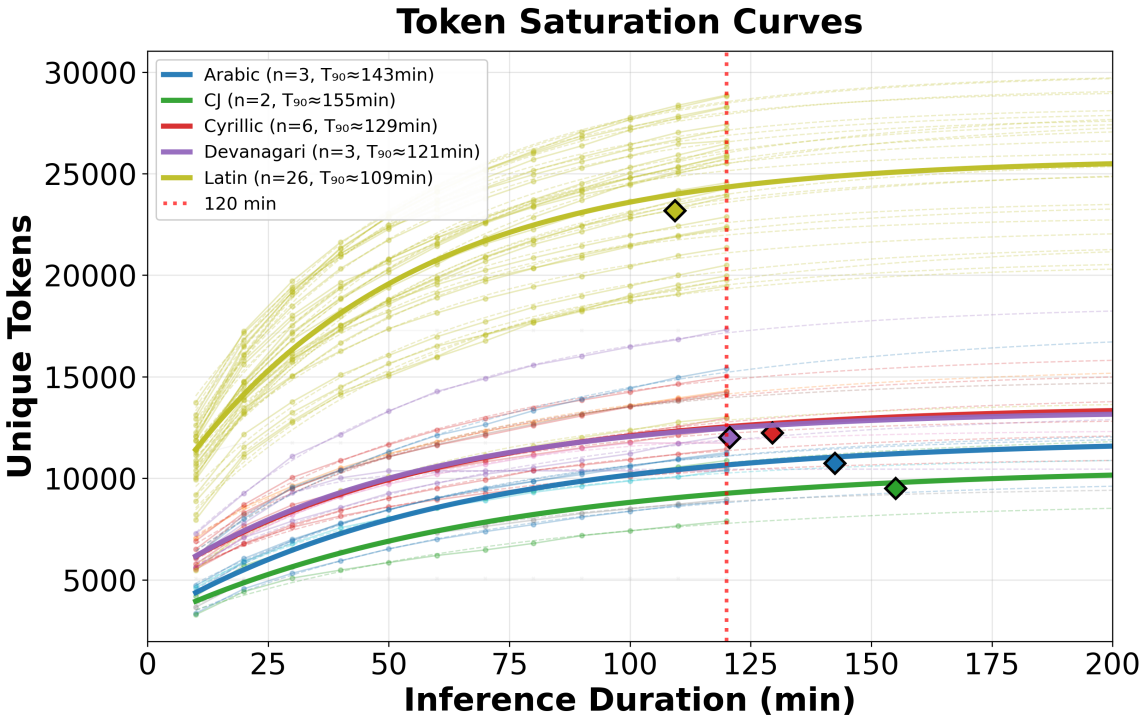

Figure 3.10: Token saturation curves across 46 languages. Thin lines show individual languages; thick lines show script median. Diamonds mark 90% saturation points (t_{90}). Red vertical dashed line indicates the collection window at 120 minutes.

thographic granularity interacts with the dynamics of sub-token activation.

3.6.4 Distribution of Sub-token Utlization Across Languages

To find out about the distributional properties of sub-token utilization across languages, we analyzed the patterns of sub-token frequency. Figure 3.11 plots these distributions on a log–log scale, grouped by writing system, with thin lines for individual languages and thick lines for script-level medians.

Across 49 languages, power-law fits ($f(r) = C\, r^{-\alpha}$) yielded a mean exponent of $\bar{\alpha} = 1.71$ ($\sigma = 0.26$), steeper than the canonical Zipf value ($\alpha \approx 1.0$) typical of natural text. The slopes showed no relationship to training scale ($r = -0.07$, $p = 0.63$ overall; $r = 0.15$, $p = 0.44$ for

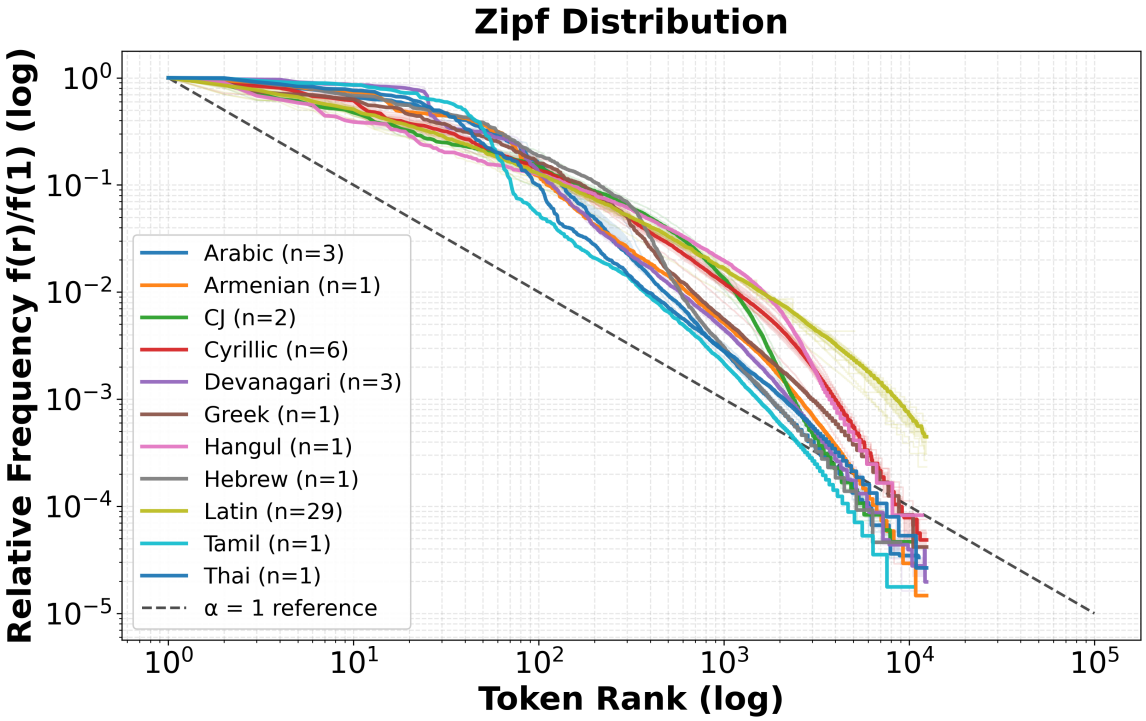

Figure 3.11: Rank-frequency distributions on a log-log scale, grouped by writing system. Thin lines show individual languages; thick lines show script-level medians with IQR ribbons. The dashed reference line indicates canonical Zipf behavior ($\alpha = 1$).

Latin scripts), suggesting that frequency concentration is not governed by data size.

To better capture curvature in the rank–frequency tails, we also fit the Zipf–Mandelbrot model ($f(r) = C(r + \beta)^{-\alpha}$). It provided a consistently better fit across languages, with β values between 6 and 20, indicating a heavier head dominated by a small subset of subtokens. This shift quantifies the model's tendency to reuse a restricted high-probability inventory, reflecting decoder confidence and BPE tokenization bias.

Systematic cross-script differences remained prominent. CJ (Chinese–Japanese) and Cyrillic languages showed the steepest decay ($\alpha \approx 2.1$ and 2.0), followed by Devanagari and Arabic ($\alpha \approx 1.9$), while Latin-script languages were flatter ($\alpha \approx 1.54$). These results suggest that sub-token utilization is shaped primarily by orthographic and segmentation structure

rather than by corpus size.

Overall, Whisper's distributions are Zipf-like but exhibit consistent heavy-head deviation, reflecting structural biases in multilingual tokenization and decoder probability allocation.

3.6.5 Distribution of Sub-token Length Across Languages

As a controlled case study to isolate the effect of training data scale on sub-token granularity while holding orthography constant, we examine frequency-weighted mean sub-token length in Latin-script languages. Figure 3.12 plots these values against training hours (log) for the 29 Latin-script languages.

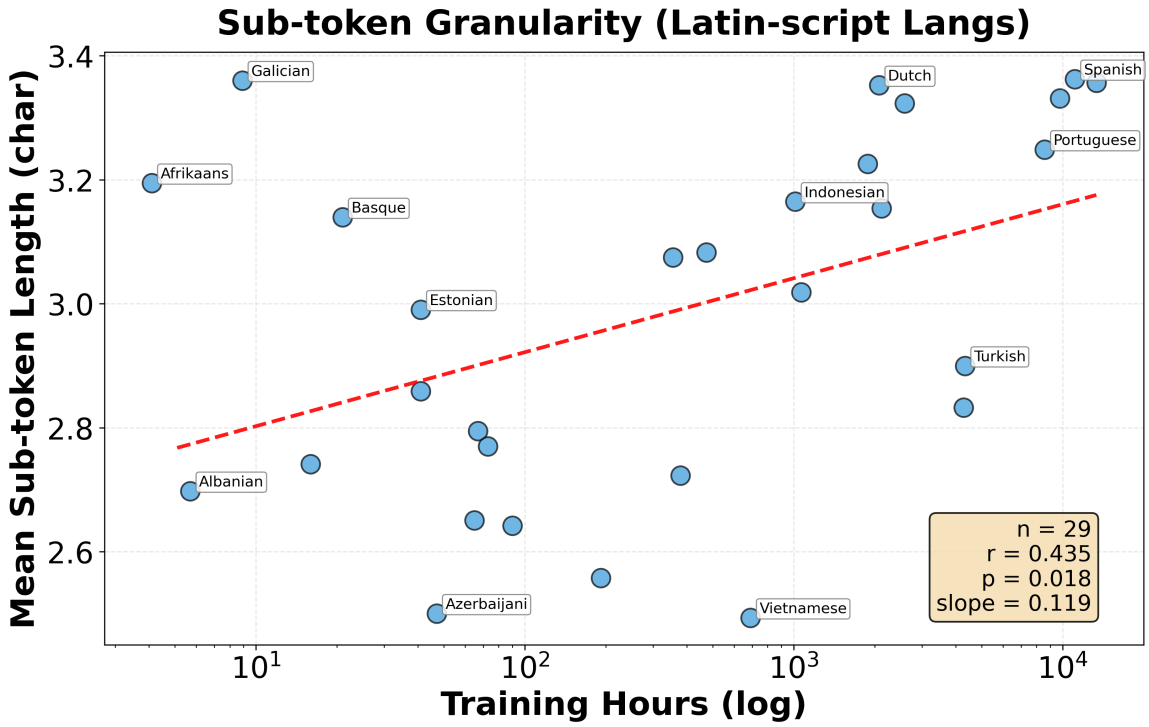

Figure 3.12: Mean sub-token length (frequency-weighted) as a function of training hours for Latin-script languages (n=29). A positive correlation indicates that higher-resource languages tend to have slightly longer sub-tokens.

Within the Latin-script languages, training hours showed a modest but statistically significant positive correlation with sub-token length ($r = 0.44$, $p = 0.018$), with a slope of 0.12 characters per $\log_{10}$ hour.

3.7 What Sub-token Patterns Reveal About Multilingual Fairness

3.7.1 Resource-Driven Decoding Disparities

The analysis consistently reveals that higher resource languages benefit from more favorable decoding characteristics. Specifically, the Whisper ASR model exhibits higher average confidence in its chosen sub-tokens for higher resource languages, and their predictive distributions are marked by lower entropy, indicating greater certainty in its predictions. Furthermore, the correct sub-token is more frequently highly ranked among the candidates considered during beam search for these languages.

These advantages can be attributed to the extensive exposure to diverse linguistic phenomena afforded by larger training datasets (Radford et al., 2023). Such exposure likely enables the formation of more robust and well-calibrated sub-token representations. The generally higher alternate-candidate diversity observed for higher resource languages also suggests that increased training data allows the model to consider a richer and more varied set of plausible alternatives during the decoding process, potentially contributing to improved overall transcription accuracy.

The identification of language-specific weaknesses, such as a consistent tendency for correct sub-tokens to be ranked lower or for predictive distributions to exhibit high entropy in low resource languages, can directly inform the design of language-specific adapters or more sophisticated parameter-efficient fine-tuning strategies (Song et al., 2024; Pfeiffer et al., 2021). Furthermore, decoder-internal metrics, including dynamic measures of confidence and entropy, could potentially serve as valuable signals for adaptive adjustments to decoding algorithms. Finally, the sub-token usage patterns unveiled by PCA and the diversity metric may illuminate critical deficiencies in current vocabulary coverage and suggest novel data augmentation techniques.

3.7.2 The Influence of Linguistic Typology on Sub-token Usage

Beyond the volume of training data, both the PCA and t-SNE analyses of sub-token usage (Figures 3.6 and 3.7) illustrate the strong influence of linguistic structure on the model's internal representations. Typological clusters observed across both methods—such as Lithuanian–Latvian, Finnish–Estonian, and Norwegian–Swedish—underscore that shared morpho-syntactic or phonological features drive stable sub-token patterns.

PCA and t-SNE offer complementary perspectives. PCA captures broad linear trends, with Romance languages forming a cohesive family cluster, while t-SNE emphasizes local neighborhoods where fine-grained token frequency differences separate closely related languages (e.g., Spanish, Portuguese, and French). Typological relationships thus operate at multiple scales: family-level similarities captured by linear methods, and language-specific idiosyncrasies revealed by non-linear analysis.

Clustering patterns also group higher- and lower-resource languages by typological relatedness. Inherent properties such as agglutinative morphology can shape representational space, partly mitigating data scarcity disadvantages when low-resource languages belong to related families. Typological relatedness therefore conditions sub-token representations and may support cross-lingual transfer.

At the same time, unrelated low-resource languages often cluster together despite their differences, suggesting Whisper's decoder treats them similarly due to shared scarcity rather than linguistic similarity. The persistence of this effect in both PCA and t-SNE indicates that data scarcity shapes representations in ways independent of the analysis method.

Finally, Estonian and Basque consistently emerge as outliers, reflecting idiosyncratic sub-tokenization patterns that may require targeted investigation and specialized handling in multilingual ASR systems.

3.7.3 Sub-token Prediction Accuracy versus Global Performance

An intriguing aspect of our findings is the observed incongruity between local sub-token prediction and global Word Error Rate (WER) for particular languages. For instance, languages like Estonian and Azerbaijani, despite demonstrating remarkably low average

ranks for their gold sub-tokens (signifying that the correct orthographic units are included as high-probability candidates by the acoustic model at a local, per-step level) do not invariably achieve the lowest overall WERs in our analyzed set.

This phenomenon highlights the inherent complexities of the end-to-end ASR decoding pipeline. While the model may possess robust local "knowledge" of correct sub-units, the ultimate transcription quality is a cumulative function of global sequence optimization during beam search, the effective mitigation of cascading errors arising from any single misstep, and the nuanced handling of intricate linguistic features that span multiple tokens. Consequently, for such languages, interventions aimed solely at further refining local sub-token prediction accuracy might yield diminishing returns on WER improvement. Strategies that enhance the model's capacity for global context modeling or its specific handling of overarching linguistic complexities may prove more fruitful.

3.7.4 Acoustic Saturation and Corpus Design

The independence of sub-token discovery from training scale has direct implications for cross-lingual evaluation design. If pre-training disparity does not determine the breadth of a model's active vocabulary during inference, then evaluation corpora need not be proportionally scaled to match training resources. Instead, our findings suggest that *acoustic saturation time* (AST), the duration beyond which additional audio yields minimal new sub-token activation, provides a more principled criterion for considerations of phonetic representation. The consistent t_{90} values near 120 minutes across resource tiers indicate that relatively compact evaluation sets can comprehensively observe a model's representational capacity, even for low-resource languages.

This has practical consequences for resource-constrained settings. Rather than pursuing ever-larger test collections to mirror high-resource benchmarks, practitioners can target acoustically saturated samples that fully engage the model's learned vocabulary. This reframes corpus design from a question of absolute size to one of *sub-token coverage*, potentially reducing collection costs while maintaining evaluation validity.

The exponential saturation of sub-token discovery suggests that inference-time sub-token

vocabulary activation is constrained by the statistical structure of speech itself, not merely by training exposure. This finding aligns with information-theoretic perspectives on language modeling: beyond a certain duration, additional audio becomes redundant with respect to the model's learned hypothesis space (Gutierrez-Vasques et al., 2023). The convergence of t_{90} across resource tiers reinforces this interpretation: saturation reflects an intrinsic property of the model–language interaction rather than a direct scaling effect of pre-training data.

For resource planning, AST offers a practical target for data collection campaigns. Instead of aiming for arbitrary hour thresholds, practitioners can use saturation curves to identify when additional recording effort yields diminishing returns. This approach prioritizes **acoustic diversity** over volume, encouraging the collection of phonetically and demographically varied samples rather than simple duration scaling.

3.7.5 Tokenization Bias and Fairness

The Zipf–Mandelbrot distributions of sub-token frequency reveal systematic biases in how multilingual ASR models deploy their sub-token inventories. The flattened heads where a small number of tokens dominate probability mass suggest that decoder behavior is shaped as much by tokenization artifacts as by linguistic productivity. This aligns with prior findings that BPE vocabularies encode resource and morphological biases (Petrov et al., 2023; Ahia et al., 2024), but extends the analysis to inference-time dynamics.

The weak positive relationship between training hours and mean sub-token length ($r = 0.21$) further indicates that segmentation granularity varies subtly across languages. Higher-resource languages exhibit slightly longer tokens, consistent with finer-grained BPE merges learned from extensive exposure, while low-resource languages show more conservative, compact segmentation. This disparity suggests that multilingual models may not learn uniform sub-token representations: instead, tokenization depth co-varies with training scale, which could potentially affect downstream performance in ways not fully captured by error rate metrics alone.

Importantly, the steeper-than-canonical Zipf exponents ($\bar{\alpha} = 1.69$) and heavy-head dis-

tributions indicate that Whisper's sub-token space is *not* uniformly explored. A small fraction of tokens accounts for the vast majority of decoder hypotheses, leaving large portions of the vocabulary rarely activated. This concentration may have implications for model interpretability, calibration, and fairness: if certain languages rely disproportionately on a narrower sub-token subset, their decoding behavior could be more brittle or less robust to distributional shift.

The lack of correlation between training hours and sub-token discovery suggests that low-resource languages are not inherently disadvantaged in terms of sub-token diversity during inference. While such languages may still exhibit higher error rates due to limited acoustic-phonetic exposure, their sub-token hypothesis spaces are comparably broad, indicating that the model has learned a rich set of candidate representations even from minimal data. This finding challenges deficit-oriented narratives in multilingual NLP: rather than viewing low-resource languages as categorically impoverished, our analysis reveals that they activate multilingual models' vocabularies as robustly as high-resource languages. The disparities that do exist—such as segmentation granularity and error rates—may stem more from tokenization design choices and acoustic variability than from fundamental limitations in the models' learned capacities.

Limitations

Although we tried to present a comprehensive analysis, this study has several limitations. First, the initial probing experiments focused primarily on languages using the Latin script, excluding many writing systems and potentially missing script-specific tokenization effects. We also omitted languages with extremely high error rates (WER $> 60\%$ for the 20-language analysis; CER $> 30\%$ for the 49-language analysis), where Whisper frequently misidentified the language altogether, making sub-token analysis unreliable. Additionally, our 10-minute samples per language in the initial phase provide only a snapshot that may not capture the full phonological diversity or domain variation present in natural speech.

Our methodology for extracting and evaluating sub-token performance relies on assumptions about alignment between hypothesis and reference transcriptions that may not perfectly represent ground truth. While we analyze the chosen beam search path, we do not

directly probe Whisper's internal beam search heuristics or alternate paths, which might offer additional insights into model decision-making. Our resource-level categorization is based solely on Whisper's training hours rather than real-world language resources.

Furthermore, our analysis is inherently dependent on the quality and accuracy of the labels provided within the Whisper training dataset. As noted by Radford et al. (2023) themselves, there can be instances of labeling errors within this large dataset (e.g., some English audio being mislabeled as Welsh). Similar mislabelings for other languages could exist and potentially influence the model's learned representations.

All experiments are conducted using a single model, Whisper large-v2, with BPE tokenization from a shared multilingual vocabulary. While we believe the methodology is general, the specific saturation dynamics, distributional parameters, and script-level patterns reported here may not directly generalize to models with different architectures, vocabulary sizes, or tokenization strategies. The expanded analysis is based entirely on read speech from Common Voice, which limits applicability to spontaneous, conversational, or code-switched speech scenarios.

Future work should extend this analysis to more scripts, longer and more diverse audio samples, and explore how sub-token behavior correlates with specific linguistic features across typologically diverse languages. Replicating the saturation analysis with other multilingual ASR systems such as MMS (Pratap et al., 2024) would strengthen the generalizability of these findings.

Postlude

This chapter addressed **RQ2**: do sub-token decoding metrics reveal systematic differences in how a large multilingual ASR model processes languages across training-data tiers, do those differences align with typological structure or resource level, and are they reducible to training data volume?

They align with both typology and resource level—but asymmetrically. Higher-resource languages are decoded through typologically differentiated internal regimes; lower-resource languages converge into a single generic regime defined by low confidence, high entropy, and typologically disorganized token usage regardless of linguistic family. Critically, this

representational disadvantage is not reducible to insufficient data. Sub-token utilization saturates after roughly two hours of audio regardless of resource tier, and the distributional parameters that vary across languages track script and typology rather than pre-training data volume. The structure of multilingual tokenizers and the design of the shared decoding hypothesis space—choices that more data collection does not change—shape the inequity.

These are technical failures, but they have real human consequences. Chapter 4 documents those consequences directly, from the perspective of the people who live with them.

Chapter 4

USER EXPERIENCE OF ASR BIAS

Overview

This chapter addresses **RQ3**: how do speakers of non-mainstream U.S. dialect varieties describe their experiences with ASR systems, and what forms of invisible labor, emotional harm, and technology abandonment does accuracy-centric evaluation fail to capture? It is based on Liang and Wassink (2026), to be presented at the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2026). A mixed-methods survey of 215 speakers across four U.S. dialect communities documents the invisible accommodation labor, internalized self-blame, and high rates of technology abandonment that no fairness benchmark records. This chapter grows out of the SocioLinx project on bias in ASR led by Alicia Beckford Wassink, who was responsible for conceptualization, study design, and fieldwork; all project members contributed to annotation and data collection. The author's contribution was the primary writing and synthesis of the analysis.

4.1 The Missing Side of ASR Bias Evaluation

Automatic Speech Recognition (ASR) systems have become deeply embedded in everyday technologies, from smartphones and smart speakers to customer service systems and accessibility tools. These systems promise hands-free convenience and natural interaction, yet mounting evidence demonstrates substantial performance disparities across dialects and accents. For speakers of non-mainstream language varieties, technical failures occur systematically and frequently, carrying consequences that extend far beyond mere inconvenience.

Prior research has documented these performance gaps extensively. Commercial ASR systems exhibit higher word error rates for certain groups of English speakers such as African American English speakers, Spanish-influenced English speakers, and Indigenous language speakers compared to speakers of dominant English varieties (Koenecke et al., 2020; Wassink

et al., 2022; Martin and Tang, 2020). These disparities stem primarily from training data imbalances and linguistic features characteristic of non-dominant varieties that systems fail to accommodate (Tatman, 2017; Feng et al., 2024; Ngueajio and Washington, 2022). These studies have established that bias exists and quantified its magnitude through error metrics, phonetic analyses, and controlled evaluations. These performance disparities reflect broader patterns of algorithmic discrimination documented across AI systems (Noble, 2018; Benjamin, 2019).

Yet while the documentation of ASR bias continues to accumulate, significantly less attention has been paid to the human side of system bias. How do system failures shape users' lived experiences? How do users feel about and react to repeated misrecognition? What emotional toll do these failures exact over time? How do experiences vary across different dialect communities facing similar technical failures? And critically, how do these experiences shape users' relationships with voice technologies and their willingness to contribute to improving biased systems?

This study addresses these questions through a mixed-methods investigation of user experiences across four U.S. dialect communities: African American speakers in Atlanta, Georgia; Gulf Coast Creole and French speakers in Mississippi and Louisiana; Hispanic and Latine Caribbean speakers in Miami Beach, Florida, and Native American speakers in Tucson, Arizona. Through quantitative analysis of structured questionnaires and qualitative analysis of open-ended narrative responses, we move beyond performance metrics to examine the lived experiences and emotional impacts of algorithmic bias: the frustrations users articulate, the accommodations they make, their understandings of why systems fail them, and the complex decisions they face about whether to persist with, abandon, or contribute towards improving technologies that exclude them.

Our findings reveal that participants maintain high expectations for ASR performance and express overwhelming willingness to contribute to model improvement, yet most report that technologies fail to consider their cultural backgrounds and require constant adjustment to achieve basic functionality. Participants report frustration, annoyance, and feelings of inadequacy, yet the emotional impact extends beyond these momentary reactions. Qualitative analysis exposes deeper psychological costs: participants recognize that systems were

not designed for them ("this wasn't made for me"), yet internalize failures as personal inadequacy despite this critical awareness. They perform extensive invisible labor, including code-switching, hyper-articulation, and emotional management, to make failing systems functional. Their linguistic and cultural knowledge remains unrecognized by technologies that encode particular varieties as standard while rendering others marginal. This tension between understanding exclusion and continuing to accommodate failing systems illuminates the complex situation of users navigating technologies that have become increasingly unavoidable yet remain systematically unreliable for their speech.

This research contributes to fairness and accountability scholarship by recentering user experience and emotional impact in evaluations of algorithmic bias. We demonstrate that fairness assessments based on accuracy metrics alone miss critical dimensions of harm: the emotional labor of managing repeated technological rejection, the cognitive burden of constant self-monitoring, and the psychological toll of feeling inadequate in one's native language. These invisible costs are central to experiences of algorithmic exclusion yet remain uncaptured in technical evaluations. We document how users interpret and respond to systematic technological failure, revealing sophisticated critical awareness alongside internalized self-blame, and exposing forms of harm that performance metrics fail to capture. Understanding ASR bias requires attending not just to error rates but to how bias is experienced and felt by affected users.

4.2 Previous Work on Performance Gaps and Lived Experience

Our work sits at the intersection of ASR bias analysis, user experience research on algorithmic harm, sociolinguistics, and critical frameworks for understanding linguistic justice in AI systems. We review three strands of research central to our research questions: (1) documented performance disparities and their causes, (2) user experiences of algorithmic bias and technological invisibility, and (3) participatory responses to biased systems and design justice.

4.2.1 ASR Performance Disparities Across Dialects

Studies have consistently documented performance disparities in commercial ASR systems across racial, ethnic, and dialectal lines. Koenecke et al. (2020) found that major commercial systems from Amazon, Apple, Google, IBM, and Microsoft exhibited higher word error rates for African American speakers compared to white speakers, with disparities particularly pronounced for speakers of African American English (AAE). These disparities stem primarily from training data imbalances, with ASR models predominantly trained on Standard American English from speakers who are disproportionately white (Dorn, 2019).

Recent work has expanded evaluation to multiple dialect varieties. Wassink et al. (2022) analyzed ASR errors across four ethnicity-related sociolects from the Pacific Northwest, demonstrating differential error rates: Caucasian American English exhibited lowest error rates, followed by African American English, Yakama English, and ChicanX English. Harris et al. (2024) evaluated ASR performance across four U.S. dialects (Standard American English, AAE, ChicanX English, and Spanglish), finding intersectional gender-dialect biases where systems performed worst for female speakers of minority dialects. Martin and Tang (2020) showed how ASR systems fail to recognize the habitual "be" construction characteristic of AAE, a morphosyntactic feature with systematic grammatical function that systems misinterpret or ignore. Choe et al. (2022) examined how speakers' native language phonologies predict specific categories of ASR errors across World Englishes. Approaches have also focused on different ways to develop systematic methods for measuring and analyzing ASR bias. For example, Feng et al. (2021) analyzed specific phonemes to identify sources of bias, while Liu et al. (2021) developed fairness measurement frameworks using demographically diverse datasets. However, critical analysis of this research literature itself reveals persistent conceptual limitations, including misconceptions about accent as something only certain speakers possess (Prinos et al., 2024).

This body of work establishes that performance disparities are substantial, systematic, and attributable to specific phonological or morphosyntactic features characterizing non-dominant varieties. However, these studies focus primarily on *documenting* failure rates rather than examining *user experiences* of failure or how communities respond when faced

with persistent technological exclusion—the gaps our work addresses.

4.2.2 User Experiences of Algorithmic Bias

Emerging research examines how users experience algorithmic bias emotionally and psychologically. Mengesha et al. (2021) investigated African American users' experiences with ASR errors through interviews and surveys, finding that participants perceived errors as culturally insensitive and felt the technology was not designed for them. Their qualitative work documented frustration, resignation, and perceptions of technological exclusion, providing crucial evidence that ASR bias creates psychological impacts beyond mere inconvenience. However, questions remain about how these feelings of exclusion interact with users' ongoing decisions to use, accommodate, or abandon failing technologies. Recent work has extended this analysis to AI voice services more broadly. Michel et al. (2025) examined accent bias and digital exclusion in synthetic AI voice services (Speechify and ElevenLabs), finding parallel patterns of users feeling unrepresented by technologies that fail to accommodate their speech patterns.

Research on intersectional invisibility in organizational contexts provides a theoretical framework for understanding these experiences. Bhattacharyya and Berdahl (2023) documented how women of color in workplaces experience systematic oversight and misrecognition at the intersection of multiple marginalized identities, identifying four forms of invisibility: erasure (being overlooked), homogenization (being confused with others), whitening (having cultural identity ignored), and exoticization (being treated as exotic other). Critically, these invisibility experiences create distinct affective reactions: ambiguous events (erasure, whitening) elicit shame and self-blame, while less ambiguous events (homogenization, exoticization) elicit anger directed at perpetrators. This dual pathway suggests that even when structural causes are visible, internalized responses may persist. This work builds on broader scholarship documenting how marginalized groups experience systematic misrecognition in institutional contexts, with implications for how users interpret technological failures (Rickford and King, 2016).

We theorize that similar dynamics operate in human-technology interaction. When ASR

systems consistently fail to recognize certain speakers, these failures may create *technological invisibility*—experiences of being unseen, misrecognized, or overlooked by systems. An open question is whether users maintain critical awareness that systems were not designed for them while simultaneously internalizing failure as personal inadequacy, and how these contradictory interpretations coexist in everyday technological practice.

4.2.3 Technological Labor and Participatory Tensions

When speakers of non-mainstream varieties must accommodate dominant norms to participate in technological systems, this accommodation constitutes a form of invisible labor: cognitive effort to modify natural speech, emotional management when systems fail, and temporal costs of repetition and troubleshooting (Mengesha et al., 2021). We draw upon the long history of accommodation research conducted in sociolinguistics and social psychology. Accommodation refers to the conscious or unconscious alterations a talker makes to their linguistic or paralinguistic presentation, either in response to interpersonal factors (such as the desire to enhance perceived social proximity or social attractiveness), in response to a change in topic or interlocutor, or of a desire to control the persona that they project (e.g., to convey humor or playfulness)(Giles and Powesland, 1975; Bell, 1984; Eckert, 2000; Coupland, 2002). Collectively, these are factors influencing style-shifting. Crucially for our purposes, there is neuro-linguistic research indicating that, particularly in diglossic communities where the use of different is kept separate (for example, by social norms delegating each dialect to a different social domain), bidialectal speakers may incur cognitive switching costs that monolinguals do not, reflecting the need to inhibit the activation of the variety that is inappropriate for a given setting (Kirk et al., 2022; Vorwerg et al., 2019). We use the term *invisible labor* to encompass three interrelated dimensions. First, drawing on Hochschild's (2012) concept of *emotional labor*—the work of managing one's feelings to fulfill external expectations—we recognize that ASR users must suppress frustration, self-blame, and alienation to continue engaging with failing systems. Second, extending this concept through Baker-Bell's (2020) framework of internalized linguistic racism, we identify a specifically *linguistic* dimension: the cognitive and emotional work of suppress-

ing one's natural speech variety to conform to a system's narrow expectations, what we term *linguistic labor*. Third, following Cunningham et al. (2024), we attend to the practical accommodation strategies—repetition, code-switching, hyper-articulation—that consume time and attention. This labor remains invisible to developers and privileged-dialect users for whom ASR "just works," yet represents a significant and inequitable burden for those whose speech falls outside training distributions. Pospisil et al. (2022) find that about 76 percent of those who find the labor and concerns to be excessive will never return to the use of voice assistants. Our study operationalizes this concept through systematic documentation of coping strategies, examining what accommodations users report making and at what frequency across different dialect communities. The accommodation burden placed on speakers of non-dominant varieties reflects broader patterns of linguistic discrimination, wherein certain ways of speaking are systematically devalued while others are treated as neutral or standard (Lippi-Green, 2012).

Design justice frameworks argue that equitable technologies require centering marginalized communities throughout design processes (Costanza-Chock, 2020). Costanza-Chock articulates principles emphasizing that those most affected by design decisions should lead design processes, that communities hold expertise about their own experiences, and that design should dismantle rather than reproduce structural inequalities. Applied to ASR, these principles suggest that developing dialect-inclusive systems requires sustained partnership with speakers of underrepresented varieties, not merely collecting their voice data.

Yet tensions emerge when communities experiencing bias are asked to contribute to fixing biased systems. Harrington et al. (2019) examine community-based collaborative design with Black and Brown communities, identifying tensions between researcher and community goals and emphasizing the importance of compensation, shared decision-making authority, and recognition of community expertise. When voice sample collection occurs without these protections, questions arise about equitable participation in AI development. Recent work has begun addressing these tensions in ASR contexts specifically. Hutchinson et al. (2025) document participatory approaches to developing speech technologies for Australian Aboriginal English, highlighting both opportunities and risks when developing ASR for marginalized language communities. A critical question emerges: if users recognize that

ASR was not designed for them and experience repeated failures, what motivates continued willingness to participate in system improvement? This question sits at the intersection of critical awareness and ongoing technological engagement. Our study examines whether communities experiencing significant harm nevertheless remain willing to contribute to model improvement, and how this willingness relates to their experiences of exclusion and their understanding of why systems fail them.

4.2.4 Positioning This Work

This work reframes the study of ASR bias by shifting attention from system-centric performance metrics to the social and experiential dimensions of human–technology interaction. Rather than asking only how recognition accuracy varies across dialects, we ask how persistent misrecognition is understood, navigated, and emotionally processed by users whose speech falls outside dominant training distributions. By situating ASR failures within users' everyday practices, expectations, and interpretive frameworks, this study treats bias not solely as a technical property of models, but as a lived condition shaped by design choices, institutional norms, and unequal distributions of accommodation labor. In doing so, we bridge technical research on ASR disparities with human-centered and justice-oriented approaches to algorithmic accountability, emphasizing the importance of user experience as a central site for evaluating fairness in voice technologies.

4.3 Study Design

4.3.1 Research Design

Our study investigates user experiences with English Automatic Speech Recognition (ASR) technologies across four geographically and linguistically diverse communities in the United States. Our methods combine structured questionnaires with open-ended responses to capture both quantitative patterns and qualitative experiences of ASR system failures and their psycho-emotional impacts.

Data collection occurred between July, 2023 and August, 2025 for four sites: Atlanta, Georgia; the Gulf Coast region (Mississippi and Louisiana); Miami Beach, Florida; and

Tucson, Arizona. These communities represent systematically underrepresented varieties with well-documented linguistic structures distinct from Standard American English (Green, 2002; Blodgett et al., 2016; Fought, 2003; Leap, 1993; Klingler, 2003).

4.3.2 Participants

We recruited a total of 494 participants across the four sites; of these, 215 of them met our inclusion criteria (target linguistic and demographic background) and provided near-complete survey responses. These 215 participants form the basis of this study. All participants self-identified as speakers of underrepresented dialects in the United States and reported regular use of voice-enabled technologies. Eight members of the research team (2-6 per location) conducted participant recruitment. Recruitment strategies varied by site to reach distinct dialect communities: Atlanta participants (n=86) were recruited at local community centers and a conference of black engineers; Gulf Coast participants (n=61) were recruited from the community by word-of-mouth; Miami Beach participants (n=43) were recruited from the student and staff body at Florida International University; and Tucson participants (n=25) were recruited through the Pascua Yaqui Cultural Center and the intertribal Native American Student Affairs center at the University of Arizona.

Table 4.1 presents an overview of the demographic composition of the participants. The Atlanta sample (n=86) consisted primarily of Black/African American participants from the Southeast. The Gulf Coast sample (n=61) included Black/African American participants and French Creole speakers from Mississippi and Louisiana. The Miami Beach sample (n=43) included Caribbean Hispanic/Latine participants, Black Caribbean, and Black/African American participants from Florida. The Tucson sample (n=25) consisted primarily of Native American participants from Arizona, including speakers of Yaqui, Navajo, Tohono O'odham, and other Indigenous backgrounds.

Table 4.1: Participant Demographics by Research Site

Characteristic	**Atlanta** (n=86)	**Gulf Coast** (n=61)	**Miami Beach** (n=43)	**Tucson** (n=25)
Gender Identity				
Women	48 (56%)	20 (33%)	31 (72%)	19 (76%)
Men	37 (43%)	38 (62%)	11 (26%)	6 (24%)
Non-binary	1 (1%)	1 (2%)	0 (0%)	0 (0%)
Trans woman	0 (0%)	1 (2%)	0 (0%)	0 (0%)
Prefer not to state	0 (0%)	1 (2%)	1 (2%)	0 (0%)
Age (years)				
Range	19–60	18–68	18–63	18–54
Mean (SD)	29.9 (8.5)	33.6 (12.1)	23.4 (11.2)	25.8 (10.2)

4.3.3 Data Collection

Questionnaire Design

We developed a questionnaire investigating participants' experiences with voice-enabled technologies, their perceptions of ASR system performance, and the psycho-emotional impacts of recognition failures. The questionnaire included both structured multiple-choice items and open-ended narrative prompts.

Key questionnaire topics included:

- Demographic information (age, location, ethnicity, gender identity)
- Self-description of speech characteristics (Q1) and speech challenges (Q4)
- Types of voice-enabled technologies used (Q5–Q7) and purposes of use (Q6a, Q8a)
- Frequency of use for different purposes (Q12a–f: 5-point Likert scale)

- Discontinued use and reasons (Q10: open-ended)
- Memorable experiences with ASR technologies (Q11: open-ended)
- Specific instances of dialect or accent-related recognition failures (Q13: open-ended)
- Types of challenges experienced (Q14: multiple-choice with options including cultural/ethnic/racial exclusion, contextual misunderstanding, physical effort requirements, self-consciousness)
- Coping strategies employed when systems fail (Q15a: multiple-choice with options including repetition, speech modification, enunciation, speaking louder/slower, imitation, manual task completion, asking others for help, technology abandonment)
- Emotional responses to system failures (Q16: open-ended)
- Expectations for ASR accuracy (Q17: 5-point scale from "Always [100%]" to "Never [0–20%]")
- Willingness to provide voice samples for system improvement (Q18a: Yes/Maybe/No)
- Willingness to pay for improved services (Q19: multiple payment tier options including compensation preference)

Questionnaire items were developed using a two-part process. Prompts originally prepared by the second author were reviewed and updated in collaboration with a UX researcher (a member of the team) with experience using culturally-informed community-based participatory research design frameworks for fairness in technology (CBPR) (Harrington et al., 2019). The questionnaire was then brought to the Seattle Black Community Advisory Board (B-CAB), who provided further feedback on item wording, overall content, and impact of the questionnaire as well as the alignment of the project with the B-CAB's own goals to advance technological fairness in the local community. B-CAB feedback was incorporated into the questionnaire.

The full questionnaire can be found in the appendix. The Atlanta collection was the first to occur. Afterward, the interview and survey were modified slightly to increase the likelihood that question wording would elicit the information of interest. We will mention below where the Atlanta questionnaire differed slightly from the other locations' surveys.

Data Collection Procedures

Study procedures were approved by the University of Washington Human Subjects Division (Review#: 00019648). Two procedures were used for the collection of data. First, a semi-structured interview was delivered in-person and audio-recorded. Following consenting procedures, respondents were invited to respond verbally to an interviewer who delivered all questions orally. Second, we offered the choice to all respondents to participate online via a self-paced survey created using the JotForm platform (jotform.com). The survey appeared only after online consent was given. The survey included radio-button, multiple-choice, short-form text entry and long-form text entry as well as audio widget response options, as appropriate to the question. All participants, regardless of collection method, responded to the same questionnaire, enabling cross-site comparison for both quantitative and qualitative analyses.

4.3.4 Analytical Approach

Quantitative Analysis

We report descriptive statistics (frequencies, percentages) for all structured questionnaire items to characterize response patterns within and across sites. For multiple-selection items (Q14: challenges experienced; Q15a: coping strategies employed), participants could select all applicable options, resulting in non-mutually exclusive categories. Percentages therefore reflect the proportion of participants who endorsed each option and may sum to greater than 100%. Comparisons across sites are reported descriptively given the non-probability sampling approach and the distinct demographic characteristics of each dialect community. Where notable differences emerged between sites, we interpret these in light of the specific linguistic and cultural contexts of each community.

Qualitative Analysis

We conducted qualitative analysis of open-ended responses to three key questions: memorable experiences with ASR technologies (Q11), specific instances of dialect-related recognition challenges (Q13), and emotional responses to system failures (Q16). Several major themes emerged from this analysis, representing distinct dimensions of how participants experience and respond to ASR bias. These themes capture both surface-level impacts (frustration with errors, need for repetition) and deeper psychological dimensions (internalized blame despite critical awareness of exclusion, feelings of linguistic and cultural invisibility).

The qualitative analysis approach taken was an adaptation of thematic content analysis as established within Grounded Theory (Glaser and Strauss, 1967). Audio-recorded responses to survey questions were automatically transcribed, then human-corrected. Corrected transcripts were next subjected to iterative coding passes to enable highlighting of common themes arising across the dialect databases and themes related to the research questions of the study. In the first pass, two team members per dialect coded utterances using a list of 14 generalized tags defined in a conversational coding guide designed to address the research questions of the study. Inter-coder reliability was assessed using Krippendorff's alpha ($\alpha \geq 0.80$). Several steps were made to improve inter-coder agreement. Weekly team meetings were used to resolve ambiguous cases, discuss code definitions, highlight interesting responses and only if necessary, update the codebook with parsimony. Coders then re-coded their materials, if the threshhold for agreement was not yet achieved. In this paper, we focus on responses coded as 'feelings' (FEL), 'presuppositions' (PSP), 'narratives' (NARR), and 'personifications' (PERS).

Participant quotes are identified by site and unique participant ID (e.g., [Atlanta] CD24-055) to enable verification while protecting anonymity. When quotes contained implicit identifying information, minor details were altered to preserve confidentiality without changing the meaning substantively.

4.4 Quantitative and Qualitative Findings of ASR Bias Harm

We present findings below from our investigation of user experiences with ASR technologies. Since some participants could choose to skip some questions, response rates varied by question. Therefore, sample sizes differ slightly across questions and are reported accordingly in the respective tables.

4.4.1 Quantitative Findings

Types of Challenges Experienced

Participants identified multiple types of challenges when using ASR technologies (Table 4.2). Cultural and ethnic exclusion was reported by a majority of participants overall (53.5%), with notably higher rates among Tucson participants (72.0%) and Miami Beach participants (62.8%) compared to Gulf Coast (52.5%) and Atlanta (44.2%) participants.

Contextual misunderstanding—voice assistants taking speech out of context or performing unintended actions—was reported by 51.2% of participants overall, with 76.0% of Tucson participants and 69.8% of Miami Beach participants reporting the issue, compared to 47.5% of Gulf Coast and 37.2% of Atlanta participants.

Physical effort requirements showed a similar pattern, with Tucson (60.0%) and Miami Beach (39.5%) participants reporting higher rates than Gulf Coast (37.7%) and Atlanta (27.9%) participants. Self-consciousness about speech when interacting with voice technologies was reported by 44.0% of Tucson participants, compared to 36.1% of Gulf Coast, 25.6% of Miami Beach, and 17.4% of Atlanta participants.

Coping Strategies

When ASR systems failed to recognize their speech, participants reported employing a variety of coping strategies (Table 4.3). Repetition was the most common strategy overall (82.8%), with consistently high rates across all sites ranging from 75.4% in the Gulf Coast to 87.2% in Atlanta.

Speech modification was a prevalent coping mechanism. In Atlanta, where participants were offered a more open-ended response option, 58.1% reported changing the way they

Table 4.2: Types of Challenges Experienced by Site (Q14; Multiple Selections Allowed)

	Atlanta (n=86)		Gulf Coast (n=61)		Miami Beach (n=43)		Tucson (n=25)		Total (N=215)	
Challenge Type	n	%	n	%	n	%	n	%	n	%
Cultural/ethnic/racial exclusion	38	44.2	32	52.5	27	62.8	18	72.0	115	53.5
Contextual misunderstanding	32	37.2	29	47.5	30	69.8	19	76.0	110	51.2
Physical effort required	24	27.9	23	37.7	17	39.5	15	60.0	79	36.7
Self-conscious about speech	15	17.4	22	36.1	11	25.6	11	44.0	59	27.4

speak or code-switching. At other sites, where more granular options were available, participants reported slowing down speech (Gulf Coast 68.9%, Miami Beach 72.1%, Tucson 76.0%), speaking louder (Gulf Coast 59.0%, Miami Beach 83.7%, Tucson 88.0%), enunciating more (Gulf Coast 47.5%, Miami Beach 83.7%, Tucson 84.0%), and changing words or phrases (Gulf Coast 54.1%, Miami Beach 60.5%, Tucson 72.0%).

Technology abandonment strategies were also prevalent. Complete discontinuation of ASR use was reported by 49.8% of participants overall, with Tucson showing the highest rate (80.0%), followed by Miami Beach (55.8%), Atlanta (47.7%), and Gulf Coast (36.1%). Manual task completion was common in the Gulf Coast (59.0%), Miami Beach (76.7%), and Tucson (80.0%). Social workarounds included imitating the voice assistant's speech patterns (26.5% overall) and asking another person to operate the device (21.9% overall).

Expectations for ASR Performance

Most participants expected ASR systems to understand their speech correctly at least 75% of the time (Table 4.4). Across all sites, 53.7% of participants expected ASR to work "often" (75% of the time), and 27.6% expected it to work "always" (100% of the time). Only 15.0% expected ASR to work only "sometimes" (50%), and 3.7% expected it to work "rarely" (25%) or never. These high expectations—with over 80% anticipating at least 75% accuracy—contrast sharply with the challenges and coping strategies reported above,

Table 4.3: Coping Strategies Employed by Site (Q15a; Multiple Selections Allowed)

	Atlanta (n=86)		Gulf Coast (n=61)		Miami Beach (n=43)		Tucson (n=25)		Total (N=215)	
Strategy	n	%	n	%	n	%	n	%	n	%
Repeat utterance	75	87.2	46	75.4	36	83.7	21	84.0	178	82.8
Code-switch/modify speech[a]	50	58.1	–	–	–	–	–	–	–	–
Change words/phrases[b]	–	–	33	54.1	26	60.5	18	72.0	77	59.7
Slow down speech[b]	–	–	42	68.9	31	72.1	19	76.0	92	71.3
Speak louder[b]	–	–	36	59.0	36	83.7	22	88.0	94	72.9
Enunciate more[b]	–	–	29	47.5	36	83.7	21	84.0	86	66.7
Stop using technology	41	47.7	22	36.1	24	55.8	20	80.0	107	49.8
Complete task manually[b]	–	–	36	59.0	33	76.7	20	80.0	89	69.0
Imitate voice assistant	14	16.3	16	26.2	20	46.5	7	28.0	57	26.5
Ask another person	9	10.5	13	21.3	16	37.2	9	36.0	47	21.9

[a]Atlanta used an earlier survey version with this broader category instead of granular speech modification options.

[b]Option not available for Atlanta; percentage calculated from Gulf Coast, Miami Beach, and Tucson only (n=129).

suggesting a substantial gap between user expectations and actual ASR performance for these dialect communities.

Willingness to Contribute Voice Samples

Despite experiencing significant challenges with ASR systems, a substantial majority of participants (75.5%) expressed willingness to contribute voice samples to help improve recognition accuracy for diverse dialects (Table 4.5). An additional 17.9% indicated they might contribute, while only 6.6% declined. Willingness was highest in Miami Beach (88.1%) and Gulf Coast (85.2%), followed by Tucson (76.0%) and Atlanta (61.9%).

Table 4.4: Expectations for ASR Accuracy by Site (Q17)

	Atlanta (n=85)		Gulf Coast (n=61)		Miami Beach (n=43)		Tucson (n=25)		Total (N=214)	
Expected Accuracy	n	%	n	%	n	%	n	%	n	%
Always (100%)	28	32.9	15	24.6	9	20.9	7	28.0	59	27.6
Often (75%)	43	50.6	32	52.5	28	65.1	12	48.0	115	53.7
Sometimes (50%)	12	14.1	11	18.0	4	9.3	5	20.0	32	15.0
Rarely (25%)	2	2.4	3	4.9	1	2.3	1	4.0	7	3.3
Never (0–20%)	0	0.0	0	0.0	1	2.3	0	0.0	1	0.5

Willingness to Pay for Improved Services

A majority of participants (63.3%) expressed willingness to pay for improved ASR services that better recognize diverse dialects (Table 4.6). Gulf Coast participants showed the highest willingness (81.5%), followed by Atlanta (58.8%), Tucson (56.0%), and Miami Beach (53.5%). Among those willing to pay, preferences also varied by site.

4.4.2 Qualitative Findings: Lived Experiences Across Four Communities

Complementing the quantitative results, we present analysis of open-ended narratives which revealed recurring patterns in how users interpret, respond to, and make sense of ASR failures. Rather than simply expressing frustration, participants' accounts exposed the sophisticated strategies they employ to navigate exclusionary systems, the emotional labor required to maintain relationships with failing technologies, and the critical awareness they hold about why these systems fail them.

Recognition of Exclusionary Design

Across all sites, participants articulated explicit awareness that voice technologies were not designed with their communities in mind. Whereas the causal and technical dimen-

Table 4.5: Willingness to Provide Voice Samples by Site (Q18a)

	Atlanta (n=84)		Gulf Coast (n=61)		Miami Beach (n=42)		Tucson (n=25)		Total (N=212)	
Response	n	%	n	%	n	%	n	%	n	%
Yes	52	61.9	52	85.2	37	88.1	19	76.0	160	75.5
Maybe	26	31.0	6	9.8	2	4.8	4	16.0	38	17.9
No	6	7.1	3	4.9	3	7.1	2	8.0	14	6.6

sions of this awareness are examined in §4.4.2, we focus here on how exclusion is experienced phenomenologically—as affect, identity disruption, and lived certainty derived from repeated failures. A Navajo speaker from Tucson captured this sentiment directly: "Because Navajo is my first language ... I get this automatic feeling of inferiority if that makes sense. Because ... in my mind I'm thinking this is not something that was made for me automatically. And I'm stepping into a world of what other people are doing and not me because I'm native." (Tucson CD24-606). This participant's elaboration reveals how exclusion becomes internalized through accumulated interactions with systems that consistently fail to recognize one's voice.

A Gulf Coast participant made the systemic nature of this exclusion explicit: "It's gonna be for everybody and that it needs to be based on everybody's speech and everybody's culture. But I don't take stuff that deep. You know I just look at stuff different, like I know that places weren't created for everyone. Yeah most things aren't inclusive" (Gulf Coast CD24-546). This statement reveals a paradox: the participant demonstrates sophisticated critical analysis—recognizing that technologies marketed as universal exclude their community—while simultaneously describing this exclusion as unremarkable, something not worth dwelling on because it reflects a broader pattern of non-inclusive design.

An Atlanta participant made the racial dimensions of this exclusion explicit: "It is very

Table 4.6: Willingness to Pay for Improved ASR Services by Site (Q19)

	Atlanta (n=85)		Gulf Coast (n=54)		Miami Beach (n=43)		Tucson (n=25)		Total (N=207)	
Response	n	%	n	%	n	%	n	%	n	%
Prefer 1-time payment	26	30.6	10	18.5	13	30.2	5	20.0	54	26.1
<$10/month	15	17.6	6	11.1	5	11.6	6	24.0	32	15.5
$10–20/month	4	4.7	13	24.1	4	9.3	2	8.0	23	11.1
$20–30/month	3	3.5	12	22.2	1	2.3	1	4.0	17	8.2
>$30/month	2	2.4	3	5.6	0	0.0	0	0.0	5	2.4
Not willing to pay	35	41.2	10	18.5	20	46.5	11	44.0	76	36.7

frustrating because for technologies designed to make life better don't work for bi-lingual people of color. Whereas other non-[people of color] around me have no trouble with the technology" (Atlanta CD24-188). This comparative observation highlights how exclusion becomes visible through contrast with others' seamless experiences.

These statements demonstrate that users do not simply experience ASR failures as random technical glitches. Rather, they understand these failures as evidence of systematic exclusion from the design process, reflecting broader patterns of whose voices, languages, and ways of speaking are centered in technological development.

Internalized Blame and Emotional Burden

Despite recognizing systemic exclusion, many participants described internalizing ASR failures as personal inadequacy. This pattern was particularly striking given participants' simultaneous awareness that the problem was not theirs to solve. A Tucson participant articulated this tension: "It's my problem that this thing's not working. Never. My thought is never like there's something wrong with this you know. Like this needs to be fixed or this is not perfect you know. That thought never entered enters my mind. I'm always like what

am I doing wrong. How can I say it to make this thing work you know. And so I think that that's the part that gets most people. Yeah it's kind of like what did I do wrong when it's really not your fault" (Tucson CD24-606).

This participant's account reveals the psychological process of misplaced responsibility: despite knowing that the system is flawed, the immediate emotional response is self-blame. The phrase "what am I doing wrong when it's really not your fault" captures the gap between the critical understanding and the unavoidable emotional experience.

Another Tucson participant described feelings of isolation: "It kinda makes you frustrated actually. ... Like I said, most people don't have problem with it, you know. Like why am I the only one struggling with this" (Tucson CD24-611). This statement illustrates how ASR failures create the illusion of individual inadequacy in addition to collective exclusion, particularly when failures are not publicly visible or discussed.

The most succinct expression of this emotional burden came from an Atlanta participant: "I feel inadequate speaking my own first language" (Atlanta CD24-198). This statement captures a profound form of linguistic alienation—the internalized sense that one's native language variety, the language of intimate relationships and cultural identity, is somehow deficient because it fails to conform to the narrow linguistic models encoded in voice recognition systems.

Strategic Accommodation to System Failure

Participants across all sites described extensive behavioral modifications required to achieve even basic functionality with voice technologies. These adaptations ranged from phonetic adjustments to code-switching, representing invisible labor that users perform to accommodate systems that do not accommodate them.

A Gulf Coast participant described the mental and linguistic accommodation required: "I had to adjust my pronunciation, speaking more slowly and deliberately, almost like I was 'talking to a foreigner.' It worked, but it felt awkward, like I was compromising my natural way of speaking to accommodate the technology" (Gulf Coast CD24-425). The comparison to "talking to a foreigner" reverses the expected relationship—rather than the technology

adapting to diverse users, users must treat the technology as a non-native speaker requiring careful and modified input.

Another Gulf Coast participant made the racialized nature of this accommodation explicit: "Yeah, I put on my white voice when I use it. Like I put on my voice when I talk to Siri, like when I want it to be immediate" (Gulf Coast CD24-546). This participant's casual reference to deploying a "white voice" reveals both the normalization of code-switching as necessary for technological interaction and the explicit awareness that the system recognizes certain racialized varieties of English while failing to recognize others.

An Atlanta participant expressed the desire for authenticity denied by these constant accommodations: "Every time I use my natural dialect I have problems. I have learned to adjust my dialect for voice recognition systems. I wish I could just speak normally" (Atlanta CD24-055). The phrase "speak normally" highlights how technological failures redefine speakers' natural, authentic speech as abnormal or problematic.

A Tucson participant described the cognitive burden of constant self-monitoring: "I find myself thinking a lot more about how I'm talking than I would in a daily day to day life" (Tucson CD24-629). This statement reveals how voice technologies introduce extra cognitive load, forcing users to monitor and adjust their speech in real-time rather than focusing on communicative goals.

Participants' accounts suggest that what technical evaluations measure as "minor performance degradation" translates, from users' perspectives, into substantial ongoing labor, requiring constant adjustment of linguistic behavior to achieve functionality that other users access seamlessly.

Cultural and Linguistic Exclusion

Participants described specific ways their linguistic and cultural knowledge remains unrecognized by voice technologies, requiring either abandonment of these resources or creation of workarounds. A participant of Guyanese heritage in Miami Beach proposed a solution: "We need a Caribbean Alexa... Because you know in the Caribbean we have our own Creole. And so an Alexa Caribbean is going to know the Caribbean Creole because if you ask Alexa

something that is Caribbean based she doesn't have that culture nor ethnicity that she can relate to" (Miami Beach CD24-569). This participant's proposal anthropomorphizes the problem effectively—the system lacks not just linguistic data but cultural knowledge and context necessary for genuine comprehension.

A Miami Beach participant described how accent becomes marked as non-normative: "When I call someone I have to use non-Miami accent, normal people talk" (Miami Beach CD24-600). The equation of "non-Miami accent" with "normal people talk" reveals how ASR systems encode particular varieties as standard while rendering others abnormal, creating a hierarchy of acceptable speech.

Another Gulf Coast participant identified the technical gap: "Siri doesn't take into account different dialects (ex. AAVE, Ebonics, French accents, slang)" (Gulf Coast CD24-503), naming specific varieties and features excluded from the system's model of language.

A Gulf Coast participant described the specific linguistic features that remain unrecognized: "Voice to text doesn't recognize or know what to do with common slang terms like 'chile.' I have had issues where voice to text doesn't understand when I am trying to play a song that has a title that is slang. If I do not speak super proper Siri doesn't understand me. Most of my names with black cultural roots aren't understood when I ask Siri to call them so Siri will call the wrong person" (Gulf Coast CD24-515). This account catalogs the cumulative exclusions—slang terms, song titles reflecting cultural tastes, personal names—that together render a person's linguistic and social world invisible to the system.

Critical Awareness of Systemic Failures

Beyond the experiential recognition of exclusion described in §4.4.2, many participants moved from describing how exclusion feels to diagnosing why it occurs, locating problems in inadequate training data, narrow linguistic models, or design processes that exclude diverse communities. A Gulf Coast participant analyzed the technical limitation: "I will realize that it is not intelligent enough, and it needs to constantly improve and update the language habits and database" (Gulf Coast CD24-495). This participant identifies insufficient training data as the source of failures rather than accepting user blame.

Some participants contextualized ASR failures within broader technological exclusion. The Gulf Coast participant who noted "It's gonna be for everybody but it's not based on everybody" (Gulf Coast CD24-546, also quoted in §4.4.2) moved beyond describing how exclusion feels to articulating a structural critique: that claimed universality masks particular, exclusionary design choices.

This critical awareness did not necessarily translate into expectations that systems would not change. As one participant noted: "I know that it will get better" (Gulf Coast CD24-537), expressing faith in future improvement that allows acceptance of present failures. Another stated: "it's easy to feel frustrated but sometimes I just have to tell myself to, like you know, relax, you know, it's learning" (Miami Beach CD24-558), framing the persistent failures as temporary growing pains rather than unsolvable problems.

4.5 What Error Rates Cannot Measure

4.5.1 The Expectations-Experience Gap

Over 80% of participants expected ASR systems to understand their speech at least 75% of the time, yet more than half reported that technologies fail to consider their cultural backgrounds. This gap between expectations and experience has significant implications. Marketing narratives that promise "natural" voice interaction may create expectations that systems will accommodate linguistic diversity. When systems fail, users who have internalized these promises may attribute failures to their own speech rather than to system limitations. The participant who described always thinking "what am I doing wrong" despite knowing "it's really not your fault" (Tucson CD24-606) captures this dynamic precisely.

This expectations gap also explains the substantial rates of technology abandonment we observed. When systems consistently fail to meet reasonable expectations, users face a choice between continued accommodation labor and discontinuation. That 80.0% of Tucson participants and 55.8% of Miami Beach participants reported stopping use of ASR technologies suggests that for many users, the costs of accommodation eventually outweigh the benefits of use.

4.5.2 The Paradox of Critical Awareness and Internalized Blame

One of the most significant findings is the coexistence of sophisticated critical awareness with internalized self-blame. Participants articulated clear understanding that systems were "not made for" them, that training data excludes their communities, and that their failures reflect design choices rather than personal deficiencies. Yet this critical awareness did not prevent emotional responses of inadequacy, frustration directed inward, and self-monitoring of speech.

This paradox extends theoretical work on intersectional invisibility (Bhattacharyya and Berdahl, 2023) into human-technology interaction. Bhattacharyya and Berdahl found that ambiguous invisibility experiences (erasure, whitening) elicit shame and self-blame even when structural causes are recognized. Our findings suggest similar dynamics operate with ASR: repeated system failures create cumulative experiences of technological invisibility that generate internalized responses despite intellectual understanding of systemic causes. The participant who described feeling "inadequate speaking my own first language" (Atlanta CD24-198) exemplifies how technological exclusion can create profound linguistic alienation.

4.5.3 Invisible Labor and Unequal Burden

The coping strategies participants reported—repetition (82.8%), slowing speech (71.3%), speaking louder (72.9%), enunciating more (66.7%)—represent substantial invisible labor. This labor is invisible in two senses: it goes unrecognized in technical evaluations that measure only accuracy, and it remains unseen by users for whom systems "just work."

The participant who described deploying a "white voice" when using Siri (Gulf Coast CD24-546) makes explicit what remains implicit in aggregate statistics: from participants' perspectives, accommodation to ASR systems requires performing linguistic identities that align with dominant speech norms while suppressing natural speech patterns. This linguistic accommodation—the invisible labor defined in §4.2.3—represents a form of everyday discrimination in which access to technology is contingent on abandoning authentic self-expression (Lippi-Green, 2012).

Notably, this burden falls unequally even among marginalized communities. Tucson

participants (primarily Native American speakers) reported the highest rates of challenges across all categories and the highest rates of technology abandonment (80.0%), suggesting that speakers of Indigenous languages and varieties who rarely feature in accessibility and fairness scholarship face particularly severe exclusion from voice technologies.

4.5.4 Implications for Participatory System Development

The finding that 75.5% of participants expressed willingness to contribute voice samples despite experiencing significant harm raises important questions for participatory approaches to system development. An Atlanta participant who described feeling "embarrassed, sad, and defeated" by ASR failures—and who reported that the technology "makes me feel like I don't speak well or struggle with basic English"—nevertheless expressed willingness to contribute voice samples, adding: "if y'all reach out later, I would be open to doing it" (Atlanta CD24-163). Yet this same participant insisted "I feel I should be compensated," rejecting the premise that improving a system that has caused harm should come at further personal cost. This response encapsulates a tension that ran through our data: willingness to contribute coexists not with naïve optimism but with clear-eyed recognition that such contributions constitute labor deserving fair exchange.

Participants' motivations for this willingness appeared to vary. Some expressed altruistic investment in future users: the same Atlanta participant stated "I just want the device to work for people like me" (CD24-163), locating motivation in collective benefit rather than personal gain. Others framed continued engagement as pragmatic patience—"I know that it will get better" (Gulf Coast CD24-537)—or active emotional management: "it's easy to feel frustrated but sometimes I just have to tell myself to, like you know, relax, you know, it's learning" (Miami Beach CD24-558). This last statement is particularly revealing: the participant performs emotional labor, in the sense of Hochschild (2012), to sustain engagement with a system that repeatedly fails them, reframing persistent exclusion as a temporary developmental stage.

This willingness to contribute represents an additional layer of invisible labor layered atop the accommodation burden documented in §4.4.2. Communities already bearing invol-

untary costs—code-switching, repetition, self-monitoring—are asked to take on voluntary labor to fix the systems imposing those costs. That 75.5% expressed willingness while a majority (63.3%) were also willing to pay for improved services underscores the depth of investment these communities maintain in voice technologies despite their exclusionary design. These findings support calls for design justice frameworks (Costanza-Chock, 2020) that center affected communities not merely as data sources but as partners with authority over development processes. The willingness our participants expressed should not be treated as a free resource to be extracted; it should be met with equitable compensation, shared decision-making, and accountability for the harms that generated the need for improvement in the first place.

4.5.5 Cross-site Comparison

The communities sampled in this study (Atlanta, Tucson, the Gulf Coast, and Miami Beach) and the ethnicities targeted within these (African American, Native American, Mississippi French/French Creole and Hispanic) were selected for their distinct histories, relationships with other communities in the region, and socio-cultural experiences. We expected that cross-community differences would emerge, and this proved to be true. The African American and Native American communities showed greatest awareness of the burden to code-switch–both for themselves and their communities. These two groups tended to display the highest levels of individual linguistic awareness (e.g., explaining how they made stylistic adjustments away from their home dialect). Ongoing research in our laboratory further explores the possibility of an association between the burden to code-switch and experiences of invisibility. On the other hand, the Miami Beach speakers tended to comment on the experiences of peers rather than make comments that placed them within their broader community.

As was discussed above, all three of these groups discussed discontinuing the use of tech that did not satisfy their expectations. The group of which we had uncertain predictions was the Gulf Coast group, which is the least-studied of the speech communities included in this research. We were interested to find that these participants expressed the most self-

blame. Although recognizing the failure of technology to work consistently for them, they were the group most likely to display acceptance and continued use.

4.6 Summary

This study demonstrates that understanding ASR bias requires moving beyond accuracy metrics to examine lived experiences of technological exclusion. Our findings reveal that critical awareness of systemic bias coexists with feelings of invisibility and internalized blame: participants who clearly articulated that systems were not made for them nevertheless described feeling inadequate speaking their native languages. This paradox, knowing the system is at fault while still blaming oneself, suggests that intellectual understanding provides incomplete protection against the psychological toll of repeated technological rejection.

The accommodation strategies participants reported represent substantial invisible labor that current fairness metrics fail to capture. Code-switching, hyper-articulation, repetition, and constant self-monitoring constitute an unequal tax on users whose speech falls outside narrow training distributions. Access to voice technology is often contingent on suppressing authentic linguistic identity. This burden falls most heavily on already marginalized communities.

Our study contributes to growing recognition that technical fairness metrics alone cannot capture the full dimensions of algorithmic harm. Voice technologies create worlds that include some speakers while excluding others, and the invisible costs of exclusion extend far beyond what error rates can measure. Yet our findings also reveal grounds for optimism: despite experiencing significant harm, participants overwhelmingly expressed willingness to contribute to model improvement and pay for improved services. This finding suggests that at least some in affected communities remain invested in shaping voice technologies rather than abandoning them entirely. Realizing this potential requires moving beyond extraction-based approaches that treat marginalized speakers as data sources toward participatory frameworks that recognize community expertise, ensure equitable compensation, and grant meaningful authority over how systems are developed and deployed. The path toward inclusive voice technologies runs through the communities currently excluded from them.

Postlude

This chapter addressed **RQ3**: how do speakers of non-mainstream U.S. dialect varieties describe their experiences with ASR systems, and what forms of invisible labor, emotional harm, and technology abandonment does accuracy-centric evaluation fail to capture? The findings establish that standard fairness metrics systematically understate the harm they are supposed to measure: the accommodation labor is real and ongoing, the self-blame is real despite structural awareness, and the abandonment is substantial and unrecorded in any training or evaluation dataset.

This chapter established that the people who live with the outputs of these design choices bear real and measurable costs that none of the foregoing metrics capture. What these costs add up to, and what addressing them would require, is the subject of the conclusion.

Chapter 5

CONCLUSION

We have presented three studies examining ASR inequity at different levels of the speech recognition pipeline: training data and adaptation for endangered languages, internal decoder representations and the relationship between training scale and sub-token utilization in a large multilingual model, and the lived experiences of speakers whose speech is systematically misrecognized. In what remains, we summarize the main findings, draw out some general principles, and outline directions for future work.

5.1 What the Studies Found

Chapter 2 asks whether existing multilingual models can be adapted to endangered fieldwork languages with the data that field linguists actually have. Adapter fine-tuning on as little as sixty minutes of transcribed audio produces usable ASR, and this approach outperforms full fine-tuning when training data falls below roughly one hour. Both methods converge at around that threshold. The more interesting finding comes from the phonological error analysis: even at the maximum data condition tested, both models had learned to handle the majority of phonological categories while failing on tone, nasality, and consonant and vowel length. These are usually the features that distinguish the languages studied from the typologically mainstream languages that tend to dominate pre-training corpora. Aggregate character error rates improved substantially, whereas the hardest cases did not as much. This is evidence that the residual failures are representational, not simply a consequence of insufficient data.

Chapter 3 takes up the representational question directly by probing an E2E ASR model's decoder at the sub-token level. Five metrics applied to the beam search path—rank of the correct sub-token, confidence, predictive entropy, alternate-candidate diversity, and usage clustering—across twenty languages point to the finding that lower-resource lan-

guages receive less favorable treatment at multiple measured dimensions of decoding. The clustering analysis further reveals that higher-resource languages form typologically coherent groups in sub-token usage space, while lower-resource languages form a cluster defined not by linguistic similarity but by the model's generic, undifferentiated treatment of languages it has seen little of. This generic decoding regime is invisible to WER. Scaling the analysis to forty-nine languages across diverse scripts then asks whether this representational disadvantage is simply a downstream effect of training data volume. The answer is no. Sub-token vocabulary activation saturates quickly across all languages, typically within two hours of audio, following trajectories that are consistent regardless of resource tier. The parameters that differ across languages after saturation track orthography and typology rather than pre-training hours, demonstrating that the inequity is structural rather than a matter of scale.

Chapter 4 shifts from what the model does internally to what users experience externally. Based on a survey of speakers from underrepresented dialect communities in the U.S., the study documents real and ongoing costs of technology failures that no benchmark records: accommodation labor in the form of code-switching and hyper-articulation; internalized self-blame despite the awareness that the systems were not designed for their communities; and varying rates of technology abandonment. The most striking finding is the coexistence of the critical structural awareness with emotional self-blame. Participants who could articulate exactly why systems fail them still described feeling inadequate speaking their native language. The communities that experience the most severe failures are in danger of being even less included in the data that is supposed to improve the systems because abandonment of the technology could further remove their real voice from the feedback loop.

5.2 Three Principles for Equitable ASR

These findings suggest three general principles for research on multilingual and multidialectal ASR fairness.

First, aggregate error metrics systematically underreport inequity. CER and WER average over error types in ways that obscure failures that could be concentrated in phonologically or typologically distinctive categories. Category-level evaluation that decomposes

errors by phonological feature class, by script, or by meaningful linguistic features reveals failure patterns that matter for practitioners but that aggregate numbers conceal. Routine evaluation should include this kind of decomposition, not treat it as an optional addition.

Second, training data volume and representational equity are not the same thing. The sub-token saturation finding provides evidence that the structure of a model's multilingual representations is shaped by a combination of factors such as tokenizer design, orthographic properties, and the architecture of the shared hypothesis space, in addition to how much audio from a language was included during pre-training. Scaling data volume addresses some problems but not the design choices that produce the generic decoding regime. Addressing internal representational inequity requires intervention at the level of tokenizer design and decoder architecture, not only data collection.

Third, accuracy metrics alone do not measure the harm that biased systems produce. The accommodation labor, the self-blame, and the abandonment that our survey in Chapter 4 documents are not incidental to ASR bias; they are its primary consequences for affected users. Fairness evaluation that stops at error rates measures the failure of the system but not the cost of that failure to the people who live with it. User experience needs to be an essential component of fairness assessment.

5.3 Future Directions

One direction this work points toward is linguistically informed tokenizer design. The saturation findings show that sub-token utilization is shaped by orthographic and typological structure in addition to training data volume, which means that the multilingual hypothesis space encoded in the tokenizer is itself a source of inequity. Current BPE vocabularies are built to minimize reconstruction loss over training corpora, with no constraint on linguistic coverage. A tokenizer designed with phonological feature relevance as an explicit objective could address the root cause of the generic decoding regime rather than treating it as an inevitable consequence of low-resource status. This is an engineering problem, but linguistic knowledge could help.

A second direction is linguistically fair training data curation. Despite promising results of adapting multilingual models to underrepresented and typologically rare languages, a

more fundamental intervention for the linguistic feature failures would be designing pre-training corpora with explicit phonological and typological coverage targets, ensuring that varieties such as tonal languages, nasal-vowel-contrasting languages, and morphologically complex languages receive sufficient pre-training exposure to allow the model to develop distinct internal representations for the features they encode. This endeavor will involve decisions about what phonological phenomena to prioritize, how to sample within languages to maximize feature coverage, and how to weight training signal so that typologically rarer features are not left behind. Bringing typological frameworks into pre-training data design is a concrete and tractable research program.

A third direction is participatory evaluation design. The user experience study demonstrates that accuracy metrics systematically miss the harm that biased ASR systems produce, and that affected communities are not disengaged from the project of improving those systems. What is missing is a principled framework for incorporating user experience into evaluation as a primary criterion rather than a qualitative supplement. Developing such a framework requires sustained participatory design with the communities it is meant to serve, such as collaboratively defining what adequate performance means, what failure costs, and what accountability mechanisms would make community participation equitable rather than extractive.

5.4 Closing

Speech recognition has the potential to bring technology to languages and communities that have never had it, but only if we can steer the models away from reproducing the inequities already visible in their outputs. Making that potential real requires understanding where disparities live inside these systems, what linguistic structures they fail to represent, and what those failures cost the people on the other side of the microphone.

We hope that the methods and findings presented here contribute to building speech recognition that is fair to all speakers.

BIBLIOGRAPHY

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. User-friendly Automatic Transcription of Low-resource Languages: Plugging ESPnet into Elpis. In Antti Arppe, Jeff Good, Atticus Harrigan, Mans Hulden, Jordan Lachler, Sarah Moeller, Alexis Palmer, Miikka Silfverberg, and Lane Schwartz, editors, *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 51–62, Online, March 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.computel-1.7/.

Vaibhav Aggarwal, Shabari S Nair, Yash Verma, and Yash Jogi. Adopting Whisper for Confidence Estimation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, April 2025. doi: 10.1109/ICASSP49660.2025.10888254. URL https://ieeexplore.ieee.org/document/10888254/.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A. Smith. MAGNET: Improving the Multilingual Fairness of Language Models with Adaptive Gradient-Based Tokenization. *Advances in Neural Information Processing Systems*, 37:47790–47814, December 2024. doi: 10.52202/079017-1514. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5572bc595de865c1450868fd5391e9c5-Abstract-Conference.html.

Antonios Anastasopoulos and David Chiang. Leveraging Translations for Speech Transcription in Low-resource Settings. In *Interspeech 2018*, pages 1279–1283. ISCA, September 2018. doi: 10.21437/Interspeech.2018-2162. URL https://www.isca-archive.org/interspeech_2018/anastasopoulos18_interspeech.html.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Ahmed Adel Attia, Dorottya Demszky, Tolulope Ogunremi, Jing Liu, and Carol Espy-Wilson. Continued Pretraining for Domain Adaptation of Wav2vec2.0 in Automatic Speech Recognition for Elementary Math Classroom Settings, May 2024. URL http://arxiv.org/abs/2405.13018. arXiv:2405.13018 [cs] version: 1.

Peter K. Austin and Julia Sallabank. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press, March 2011. ISBN 978-1-139-50083-8. Google-Books-ID: 0XZRauYgO6AC.

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. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, December 2021. URL http://arxiv.org/abs/2111.09296. arXiv:2111.09296 [cs, eess].

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, and Michael Auli. Unsupervised Speech Recognition. In *Advances in Neural Information Processing Systems*, volume 34, pages

27826–27839. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ea159dc9788ffac311592613b7f71fbb-Abstract.html.

Junwen Bai, Bo Li, Qiujia Li, Tara N. Sainath, and Trevor Strohman. Efficient Adapter Finetuning for Tail Languages in Streaming Multilingual ASR. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10841–10845, April 2024. doi: 10.1109/ICASSP48485.2024.10447399. URL https://ieeexplore.ieee.org/abstract/document/10447399?casa_token=Hx-_zv8-9WoAAAAA:ofTZ2KIvaMHrv1XBG11yHmvACc59RTZxCzgeoXDn3kj5GcAPO4L0XcXJEgaEVXh1HATzDVha.

April Baker-Bell. *Linguistic Justice: Black Language, Literacy, Identity, and Pedagogy*. Routledge, New York, April 2020. ISBN 978-1-315-14738-3. doi: 10.4324/9781315147383.

Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4):923–934, December 2024a. ISSN 1381-2416, 1572-8110. doi: 10.1007/s10772-024-10141-5. URL https://link.springer.com/10.1007/s10772-024-10141-5.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. Probing Whisper Predictions for French, English and Persian Transcriptions. In Mourad Abbas and Abed Alhakim Freihat, editors, *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 129–138, Trento, October 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.icnlsp-1.15/.

Russell Barlow. Documentation of Ulwa, an endangered language of Papua New Guinea, 2018a. URL http://hdl.handle.net/2196/00-0000-0000-000F-CB61-A.

Russell Barlow. A Grammar of Ulwa. May 2018b. URL http://hdl.handle.net/10125/62506.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Us-

ing Data Augmentation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.42. URL https://aclanthology.org/2023.acl-long.42/.

Allan Bell. Language style as audience design. *Language in Society*, 13(2):145–204, June 1984. ISSN 0047-4045, 1469-8013. doi: 10.1017/S004740450001037X. URL https://www.cambridge.org/core/product/identifier/S004740450001037X/type/journal_article.

Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley, July 2019. ISBN 978-1-5095-2639-0. Google-Books-ID: nPy9uwEACAAJ.

Barnini Bhattacharyya and Jennifer L. Berdahl. Do you see me? An inductive examination of differences between women of color's experiences of and responses to invisibility at work. *Journal of Applied Psychology*, 108(7):1073–1095, 2023. ISSN 1939-1854. doi: 10.1037/apl0001072.

Steven Bird. Sparse Transcription. *Computational Linguistics*, 46(4):713–744, December 2020. doi: 10.1162/coli_a_00387. URL https://aclanthology.org/2020.cl-4.1/.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120/.

Hennie Brugman and Albert Russel. Annotating Multi-media / Multi-modal resources with ELAN. 2004.

Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Rep-

resentation, April 2022. URL http://arxiv.org/abs/2204.00540. arXiv:2204.00540 [cs].

June Choe, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao, and Nicole Holliday. Language-specific Effects on Automatic Speech Recognition Errors for World Englishes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7177–7186, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.628.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, March 2020. ISBN 978-0-262-35686-2. doi: 10.7551/mitpress/12255.001.0001. URL https://direct.mit.edu/books/oa-monograph/4605/Design-JusticeCommunity-Led-Practices-to-Build-the.

Rolando Coto-Solano. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri. In Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors, *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.americasnlp-1.20. URL https://aclanthology.org/2021.americasnlp-1.20/.

Nikolas Coupland. Language, situation, and the relational self: theorizing dialect-style in sociolinguistics. In John R. Rickford and Penelope Eckert, editors, *Style and Sociolin-*

guistic Variation, pages 185–210. Cambridge University Press, Cambridge, 2002. ISBN 978-0-521-59191-1. doi: 10.1017/CBO9780511613258.012.

Mingyu Cui, Daxin Tan, Yifan Yang, Dingdong Wang, Huimeng Wang, Xiao Chen, Xie Chen, and Xunying Liu. Exploring SSL Discrete Tokens for Multilingual ASR, September 2024. URL http://arxiv.org/abs/2409.08805. arXiv:2409.08805 [cs].

Xiaodong Cui, Brian Kingsbury, Jia Cui, Bhuvana Ramabhadran, Andrew Rosenberg, Mohammad Sadegh Rasooli, Owen Rambow, Nizar Habash, and Vaibhava Goel. Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA babel program. In *Interspeech 2014*, pages 2103–2107. ISCA, September 2014. doi: 10.21437/Interspeech.2014-477. URL https://www.isca-archive.org/interspeech_2014/cui14b_interspeech.html.

Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé Iii, Christina Harrington, and Hanna Wallach. Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12826–12833, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.761. URL https://aclanthology.org/2024.findings-acl.761/.

Mitchell DeHaven and Jayadev Billa. Improving Low-Resource Speech Recognition with Pretrained Speech Models: Continued Pretraining vs. Semi-Supervised Training, July 2022. URL http://arxiv.org/abs/2207.00659. arXiv:2207.00659 [cs].

Rachel Dorn. Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English. In Venelin Kovatchev, Irina Temnikova, Branislava Šandrih, and Ivelina Nikolova, editors, *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/issn.2603-2821.2019_003. URL https://aclanthology.org/R19-2003/.

C.M. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. Embedding

Structure Matters: Comparing Methods to Adapt Multilingual Vocabularies to New Languages. In Duygu Ataman, editor, *Proceedings of the 3rd Workshop on Multilingual Representation Learning (MRL)*, pages 268–281, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.20. URL https://aclanthology.org/2023.mrl-1.20/.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2019: TTS without T, July 2019. URL http://arxiv.org/abs/1904.11469. arXiv:1904.11469 [cs].

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Napo Quichua, 2024. URL https://www.ethnologue.com/language/qvo. Edition: 27.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Tena Lowland Quichua, 2025. URL https://www.ethnologue.com/language/quw. Edition: 26.

Penelope Eckert. *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Wiley-Blackwell, Malden, Mass., 2000. ISBN 978-0-631-18604-5.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying Bias in Automatic Speech Recognition, April 2021. URL http://arxiv.org/abs/2103.15122. arXiv:2103.15122 [cs, eess].

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567, March 2024. ISSN 0885-2308. doi: 10.1016/j.csl.2023.101567. URL https://www.sciencedirect.com/science/article/pii/S0885230823000864.

Carmen Fought. *Chicano English in Context*. Palgrave Macmillan, January 2003. ISBN 978-0-333-98637-0. Google-Books-ID: AzskWeEaWrEC.

Howard Giles and Peter F. Powesland. *Speech style and social evaluation*. Speech style

and social evaluation. Academic Press, Oxford, England, 1975. ISBN 978-0-12-283750-0. Pages: viii, 218.

Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 1967. ISBN 978-0-202-30260-7. Google-Books-ID: oUxEAQAAIAAJ.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, March 1992. doi: 10.1109/ICASSP.1992.225858. URL https://ieeexplore.ieee.org/document/225858.

Lisa J. Green. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge, 2002. ISBN 978-0-521-89138-7. doi: 10.1017/CBO9780511800306. URL https://www.cambridge.org/core/books/african-american-english/1AE59657F9CF1BBC3A2BF2B9BB29D1D0.

Karolina Grzech. Upper Napo Kichwa: a documentation of linguistic and cultural practices, 2020. URL http://hdl.handle.net/2196/00-0000-0000-000C-F5FB-A.

Frantisek Grézl, Martin Karafiát, and Karel Veselý. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658, May 2014. doi: 10.1109/ICASSP.2014.6855089. URL https://ieeexplore.ieee.org/abstract/document/6855089?casa_token=6ZGIk_00IdMAAAAA:ZLHAUKxXLNL5-ihR9ws6FDUQhfkGq-T4VWiPHXmbfP2njmICzvMdqjJUB8qPm6ptDeVmDruw.

Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyên, Maxime Fily, Guillaume Jacques, and Alexis Michaud. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource setting. pages 4905–4909, 2022a. doi: 10.21437/Interspeech.2022-11314. URL https://www.isca-archive.org/interspeech_2022/guillaume22_interspeech.html.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.computel-1.21. URL https://aclanthology.org/2022.computel-1.21.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, July 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. Recent Advances in Discrete Speech Tokens: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4184–4204, April 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3643619. URL https://ieeexplore.ieee.org/document/11298521/.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. Languages Through the Looking Glass of BPE Compression. *Computational Linguistics*, 49(4):943–1001, December 2023. doi: 10.1162/coli_a_00489. URL https://aclanthology.org/2023.cl-4.5/. Place: Cambridge, MA.

Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):216:1–216:25, November 2019. doi: 10.1145/3359318. URL https://doi.org/10.1145/3359318.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. Modeling Gender and Dialect Bias in Automatic Speech Recognition. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 15166–15184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.890. URL https://aclanthology.org/2024.findings-emnlp.890/.

Nikolaus P. Himmelmann. Documentary and descriptive linguistics. *Linguistics*, 36(1): 161–196, January 1998. ISSN 1613-396X. doi: 10.1515/ling.1998.36.1.161. URL https://www.degruyter.com/document/doi/10.1515/ling.1998.36.1.161/html.

Nikolaus P Himmelmann and John U Wolff. *Toratán (Ratahan)*, volume 130. Lincom Europa, 1999.

Arlie Russell Hochschild. *The Managed Heart: Commercialization of Human Feeling*. University of California Press, 1 edition, 2012. ISBN 978-0-520-27294-1. URL https://www.jstor.org/stable/10.1525/j.ctt1pn9bk.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, May 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, June 2021. URL http://arxiv.org/abs/2106.07447. arXiv:2106.07447 [cs, eess].

Ben Hutchinson, Celeste Rodríguez Louro, Glenys Collard, and Ned Cooper. Designing Speech Technologies for Australian Aboriginal English: Opportunities, Risks and Participation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 108–124, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732010. URL https://dl.acm.org/doi/10.1145/3715275.3732010.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Towards Building ASR Systems for the Next Billion Users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10813–10821, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i10.21327. URL https://ojs.aaai.org/index.php/AAAI/article/view/21327. Number: 10.

Hui Jiang. Confidence measures for speech recognition: A survey. *Speech Communication*, 45(4):455–470, April 2005. ISSN 0167-6393. doi: 10.1016/j.specom.2004.12.004. URL https://www.sciencedirect.com/science/article/pii/S0167639305000051.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.87. URL https://aclanthology.org/2023.acl-short.87/.

Austin Jones, Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzic, and Keith Langston. Comparing Kaldi-Based Pipeline Elpis and Whisper for Čakavian Transcription. In Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Saliha Muradoglu, Eric Le Ferrand, Elena Klyachko, Ekaterina Vylomova, Tatiana Shavrina, and Francis Tyers, editors, *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 61–68, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fieldmatters-1.8. URL https://aclanthology.org/2024.fieldmatters-1.8/.

Anthony Jukes. Documentation of Toratán (Ratahan), 2010. URL http://hdl.handle.net/2196/00-0000-0000-0001-56FB-9.

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, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages, November 2025. URL `http://arxiv.org/abs/2511.09690`. arXiv:2511.09690 [cs].

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Interspeech 2021*, pages 1529–1533. ISCA, August 2021. doi: 10.21437/Interspeech.2021-2062. URL `https://www.isca-archive.org/interspeech_2021/khare21_interspeech.html`.

Jiyeon Kim, Mehul Kumar, Dhananjaya Gowda, Abhinav Garg, and Chanwoo Kim. Semi-supervised transfer learning for language expansion of end-to-end speech recognition models to low-resource languages, November 2021. URL `http://arxiv.org/abs/2111.10047`. arXiv:2111.10047 [eess].

Neil W. Kirk, Mathieu Declerck, Ryan J. Kemp, and Vera Kempe. Language control in regional dialect speakers – monolingual by name, bilingual by nature? *Bilingualism: Language and Cognition*, 25(3):511–520, May 2022. ISSN 1366-7289, 1469-1841. doi: 10.1017/S1366728921000973. URL `https://www.cambridge.org/core/product/identifier/S1366728921000973/type/journal_article`.

Thomas Klingler. *If I Could Turn My Tongue Like That: The Creole Language of Pointe Coupee Parish, Louisiana*. LSU Press, August 2003. ISBN 978-0-8071-5590-5. Google-Books-ID: Q4B2kWdViU4C.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):

7684–7689, April 2020. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681, June 2024. doi: 10.1145/3630106.3658996. URL http://arxiv.org/abs/2402.08021. arXiv:2402.08021 [cs].

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. DecoderLens: Layerwise Interpretation of Encoder-Decoder Transformers. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4764–4780, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.296. URL https://aclanthology.org/2024.findings-naacl.296/.

William Leap. *American Indian English*. University of Utah Press, 1993. ISBN 978-0-87480-416-4. Google-Books-ID: pL55AAAAIAAJ.

Gina-Anne Levow, Emily P. Ahn, and Emily M. Bender. Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:96–106, March 2021. doi: 10.33011/computel.v1i.967. URL https://journals.colorado.edu/index.php/computel/article/view/967.

Siyu Liang and Gina-Anne Levow. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In Éric Le Ferrand, Elena Klyachko, Anna Postnikova, Tatiana Shavrina, Oleg Serikov, Ekaterina Voloshina, and Ekaterina Vylomova, editors, *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-282-4. URL https://aclanthology.org/2025.fieldmatters-1.3/.

Siyu Liang and Alicia Beckford Wassink. "This Wasn't Made for Me": Recentering User Experience and Emotional Impact in the Evaluation of ASR Bias. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2026)*, Montreal, Canada, June 2026. Association for Computing Machinery (ACM).

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31225–31235, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1591/.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. The Limits of Data Scaling: Sub-token Utilization and Acoustic Saturation in Multilingual ASR. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC 2026)*, Palma de Mallorca, Spain, 2026. European Language Resources Association (ELRA).

Rosina Lippi-Green. *English with an Accent: Language, Ideology and Discrimination in the United States*. Routledge, London, 2 edition, March 2012. ISBN 978-0-203-34880-2. doi: 10.4324/9780203348802.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions, November 2021. URL http://arxiv.org/abs/2111.09983. arXiv:2111.09983 [cs, eess].

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131, Dubrovnik, Croatia, May 2023.

Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.10. URL https://aclanthology.org/2023.eacl-main.10/.

Julia Mainzinger and Gina-Anne Levow. Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke. In Xiyan Fu and Eve Fleisig, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 170–176, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-srw.16. URL https://aclanthology.org/2024.acl-srw.16.

Joshua L. Martin and Kevin Tang. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Interspeech 2020*, pages 626–630. ISCA, October 2020. doi: 10.21437/Interspeech.2020-2893. URL https://www.isca-archive.org/interspeech_2020/martin20_interspeech.html.

Stuart McGill. Cicipu documentation, 2012. URL http://hdl.handle.net/2196/00-0000-0000-0001-7D83-D.

Stuart McGill. Cicipu. *Journal of the International Phonetic Association*, 44(3):303–318, December 2014. ISSN 0025-1003, 1475-3502. doi: 10.1017/S002510031400022X. URL https://www.cambridge.org/core/journals/journal-of-the-international-phonetic-association/article/cicipu/A101A2D46CEF24F5D21DAE013FCEFD01.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*, 4, November 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.725911. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.725911/full.

Yajie Miao, Florian Metze, and Shourabh Rawat. Deep maxout networks for low-resource speech recognition. In *2013 IEEE Workshop on Auto-*

matic Speech Recognition and Understanding, pages 398–403, December 2013. doi: 10.1109/ASRU.2013.6707763. URL https://ieeexplore.ieee.org/abstract/document/6707763?casa_token=Rln-ji6EfHYAAAAA:EIEZKWsN3IK5l8_aRCZANtKcmLZ9jkfCM05DxF6byxAeHP72w5vATSYlWIwDiIP5AnqxTJYM.

Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. "It's not a representation of me": Examining Accent Bias and Digital Exclusion in Synthetic AI Voice Services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–245, Athens Greece, June 2025. ACM. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732018. URL https://dl.acm.org/doi/10.1145/3715275.3732018.

Benjamin Milde and Arne Köhn. Open Source Automatic Speech Recognition for German, July 2018. URL http://arxiv.org/abs/1807.10311. arXiv:1807.10311 [cs].

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.38. URL https://aclanthology.org/2021.naacl-main.38/.

Mikel K. Ngueajio and Gloria Washington. Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. volume 13518, pages 421–440. 2022. doi: 10.1007/978-3-031-21707-4_30. URL http://arxiv.org/abs/2211.09511. arXiv:2211.09511 [cs, eess].

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, February 2018. ISBN 978-1-4798-3724-3.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148, March 2023. ISSN 0306-4573. doi: 10.1016/j.ipm.2022.103148. URL https://www.sciencedirect.com/science/article/pii/S0306457322002497.

Erin O'Rourke and Tod D. Swanson. Tena Quichua. *Journal of the International Phonetic Association*, 43(1):107–120, 2013. ISSN 0025-1003. URL https://www.jstor.org/stable/26351942.

Naomi Elizabeth Palosaari. *Topics in Mocho' phonology and morphology.* Ph.D., The University of Utah, United States – Utah, 2011. URL https://www.proquest.com/docview/863689180/abstract/E7686087339D442BPQ/1.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, April 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://ieeexplore.ieee.org/document/7178964.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 36963–36990, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800/.

Leena G. Pillai, Kavya Manohar, Basil K. Raju, and Elizabeth Sherly. Multistage Fine-tuning Strategies for Automatic Speech Recognition in Low-resource Languages, November 2024. URL http://arxiv.org/abs/2411.04573. arXiv:2411.04573 [cs].

Bettina Pospisil, Thilo Sauter, Albert Treytl, Edith Huber, and Walter Seböck. "Totally Unnecessary" or "Simply Convenient" – About Users and Non-Users of Voice Assistants. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–7, July 2022. doi: 10.1109/HSI55341.2022.9869441. URL https://ieeexplore.ieee.org/document/9869441/.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and others. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

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. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024. ISSN 1533-7928. URL http://jmlr.org/papers/v25/23-1318.html.

Kerri Prinos, Neal Patwari, and Cathleen A. Power. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1245–1254, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658969. URL https://dl.acm.org/doi/10.1145/3630106.3658969.

Agnieszka Przezdziak. *Optimizing Speech Recognition for Low-Resource Languages: Northern Sotho*. 2024. URL https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-533377.

Jaime Pérez González. Documentation of Mocho' (Mayan): Language Preservation through

Community Awareness and Engagement, 2018. URL http://hdl.handle.net/2196/00-0000-0000-0010-7985-9.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR, July 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

John R. Rickford and Sharese King. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4): 948–988, 2016. ISSN 1535-0665. URL https://muse.jhu.edu/pub/24/article/641206.

Sanjay Rijal, Shital Adhikari, Manish Dahal, Manish Awale, and Vaghawan Ojha. Whisper Finetuning on Nepali Language, November 2024. URL http://arxiv.org/abs/2411.12587. arXiv:2411.12587 [cs].

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu. In Kengatharaiyer Sarveswaran, Ashwini Vaidya, Bal Krishna Bal, Sana Shams, and Surendrabikram Thapa, editors, *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 202–207, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics. URL https://aclanthology.org/2025.chipsal-1.20/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.96. URL https://aclanthology.org/2021.eacl-main.96/.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR, June 2024. URL http://arxiv.org/abs/2406.06619. arXiv:2406.06619 [eess].

Monorama Swain, Anna Katrine Van Zee, and Anders Søgaard. On Mitigating Performance Disparities in Multilingual Speech Recognition. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5647–5655, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.323. URL https://aclanthology.org/2024.emnlp-main.323/.

Rachael Tatman. Gender and Dialect Bias in YouTube's Automatic Captions. In Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors, *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://aclanthology.org/W17-1606/.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. Fully Convolutional ASR for Less-Resourced Endangered Languages. In Dorothee Beermann, Laurent Besacier, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130, Marseille, France, May 2020. European Language Resources association. ISBN 979-10-95546-35-1. URL https://aclanthology.org/2020.sltu-1.17/.

Nick Thieberger. *The Oxford Handbook of Linguistic Fieldwork*. OUP Oxford, 2012. ISBN 978-0-19-957188-8. Google-Books-ID: 86AE2_0nPbkC.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10094667. URL https://ieeexplore.ieee.org/document/10094667/.

Geoffroy Vanderreydt, François Remy, and Kris Demuynck. Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition. In *Interspeech 2022*, pages 3053–3057. ISCA, September 2022. doi: 10.21437/Interspeech.2022-10744. URL https://www.isca-archive.org/interspeech_2022/vanderreydt22_interspeech.html.

Constanze C. Vorwerg, Sumanghalyah Suntharam, and Marie-Anne Morand. Language control and lexical access in diglossic speech production: Evidence from variety switching in speakers of Swiss German. *Journal of Memory and Language*, 107:40–53, August 2019. ISSN 0749596X. doi: 10.1016/j.jml.2019.03.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0749596X19300294.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140: 50–70, May 2022. ISSN 0167-6393. doi: 10.1016/j.specom.2022.03.009. URL https://www.sciencedirect.com/science/article/pii/S0167639322000486.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Michael Wroblewski. Amazonian Kichwa Proper: Ethnolinguistic Domain in Pan-Indian Ecuador. *Journal of Linguistic Anthropology*, 22(1):64–86, 2012. ISSN 1548-1395. doi: 10.1111/j.1548-1395.2012.01134.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1548-1395.2012.01134.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1548-1395.2012.01134.x.

G. Thimmaraja Yadava and H S Jayanna. Development and comparison of ASR models using kaldi for noisy and enhanced kannada speech data. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1832–1838, September 2017. doi: 10.1109/ICACCI.2017.8126111. URL http://ieeexplore.ieee.org/document/8126111/. Conference Name: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) ISBN: 9781509063673 Place: Udupi.

Anna Zee, Marc Zee, and Anders Søgaard. Group Fairness in Multilingual Speech Recognition Models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.143. URL https://aclanthology.org/2024.findings-naacl.143/.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186, May 2022. doi: 10.1109/ICASSP43922.2022.9746682. URL https://ieeexplore.ieee.org/abstract/document/9746682?casa_token=l7X2TcXPHBMAAAAA:SNvfG71x9jZAzNOMWkVOll5sRRVKLRC_N2raSa2PjL24PJkZOuUJt163UItbq7l_-j_Sowk2.

Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 215–219, May 2014. doi: 10.1109/ICASSP.2014.6853589. URL https://ieeexplore.ieee.org/abstract/document/6853589?casa_token=BHjd-jDajQsAAAAA:sR1ejuV3fkI76ANBKPKr8bXawwK4HRI5y7HVbPc1_eXPZLvSANV8BoRp8_K38MyfcCWjjt82.

Judit Ács. Exploring BERT's Vocabulary, February 2019. URL https://juditacs.github.io/2019/02/19/bert-tokenization-stats.html.

Appendix A

SUPPLEMENTARY MATERIALS: FIELDWORK ASR ADAPTATION

A.1 Training Details

The models were trained on an NVIDIA T4 GPU, with training times ranging from approximately 1 to 60 minutes per model. The hyper-parameters were defined as follows:

- **Learning rate:** MMS: 1e-3; XLS-R: 3e-4
- **Maximum epochs:** 30
- **Best model metric:** Character Error Rate (CER)
- **Early stopping:** 3 epochs
- **Early stopping threshold:** 0.003

A.2 Dataset Details

Table A.1 provides detailed information on the genres and types of content present in the datasets used in this study, along with key linguistic references and citations to the original documentation archives. Table A.2 summarizes the total archived hours of audio recordings for each language and the amount of data remaining after the cleaning and preprocessing steps described earlier in the manuscript.

A.3 Error Rates

We consider four phonological categories:

$$C \in \{\text{Tone, Nasality, V\_length, C\_length}\}.$$

Over the entire dataset, we record:

- S_C: the total *substitution* errors for category C,
- D_C: the total *deletion* errors for category C,
- I_C: the total *insertion* errors for category C,
- L_C: the total *reference tokens* exhibiting category C (e.g. `tone_labels` for tone, `total_vowels` for vowel length, etc.).

We then define the total errors and error rate for each category C as follows:

$$E_C = S_C + D_C + I_C \quad \text{and} \quad \text{ErrorRate}_C = \frac{E_C}{L_C}.$$

For example, if C = tone then L_C = tone_labels (the number of reference tokens with at least one tone diacritic). Similarly, if C = vowel_length then L_C = total_vowels (the total vowel tokens in the reference).

Language	**Genres of content**	**Phonological Description**	**Documentation**
Cicipu	Greetings, conversations, hortative discourse, narratives, procedural discourse, ritual discourse, elicitation activities	McGill (2014)	McGill (2012)
Mocho'	Biographical and non-biographical narratives (historical events, myths, local beliefs, traditional building, witchcraft), prayer, conversation, elicitation sessions, text translation	Palosaari (2011)	Pérez González (2018)
Toratán	Conversational data, elicitation sessions, narratives (personal history, folk tales)	Himmelmann and Wolff (1999)	Jukes (2010)
Ulwa	Conversational data, traditional stories, personal stories, traditional singing and dancing video	Barlow (2018b)	Barlow (2018a)
Upper Napo Kichwa	Grammatical elicitation, life interviews	Wroblewski (2012), O'Rourke and Swanson (2013)	Grzech (2020)

Table A.1: Details of depository content for languages used in Chapter 2, related linguistic work referenced, and original documentation citations.

Language	**Total Hours**	**Cleaned Hours**
Cicipu	5.66	3.09
Mocho'	7.26	4.21
Toratán	22.84	11.15
Ulwa	3.25	2.83
Upper Napo Kichwa	13.19	6.97

Table A.2: Total archived and cleaned hours of audio for all languages used in Chapter 2.

Appendix B

SURVEY QUESTIONNAIRE: USER EXPERIENCE OF ASR BIAS

B.1 Questionnaire Design

B.1.1 Complete Questionnaire

The following presents the complete questionnaire administered to participants. Question types are indicated as follows: [O] = Open-ended response; [MC] = Multiple choice; [MA] = Multiple selection (select all that apply); [Likert] = 5-point Likert scale. The Atlanta site used an earlier version of the survey with some different response options, as noted below.

Demographics and Speech Background

Q1 [O]: How would you describe the way that you speak?

Q2 [O]: Would you be willing to share how you self-identify in terms of ethnicity or cultural background? If so, what term(s) do you use?

Q3 [MC]: Please indicate your age range: *Under 30 / 30–50 / Over 50*

Q4 [O]: Do you have any speech challenges (impairments or disabilities) e.g. stuttering, stammering, slurring, etc. that might affect your ability to use voice recognition technologies effectively? Please describe.

Voice Technology Use

Q5 [MA]: You are now going to see some different types of Small and Large Vocabulary technologies. Please select the ones that you use or have used. *Options: Stand-alone Smart Speakers (e.g., Amazon Echo, Google Home, Siri HomePod); Virtual Assistants on Mobile Phones, Tablets or Smart TVs; Voice Commands in Smart Home Appliances; Voice Commands in Car Infotainment Systems; Voice-controlled Gaming Devices; Voice Biometrics for Authentication; Voice-enabled Wearable Devices*

Q6a [O]: If you selected any of the Small or Large Vocabulary uses above: What are some tasks you use this tech for?

Q6b [O]: If there are any other tasks you use your Small or Large Vocabulary devices

for, please specify.

Q7 [MA]: You are now going to see some different types of Conversational technologies. Please select the ones that you use or have used. *Options: Video conferencing apps with auto-generated transcription (e.g., captions and transcripts on Zoom, Microsoft Teams, etc.); Voice-to-text/dictation features; Automated phone systems (IVR)*

Q8a [O]: If you selected any of the Conversational uses above: What are some tasks you use your tech for?

Q8b [O]: If there are any other tasks you use your Conversational devices for, please specify.

Q9 [O]: If there are any other kinds of voice recognition technologies you have used, please specify.

Q10 [O]: Are there any devices from the list we just talked about that you used, but stopped using? Can you tell us about what led you to stopping?

Experiences with Voice Recognition

Q11 [O]: Can you tell us about your most memorable experience with voice recognition technology? What was the technology and how did you use it? Did you accomplish what you wanted through using it?

Q12a–f [Likert]: How often do you use voice recognition technology for the following purposes? *Scale: 1 = Very frequently to 5 = Never.* (a) To be productive and manage tasks; (b) Using voice commands as shortcuts for accessing apps on my device; (c) To keep up with information, news, weather; (d) To play around and discover new capabilities of voice technology; (e) To support me in accessibility (captioning speech); (f) Other (please specify).

Dialect and Accent Challenges

Q13 [O]: Can you share a time where you had a challenge with your way of speaking (dialect, accent or language) being understood by voice recognition systems?

Q14 [MA]: Which, if any of the following reasons describe the challenge(s) that you identified above, or other challenges that you have experienced? Select all that apply.

Options for the Gulf Coast, Miami Beach, and Tucson: Voice technologies don't seem to take into consideration my cultural/ethnic/racial background; Voice assistants take what

I say out of context and either don't work or do something I didn't want; It takes too much physical effort for me to speak clear enough to be understood; I feel self-conscious about how I speak when talking to voice technologies

Additional options for Atlanta only: Voice technologies force me to change the way I talk to be understood; Voice technologies force me to repeat my speech over and over to be understood; When voice assistants/smart devices don't recognize my speech, I still have to manually complete a task; Voice technologies don't work for my accent/dialect most or all of the time; Voice technologies don't seem useful for my day-to-day lifestyle

Q14b [O]: If there are any other challenges or issues you have experienced, please specify.

Coping Strategies

Q15a [MA]: If you have experienced challenges in using voice recognition, are there things you do to fix or deal with the challenge? Select all that apply.

Options for Gulf Coast, Miami Beach, and Tucson: Repeat what I said; Change the words or phrases I use; Slow down my speech; Enunciate more; Speak louder; Imitate the voice assistant; Get another person to control the device for me; Manually complete the task instead; I just stop using it

Options for Atlanta: Repeat what I said; Change the way that I speak/code-switch; Imitate the voice assistant; Get another person to control the device for me; I just stop using it

Q15b [O]: If you said above that there are other ways you fix or deal with challenges in using voice recognition, please specify.

Emotional Responses

Q16 [O]: How does it make you feel when your voice assistant doesn't understand your speech/makes frequent mistakes?

Expectations and Willingness

Q17 [MC]: How often do you expect voice recognition technologies to be able to understand your speech dialect or accent correctly? *Options: Always (100% every time I use my device); Often (75% — most times I use my device); Sometimes (50% — every now and again); Rarely (25% occasionally); Never (0–20%)*

Q18a [MC]: Would you be willing to provide voice samples or feedback to help improve

voice recognition technologies' accuracy for different dialects and accents? *Options: Yes, I would be interested in contributing; Maybe, I would need more information; No, I would not be interested in contributing*

Q19 [MC]: Finally, would you be willing to pay for a product/service that provides improved voice recognition for diverse dialects and helps in undoing bias? If yes, which describes how much you'd be willing to pay? *Options: Not willing to pay—I feel I should be compensated; Prefer 1-time payment; <$10/month; $10–20/month; $20–30/month; More than $30/month*

Voice Sample Collection (Optional)

Q20: Would you like to contribute your voice to our database? [If yes, participants were directed to a voice recording interface.]

Q21 [O]: Please share any thoughts or comments on participating in this exercise. What questions might you have for our team regarding the study?