




2023

Automatic Transcription of Northern Prinmi Oral Art: Approaches and Challenges to Automatic Speech Recognition for Language Documentation

Connor Bechler

University of Kentucky, cbechler2@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0009-0001-9841-4820>

Digital Object Identifier: <https://doi.org/10.13023/etd.2023.202>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Bechler, Connor, "Automatic Transcription of Northern Prinmi Oral Art: Approaches and Challenges to Automatic Speech Recognition for Language Documentation" (2023). *Theses and Dissertations--Linguistics*. 51.

https://uknowledge.uky.edu/ltt_etds/51

This Master's Thesis is brought to you for free and open access by the Linguistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Linguistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Connor Bechler, Student

Dr. Josef Fruehwald, Major Professor

Dr. Kevin McGowan, Director of Graduate Studies

AUTOMATIC TRANSCRIPTION OF NORTHERN PRINMI ORAL ART:
APPROACHES AND CHALLENGES TO AUTOMATIC SPEECH RECOGNITION
FOR LANGUAGE DOCUMENTATION

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Arts in the
College of Arts and Sciences
at the University of Kentucky

By

Connor Nathaniel Bechler

Lexington, Kentucky

Director: Dr. Josef Fruehwald, Assistant Professor of Linguistics

Lexington, Kentucky

2023

Copyright © Connor Nathaniel Bechler 2023
<https://orcid.org/0009-0001-9841-4820>

ABSTRACT OF THESIS

AUTOMATIC TRANSCRIPTION OF NORTHERN PRINMI ORAL ART: APPROACHES AND CHALLENGES TO AUTOMATIC SPEECH RECOGNITION FOR LANGUAGE DOCUMENTATION

One significant issue facing language documentation efforts is the transcription bottleneck: each documented recording must be transcribed and annotated, and these tasks are extremely labor intensive (Ćavar et al., 2016). Researchers have sought to accelerate these tasks with partial automation via forced alignment, natural language processing, and automatic speech recognition (ASR) (Neubig et al., 2020). Neural network—especially transformer-based—approaches have enabled large advances in ASR over the last decade. Models like XLSR-53 promise improved performance on under-resourced languages by leveraging massive data sets from many different languages (Conneau et al., 2020). This project extends these efforts to a novel context, applying XLSR-53 to Northern Prinmi, a Tibeto-Burman Qiangic language spoken in Southwest China (Daudey & Pincuo, 2020).

Specifically, this thesis aims to answer two questions. First, is the XLSR-53 ASR model useful for first-pass transcription of oral art recordings from Northern Prinmi, an under-resourced tonal language? Second, does preprocessing target transcripts to combine grapheme clusters—multi-character representations of lexical tones and characters with modifying diacritics—into more phonologically salient units improve the model's predictions? Results indicate that—with substantial adaptations—XLSR-53 will be useful for this task, and that preprocessing to combine grapheme clusters does improve model performance.

KEYWORDS: Automatic Speech Recognition, Language Documentation, Northern Prinmi, Oral Art, XLSR-53, Transformers

Connor Nathaniel Bechler

(Name of Student)

04/28/2023

Date

AUTOMATIC TRANSCRIPTION OF NORTHERN PRINMI ORAL ART:
APPROACHES AND CHALLENGES TO AUTOMATIC SPEECH RECOGNITION
FOR LANGUAGE DOCUMENTATION

By
Connor Nathaniel Bechler

Dr. Josef Fruehwald

Director of Thesis

Dr. Kevin McGowan

Director of Graduate Studies

04/28/2023

Date

DEDICATION

To my late wife, Jaclyn Diane Vander Ploeg. You are my sweetness and my light.

ACKNOWLEDGMENTS

First, I'd like to thank my Thesis Chair Dr. Josef Fruehwald, whose detailed explanations, careful thinking, critical insights, and overarching guidance ensured this project could, despite many obstacles, reach completion. I have greatly enjoyed and appreciated our many frenetic meetings and conversations.

Second, I would like to thank the other two members of my Thesis Committee, Dr. Kevin McGowan and Dr. Rusty Barrett, for sharing their invaluable perspectives, questions, and suggestions. Dr. McGowan, thank you for contributing your incisive observations and energizing curiosity throughout this process. Dr. Barrett, thank you for your encouragement and guidance, which shaped the original project that led to this thesis.

Additionally, I would like to thank Dr. Mark Lauersdorf for his support and instruction, which have deeply benefited both this project and my academic development. While not explicitly listed here, I am also grateful to the other members of the University of Kentucky linguistics faculty and staff. My time in the MALTT program has been incredibly beneficial and enjoyable due to your contributions; thank you for making the department such a warm, supportive, and intellectually nourishing place.

I'd also like to thank the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for their support and use of the Lipscomb Compute Cluster and associated research computing resources.

Moreover, I'd like to thank Dr. Henriëtte Daudey and Gerong Pincuo, whose work introduced me to the Prinmi language and served as the foundation of this project. My gratitude also to those who contributed their voices, talents, and knowledge to the creation of the "Documentation of Northern Prinmi Oral Art" collection.

In addition to those I've thanked above, I'd like to thank my family and friends for all their material and emotional support, especially over the last six months.

To each of my peers in the MALTT program: thank you for being my friend and helping me in myriad, life-giving ways. My special thanks to Nour Kayali, Catie Mott, Angel Passarelli, Ian Schneider, John Winstead, and Ellie Wren-Hardin. You have suffered through much with me; thank you for helping shoulder my burdens. Being in community with each of you has deeply enriched my life and made Lexington home.

My immense gratitude to my parents Curt and Shari, for supporting me in all my endeavors and teaching me to love learning and love my neighbors; to Christian, for being my oldest friend, dearest antagonist, and perpetual mentor; and to Jon, Kerri, Hanna, and Ben Vander Ploeg, for welcoming me into their family and supporting me through our shared grief.

Finally, my unceasing gratitude and love to my late wife Jaclyn Vander Ploeg, whose love, friendship, and guidance over the last ten years have shaped me in innumerable ways, all for the better. I have so many stories to share with you; I miss you terribly.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	III
LIST OF TABLES	VI
LIST OF FIGURES	VII
CHAPTER 1. INTRODUCTION	1
CHAPTER 2. BACKGROUND	3
2.1 LITERATURE REVIEW	3
2.1.1 Language Documentation	3
2.1.2 Automatic Speech Recognition	4
2.1.3 Challenges for ASR for Language Documentation	6
2.1.4 Northern Prinmi	9
2.2 DOCUMENTATION CONTEXT	13
2.2.1 Description of Collection	13
2.2.2 Survey of Phones and Orthography	15
CHAPTER 3. METHODS	17
3.1 MODEL SELECTION	17
3.2 DATA PREPROCESSING	18
3.2.1 Partitioning the Dataset	18
3.2.2 Phrasal Segmentation of the Data	19
3.2.3 Preprocessing the Transcription Data	20
3.3 MODEL FINE-TUNING	23
CHAPTER 4. RESULTS	26
4.1 WER AND CER	26
4.2 SUBSTITUTION RATES	29
CHAPTER 5. DISCUSSION	30
5.1 EFFECTS OF PREPROCESSING	30
5.2 MODEL PERFORMANCE ON THE CORPUS	32
5.2.1 Model Comparison	32
5.2.2 What is a Useful Error Rate?	34
5.3 FUTURE DIRECTIONS	35
5.4 APPLICATIONS AND IMPLICATIONS	36
REFERENCES	38
VITA	42

LIST OF TABLES

Table 2.1 Orthographic Representations of Northern Prinmi Phone Inventory	16
Table 3.1 Summary of Data Partition Properties	19
Table 3.2 Fine-Tuning Parameters.....	24
Table 4.1 Model WER by Testing Set Subsection	27
Table 4.2 Model CER by Testing Set Subsection.....	27
Table 4.3 Model Substitutions by Character Class	29
Table 5.1 Prediction Label Vocabulary by Model.....	30
Table 5.2 Comparison of Closely Related Projects' WER and CER	32

LIST OF FIGURES

Figure 2.1 Map of the Pūmī language area (Daudey, 2014, p. 4).....	10
Figure 2.2 Wādū Pūmī consonant phonemes (Daudey, 2014, p. 19).....	11
Figure 2.3 Composite chart of vowels (Daudey, 2014, p. 47).....	12
Figure 3.1 Example Output and Segmentation by Preprocessing Method	23
Figure 3.2 XLSR-53 Fine-Tuning Schematic.....	24
Figure 3.3 Word and Character Error Rate Formulas.....	25
Figure 4.1 XLSR-53 Evaluation Schematic.....	26
Figure 4.2 Explanation of WER and CER Calculation.....	27

CHAPTER 1. INTRODUCTION

One significant issue facing language documentation efforts is the transcription bottleneck: each recording that is collected must be transcribed and annotated, and these tasks are extremely labor intensive (Ćavar et al., 2016). Researchers have sought to accelerate these tasks with partial automation via forced alignment, natural language processing, and automatic speech recognition (ASR) (Neubig et al., 2020).

Neural network—especially transformer-based—approaches have enabled large advances in ASR over the last decade. Models like XLSR-53 promise improved performance on under-resourced languages by leveraging massive data sets from many different languages (Conneau et al., 2020). Prior research has successfully applied the XLSR-53 model to the documentation of several low-resource languages (Coto-Solano et al., 2022; Guillaume et al., 2022; Nowakowski et al., 2023). This project extends these efforts to a novel context, applying XLSR-53 to Northern Prinmi, a Tibeto-Burman Qiangic language spoken in Southwest China (Daudey & Pincuo, 2020).

Specifically, this project aims to answer two questions.

First, is the XLSR-53 ASR model useful for first-pass transcription of documentation data from Northern Prinmi, an under-resourced tonal language? In particular, can it be successfully applied to recordings of oral art, like rituals and songs?

Second, does preprocessing target transcripts to combine grapheme clusters—multi-character representations of lexical tones and characters with modifying diacritics—into more phonologically salient units improve the model's predictions?

In Chapter 2, I provide an overview of this project's background, exploring the literature on automatic speech recognition and its application to language documentation

as well as Northern Prinmi. I then describe the particular documentation context, discussing its notable characteristics and the challenges these features pose for automatic speech recognition.

In Chapter 3, I describe my methodology, explaining how I chose XLSR-53, preprocessed the documentation collection for use with XLSR-53, and fine-tuned four XLSR-53 models on the preprocessed data.

In Chapter 4, I report my results, providing both Word and Character Error Rates for each model, as well as rates of character substitution by each model, while in Chapter 5, I discuss these results in comparison to both my expectations and previous work. I additionally provide an overview of possible future directions for my own work, as well as implications for the field more broadly.

Ultimately, I find that XLSR-53 may be useful for first-pass transcription of Northern Prinmi oral art, although my models are not very accurate. The error rates of even my best model are quite high, especially in comparison with other similar projects' models. However, compared with having no predictions at all, these predictions could still serve as a baseline for correction.

I also find that preprocessing transcription targets to more closely approximate phonologically salient units does have a positive impact on model performance. Combining tone character pairs into single predictive labels produces a substantial improvement in character error rates, whereas combining character diacritic clusters into single predictive labels has a slightly less substantial impact.

CHAPTER 2. BACKGROUND

2.1 Literature Review

2.1.1 Language Documentation

Language documentation may be broadly described as producing “a lasting, multipurpose record of a language” (Himmelman, 2006). A key element of a linguistic record being multipurpose is its possession of annotations, most crucially transcriptions and translations, as these allow both linguists and non-linguists to understand the content of recordings and navigate collections.

One significant issue facing language documentation is therefore what is referred to as the transcription bottleneck: each recording that is collected must be transcribed and annotated, and these tasks are extremely labor intensive, estimated to take anywhere between 30–120 hours of work per hour of speech data (Adams et al., 2017; Ćavar et al., 2016; Seifart et al., 2018; Shi et al., 2021). Closing the gap between quantity of recorded speech and quantity of annotated speech will greatly increase the utility of language documentation corpora for both the documented language’s communities and linguistic researchers (Seifart et al., 2018).

Researchers have sought to overcome this challenge by accelerating transcription and annotation through a variety of computational approaches, including forced alignment, natural language processing, and ASR (Neubig et al., 2020). Forced alignment and natural language processing, however, are both reliant on ASR for base transcriptions.

In the process of adapting speech and language technologies for language documentation, it is crucial that the technologies serve the purposes of documentation rather than vice versa. Bird (2020) cautions against using language technology for

documentation as an opportunity for data capture, while also calling for researchers to put more focus on community priorities, like intra-community preservation and transmission of teachings on country and other traditional knowledge. With this project, I hope to test one way a current speech and language technology, ASR, may be brought to bear for this sort of knowledge preservation, specifically examining and trying to optimize ASR performance on recordings of oral art.

2.1.2 Automatic Speech Recognition

Many ASR models have been based on hidden Markov models (HMMs). HMM-based systems tend to have three main components: an acoustic model, a lexical/pronunciation model, and a text-based language model (Besacier et al., 2014). The lexical/pronunciation model is essentially a pronunciation dictionary, the acoustic model is trained on speech from many different speakers so that it can discriminate between speech units across speakers and contexts, and the language model is trained on large amounts of text so that the overarching architecture's predictions will be constrained by the language's typical word order and collocations. This means that using HMM-based ASR systems for documentation or with endangered and low-resource languages poses a unique challenge, as these contexts often lack the required corpora and lexical resources (Adams et al., 2017).

Despite these challenges, at least one HMM-based ASR model, Kaldi (Povey et al., 2011) has been specifically adapted to the task of language documentation as part of the Endangered Language Pipeline and Inference System (Elpis) (Foley et al., 2018).

Presenting an alternative to HMMs, neural network approaches have proliferated in ASR over the last decade, coming to dominate the field (Wang et al., 2019). Beginning

in 2011, neural networks were adopted into hybrid HMM-deep neural network (DNN) models, using DNNs to perform acoustic modeling in place of Gaussian mixed models (GMMs). Since 2014, many non-HMM ASR approaches have been proposed; the methods and models most relevant to under-resourced languages include the DNN-based model Deep Speech (Hannun et al., 2014), the long short-term memory (LSTM) model Persephone (Adams et al., 2018), the attention-based encoder-decoder model ESPnet (Watanabe et al., 2018), and the transformer-based model XLSR-53 (Conneau et al., 2020).

In addition to being distinguished from prior HMM-based ASR models by their use of neural networks, these models are also distinct in being end-to-end speech recognition systems: they directly map speech input to text output without (generally) relying on lexical and syntactic-semantic models. For example, while Deep Speech uses a language model for post-processing, Persephone, ESPnet, and XLSR make their predictions solely on the bases of labeled audio data in the target language (and, in the case of XLSR, massive amounts of unlabeled multilingual audio data).

As a result, while they may require more training data overall, end-to-end neural network ASR models require substantially fewer linguistic resources in the target language than HMM-based ASR models (Wang et al., 2019). This makes them a promising candidate for use in low-resource contexts like language documentation, and both researchers and practitioners have begun to test them in a wide range of contexts (Adams et al., 2021; Boulianne, 2022; Coto-Solano et al., 2022; Guillaume, Wisniewski, Macaire, et al., 2022; Jimerson et al., 2018; Nowakowski et al., 2023; Shi et al., 2021; Zahrer et al., 2020).

One of the most used end-to-end ASR models is XLSR, a particular instance of Meta’s wav2vec 2.0 system (Conneau et al., 2020). wav2vec 2.0 is a transformer-based unsupervised-learning model which is trained on unlabeled speech to identify shared speech units across audio recordings (Baevski et al., 2020). wav2vec 2.0 models can be used for automatic transcription through fine-tuning, retraining an already trained model on novel data for a new task. Specifically, wav2vec 2.0 is designed to be fine-tuned by adding a linear character prediction layer trained on transcribed speech data with a Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006) on top of the existing model in order to perform automatic transcription.

XLSR-53 is a particular cross-lingual wav2vec 2.0 model trained on 56,000 hours of speech from 53 languages (Conneau et al., 2020). Training on cross-lingual speech data allows the model to perform significantly better on previously unseen languages than monolingual wav2vec 2.0 models. As a result of these findings, the authors argue their approach holds significant promise for work on low-resource languages; while they do not provide an explicit definition of low-resource, they provide the examples of two languages which possess 3 and 5 hours of labeled speech data. Prior research has successfully applied the XLSR model to the documentation of several languages with relatively small corpora of recorded speech, including Cook Islands Māori, Japhug, and Ainu (Coto-Solano et al., 2022; Guillaume, Wisniewski, Macaire, et al., 2022; Nowakowski et al., 2023).

2.1.3 Challenges for ASR for Language Documentation

As implied in the above sections, aside from the definitional lack of data, there are a multitude of other challenges involved in applying ASR in the context of language documentation.

One key challenge is overcoming asymmetries between the properties of linguistic records collected through language documentation fieldwork and the requirements of existing ASR models (Wisniewski et al., 2020). Due to the differing goals and the unique circumstances of each documentation project, documentation recordings are characterized by relatively few speakers (Adams et al., 2018; Boulianne, 2022; Foley et al., 2018), idiosyncratic annotation and formatting schemes (Wisniewski et al., 2020), and widely varying audio quality produced by diverse recording settings and strategies (Amith et al., 2021; Ćavar et al., 2016). In contrast, commonvoice and Multilingual LibriSpeech, two of the major multilingual speech datasets used for end-to-end ASR training, both consist entirely of single speaker read speech transcribed with consistent annotation schemes (Ardila et al., 2020; Pratap et al., 2020).

A solution to this particular problem is adapting language documentation practices to facilitate the collection of linguistic records more amicable to computational methods (Amith et al., 2021; Seifart et al., 2018; Zahrer et al., 2020). However, as noted by Wisniewski et al. (2020, p. 307), given the extremely limited resources available to language documentation researchers and consultants, “there is a potential conflict between the traditional perspective of creating a reasonably thorough and balanced record for posterity... and on the other hand, the requirement to put together data sets that lend themselves easily to Natural Language Processing.” Moreover, changing future practices does not help with applying ASR to documentation materials already produced.

Another—potentially parallel—solution is to improve methods for preprocessing documentation records, applying small transformations to bridge their differences from ideal ASR data. Several researchers have investigated the effects of specific preprocessing

choices on end-to-end ASR of language documentation recordings (Adams et al., 2017, 2018; Guillaume, Wisniewski, Macaire, et al., 2022; Wisniewski et al., 2020).

Working with the LSTM model Persephone, Adams et al. (2017, 2018) assessed if including tones as part of the input transcription and jointly predicting phonemes and tones or separately predicting tones from phonemes results in more accurate transcriptions of Chatino and Yongning Na. They segmented all multi-character phonemes and tones as single vocabulary items. They found that jointly predicting phonemes and tones (taking both as input in the same linear character sequence) had roughly the same accuracy as predicting phonemes and tones separately.

Wisniewski et al. (2020), also working with Persephone, tested a Unicode-based “grapheme cluster” segmentation scheme on the same corpus of Yongning Na as Adams et al. (2018). They segmented each character with its Unicode modifying characters into a prediction label. This means that some IPA diacritics, like “^h”, are considered independently, while others, like the nasal and syllabic diacritics, are segmented jointly with the character they modify; lexical tones, while made up of multiple characters, are also segmented together. They found that applying this segmentation method achieved very similar (albeit marginally worse) results to Adam et al.’s manually curated phoneme segmentation. However, they also note that the Persephone LSTM model requires phonemically transparent transcription to be effective, as the system fails to successfully transcribe audio from other languages when trained on orthographically transcribed recordings.

Applying XLSR-53 to an audio corpus of the Sino-Tibetan Japhug language, Guillaume et al. (2022) followed Wisniewski et al. (2020) in mapping single characters

one-to-one to prediction labels. Unlike Wisniewski, they do not mention combining Unicode modifiers with their preceding characters. They achieved WER and CER rates highly comparable to Wisniewski et al. (2020).

2.1.4 Northern Prinmi

Prinmi (also known as Pumi, or Pǔmǐ [普米] in Mandarin) is a Tibeto-Burman Qiangic language spoken by roughly 45,000 people in the southwestern Chinese provinces of Sichuan and Yunnan (Daudey & Pincuo, 2020). Northern Prinmi is the language's northern branch, with a speech community along Yunnan's northern border extending into south-west Sichuan (Daudey, 2014). The region is not exclusively, or even predominantly, settled by the Prinmi; rather, Prinmi reside alongside Tibetan, Nuosu, Naxi, Mosuo, and Han people. As a result, many speakers of Prinmi also speak other languages, and Northern Prinmi has been particularly influenced by Tibetan and the Southwestern Mandarin Chinese.

There is no widely adopted orthography for writing Prinmi, and it is not widely taught, although the Tibetan script is used to teach it in one school in Yunnan (Daudey & Pincuo, 2018). Prinmi has extensive internal variation, with several different dialect classification schemes. Ding (2014) proposes three possible dialect categories, Western, Central, and Northern. Grammars have been produced for varieties of both the Central (Ding, 2014) and Northern (Daudey, 2014) dialects, but not for any dialect group as a whole, as the dialects are not standardized. Ding notes that his grouping of varieties under the term “Northern” is tentative, as the region the term refers to contains more than half of all Prinmi speakers and is not extensively documented.

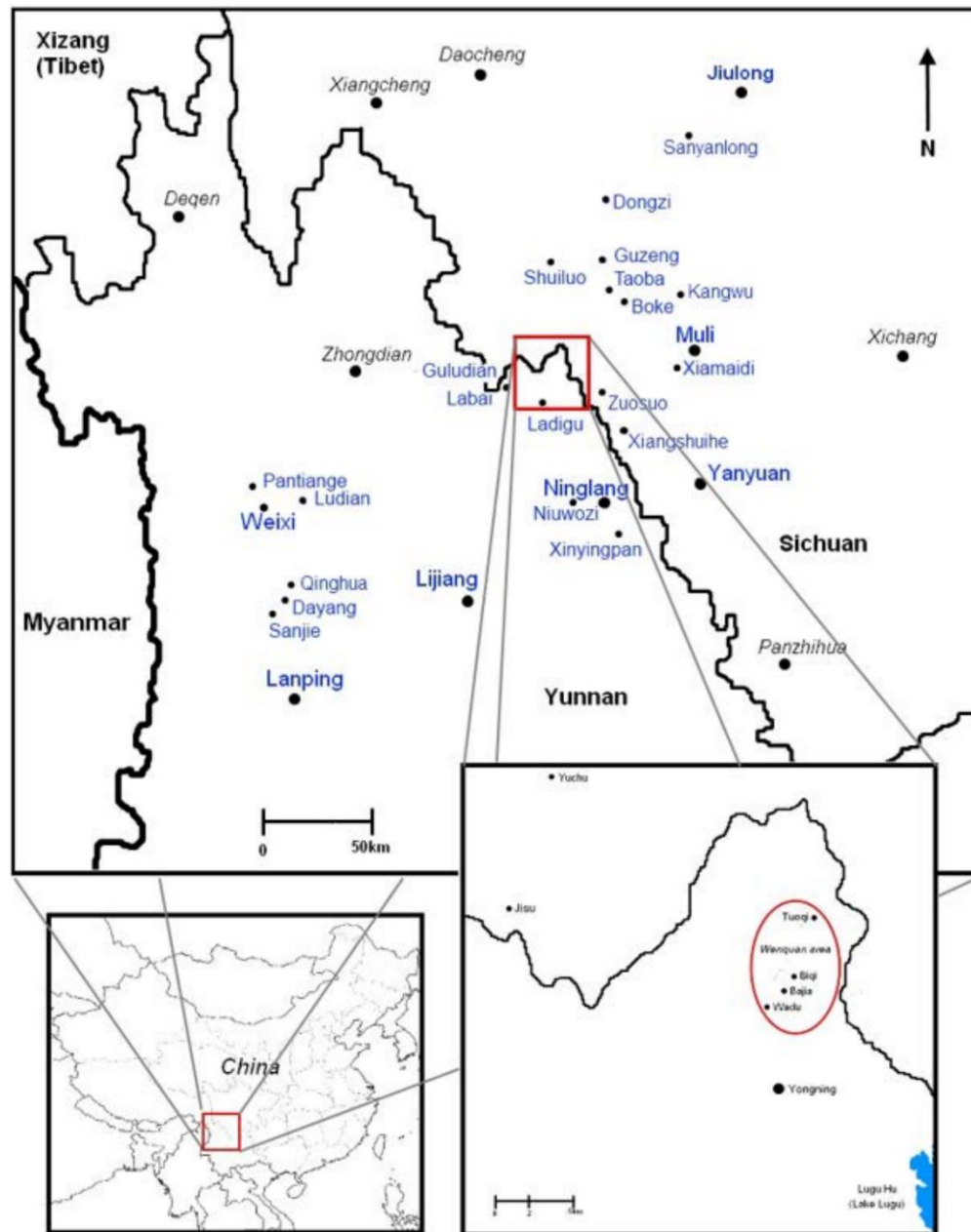


Figure 2.1 Map of the Pumi language area (Daudey, 2014, p. 4)

Due to this lack of a generalized grammar description, my phonological description of Northern Prinmi is based on the grammar of Wadu Pumi published by Henriette Daudey in 2014. Wadu Pumi is spoken in Middle Wadu Village of the Wenquan township of

Yunnan. Therefore, it will not reflect every variety of Northern Prinmi perfectly, but instead will provide a general sense for the type of features common to Northern Prinmi varieties. I am primarily interested in characterizing the range of phonetic possibilities that are produced by speakers of Wadu Pumi to identify possible phonetic classes the automatic speech recognition model will need to recognize and correctly classify to perform effective automatic transcription of Northern Prinmi data from a variety of locations.

Wadu Pumi's consonant system distinguishes between seven places of articulation—bilabial, alveolar, retroflex, alveopalatal, velar, uvular, and what Daudey refers to as cavity—and six manners of articulation—stops, affricates, fricatives, nasals, liquids, and approximates—and possesses a total of 42 phonemic consonants (2014).

	Bilabial	Alveolar	Retroflex	Alveopalatal	Velar	Uvular	Cavity
	p	t	ʈ		k	q	
Stops	p ^h	t ^h	ʈ ^h		k ^h	q ^h	
	b	d	ɖ		g		
		ts	tʂ	tɕ			
Affricates		ts ^h	tʂ ^h	tɕ ^h			
		dz	dʐ	dʑ			
Fricatives		s	ʂ	ɕ			h
		z	ʐ	ʑ			ʃ
Nasals	m	n			ŋ		
	m̥	n̥					
Liquids		l	ɭ				
		l̥	ɭ̥				
Approximants	w			j			

Figure 2.2 Wādū Pūmī consonant phonemes (Daudey, 2014, p. 19)

The place of articulation label "cavity" is used by Daudey for two fricatives that may be freely realized at the velum, uvula, or glottis. Wadu Prinmi stops and affricates contrast by voicing and aspiration, while fricatives, nasals, and liquids contrast by voicing. Palatalization is phonetically realized with bilabial and alveolar stops, nasals, and liquids, alveopalatal consonants, cavity fricatives, and velar stops and nasals in specific contexts.

Wadu Pumi has seven oral vowels, /i/, /ɯ/, /u/, /ə/, /æ/, /ɐ/, and /ɑ/; four nasal vowels, /ĩ/, /ẽ/, /õ/, and /ã/; and three diphthongs, /ej/, /ej/, and /aw/.

		Front		Central		Back	
		-round	+round	-round	+round	-round	+round
High	Oral	i			ɯ		u
	Nasal	ĩ					
Mid	Oral	ej [e] / (ej)			ə		aw [o]
	Nasal	ẽ					õ
Low	Oral	æ			ɐ		ɑ
	Nasal	ã					

Figure 2.3 Composite chart of vowels (Daudey, 2014, p. 47)

/aw/ and /ej/ developed from /o/ and /e/, respectively, and are still realized /o/ and /e/ in specific contexts and other varieties of Northern Prinmi. Daudey also notes that her main consultant considers the phoneme Daudey labels as /ɐ/ to be farther back, and therefore transcribes it as /ʌ/. Due to phonological processes, Daudey still considers the underlying form to be /ɐ/.

In terms of syllable structure, Wadu Pumi is characterized by the pattern (C_1)(G)V(G) (Daudey, 2014). Optional C_1 can be any consonant, the first G is an optional medial /w/, V is obligatory and can be any vowel, and the second G is an optional

/w/ or /j/ offglide present in diphthongs. Daudey considers palatalization to be an optional feature of the initial consonant and tone to be an optional feature of the vowel.

Daudey describes Wadu Pumi as possessing a complex tonal system with tone spreading, tone sandhi, and several phrasal intonation processes (2014). However, it only possesses four underlying lexical tones: a high-level tone, a high-falling tone, a low-level tone, and a low-rising tone. Moreover, an extra high high-level tone is realized intonationally when speakers intend to intensify a specific word. As mentioned above, syllables may also be toneless.

2.2 Documentation Context

2.2.1 Description of Collection

This study's data is specifically drawn from the corpus of Northern Prinmi oral art collected by Henriëtte Daudey and Gerong Pincuo in 2017 and hosted by the Archive of Endangered Languages (ELAR) (2018). The collection, “Documentation of Northern Prinmi oral art, with a special focus on ritual speech,” contains over 24 hours of audio and video recordings, primarily documenting rituals but also containing songs and folktales. 39 speakers—12 females and 27 males—were recorded in 12 different locations. Recordings were sampled at 48000 HZ. Daudey and Pincuo (2018) define ritual speech as “speech that is addressed to the supernatural world,” and Northern Prinmi rituals are typically delivered in a chant intonation. These rituals are linked with both indigenous Prinmi religious practices and Tibetan Buddhism. The other forms of oral art in the collection are primarily distinguished by not being addressed to supernatural entities. The vast majority of recordings in the collection are either rituals or songs, meaning that most speech in the

collection is chanted or sung rather than spoken conversationally. Some recordings also include instrumental music, and many have ambient noise. Daudey and Pincuo are analyzing this collection to characterize Northern Primi ritual speech, exploring its genre conventions and grammatical structures like exhaustive constructions (2020).

Of the 24 hours recorded, 3 hours of audio were annotated with time-aligned IPA transcriptions. Pincuo transcribed the recordings and provided Chinese translations, while Daudey glossed them, provided English translations, and time aligned them using ELAN (2018). Daudey and Pincuo state that they transcribed recordings from the Wenquan area phonemically and recordings from other regions phonetically, which I interpret respectively as “broad” and “narrow” transcription.

For the purposes of this study, I filtered the collection by transcription status and created a list of all recordings with time-aligned transcriptions. This resulted in a much smaller corpus containing 34 recordings of 15 speakers from 8 locations, for a total of 187 minutes of audio. These locations include the Jiulong, Ninglang, Shuiluo, and Yanyuan townships of Sichuan, and the Jiaze, Mudiqing, Tuodian, and Wenquan townships of Yunnan.

From ELAR's web interface, I downloaded the ELAN file for each transcribed recording. As only MP3 versions of the audio could be downloaded through the ELAR web interface, I requested and was provided with uncompressed WAV files of each transcribed recording. Obtaining WAV files was necessary for proper time-alignment between the audio and ELAN transcription tiers and provided the added benefit of higher audio quality.

Recordings range in length from 1.5 minutes to almost 15 minutes. Most recordings have one to three speakers: the participant performing the ritual or song—who typically speaks continuously with little interruption for the majority of the recording time—and one or both of the two researchers. Only one recording is described as a performance of two individuals and includes one participant chanting and another participant singing. Most speech in the corpus is non-overlapping, as very little of the audio contains conversations.

Each ELAN transcription file—in the EAF file format—includes several transcription tiers for each speaker. These vary across recordings, but all recordings include Northern Prinmi phrasal transcriptions, English phrasal translations, and Chinese phrasal translations. Some also include word-level and morpheme-level glosses in Northern Prinmi, English, and Chinese.

2.2.2 Survey of Phones and Orthography

As this corpus represents multiple varieties of Northern Prinmi—rather than solely Wadu Pumi—I conducted additional analysis of the transcriptions to identify which classes of sounds were actually realized during data collection. This analysis was conducted using regex functions in Python (Van Rossum & Drake, 2009), and identified 44 consonants, five consonant clusters, 15 vowels, four diphthongs, and five tones (although one of these is likely transcription error).

This sound inventory may differ from that identified by Daudey's 2014 grammar of Wadu Pumi for several reasons. First, the collection includes varieties spoken in other locations and regions, and so may include different phonemes altogether. Second, regardless of the phonologies involved, as mentioned above recordings made outside of Wenquan Township were transcribed narrowly rather than broadly, and so even if the

phonemes are the same these transcriptions may also include allophones. Third, even with their broad transcriptions of Wenquan Township recordings, the researchers applied a different orthographic scheme from Daudey's 2014 grammar when transcribing data for the oral art collection. I provide a table below of the likely correspondences between the corpus and grammar's orthographic representations of Northern Prinmi phones.

Table 2.1 Orthographic Representations of Northern Prinmi Phone Inventory

Consonants	Bilabial	Labiodental	Alveolar	Retroflex	Alveopalatal	Velar	Uvular	Cavity
Stops	p b		t d	ʈ ɖ		k g	q	
	p ^h		t ^h	ʈ ^h		k ^h	q ^h	
Affricates			ts dz	ʈʂ ɖʐ	tɕ dʒ			
			tʂ ^h	ʈʂ ^h	tɕ ^h			
Fricatives		f	s z	ʂ ʐ	ɕ ʒ	x		(h)
			s ^h	ʂ ^h	ɕ ^h			ɦ
Lateral fricative			ɬ					
Nasals	ɱ m		ɳ n			ŋ		
Liquids			ɹ (ɻ) r (ɹ) l					
			(l)					
Approximates	w					j		

Vowels		Front		Central		Back	
		-round	round	-round	round	-round	round
High	Oral	i		i	(u)	u	u
	Nasal	ĩ		ĩ			
Mid	Oral			ə		ʌ (ɐ)	o
	Nasal	(ẽ)		ẽ			õ
Low	Oral	æ				ɑ	
	Nasal	ã				ã	

Diphthongs	Front	Central	Back
	ei (ej)	əu	au (aw)

Tones	Description of tone	Relative Pitch
	High level	55 (44)
	High falling	51 (54)
	Low level	22
	Low rising	35 (24)

Key	
Shared representation	character
Shared but different	2018 character (2014 character)
Only in Daudey & Pincuo (2018)	2018 character
Only in Daudey (2014)	(2014 character)

CHAPTER 3. METHODS

3.1 Model Selection

Only a small corpus of Northern Prinmi written texts is available. In total, ten recordings totaling 42 minutes are transcribed in a Pangloss corpus contributed to by Daudey and Guillaume Jacques (2011), three texts are included in the appendix of Daudey’s (2014) grammar, and 36 recordings are transcribed in the oral art collection (Daudey & Pincuo, 2018). Of these, the transcribed recordings provide the most textual data, and they include around 28,000 words and 3,500 word types. This means that any Northern Prinmi language model produced from currently collected texts would be quite small.

While it is hard to pinpoint exactly what size of written corpus is necessary for a useful language model for ASR, one study (Pellegrini & Lamel, 2008) examining the relative impact of audio versus textual data for HMM ASR models used a 10,000-word corpus with 7,000 types as its smallest textual data testing condition. Although this paper was looking at Amharic, an agglutinative language, this implies 3,500 types is a very small lexicon for HMM ASR.

Another study working on ASR in a documentation context suggests that “a few tens of thousands of words... [is] an insufficient amount to train a language model according to standard workflows” (Guillaume, Wisniewski, Macaire, et al., 2022, p. 172). While some of the same authors from that study went on to successfully incorporate a small language model to improve the performance of XLSR-53 (Guillaume, Wisniewski, Galliot, et al., 2022), it is unclear how useful such a small lexicon and language model would be for HMM ASR.

Due to this lack of audio and textual data, I decided to use an end-to-end transcription model which would bypass the need for a lexicon and text corpus. I selected XLSR-53 for use with the Northern Prinmi corpus due to its successful applications in settings with very little audio data (Boulianne, 2022) and on documentation tasks (Coto-Solano et al., 2022; Guillaume, Wisniewski, Galliot, et al., 2022; Guillaume, Wisniewski, Macaire, et al., 2022; Nowakowski et al., 2023).

3.2 Data Preprocessing

3.2.1 Partitioning the Dataset

As in machine learning more generally, it is best practice to use a training set and testing set with XLSR-53; the model is trained (or fine-tuned) on the larger training set and then tested on the smaller testing set to confirm the patterns it learned during training actually generalize to other data.

I decided on a 90% training, 10% testing split, selecting 8 out of the 34 recordings to serve as the testing data, resulting in ~167 minutes of training data and ~20 minutes of testing data. The eight testing recordings were heuristically selected to maximize variance within each dataset while ensuring that both datasets were still broadly reflective of one another.

In order to maximize testing set diversity, testing set recordings were in part selected by length so that there could be more total recordings—with more distinct speakers and topics—included in the testing dataset. The average testing recording is ~2.5 minutes, while the average training recording is ~6.4 minutes. Moreover, each recording in the testing set has a different primary speaker and topic. I did not segregate speakers entirely

by set—some speakers are present in both sets—because Liu et al. (2022) suggest that speaker segregation does not improve ASR performance on under-resourced languages.

In terms of other characteristics, such as genre, region, and postal address, I sought to balance the properties of the two sets. Song recordings make up 20% of the training set minutes and ~36% of the testing set minutes. Recordings from Sichuan make up ~16% of the training data minutes and ~18% of the testing data minutes. In terms of more granular location information, seven of eight recording location postal addresses are included in the training set, while four of eight are included in the testing set, with all but one of the testing set locations also present in the training set.

Table 3.1 Summary of Data Partition Properties

	Training Set	Testing Set
% of Recordings	90%	10%
# of Recordings	26	8
Total Minutes	167	20
Average Minutes	6.4	2.5
Song Recordings % (of Total Min.)	20%	36%
Sichuan Recordings % (of Total Min.)	16%	18%
# of Postal Addresses	7	4

3.2.2 Phrasal Segmentation of the Data

After partitioning the data through the process described above, I used Python (Van Rossum & Drake, 2009) to preprocess the audio and text data and load and fine-tune the XLSR-53 ASR model. Preprocessing consisted of two steps: phrasal segmentation of the audio and transcription data, described in this subsection, and orthographic preprocessing of the transcription data, described in the subsection below.

The XLSR-53 model is built to process short audio segments of roughly one phrase at a time, using a short audio input to generate a transcription text prediction. Therefore, in

order to apply the XLSR-53 model to the corpus, each audio and transcription file required segmentation by phrase.

As the corpus is almost entirely composed of recorded rituals and songs primarily sung or chanted by one speaker, I decided to segment the audio and transcription files by the time-alignments of the primary speaker's Northern Prinmi phrase transcriptions. I wrote a Python script to accomplish this task using the `pypmi` and `librosa` python libraries (Lubbers & Torreira, 2013; McFee et al., 2022). I designed the script to iterate through the data directory for the training and testing datasets separately, using `librosa` to load each WAV file and `pypmi` to load each ELAN EAF transcription file. It identified the primary speaker's transcription tier by iterating through each recording's Northern Prinmi phrasal tiers, returning the tier with the highest number of phrasal annotations.

Using this primary tier's phrases and time stamps, it created a list of phrasal segments for each recording, each entry in the list possessing its file of origin, segment number, start time, end time, Northern Prinmi transcript, and corresponding segment of audio. The script concatenated these segment lists together into one larger list for the training and testing sets separately, converted those lists into Pandas dataframes (McKinney, 2010; The pandas development team, 2022), and then converted those dataframes into two HuggingFace datasets for further preprocessing and use with the XLSR-53 model.

3.2.3 Preprocessing the Transcription Data

In this section, I describe the methods by which I preprocessed the transcription data to both increase model efficacy and answer my second research question on how

different approaches to label segmentation preprocessing perform with XLSR-53 in the context of Northern Prinmi, a low-resource tonal language.

All transcription data was first preprocessed using a Python function adapted from von Platen (2021) to remove special characters, including punctuation. All punctuation characters were removed aside from the hyphen (“-”) character, which was used in the data to mark morpheme boundaries. I chose to leave the hyphen in the training data for all models, as previous work has shown the XLSR-53 model can perform well even while predicting punctuation (Guillaume, Wisniewski, Macaire, et al., 2022); given the hyphen’s morphological orthographic function, the trade-off in accuracy versus aid to word identification seemed worthwhile.

I adapted this Python function to also correct several tonal transcription errors that existed in the transcription data. The tone superscript sequence “⁵⁵⁵” occurred in some transcriptions, and these sequences were converted to “⁵⁵,” the closest licit tone sequence. Individual tone superscripts also occurred in some transcriptions. Unlike the “⁵⁵⁵” sequences, these could not be predictably corrected, and were instead removed. These corrections were made to avoid training any of the models to predict single or triple tone superscript sequences.

Transcription is a theoretically fraught process, deeply interconnected with phonetic and phonological arguments. Given that the XLSR-53 model is fine-tuned for ASR by learning associations between speech representations and character sequences which act as labels, the transcription conventions used to produce the model’s output labels may have a sizable effect on the model’s efficacy. In the HuggingFace tutorial for finetuning XLSR-53 (2021), von Platen notes that it is typical to use letters as labels for

CTC loss; i.e., using a one-for-one correspondence between unique characters in the transcription data and labels the model will use for prediction. The transcription data in the Northern Prinmi Oral Art corpus is written with one set of orthographic conventions, but it is possible to preprocess the transcriptions into orthographically distinct target transcription sets to test how different label segmentation schemes impact automatic transcription accuracy.

I conducted this further preprocessing of the transcription data to explore how differing transcriptions—and therefore segmentations—of lexical tones and multi-character phonemes impacted model performance. As discussed in the literature review above, differing strategies for segmenting lexical tone and grapheme clusters into predictive labels have been explored by several previous researchers (Adams et al., 2017, 2018; Guillaume, Wisniewski, Macaire, et al., 2022; Wisniewski et al., 2020).

I wrote two Python functions to perform this preprocessing. One function combines each pair of tone characters—two superscript numbers—into a one-character placeholder intended to represent a lexical tone contour. The other combines a character and all of its modifying diacritics into a single character placeholder intended to represent an individual phone; notably, this function does not combine affricates. Both functions relied on manually collected lists of the character sequences to combine, identified using regex searches of tone superscripts and diacritic characters respectively.

I applied these two functions in four different ways to the transcription data to allow for the fine-tuning of four XLSR-53 models with different transcription target labels. The first function was applied to produce a solely tone-character-combined target transcription (hereafter referred to as tones-combined) set, echoing the approach of Wisniewski et al.

(2020) in which each Unicode character that is not a modifier receives its own label. I applied the second function to produce a solely character-diacritics-combined target (hereafter referred to as cluster-combined) set, without combining tone superscripts. Both functions were applied to produce a target transcription set (hereafter referred to as the both-combined set) in which both tone superscripts and character-diacritic clusters were combined, following the approach of Adams et al. (2017, 2018) in segmenting by phoneme and considering multi-character renderings of lexical tones as single segments. The original target transcription set without combinations was also preserved to be used as a control.

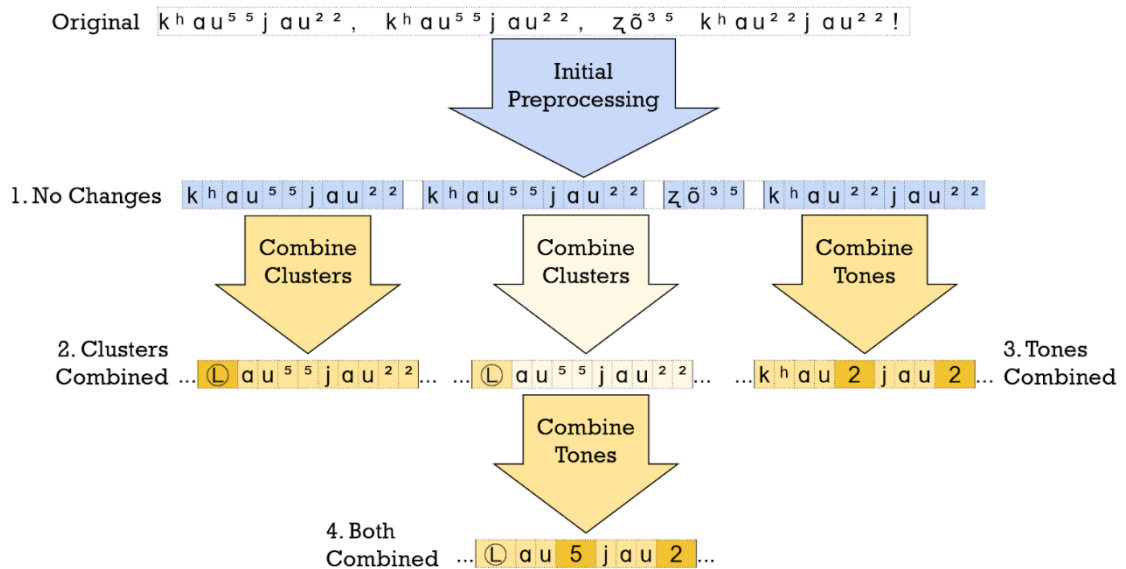


Figure 3.1 Example Output and Segmentation by Preprocessing Method

3.3 Model Fine-Tuning

I fine-tuned four separate XLSR-53 models on the target datasets generated via the preprocessing functions described in the previous function. Fine-tuning was conducted on several nodes of the Lipscomb Compute Cluster equipped with Nvidia V100 GPUs, provided by the University of Kentucky Center for Computational Sciences and

Information Technology Services Research Computing. XLSR-53 was accessed and fine-tuned through the HuggingFace Transformers Python library (Wolf et al., 2020).

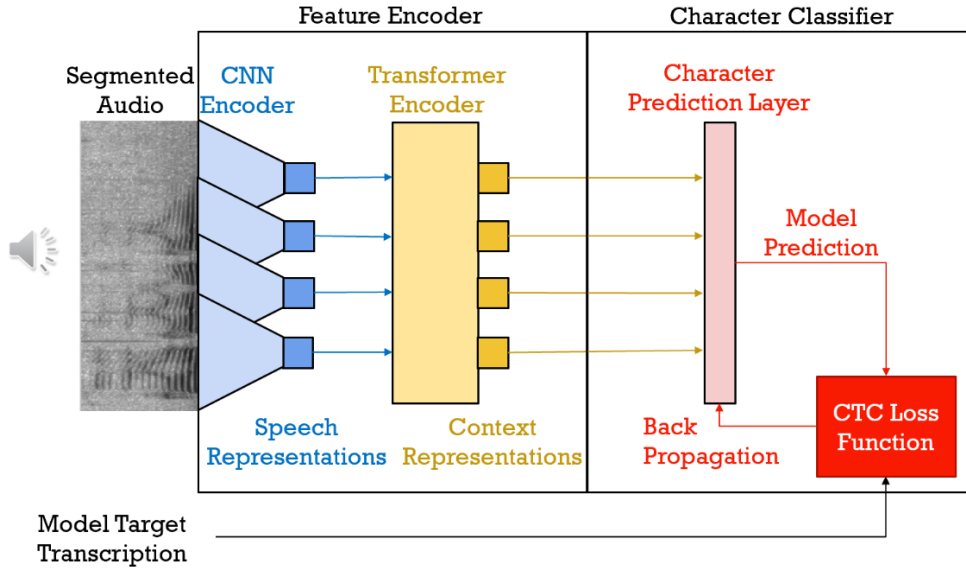


Figure 3.2 XLSR-53 Fine-Tuning Schematic

I adapted the Python script used to load and fine-tune the models from HuggingFace’s XLSR-53 fine-tuning tutorial by Patrick von Platen (2021). I decided to not change the training parameters; von Platen describes setting them heuristically without detailing a process for adapting them to a different fine-tuning training set.

Table 3.2 Fine-Tuning Parameters

Training Parameters	Value	CTC Model Parameters	Value
Per device training batch size	1	Attention dropout	0.1
Gradient accumulation steps	2	Hidden dropout	0.1
Epochs	30	Feature projection dropout	0
Mixed precision	TRUE	Mask time probability	0.05
Learning rate	3E-04	Layerdrop	0.1
Evaluation steps	100	CTC loss reduction	mean
Warmup steps	500		

Following von Platen (2021), I used word error rate (WER) as the evaluation metric during fine-tuning, calculated periodically to assess how well the process was going. WER—and character error rate (CER), which I use in the results section—are described in Figures 1 and 2 below. In these formulas, S stands for substitutions, D for deletions, I for insertions, and H for hits (correct matches). Both calculations compare a reference sentence (or string of characters) with a hypothesis sentence, aligning the two sentences and counting the number of insertions, deletions, substitutions, and hits (correct predictions).

$$WER = \frac{S + D + I}{H + S + D} \qquad CER = \frac{S + D + I}{H + S + D}$$

Figure 3.3 Word and Character Error Rate Formulas

WER evaluation during training was calculated comparing the specific model's relevant preprocessed target transcriptions as the reference with the model's direct predicted transcriptions as the hypothesis. It was calculated using the HuggingFace dataset library WER metric function.

CHAPTER 4. RESULTS

4.1 WER and CER

After fine-tuning each model, I used the Python library JiWER (Vaessen, 2022) to calculate each of the four model's WER and CER on both the full testing set and the province and genre subsections. Prior to evaluation, the predictions of the cluster-combined, tones-combined, and both-combined models were converted back to the original orthographic conventions. This decision had two core motivations. First, reconversion allowed for all error rates to be comparable to one another, calculated on the same character vocabulary. Second, to be maximally useful, the models should directly output predictions with the same orthographic conventions as the original transcriptions; calculating error from reconverted predictions evaluates the models' abilities to do this. Therefore, the WER and CER reported here are not calculated in the same fashion as the WER calculated during fine-tuning. This process is displayed in Figures 4.1 and 4.2 below.

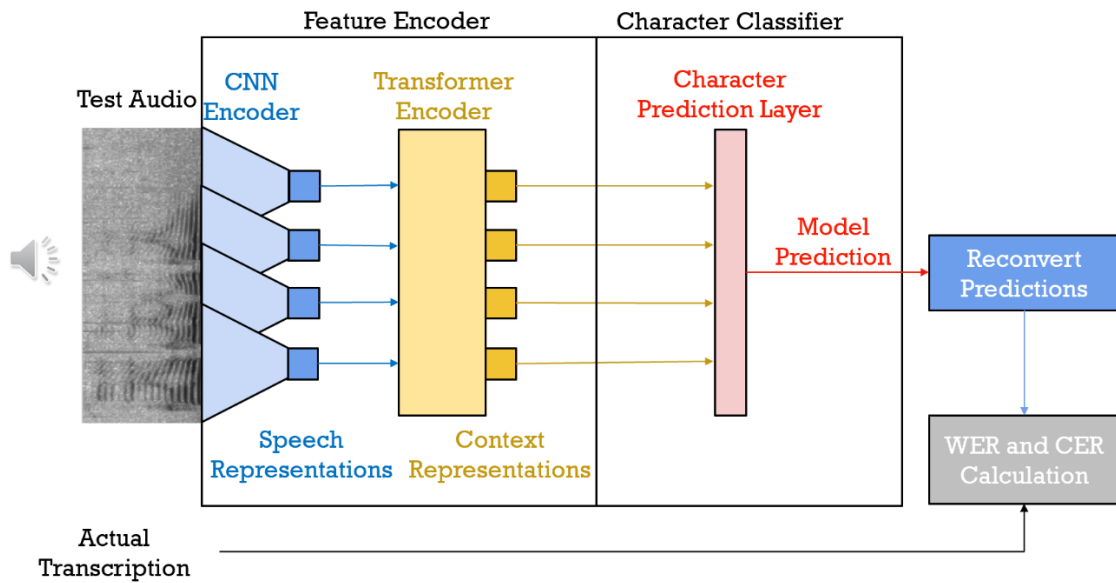


Figure 4.1 XLSR-53 Evaluation Schematic

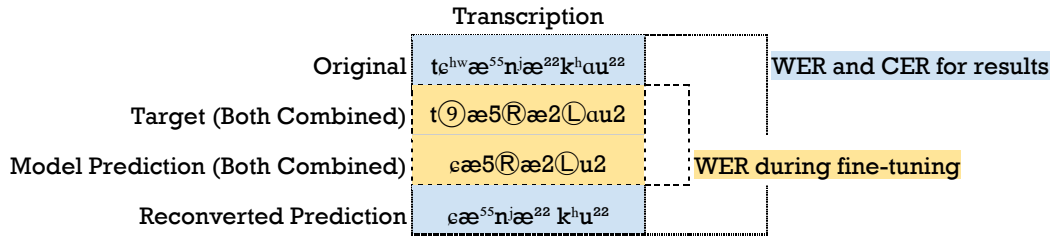


Figure 4.2 Explanation of WER and CER Calculation

The WER and CER results are displayed in Tables 4.1 and 4.2 below, with the lowest (best) score on each subsection of the testing data bolded. Red backgrounds indicate higher rates, and white backgrounds indicate lower rates.

Table 4.1 Model WER by Testing Set Subsection

Testing Set Section		% of Training Data	Model WER			
			No Changes	Cluster Combined	Tones Combined	Both Combined
Full	NA	1.003	0.979	0.997	0.987	
Yunnan	84%	1.002	0.971	0.979	0.963	
Sichuan	16%	1.005	1.009	1.064	1.075	
Rituals	80%	1.010	0.977	0.997	0.990	
Songs	20%	0.972	0.987	0.995	0.976	

Table 4.2 Model CER by Testing Set Subsection

Testing Set Section		% of Training Data	Model CER			
			No Changes	Cluster Combined	Tones Combined	Both Combined
Full	NA	0.479	0.481	0.441	0.433	
Yunnan	84%	0.433	0.435	0.389	0.378	
Sichuan	16%	0.652	0.655	0.636	0.642	
Rituals	80%	0.487	0.482	0.444	0.439	
Songs	20%	0.443	0.476	0.426	0.404	

Comparing individual models on the full test set, the model fine-tuned with combined character clusters (clusters combined) achieves the lowest WER, while the model fine-tuned with both character-diacritic clusters and tone character pairs combined

(both combined) achieves the lowest CER, and an almost as low WER. No single model performs best on every section of the data.

Combining tone character pairs is the preprocessing method which most improves model CER performance, by roughly 0.03 pts without cluster combination and 0.04 pts with cluster combination. The both-combined model has the lowest CER across every subset except for recordings from Sichuan (on which it is only slightly worse than the tones-combined model). However, it does not have consistently better WERs than the other models, and its CER rates are only slightly better than the tones-combined model, which implies the majority of its performance gains come from the combination of tones.

Preprocessing to combine character-diacritic clusters has mixed results. The two models which did not combine character clusters (the no changes and tones-combined models) performed only slightly worse than the two models which did combine clusters, about a 0.01 CER difference and a 0.02 WER difference on the full testing set. The clusters-combined model actually has a worse CER than the no changes model on the full testing set.

The models perform better on certain subsections of the testing data than others, and this patterning is not completely representative of which subsections are most reflected in the training data. All models perform substantially better on recordings from Yunnan than Sichuan, and all but one perform slightly better on songs than rituals. This second result is especially surprising considering that recordings of songs only make up 20% of the training data. The one exception to this is the clusters-combined model, which has a slightly better (lower by 0.01) WER on rituals than songs.

4.2 Substitution Rates

In order to analyze what kinds of errors each model was making, I wrote a script using the Python library Levenshtein (Bachmann, 2022) to count how many times each model incorrectly substituted one character label for another in its prediction (for example, predicting the character “t” instead of “d”). These substitutions were calculated comparing the no changes transcription data with the reconverted predictions.

I used these counts to produce substitution tables, aggregating how many times each model mistakenly chose a different label instead of the correct target label. I focused on substitutions instead of insertions and deletions because, in the case of substitutions, the relationship between the target label and the mistakenly predicted label may provide evidence for why the error occurred. I display a summary of this data in Table 4.3 below, showing each model’s percentage of substitutions for consonant, vowel, diacritic, and tone characters. Each percent is equal to the sum of times each target character of the given type was mistakenly predicted to be a different character over the number of times the target character type occurred in the testing set.

Table 4.3 Model Substitutions by Character Class

Character Class	# in Testing Set	Model			
		No Changes	Clusters Combined	Tones Combined	Both Combined
Tones	13,554	21.43%	21.76%	21.03%	20.69%
Vowels	7,658	30.77%	28.36%	31.84%	29.86%
Consonants	7,390	31.88%	33.55%	31.76%	31.53%
Diacritics	1,868	23.98%	30.35%	28.80%	24.63%

The both-combined model makes the fewest tone and consonant substitution errors, while the clusters-combined model makes the fewest vowel substitution errors, and the no changes model makes the fewest diacritic substitution errors.

CHAPTER 5. DISCUSSION

5.1 Effects of Preprocessing

These findings suggest that the default XLSR-53 approach of segmenting characters one-to-one into prediction labels is not untenable, as there are only modest gains for both WER and CER by combining characters with their diacritics. This replicates the findings of Wisniewski et al. (2020), whose one-to-one character to label segmentation also performed only marginally worse than language-specific phone segmentation. Wisniewski et al. interpret this as being a general product of machine learning systems’ abilities to learn patterns without exploring any specific mechanisms.

However, the fact that there are any gains at all from combining character-diacritic clusters is somewhat surprising, as it increases the number of labels the ASR model must choose between when transcribing a frame of audio. The following table displays each model’s number of prediction labels.

Table 5.1 Prediction Label Vocabulary by Model

	Model			
	No changes	Clusters Combined	Tones Combined	Both Combined
Prediction Labels	54	106	59	111

The clusters-combined model has almost twice as many prediction labels as the no changes model yet performs at roughly the same level. The tones-combined model has slightly more labels than the no changes model yet still performs better, and the both-combined model outperforms all other models, yet by far has the most prediction labels. While this analysis ignores the frequency of each label in the training and testing sets, it nonetheless does suggest that cluster-based labels must be providing the model with some additional information that is useful for prediction—possibly something related to

language-specific phonotactics—because otherwise adding more labels would cause the models to perform more poorly.

The question therefore becomes if the slight gains provided by language specific prediction labels more closely approximating phonemes is worth the additional time required to manually produce them. If a member of the documentation project is involved in the automatic transcription project, this question is easier to answer, as their familiarity with the transcription conventions will make producing language-and-project specific prediction labels relatively easy. Even if members of the original documentation project are not directly involved, depending on the particular documentation project and how much recording data is actually available, these modest gains may be worth the time necessary to implement more language specific prediction labels.

The positive impact of using one prediction label per lexical tone unit was a much more expected result, for several reasons. The first is that both Adams et al. (2017, 2018) and Wisniewski et al. (2020) combined lexical tone character pairs into single predictive labels in their projects and had relatively high degrees of success, which at bare minimum made it clear the approach was feasible with other neural network-based ASR models.

Moreover, at least in the case of Northern Prinmi, there are the same number of licit lexical tones as there are individual characters combined to represent those lexical tones. Northern Prinmi has low level, low rising, high level, and high falling tones, and these tones are written in the corpus respectively as ²², ³⁵, ⁵⁵, and ⁵¹, composed orthographically by combinations of ¹, ², ³, and ⁵. Therefore, combining tone character pairs does not meaningfully increase the number of prediction labels, but does make it impossible for the

model to produce illicit tones; e.g., the no changes model could produce the illicit tone digraph sequence “¹⁵,” whereas the tones combined or both-combined model could not.

This suggests that when applying XLSR-53 to tonal languages, it is worthwhile to ensure that each lexical tone receives one prediction label, even if it is typically written with multiple characters. One caveat is that Northern Prinmi has relatively few lexical tones, so it is possible the benefits of this approach would not scale to languages with substantially more tones, especially as there would be a disproportionate increase in individual tone labels compared to component characters used to write tones.

5.2 Model Performance on the Corpus

5.2.1 Model Comparison

Comparing all the XLSR-53 Northern Prinmi models with other applications of ASR to language documentation, their performance is relatively poor, especially in comparison with other implementations of the XLSR-53 model.

Table 5.2 Comparison of Closely Related Projects’ WER and CER

	Language	Model	Transcribed Audio	WER	CER
This project	Northern Prinmi	XLSR-53	3h 7m	0.987	0.433
Guillaume et al. (2022)	Japhug	XLSR-53	3h	0.267	0.086
Adams et al. (2021)	Japhug	ESPnet	2h 50m	--	0.128
Wisniewski et al. (2020)	Yongning Na	Persephone	7h 49m	--	0.186

Two overlapping research teams (Guillaume, Wisniewski, Galliot, et al., 2022; Guillaume, Wisniewski, Macaire, et al., 2022) fine-tuned an XLSR-53 model on language documentation recordings of Japhug, another Sino-Tibetan language. They fine-tuned multiple XLSR-53 models on different quantities of audio data, including a three-hour corpus to which they apply a 90/10 training/testing split. With three hours of transcribed

audio data, Guillaume et al. achieved an average CER of 0.086 and an average WER of 0.267. These scores are substantially better than my best model, which only achieved a CER of 0.433 and a WER of 0.987. Notably, however, there are several crucial differences between the two projects.

First, Japhug is not a tonal language, and its basic one-to-one character to prediction label vocabulary consists of only 44 characters, 10 fewer than Northern Prinmi. Not having to predict lexical tones simplifies the ASR process, and the smaller vocabulary size may also help. Second, Guillaume et al.'s testing set consisted solely of narratives—rather than including sung or chanted speech—a genre of speech generally easier for ASR and specifically more similar to the read speech XLSR-53 is primarily trained on.

Third, Guillaume et al. tuned their model's hyper-parameters specifically for their dataset, training the model 91 separate times to discover the optimal values. Most of the models they trained in this process performed worse than their best model, although their worst model's CER was 0.288, still quite a bit better than my best model's CER of 0.433. The final difference, and the only one which should have disadvantaged them, is that Guillaume et al. did not evaluate their models' performance on manually phrasally segmented testing data. They instead used silence detection to segment the recordings in order to simulate a fully automatic transcription pipeline.

All of these comparisons provide different possible causes for the poor performance of the Northern Prinmi models: Northern Prinmi's tonality, the corpus' number of orthographic characters and genres, and the models' lack of hyper-parameter optimization. However, there are several additional potential causes that are not made explicit by such a comparison.

Daudey’s 2014 grammar gestures at the fact that there is no standard Northern Prinmi and that each village may speak a slightly different variety. Paired with the fact that Daudey and Pincuo (2018) mention that they transcribed recordings broadly in the Wenquan area and narrowly other places, the collection may actually be representative of up to eight distinct varieties of Prinmi speech, seven of which are narrowly transcribed. It is possible, therefore, that this level of speech variety diversity is too high for XLSR-53, or at least that the differences in transcription may be confusing to the model. Indeed, depending on the differences between the broad and narrow transcriptions, it is possible that transcription differences alone could contribute a substantial amount of the models’ errors. Amith et al. (2021), working with Yoloxóchitl Mixtec, describe facing similar challenges with intralanguage variation, although notably their models fared much better, with a CER as low as 0.195 on 10 hours of training data.

The audio quality of the recordings may also be a problem, another challenge for ASR in documentation contexts named by Amith et al. and Cavar et al. (2021; 2016). All of the recordings are chanted or sung, with some of the chants being incredibly high tempo, presumably making both human and machine transcription much more difficult. Alongside the vocal quality of the participants, several of the recordings include musical instruments, an additional challenge for signal recognition.

5.2.2 What is a Useful Error Rate?

Amith et al. (2021) claim that reductions in CER are beneficial down to a CER of 0.06-0.10, as predictions made with this CER take as long to correct as to simply review. While this provides a floor for utility, the literature is less clear on a ceiling. Jimerson et al. (2018) state that a WER of 0.95 is probably not useful, while implying a WER of ~0.70

may be useful. Zahrer et al. (2020, p. 2898) describe their model’s phoneme error rate (PER, roughly the equivalent of CER) result of 0.437 as not useful, noting that any PER/CER around 0.50 translates to correcting roughly half of all characters. As my best model achieves a CER of 0.433 and a WER of 0.987, there is definite room for improvement.

However, looking solely at error rate metrics may be misleading. Guillaume, Wisniewski, and Macaire et al. (2022) conducted a qualitative analysis of the correction process, passing their model’s predictions to the documentation linguist who transcribed the original data, and found that the linguist made fewer corrections to the predictions than the predictions’ CER suggested. While they were working with a relatively low CER (0.074), it is possible this observation could generalize to other contexts.

5.3 Future Directions

Most of the issues described could be targeted and potentially ameliorated in future work. The lack of hyper-parameter optimization is the most immediately obvious, requiring the fewest modifications to the project’s methods. Guillaume, Wisniewski, and Macaire et al. (2022) saw a difference of 0.20 points in CER between their best and worst optimized models, a huge improvement obtained solely through searching the parameter space.

Training on different partitions of the data could also improve model performance, as the model performed substantially better on specific recordings and genres of recordings than others. Training solely on the recordings from Wenquan which are all transcribed broadly could increase model consistency, and training on the less acoustically challenging recordings could provide the model better baseline speech recognition performance.

Assuming that the models' performance can be significantly improved, I would also like to create an easier XLSR-53 pipeline for implementing it with other documentation projects. This would involve building a front end or command line interface to conduct the steps I describe in this project: preprocessing audio and transcripts from ELAN files, partitioning the data, segmenting prediction labels from the transcriptions, fine-tuning the model on the partitioned and preprocessed data, and evaluating the model's performance on the testing data. Moreover, to actually make it useful, I would also need to integrate speaker diarization and silence recognition for preprocessing untranscribed audio recordings so the model could actually be applied.

5.4 Applications and Implications

While not resulting in a high-performance, immediately applicable ASR model for actual documentation use, this attempt does illustrate the increasing accessibility of powerful and innovative speech and language technologies to non-specialists. It also demonstrates that there is still significant work to be done, both in increasing accessibility even further and in tailoring these tools to better suit work on language documentation.

Substantial work has been done to make language technology better serve the needs of documentation workers and communities, with the development of toolsets and pipelines like ELPIS (Foley et al., 2018) and ESPnet (Watanabe et al., 2018). However, to my knowledge, transformer-based models have not yet been similarly integrated. While HuggingFace makes these models much more accessible, the HuggingFace API is far from being optimally designed for the needs of language documentation practitioners. Work to

integrate transformer-based models into either an existing or new non-specialist documentation pipeline would likely be a valuable contribution.

In this instance, even leveraging a recent and high performing transformer-based ASR model, automatic transcription of culturally significant genres for documentation—teachings on country and traditional knowledge conveyed through songs and rituals (Bird, 2020)—was still a major challenge. General computational improvements to the underlying models may help alleviate these problems, but more specific work on ASR for non-conversational or read speech would also be beneficial. Moreover, less described languages may have linguistic features that are highly distinct from the languages for which most ASR methods are developed (Guillaume, Wisniewski, Macaire, et al., 2022), so special attention must be paid to preprocessing and adapting the ASR model to the specific needs of each language.

Overarchingly, there are many reasons for optimism in this area; computational tools continue to improve, and many vital conversations are now taking place between ASR specialists, documentation workers, and language communities (Amith et al., 2021; Bird, 2020; Guillaume, Wisniewski, Galliot, et al., 2022; Guillaume, Wisniewski, Macaire, et al., 2022; Zahrer et al., 2020). This attempt sought to synthesize some of these diverse perspectives, and in the process, made manifest many of this approach’s promises and perils.

REFERENCES

- Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., & Michaud, A. (2018). *Evaluating phonemic transcription of low-resource tonal languages for language documentation*. 3356. <https://shs.hal.science/halshs-01709648>
- Adams, O., Cohn, T., Neubig, G., & Michaud, A. (2017). Phonemic Transcription of Low-Resource Tonal Languages. *Proceedings of the Australasian Language Technology Association Workshop 2017*, 53–60. <https://aclanthology.org/U17-1006>
- Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., & Hill, N. (2021). *User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis* (arXiv:2101.03027). arXiv. <https://doi.org/10.48550/arXiv.2101.03027>
- Amith, J. D., Shi, J., & Castillo Garcia, R. (2021). End-to-End Automatic Speech Recognition: Its Impact on the Workflow for Documenting Yoloxóchtli Mixtec. *First Workshop on NLP for Indigenous Languages of the Americas. 11 June 2021*. <https://www.aclweb.org/anthology/2021.americasnlp-1.8.pdf>. <https://par.nsf.gov/biblio/10281120-end-end-automatic-speech-recognition-its-impact-workflow-documenting-yoloxochitl-mixtec>
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common Voice: A Massively-Multilingual Speech Corpus* (arXiv:1912.06670). arXiv. <http://arxiv.org/abs/1912.06670>
- Bachmann, M. (2022). *Levenshtein* (0.20.9) [Python]. <https://github.com/maxbachmann/Levenshtein>
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460. <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>
- Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56, 85–100. <https://doi.org/10.1016/j.specom.2013.07.008>
- Bird, S. (2020). Decolonising Speech and Language Technology. *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519. <https://doi.org/10.18653/v1/2020.coling-main.313>
- Boulianne, G. (2022). Phoneme transcription of endangered languages: An evaluation of recent ASR architectures in the single speaker scenario. *Findings of the Association for Computational Linguistics: ACL 2022*, 2301–2308. <https://doi.org/10.18653/v1/2022.findings-acl.180>
- Ćavar, M., Ćavar, D., & Cruz, H. (2016). Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4004–4011. <https://aclanthology.org/L16-1632>

- Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). *Unsupervised Cross-lingual Representation Learning for Speech Recognition* (arXiv:2006.13979). arXiv. <https://doi.org/10.48550/arXiv.2006.13979>
- Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., Koka'ua, L., Tanveer, S., & Feldman, I. (2022, June 20). Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. *Language Resources and Evaluation (LREC) Conference 2022*. <https://mro.massey.ac.nz/handle/10179/17252>
- Daudey, G. H. (2014). *A grammar of Wadu Pumi*. La Trobe University.
- Daudey, H., & Pincuo, G. (2018). *Documentation of Northern Prinmi oral art, with a special focus on ritual speech*. Endangered Languages Archive. Handle: <http://hdl.handle.net/2196/00-0000-0000-0010-8820-B>
- Daudey, H., & Pincuo, G. (2020). 'Pour out libation to all the gods!': Exhaustive constructions in Northern Prinmi ritual speech. *Linguistics of the Tibeto-Burman Area*, 43(1), 2–18. <https://doi.org/10.1075/ltba.19011.dau>
- Ding, P. S. (2014). *A Grammar of Prinmi: Based on the Central Dialect of Northwest Yunnan, China*. BRILL. <http://ebookcentral.proquest.com/lib/kentucky-ebooks/detail.action?docID=1730292>
- Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvíl, F., Maxwell-Smith, Z., Nash, D., Olsson, O., Richards, M., San, N., Stoakes, H., Thieberger, N., & Wiles, J. (2018). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 205–209. <https://doi.org/10.21437/SLTU.2018-43>
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. <https://doi.org/10.1145/1143844.1143891>
- Guillaume, S., Wisniewski, G., Galliot, B., Nguyễn, M.-C., Fily, M., Jacques, G., & Michaud, A. (2022). *Plugging a neural phoneme recognizer into a simple language model: A workflow for low-resource settings*. 4905. <https://doi.org/10.21437/Interspeech.2022-11314>
- Guillaume, S., Wisniewski, G., Macaire, C., Jacques, G., Michaud, A., Galliot, B., Coavoux, M., Rossato, S., Nguyễn, M.-C., & Fily, M. (2022). Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 170–178. <https://doi.org/10.18653/v1/2022.computel-1.21>
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). *Deep Speech: Scaling up end-to-end speech recognition* (arXiv:1412.5567). arXiv. <http://arxiv.org/abs/1412.5567>
- Himmelman, N. P. (2006). Language documentation: What is it and what is it good for. *Essentials of Language Documentation*, 178(1).

- Jacques, G., & Daudey, H. (2011). Pumi Corpus. *Pangloss Collection (Cocoon Platform)*. <https://pangloss.cnrs.fr/corpus/Pumi?lang=en&seeMore=true>
- Jimerson, R., Simha, K., Ptucha, R., & Prud'hommeaux, E. (2018). Improving ASR Output for Endangered Language Documentation. *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*. <https://doi.org/10.21437/SLTU.2018-39>
- Liu, Z., Spence, J., & Prud'hommeaux, E. (2022). *Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation* (arXiv:2208.12888). arXiv. <https://doi.org/10.48550/arXiv.2208.12888>
- Lubbers, M., & Torreira, F. (2013). *pypmi-ling: A Python module for processing ELANs EAF and Praats TextGrid annotation files*. <https://pypi.python.org/pypi/pypmi-ling>
- McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktorandreevichmorozov, Choi, K., Moore, J., ... Kim, T. (2022). *librosa/librosa: 0.9.2*. Zenodo. <https://doi.org/10.5281/zenodo.6759664>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Neubig, G., Rijhwani, S., Palmer, A., MacKenzie, J., Cruz, H., Li, X., Lee, M., Chaudhary, A., Gessler, L., Abney, S., Hayati, S. A., Anastasopoulos, A., Zamaraeva, O., Prud'hommeaux, E., Child, J., Child, S., Knowles, R., Moeller, S., Micher, J., ... Littell, P. (2020). *A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization* (arXiv:2004.13203). arXiv. <https://doi.org/10.48550/arXiv.2004.13203>
- Nowakowski, K., Ptaszynski, M., Murasaki, K., & Nieuważny, J. (2023). Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2), 103148. <https://doi.org/10.1016/j.ipm.2022.103148>
- Pellegrini, T., & Lamel, L. (2008). *Are Audio or Textual Training Data More Important for ASR in Less-Represented Languages?* The first International Workshop on Spoken Languages Technologies for Under-resourced languages.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., & others. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, CONF*.
- Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*, 2757–2761. <https://doi.org/10.21437/Interspeech.2020-2826>
- Seifart, F., Evans, N., Hammarström, H., & Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, 94(4), e324–e345. <https://doi.org/10.1353/lan.2018.0070>
- Shi, J., Amith, J. D., García, R. C., Sierra, E. G., Duh, K., & Watanabe, S. (2021). *Leveraging End-to-End ASR for Endangered Language Documentation: An*

- Empirical Study on Yolox\ 'ochitl Mixtec* (arXiv:2101.10877). arXiv. <https://doi.org/10.48550/arXiv.2101.10877>
- The pandas development team. (2022). *pandas-dev/pandas: Pandas* (v1.5.1). Zenodo. <https://doi.org/10.5281/zenodo.7223478>
- Vaessen, N. (2022). *JiWER* (2.5.1) [Python]. <https://github.com/jitsi/jiwer>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
- von Platen, P. (2021). *Fine-tuning XLSR-Wav2Vec2 for Multi-Lingual ASR with 🧠 Transformers* [Jupyter Notebook]. HuggingFace. [https://github.com/patrickvonplaten/notebooks/blob/master/Fine Tune XLSR Wav2Vec2 on Turkish ASR with %F0%9F%A4%97 Transformers.ipynb](https://github.com/patrickvonplaten/notebooks/blob/master/Fine_Tune_XLSR_Wav2Vec2_on_Turkish_ASR_with_%F0%9F%A4%97_Transformers.ipynb)
- Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. *Symmetry*, 11(8), Article 8. <https://doi.org/10.3390/sym11081018>
- Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplín, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T. (2018). *ESPnet: End-to-End Speech Processing Toolkit* (arXiv:1804.00015). arXiv. <https://doi.org/10.48550/arXiv.1804.00015>
- Wisniewski, G., Michaud, A., & Guillaume, S. (2020). *Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?* 306. <https://shs.hal.science/hal-02513914>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. <https://doi.org/10.48550/arXiv.1910.03771>
- Zahrer, A., Zgank, A., & Schuppler, B. (2020). Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2893–2900. <https://aclanthology.org/2020.lrec-1.353>

VITA

Connor Nathaniel Bechler was born in Newton, Kansas, but spent his childhood in Hudsonville, Michigan, graduating from Hudsonville High School in 2016. He attended Calvin University from 2016 to 2020, graduating with a Bachelor of Arts degree in Linguistics, English Literature, and Mandarin Chinese. Over the last two years, he has served as a teaching assistant at the University of Kentucky. Upon publication of this thesis, he will graduate with a Master of Arts degree in Linguistic Theory and Typology from the University of Kentucky.