Child Support: Leveraging Lexifiers Resources to Support Creoles ASR

Éric Le Ferrand
SUNY Buffalo
ericlefe@buffalo.edu

Fabiola Henri
SUNY Buffalo
fabiolah@buffalo.edu

Abstract

Creole languages emerged from colonial contact and the slave trade. Although they inherit the bulk of their vocabulary from a "lexifier" language, they remain classic low-resource languages, presenting significant challenges for speech technology. This paper explores how the abundant resources of a lexifier can be leveraged for Creole-specific tools, focusing on Automatic Speech Recognition (ASR). Specifically, we use an artificial dataset generated a French-trained Text-to-Speech (TTS) model and French datasets to pre-finetune ASR models for two French-based Creoles. Our results demonstrate that a two-stage training setup where models are first trained on artificial datasets leads to substantial performance boost for transcribing Creole languages. Additionally, this approach serves as a viable first step for ASR development in zero-resource scenarios.

1 Introduction

Creole languages are typically characterized by a large portion of their vocabulary inherited from a lexifier language. While the European input to Creoles comes from a mix of regional and non-standard varieties spoken by colonists and traders, rather than the standardized forms on which most ASR systems are trained, we believe that ASR resources available for the lexifiers can be, to some extent leveraged for fine-tuning models for Creole languages, at least at the lexical level.

The advent of Transformer-based architectures has streamlined the development of ASR for any language, provided a baseline of high-quality data is available. Increasingly, ASR models for indigenous languages worldwide are achieving performance levels previously reserved for only the most highly documented languages a decade ago. Nevertheless, data scarcity remains a central challenge; most indigenous languages possess either vast quantities of untranscribed speech or raw text primarily intended for educational or artistic use. Creole languages, however, hold a distinct advantage: their lexicons are derived from lexifiers that are typically high-resource. These existing resources can be strategically exploited as data substitutes to build robust language technologies for Creoles.

In this paper, we explore two strategies for synthesizing data from lexifier languages to supplement limited resources and enhance Creole ASR performance. Our first approach involves generating synthetic speech by processing raw Creole text through a TTS system trained on the lexifier. In our second approach, we adapt an existing lexifier speech dataset by mapping its orthographic transcriptions to Creole writing systems. We evaluate these datasets through two training paradigms: (1) Direct Substitution, where a model is trained solely on synthetic data and tested on Creole, and (2) Sequential Pre-training, where a model is pre-trained on the artificial dataset before being fine-tuned on authentic Creole data. These strategies are evaluated in two case studies: Haitian Creole and Mauritian Creole. As an additional contribution, we are releasing a formatted corpus for Mauritian Creole to facilitate future research on ASR.

2 Background

The *transcription bottleneck* (Himmelmann, 1998) has long been recognized as a primary obstacle in language documentation, with ASR proposed as a potential solution (Prud'hommeaux et al., 2021). Recently, the emergence of Transformer-based architectures (Vaswani et al., 2017) has significantly reduced the data requirements for model training (Conneau et al., 2021; Hsu et al., 2021; Barrault et al., 2023; Radford et al., 2023; Pratap et al., 2024). This shift has facilitated the development of numerous ASR models specifically tailored for endangered languages (Tsoukala et al., 2023; Seo et al., 2024; Jones et al., 2024; Daul et al., 2026).

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 111–117

	baseline	TTS French	Mapped French
tokens	49526	78908	160700
types	3273	13641	8186

Table 1: Token type count for Haitian Creole

	baseline	TTS French	Mapped French
tokens	35781	53460	160700
types	2762	6544	7731

Table 2: Token type count for Mauritian Creole

The shared linguistic features between a creole and its lexifier have frequently been leveraged for computational tasks. In machine translation, research indicates that models pretrained on lexifiers yield superior performance when applied to their descendant creoles (Lin et al., 2023; Ayasi, 2025). Similarly, in ASR development, it is common practice to favor foundational models pretrained on the lexifier over general multilingual models for subsequent fine-tuning (Macaire et al., 2022; Havard et al., 2025), even with End2End architectures trained to produce text in the lexifier (Le Ferrand and Prud'hommeaux, 2024).

To address data scarcity, TTS has become a well-established, albeit constrained, method for data augmentation (Ueno et al., 2021; Laptev et al., 2020; Gokay and Yalcin, 2019). While pooling the data in a single set is usually the method exploited. Widely different kind of dataset might cause the model training to fail. Recent research suggests that a two-stage training protocol—rather than a simple pooling of all datasets—is significantly more efficient for model convergence Tapo et al. (2024); Le Ferrand et al. (2025); Sung et al. (2025).

Orthographic mapping has emerged as a remarkably effective strategy for enhancing ASR performance. This approach has been extensively documented in recent literature, particularly within the context of South and East Asian languages, where reconciling divergent writing systems is often a prerequisite for cross-lingual transfer (Khare et al., 2021; Lee et al., 2025; Sung et al., 2025).

3 Data

3.1 Creoles languages

We focus on two French-lexified Creoles for which exploitable data is available, namely Mauritian and Haitian Creoles. Beyond the lexicon, Mauritian and Haitian differ substantially from Standard French in their grammatical and structural properties. For instance, definite determiners are typically postposed in both languages (e.g., *liv la* 'the book'), in contrast to preposed determiners in French, and their pronominal systems are largely built on forms historically related to French strong (tonic) pronouns rather than clitic forms. At the phonological level, however, the divergence from Standard French is more limited: the phonotactic systems remain broadly comparable, both languages retain nasal vowels (albeit with some restructuring), and Mauritian simplifies certain segments of French origin, such as the reduction or loss of palato-alveolar fricatives (/ʃ/, /ʒ/). While they draw on earlier, non-standard varieties of French that also contributed to the development of Québécois Frenches, their emergence reflects distinct contact-driven processes in colonial settings. Both Mauritian and Haitian remain in close and ongoing contact with Standard French. Mauritian Creole is additionally in contact with other languages, including English and Bhojpuri.

3.2 Original datasets

For Haitian Creole, we used a subset of the CMU dataset[1]. We use 5h for the train and 2h for the test. For the textual data, we used a subset of Kreyol-MT (Robinson et al., 2024). For Mauritian Creole, we curated a subset of field linguistic recordings extracted from PARADISEC public archives[2]. We used 2h30 for the train and 38min for the test. The formatted data for Mauritius Creole is publicly available[3]. For the textual data we used a subset of KreolMorisienMT (Dabre and Sukhoo, 2022).

3.3 Artificial Datasets

Our methodology involved the construction of two synthetic datasets. The first, TTS_French, consists of 10 hours of speech generated by passing cleaned Creole text through the MMS-TTS French model (Pratap et al., 2024). To ensure data quality, we filtered the source text to remove special characters and numerical values. The second dataset, mapped_French, utilizes 10 hours of audio randomly sampled from the Corpus de Français Parlé de nos Régions (CFPR) (Avanzi et al., 2016). To align this French audio with Creole orthography, we performed a two-step conversion: first, gener-

[1] http://www.speech.cs.cmu.edu/haitian/

[2] https://catalog.paradisec.org.au/

[3] https://huggingface.co/datasets/eleferrand/Morisyen_Corp_ASR

French	Mauritius	Haiti
et d'ailleurs	et daye	èt daye
je sais qu'après	ze se kapre	je sè kaprè
mais je les connais	me ze le kone	mè je lè kònè
j'ai pas le permis	ze pa le permi	jè pa le pèrmi
ça fait longtemps	sa fe lontan	sa fè lontan

Table 3: Examples of othrography mapping between French and Creoles

ating phonetic transcriptions via charsiu-g2p (Zhu et al., 2022), and subsequently applying a mapping table to translate those phonemes into their respective Creole graphemes. Examples of transcription conversion can be found Table 3. It is important to mention that such mapping does not produce correct creoles structures but will generally produce accurate lexical forms. The mapping tables can be found in Table 4. Additional information on all collections can be found in Table 1 for Haitian and Table 2 for Mauritian.

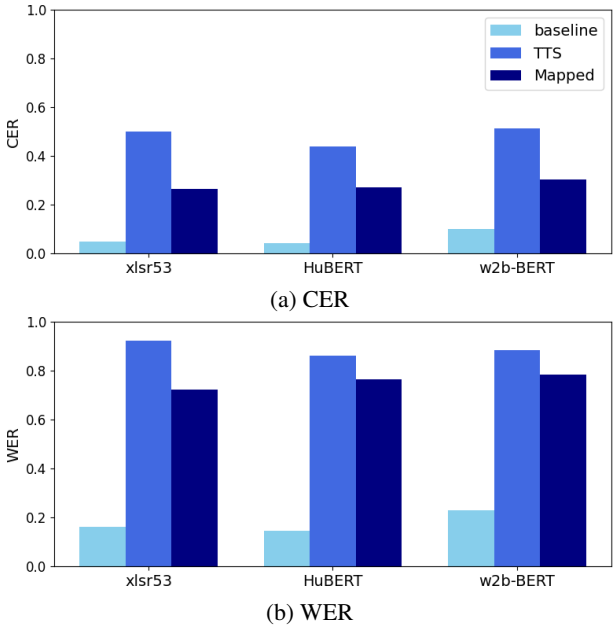
4 Methods

First, for each creole, we train three models, (1) a baseline model trained on the training set of the creole, (2) a model we call TTS_French trained on the 10h of synthetic speech generated from the text in Creole, and (3) Mapped_French, a model trained on the French data, which transcriptions are mapped to the creole writing system. Each model is then tested on the Creole testing set.

In a second phase, for each creole, we take the TTS_French model and the Mapped_French model and we keep training them with the original training sets in Creole. We test the resulting model in the original test sets.

We explore these configurations with 3 pretrained acoustic models: XLSR53 (Conneau et al., 2021) a multilingual model based on wav2vec architecture, HuBERT-large, a monolingual model trained on English (Hsu et al., 2021) and wav2vec-BERT a monolingual model trained on English (Barrault et al., 2023). Each model has been trained for 30 epochs with a batch size of 16. For the model configuration, attention dropout, hidden dropout, feature projection dropout and layerdrop are set to 0.0, mask time probability to 0.05, and the CTC loss reduction method takes the mean over a batch. We set the learning rate at 0.0003 and optimized with AdamW. Features encoder is left unfrozen and zero_to_infinity is set to True. The models are trained on a single 48GB A100 GPU. Fine-tuning for each model takes approximately 60 minutes for the Creole languages and 2h for the French-based datasets. Decoding is systematically done with a trigram language model trained on the training set of each creole with kenlm[4].

5 Results



(a) CER

(b) WER

Figure 1: Baseline results for Haitian Creole

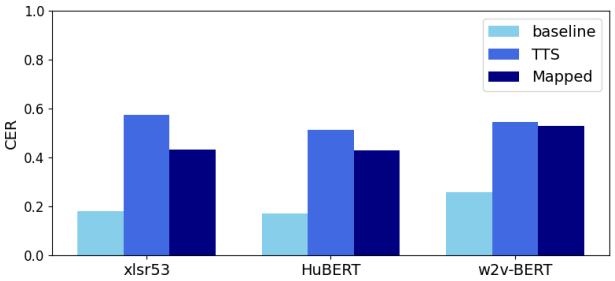
Baseline results for Haitian Creole can be found in Figure 1 and Mauritius Creole in Figure 2. Performance across the various models remains relatively consistent, with no substantial deviations observed between architectures. Notably, Haitian Creole outperforms Mauritian Creole, likely due to a more rigorously curated corpus with fewer transcription inconsistencies and less frequent code-switching. Analysis of the Word Error Rate (WER) reveals that while models trained on synthetic data face challenges, the Haitian Creole "Mapped" model achieves a WER of 0.7. It is also worth noting that Mapped models generally outperform TTS-based models. Furthermore, both approaches correctly predict nearly half of the characters. While these models are not yet performing well, they provide a viable initial transcription layer for scenarios where no aligned data is available. Small improvement is still noticed for XLSR for the mapped model which show that this model is the most reli-

[4]https://github.com/kpu/kenlm

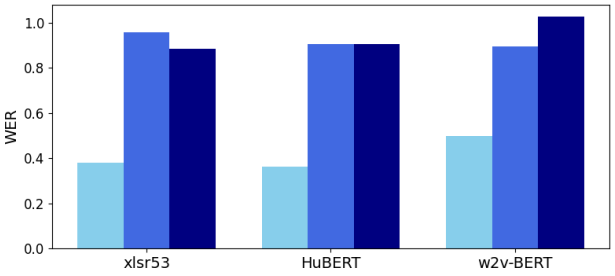
Phoneme	i	e	ɛ	a	ɔ	o	u	y	ã	ɑ̃	ɛ̃	ɔ̃	ʃ	ʒ	ɲ	ŋ	ʁ	w	j
Haitian	i	e	è	a	ò	o	ou	i	an	an	en	on	ch	j	ny	ng	r	w	y
Mauritian	i	e	e	a	o	o	ou	i	an	an	en	on	s	z	ny	ng	r	w	y

Table 4: Mapping tables between French phonemes and corresponding graphemes in both Creoles. Missing values have identical phonemes and graphemes mapping (e.g. /l/ /m/ /a/...)

able in this context.

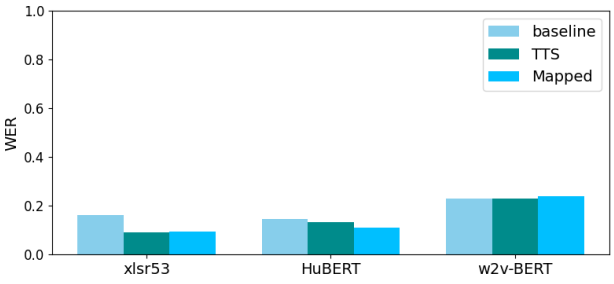


(a) CER

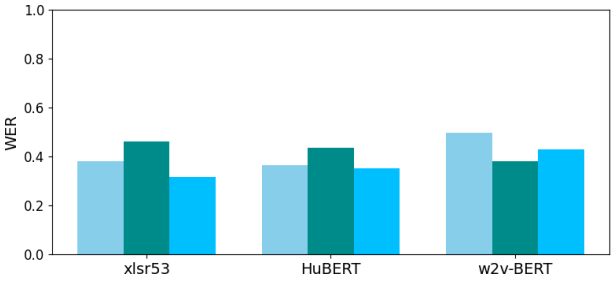


(b) WER

Figure 2: Baseline results for Mauritian Creole



(a) Haitian



(b) Mauritian

Figure 3: Word Error Rate for the 2 phase training models

The results for the 2 phase training experiments can be found in Figure 3. The two-stage training process, transitioning from artificial to real-world data, successfully enhanced performance in 75% of cases for both languages. Among the tested architectures, XLSR-53 stood out as the top performer, achieving substantial improvements of 40% for Haitian Creole and 20% for Mauritian Creole. While the mapped dataset yielded the most consistent results across both languages, the use of TTS data proved less reliable, leading to a performance decline in two out of three models for Mauritian Creole.

To evaluate the consistency of these findings, we assessed performance across two additional testing sets. For Mauritian Creole, the second author recorded 10 minutes of audio from a children's book, while for Haitian Creole, we put together data from 3 sources to reach 8min of speech: the little data available in a few recording available on gitHub[5], CommonVoice[6], and the recording of a children book[7]. The performance metrics for these external datasets are detailed in Figure 4.

While out-of-domain error rates are notably higher for Haitian Creole, overall performance remains strong—with the exception of wav2vec-bert, which underperforms significantly on this language. The initial findings hold true: preliminary fine-tuning on lexifier-derived datasets, particularly the mapped version, consistently improves results. This confirms that the method is robust across both in-domain and out-of-domain scenarios.

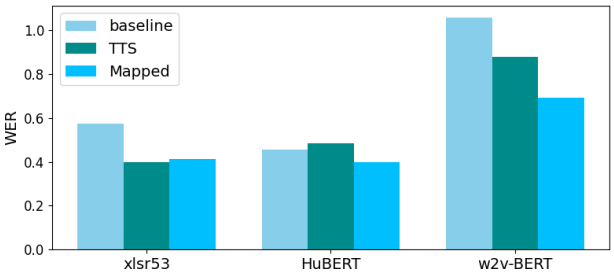
6 Conclusion

This paper investigates strategies for leveraging lexifier resources to support ASR for Creole languages. Focusing on Haitian and Mauritian Creole as case

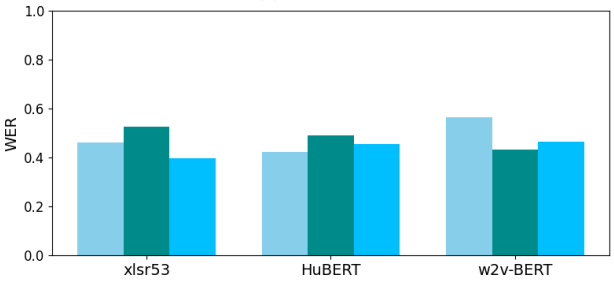
[5]https://github.com/KerlinMichel/KreyolTranskripsyon/tree/main

[6]https://mozilladatacollective.com/datasets/cmn1pz91w00v3o107hknri5xy

[7]https://www.youtube.com/watch?v=QhtGolDZsKY



(a) Haitian



(b) Mauritian

Figure 4: Word Error Rate for the 2 phase training models on Extra testing data.

studies, we examine two scenarios involving models trained on synthetic data: one utilizing a French TTS system applied to Creole text, and another employing French data where the orthography has been mapped to the Creole writing system.

Experimental results demonstrate that: (1) using models trained solely on synthetic data, roughly half of the characters are correctly transcribed; and (2) performance improves substantially across both in-domain and out-of-domain testing when models undergo preliminary training on synthetic data followed by fine-tuning on authentic datasets. This boost is particularly pronounced when using the mapped dataset configuration.

In future research, we plan to investigate whether these findings extend to English- and Portuguese-based lexifiers. Furthermore, we intend to employ this methodology to semi-automatically generate a large-scale speech dataset encompassing several underresourced French-based Creoles.

Limitations

We acknowledge several limitations to the current study: (1) Our focus was restricted to two languages and a single lexifier, which may constrain the generalizability of the findings to other creoles. (2) We utilized only one TTS model. While robust, employing alternative architectures could yield different outcomes. (3) While the French dataset used consists of fieldwork data well-suited to our objectives, the use of different source corpora might influence the results. (4) Finally, although our in-domain test sets are relatively large, we recognize that our out-of-domain testing remains limited in scope.

Acknowledgements

This study has been conducted as part of the NSF DLI-DEL project Award Number 2450839.

References

Mathieu Avanzi, Marie-José Béguelin, and Federica Diémoz. 2016. *Corpus de français parlé et français parlé des corpus*. Revue Corpus.

Ananya Ayasi. 2025. Krey-all wmt 2025 creolemt system description: Language agnostic strategies for low-resource translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1158–1165.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Raj Dabre and Aneerav Sukhoo. 2022. Kreol-MorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Massimo Marie Daul, Alessio Tosolini, and Claire Bowern. 2026. Linguistically informed tokenization improves asr for underresourced languages. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 31–37.

Ramazan Gokay and Hulya Yalcin. 2019. Improving low resource turkish speech recognition with data augmentation and tts. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, pages 357–360. IEEE.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Speech technologies with fieldwork recordings: the case of haitian creole. In *Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 40.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Austin Jones, Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzić, and Keith Langston. 2024. Comparing kaldi-based pipeline elpis and whisper for čakavian transcription. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 61–68.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Éric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. Faithful transcription: Leveraging bible recordings to improve asr for endangered languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 333–342.

Éric Le Ferrand and Emily Prud'hommeaux. 2024. Automatic transcription of grammaticality judgements for language documentation. In *The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 33.

Sangmin Lee, Woojin Chung, and Hong-Goo Kang. 2025. Lama-ut: Language agnostic multilingual asr through orthography unification and language-specific transliteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24393–24401.

Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. 2023. Low-resource cross-lingual adaptive training for nigerian pidgin. In *Proc. Interspeech 2023*, pages 3954–3958.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

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

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Krey\ol-mt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.

Jean Seo, Minha Kang, Sungjoo Byun, and Sangah Lee. 2024. Manwav: The first manchu asr model. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 6–11.

Hung-Yang Sung, Chien-Chun Wang, Kuan-Tang Huang, Tien-Hong Lo, Yu-Sheng Tsao, Yung-Chang Hsu, and Berlin Chen. 2025. Clift-asr: A cross-lingual fine-tuning framework for low-resource taiwanese hokkien speech recognition. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 176–183.

Allahsera Tapo, Éric Le Ferrand, Zoey Liu, Christopher Homan, and Emily Prud'hommeaux. 2024. Leveraging speech data diversity to document indigenous heritage and culture. In *Proc. Interspeech 2024*, pages 5088–5092.

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

Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2021. Data augmentation for asr using tts via a discrete representation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 68–75. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Proc. Interspeech 2022*, pages 446–450.