Faithful Transcription: Leveraging Bible Recordings to Improve ASR for Endangered Languages

Éric Le Ferrand[1,2], **Cian Hauser**[3], **Joshua Hartshorne**[4], **Emily Prud'hommeaux**[2]
[1]University at Buffalo, [2]Boston College
[3]University of Chicago, [4]MGH Institute of Health Professions

Abstract

While automatic speech recognition (ASR) now achieves human-level accuracy for a dozen or so languages, the majority of the world's languages lack the resources needed to train robust ASR models. For many of these languages, the largest available source of transcribed speech data consists of recordings of the Bible. Bible recordings are appealingly large and well-structured resources, but they have notable limitations: the vocabulary and style are constrained, and the recordings are typically produced by a single speaker in a studio. These factors raise an important question: to what extent are Bible recordings useful for developing ASR models to transcribe contemporary naturalistic speech, the goal of most ASR applications? In this paper, we use Bible recordings alongside contemporary speech recordings to train ASR models in a selection of under-resourced and endangered languages. We find that models trained solely on Bible data yield shockingly weak performance when tested on contemporary everyday speech, even when compared to models trained on other (non-Bible) out-of-domain data. We identify one way of effectively leveraging Bible data in the ASR training pipeline via a two-stage training regime. Our results highlight the need to re-assess reported results relying exclusively on Bible data and to use Bible data carefully and judiciously.

1 Introduction

Over the past decade, advances in automatic speech recognition (ASR) have yielded near-human levels of performance on popular ASR benchmarks for English and a handful of politically and economically dominant languages (Hinton et al., 2012; Amodei et al., 2016; Baevski et al., 2020; Hsu et al., 2021). Languages with fewer resources have benefited from these advances, as well, particularly via the approach of fine-tuning from multilingual pretrained models (Conneau et al., 2021; Radford et al., 2023; Pratap et al., 2024). The majority of the world's languages, however, have insufficient transcribed audio to train robust ASR models.

The problem is especially dire for endangered languages, which often lack a written tradition and a population of speakers who are able and willing to share their language with outsiders (Himmelmann, 2018; Eberhard et al., 2025). For many such languages, the only freely available large transcribed speech datasets are recordings of the Bible produced by Christian missionary organizations. The New Testament totals more than 30 hours of audio, and has a unique index – book, chapter, verse – for each utterance. These features make the Bible a particularly attractive resource for developing speech technologies for endangered and other under-resourced languages (Black, 2019; Meyer et al., 2022; Pratap et al., 2024).

Prior ASR research using exclusively Bible recordings for both training and testing often reports very impressive results (Pratap et al., 2024). Although it is sometimes noted that these recordings are produced by a single speaker reading text from a restricted domain in a recording studio, it is rarely acknowledged that such conditions often result in high accuracy regardless of the language or architecture. Very little of this prior work explores whether these models generalize well to fieldwork or any other speech data produced by a variety of speakers in more natural contemporary settings. In this paper, we attempt to address this question with experiments in which we train and test ASR models with and without Bible data under a variety of conditions across across 5 endangered languages and 4 widely spoken but under-resourced languages.

We replicate earlier findings that models trained on Bible recordings yield high accuracy on test data from that same Bible corpus, but we find that including Bible data directly in training, by itself or in combination with in-domain data, yields dismal results on speech produced in everyday contem-

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

porary settings, far worse than training solely on a small amount of in-domain data. For 5 of the 9 languages, we have non-Bible data from multiple sources. In these cases, we can compare ASR performance of models trained on Bible data with models trained on naturalistic contemporary speech from a different domain or in a different style than the target data (e.g., interviews vs. narratives). We find again that models trained on Bible datasets – despite their impressive size and high quality audio – fare poorly against models trained on much smaller non-Bible out-of-domain data. Our only successful method for using Bible data is to include it in an initial continued pre-training stage followed by fine-tuning on data from the non-Bible domain.

Our results demonstrate the challenges and limitations associated with using Bible data for ASR model training. While we are not permitted to release the Bible ASR datasets, a secondary contribution of our work is the code we used to extract the data from `Bible.is` and to align the audio with the Bible texts. We will also release, with permission, all the non-Bible datasets as ASR-ready Hugging Face datasets.

2 Prior work

There is a growing interest in developing ASR systems for endangered and Indigenous languages, often motivated by the demand in the linguistics research and speaker communities for tools that can facilitate transcription of audio recordings for documentary and educational purposes. This research typically focuses on a single language of interest (Gupta and Boulianne, 2020a; Sikasote and Anastasopoulos, 2022; Shi et al., 2021; Zahrer et al., 2020; Ćavar et al., 2016), but there is some work exploring a larger number of languages spanning multiple families in order to explore the impact of data collection methods and ASR architectures (Adams et al., 2018; Jimerson et al., 2023; Liu et al., 2024b; Wisniewski et al., 2020).

The Bible is widely used in NLP research for a variety of tasks beyond ASR (Gupta and Boulianne, 2020b; Meyer et al., 2022; Black, 2019; Pratap et al., 2024; Adams et al., 2019; Rista and Kadriu, 2021; Nzenwata and Ogbuigwe, 2024; Liu et al., 2024a), such as POS tagging (Agić et al., 2015), LLM adaptation (Ebrahimi and Kann, 2021), typological exploration (McCarthy et al., 2020; Kann, 2024), and machine translation (Mayer and Cysouw, 2014; Liu et al., 2021; Domingues et al., 2024). Our work was inspired in part by recent findings in MT showing that combining Bible data with fieldwork yields disappointing results in low-resource settings for Indigenous languages. Domingues et al. (2024) caution against using Bible data for endangered and Indigenous MT, citing two main concerns. First, the Bible is associated with colonial histories marked by displacement, forced assimilation, and cultural suppression. Second, the careless use of this data can increase MT hallucinations due to its narrow domain, vocabulary, and style, findings that were previously noted by Mayer and Cysouw (2014).

3 Data

We collected data for five endangered Indigenous languages from two sources: FormosanBank (Amis, Tsou, and Rukai) (Mohamed et al., 2024; Hartshorne et al., 2024)[1] and the AmericasNLP-2022 ASR Shared Task (Quechua and Bribri). (Ebrahimi et al., 2023).[2] We additionally acquired or assembled ASR-compatible datasets for four widely-spoken but under-resourced languages: Fongbe (Laleye et al., 2016), Iban (Juan et al., 2015), Bambara (Tapo et al., 2024), and Swahili (Gelas et al., 2012). For all 9 languages, we extracted and aligned the full New Testament from the Faith Comes by Hearing (`Bible.is`) website.[3] Table 1 provides basic information about each of the corpora, including source, speech collection method, number of tokens and types, duration of audio in minutes, and the type:token ratio.

Amis, Tsou, and Rukai are endangered Indigenous languages of Taiwan belonging to the Formosan branch of the Austronesian language family. FormosanBank includes multiple speech corpora per language. For each of the three languages, we use the ePark corpus (Indigenous Language Research and Development Foundation, 2023b) (denoted in this paper as **Corpus A**), which here consists of recordings of spontaneously delivered narratives; and the ILRDF corpus (Indigenous Language Research and Development Foundation, 2023a) (denoted in this paper as **Corpus B**) which contains recordings of a large number of speakers reading example utterances demonstrating the usage of individual dictionary entries. From each corpus, we select 4 hours of audio, a reasonable size for a typi-

[1]https://github.com/FormosanBank/FormosanBank
[2]http://turing.iimas.unam.mx/americasnlp/2022_st.html
[3]https://www.faithcomesbyhearing.com/

cal fieldwork corpus and comparable to one of the two AmericasNLP languages.

For Swahili and Bambara, we acquired non-Bible corpora from two distinct sources. For Swahili, **Corpus A** is a dataset collected as part of the ALFFA project, which consists of both read speech and transcribed web news broadcasts. **Corpus B** is Common Voice Delta Segment 12.0.[4] For Bambara, **Corpus A** is a recently collected dataset of recordings of Griot storytellers.[5] **Corpus B** was created from fieldwork recordings from the 1980s provided privately to us by a colleague. While not currently publicly available, this corpus will be released as a Hugging Face dataset.

The two languages from the AmericasNLP-2022 shared task on ASR are Bribri and Quechua. The Bribri corpus consists of fieldwork recordings made in the 2010s extracted from the Pandialectical Corpus of the Bribri Language[6], and the Quechua data is derived from radio broadcast conversations in the Siminchik dataset (Cardenas et al., 2018). The Iban data consists of manually transcribed recordings from Malaysian radio and television programs, while the Fongbe data consists entirely of read speech. For these four languages, only one corpus is available, which we denote as **Corpus A**.

All Bible recordings and transcripts were extracted from the Faith Comes by Hearing (`Bible.is`) website. Information about the somewhat laborious extraction and alignment process are included in Appendix A.3. We note that the scripts originally provided with the CMU Wilderness dataset (Black, 2019) are no longer compatible with the current structure of the `Bible.is` website.

The datasets for Amis, Tsou, Rukai, Swahili, Bambara, Iban and Fongbe were split into training, development, and test sets using a 70/10/20 ratio. The AmericasNLP datasets for Quechua and Bribri, which were substantially smaller, were partitioned into 80/20 train/test splits. A development set and a test set was created for each Bible corpus comparable in size to that of the non-Bible corpus or corpora for that language.

4 Methods

All experiments are conducted with XLSR-53 (Conneau et al., 2021), a multilingual speech model based on the wav2vec2 2.0 architecture. While

[4]https://commonvoice.mozilla.org/ksf/datasets
[5]https://github.com/RobotsMali-AI/jeli-asr
[6]http://bribri.net

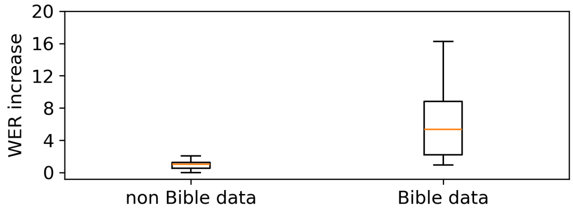

Figure 1: Relative degradation of WER between in-domain and out-of-domain of models tested on non Bible data (left) and Bible data (right)

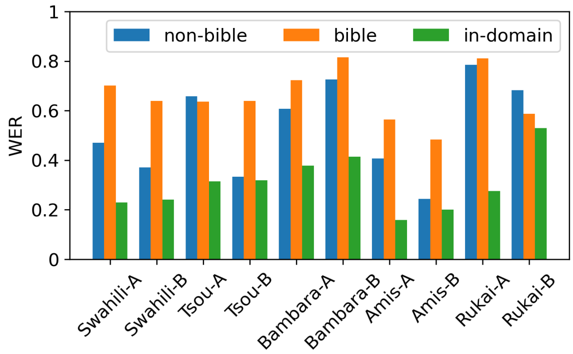

Figure 2: WER for models trained exclusively on Bible data (bible); exclusively on the non-Bible data from the other source (non-bible); and exclusively on the non-Bible data from the same source (in-domain).

there are a number of ASR architectures that we could have chosen, we found XLSR-53 to provide the best balance between efficiency, resource use, and accuracy.[7] Following a popular tutorial,[8] we train a CTC layer for 30 epochs, select the best model with the validation set, and then decode with a trigram LM trained on the transcripts of the acoustic training data. We explore different experimental configurations to probe the utility of the Bible for ASR training.

Bible versus alternative sources: For the languages with two available non-Bible datasets, we fine-tune XLSR-53 separately on each of the two sources and on the Bible data, resulting in three distinct models. Each model is then evaluated on test sets from all three corpora. For the languages which have only one non-Bible corpus, we do not have an alternative source, so we instead fine-tune separately with the one source and the Bible data, resulting in 2 distinct models, each of which is evaluated on both test sets.

[7]We found the performance of Whisper to be inconsistent across languages and domains, supporting prior work found Whisper to be particularly fragile to cross-domain ASR (Tapo et al., 2024).
[8]https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

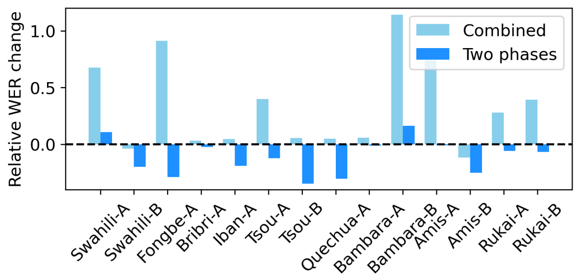

Figure 3: Relative change in WER according to whether the Bible and non-Bible datasets are combined into a single training set or used in two-stage training, with testing on **in-domain** data.

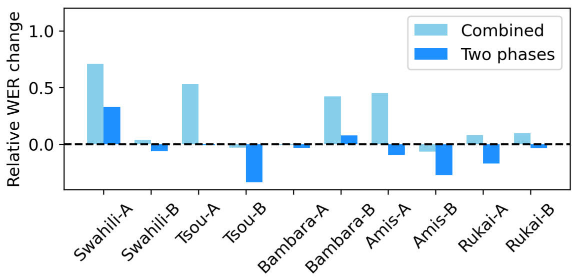

Figure 4: Relative change in WER according to whether the Bible and non-Bible datasets are combined into a single training set or used in two-stage training, with testing on **out-of-domain** data.

Bible as a complementary data source: For the languages with two non-Bible sources, we combine Bible data with the other sources in three ways. (1) Train a single model on the full 30+ hours of Bible data together with each of the two non-Bible sources. (2) Train a single model on 5 randomly selected hours of the Bible data together with each of the two non-Bible sources. (3) Apply a curriculum-style approach by first continuing speech representation training on the Bible data, then fine-tuning to each of the two non-Bible sources (Kunze et al., 2017; Khare et al., 2021; San et al., 2024). In all cases, we evaluate the resulting models on all three test sets. For languages with only one non-Bible corpus, we carry out both combined trainings (full Bible, 5h of Bible) and two-stage training, and we evaluate on the Bible and non-Bible test sets.

5 Results

Detailed WER results for all experiments are provided in Appendix A.3. Figure 1 shows the relative degradation in WER from in-domain testing (when the model has been trained on data from the same source as the test set) to out-of-domain testing (when the model has been trained on data from a different source from the test set). While the performance degradation on non-Bible test sets is minimal, we see steep declines in accuracy when training on the Bible and testing on other data. This indicates that training and testing exclusively on Bible data may significantly overestimate the model's ability to generalize to new data.

Figure 2 illustrates how models trained on Bible and non-Bible datasets perform when evaluated on various non-Bible test sets, either from the same dataset (in-domain) or a different one (out-of-domain). As expected, in-domain evaluations have the lowest WER for each dataset. Models trained on non-Bible data exhibit a moderate increase in WER when tested out-of-domain, which aligns with typical domain shift behavior. In contrast, models trained on the Bible almost systematically yield the worst performance, underscoring their limited generalization to everyday speech. It is also important to note that although the Bible datasets are 5 to 10 times larger than the non-Bible datasets, models trained on the Bible perform worse for nearly all languages. Increasing the amount of training data does not always lead to better performance, particularly when that data does not align well with the target use case.

Figures 3 and 4 (as well as Tables 2-10) illustrate the impact of different strategies for including Bible data on in-domain and out-of-domain performance. The results show that combining Bible and non-Bible data into a single training set almost systematically degrades WER across both in-domain and out-of-domain evaluations. Adding only a small amount of additional Bible data (5 hours, typically doubling the size of the training corpus) also consistently degrades performance. In contrast, using a two-stage training approach – where the model is first trained on Bible data and then fine-tuned on non-Bible data – leads to improved WER, particularly for in-domain testing.

Several factors may shed light on the unusual behavior of models trained on Bible data. First, the type-token ratio of the Bible texts is significantly lower than that of other sources, indicating a limited lexical diversity (see Table 1). Second, the vocabulary overlap between the training data of the Bible and the test data of the non-Bible sources is low for all languages and typically much lower than the vocabulary overlap between the two non-Bible sources. Figure 5 shows the relationship between train/test vocabulary overlap and WER. We see

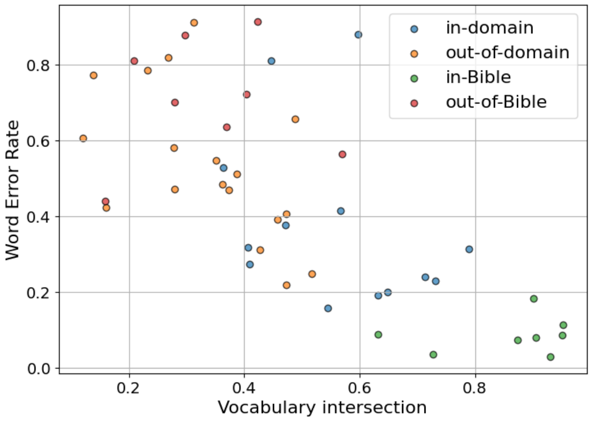

Figure 5: Correlation between WER and vocabulary overlap between the data used to train a model and the data of the test set being evaluated with that model (coeff= -0.711 p<0.001). WER is lower when vocabulary overlap is higher.

much that WER is higher when vocabulary overlap is low, and that vocabulary overlap across non-Bible sources (orange markers) is often within the range of in-domain overlap (blue and green markers), while overlap between Bible training data and non-Bible testing data (red markers) is uniformly low. Third, in the `Bible.is`recordings, the speech has a distinctive theatrical manner, which differs substantially from the more natural, ecologically collected speech of the non-Bible sources. Lastly, although the exact number of speakers per Bible dataset is not provided by `Bible.is`, we and others (Black, 2019; Pratap et al., 2024) observe that it is typically a few individuals and often a single male speaker.

6 Discussion and Conclusions

While datasets derived from the Bible, by virtue of their sheer size, ought to offer potential utility for building ASR models, our results demonstrate the need to use this data judiciously. Relying entirely on Bible recordings to train ASR models yields disappointing results with contemporary speech, as does combining Bible data in varying quantities with in-domain training data. Only by incorporating Bible data into a two-stage training scheme were we able to improve performance over fully in-domain training, even with small datasets.

In our future work, we plan to include more languages, carry out more experiments with alternative sources, and to explore additional methods for integrating Bible data. In the meantime, we hope that the results presented here will discourage other researchers from making broad claims about ASR model performance for endangered and low-resource languages based solely on models trained and tested with Bible data.

Limitations

While we demonstrated the negative effect that careless training on Bible data can have on ASR performance, the underlying reasons behind these results are still unclear. This data has many unusual features, including limited speaker diversity, a restricted vocabulary, content distinct from contemporary issues, and a style that may not reflect current speech and language patterns. Identifying which – or which combination – of these factors is beyond the scope of this short focused paper. The goal of our work here is not to determine why Bible data has this effect but rather to show that this effect holds regardless of the language or the domain of the target dataset. We note that additional experiments could be carried out for the languages for which we have multiple non-Bible data sources to see whether the results obtained on the different data combination strategies for Bible data are also valid with the non-Bible datasets. We again leave these for future work.

Ethical Considerations

All of the data used for our experiments is publicly available. The endangered language data from Indigenous communities (Rukai, Tsou, Amis, Bribri, Quechua) was made available for research purposes with the consent of representatives of these communities. We do note in the paper the ethical dilemma of using Bible data to support language reclamation efforts of Indigenous communities, and we reiterate our acknowledgment of that concern here. Nevertheless, while missionary and colonial histories have caused harm to many Indigenous communities around the world, it remains difficult to draw a clear picture of the ethical implications surrounding the use of Bible translations today. The versions used in our study are all relatively recent – the oldest dating from 2007 – and were collected with the participation of local religious organizations with Indigenous membership, such as the Bible Society of Taiwan or the Sociedad Bíblica de Costa Rica, which suggests a degree of local agency in their production and distribution.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #2319296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, pages 96–108.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, and 1 others. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP*, 2:21.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in ai tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL)*, pages 283–293.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, and 1 others. 2023. Findings of the second americasnlp competition on speech-to-text translation. In *NeurIPS 2022 Competition Track*, pages 217–232. PMLR.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2521–2527, Marseille, France. European Language Resources Association.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language cree. In *Proceedings of the 1st joint workshop on spoken language technologies for under-resourced languages (SLTU) and collaboration and computing for under-resourced languages (CCURL)*, pages 362–367.

Joshua K. Hartshorne, Éric Le Ferrand, Li-May Sung, and Emily Prud'hommeaux. 2024. Formosanbank and why you should use it. In *Architectures and Mechanisms in Language Processing (AMLaP) Poster*.

Nikolaus P. Himmelmann. 2018. Meeting the transcription challenge. In *Reflections on language documentation 20 years after Himmelmann 1998*, pages 33–40. University of Hawai'i Press, Honolulu.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and 1 others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel-rahman 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.

Indigenous Language Research and Development Foundation. 2023a. Online dictionary of aboriginal languages. https://e-dictionary.ilrdf.org.tw.

Indigenous Language Research and Development Foundation. 2023b. Yuanzhumin yuyan leyuan (epark). https://web.klokah.tw/.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1008–1016.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban. In *Proceedings of INTERSPEECH*, Dresden, Germany.

Amanda Kann. 2024. Massively multilingual token-based typology using the parallel Bible corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.

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.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer learning for speech recognition on a budget. *ACL 2017*, page 168.

Frejus A. A. Laleye, Laurent Besacier, Eugene C. Ezin, and Cina Motamed. 2016. First Automatic Fongbe Continuous Speech Recognition System: Development of Acoustic Models and Language Models. In *Federated Conference on Computer Science and Information Systems*.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of Bibles in low-resource machine translation. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.

Zoey Liu, Nitin Venkateswaran, Éric Le Ferrand, and Emily Prud'hommeaux. 2024a. How important is a language model for low-resource asr? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 206–213.

Zoey Liu, Nitin Venkateswaran, Éric Le Ferrand, and Emily Prud'hommeaux. 2024b. How important is a language model for low-resource ASR? In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 206–213.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, and Colin Leong. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proceedings of Interspeech 2022*, pages 2383–2387.

Wael Mohamed, Éric Le Ferrand, Li-May Sung, Emily Prud'hommeaux, and Joshua Hartshorne. 2024. Formosanbank. https://ai4commsci.gitbook.io/formosanbank.

Uchenna Nzenwata and Daniel Ogbuigwe. 2024. Automatic speech recognition for the ika language. *arXiv preprint arXiv:2410.00940*.

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

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

Amarildo Rista and Arbana Kadriu. 2021. Casr: A corpus for albanian speech recognition. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 438–441. IEEE.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 100–112.

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

Claytone Sikasote and Antonios Anastasopoulos. 2022. Bembaspeech: A Speech Recognition Corpus for the Bemba Language. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

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.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 306–315.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from Muyu. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2893–2900, Marseille, France. European Language Resources Association.

A Appendix

A.1 Acquiring Bible Data

Collecting data from the Bible came with its own set of challenges. The website `Bible.is` provides spoken Bible recordings in over 1,000 languages, with audio and text. Each audio file is 5-10 minutes and comes with an associated transcription. We downloaded the audio in MP3 format, then used ffmpeg to convert the audio to 16kHz mono WAV format. We used BeautifulSoup to parse and extract individual chapters.

The text then needed to be aligned to the audio. Since we only needed sentence-level alignment—not word- or phone-level—we used the Estonian model from the MAUS aligner, which proved the most robust across different languages. One recurring issue was that each chapter often included an introductory segment (e.g., the name of the book and chapter number) that was not present in the transcription, which disrupted the alignment. To address this, we padded the beginning of each transcript with the expected introductory text, automatically generated for most languages.

Rukai presented a unique challenge: each chapter began with a 9 to 15-second spoken introduction not found in the text. To handle this, we used an ASR model trained on Amis, a related language. For each Rukai chapter, we transcribed the first 20 seconds using the Amis ASR model, then performed a sliding alignment between this ASR-generated transcription and the gold-standard transcript using Levenshtein distance. We extracted all characters in the ASR transcription that appeared before the best match and prepended them to the transcript. We then force-aligned the audio using the modified transcript and discarded the aligned segments that corresponded to the introduction.

A.2 Dataset Information

Table 1 provides information about each of the datasets used in the paper, including the source, the style or content, the tokens, types, minutes, and token:type ration.

A.3 Word Error Rates

In Tables 2-10, we provide the WER for each combination of model and test set for each language. Each row represents a model trained on the data in the first field of that row. Each column represents the test set indicated in the first field of that column.

Language	Source	Style	Tokens	Types	Minutes	TT Ratio
Amis	`Bible.is`	read speech	208742	9990	2355	0.04
	ePark (A)	fieldwork	18981	5449	240	0.28
	ILRDF (B)	read speech	17608	4984	240	0.28
Bambara	`Bible.is`	read speech	196900	3821	1094	0.019
	Griots (A)	performance	93527	5571	409	0.05
	Fieldwork (B)	conversation	80900	3723	334	0.04
Rukai	`Bible.is`	read speech	191127	14944	2646	0.078
	ePark (A)	fieldwork	14685	7118	240	0.48
	ILRDF (B)	read speech	16907	6036	240	0.35
Swahili	`Bible.is`	read speech	137445	17917	1366	0.13
	ALFFA (A)	fieldwork	102109	14310	660	0.14
	CommonVoice (B)	read speech	66327	14232	660	0.21
Tsou	`Bible.is`	read speech	136732	8390	1992	0.061
	ePark (A)	fieldwork	16848	5784	240	0.343
	ILRDF (B)	read speech	17649	5674	240	0.321
Bribri	`Bible.is`	read speech	211038	10770	1495	0.051
	AmericasNLP (A)	fieldwork	6182	1447	40	0.234
Fongbe	`Bible.is`	read speech	259650	3195	1420	0.01
	ALFFA (A)	fieldwork	55506	1509	420	0.27
Iban	`Bible.is`	read speech	176208	4134	1407	0.02
	OpenSLR (A)	radio show	59576	4197	420	0.07
Quechua	`Bible.is`	read speech	102943	28987	1376	0.281
	AmericasNLP (A)	conversations	19572	7728	224	0.394

Table 1: For each language and each corpus, we show the number of tokens, types, and minutes of audio, along with the type:token ratio (TT Ratio), which is a measure of vocabulary diversity, where values near 0 suggest a limited vocabulary and values near 1 indicate a more diverse vocabulary.

Amis	Source A	Source B	Bible
Source A	0.200	0.243	0.392
Source B	0.407	0.158	0.511
Bible	0.565	0.484	0.030
Bible small	0.578	0.458	0.126
Source A+Bible	0.348	0.353	0.186
Source B+Bible	0.380	0.140	0.109
2phase Bible+A	0.198	0.220	0.251
2phase Bible+B	0.296	0.118	0.284

Table 2: Word Error Rates for Amis

Tsou	Source A	Source B	Bible
Source A	0.319	0.334	0.473
Source B	0.659	0.314	0.775
Bible	0.636	0.639	0.086
Bible small	0.868	0.855	0.391
Source A+Bible	0.446	0.511	0.765
Source B+Bible	0.639	0.331	0.781
2phase Bible+A	0.280	0.331	0.252
2phase Bible+B	0.438	0.205	0.298

Table 3: Word Error Rates for Tsou

Bambara	Source A	Source B	Bible
Source A	0.414	0.726	0.486
Source B	0.608	0.378	0.548
Bible	0.723	0.815	0.184
Bible small	0.746	0.832	0.266
Source A+Bible	0.438	0.720	0.627
Source B+Bible	0.863	0.809	0.895
2phase Bible+A	0.409	0.703	0.318
2phase Bible+B	0.655	0.439	0.398

Table 4: Word Error Rates for Bambara

Rukai	Source A	Source B	Bible
Source A	0.530	0.682	0.582
Source B	0.785	0.275	0.424
Bible	0.811	0.588	0.075
Bible small	0.787	0.578	0.159
Source A+Bible	0.678	0.737	0.708
Source B+Bible	0.863	0.383	0.464
2phase Bible+A	0.499	0.566	0.309
2phase Bible+B	0.758	0.257	0.259

Table 5: Word Error Rates for Rukai

Swahili	Source A	Source B	Bible
Source A	0.241	0.371	0.220
Source B	0.471	0.229	0.313
Bible	0.701	0.639	0.113
Bible small	0.678	0.599	0.152
Source A+Bible	0.404	0.633	0.566
Source B+Bible	0.488	0.220	0.345
2phase Bible+A	0.267	0.492	0.176
2phase Bible+B	0.442	0.184	0.207

Table 6: Word Error Rates for Swahili

Bribri	Source A	Bible
Source A	0.881	0.913
Bible	0.880	0.089
Bible small	1.028	0.343
Source A+Bible	0.908	0.911
2phase Bible+A	0.858	0.537

Table 7: Word Error Rates for Bribri

Quechua	Source A	Bible
Source A	0.812	0.821
Bible	0.915	0.081
Bible small	0.938	0.213
Source A+Bible	0.851	0.712
2phase Bible+A	0.566	0.220

Table 8: Word Error Rates for Quechua

Fongbe	Source A	Bible
Source A	0.270	0.590
Bible	0.608	0.155
Bible small	0.841	0.312
Source A+Bible	0.517	0.171
2phase Bible+A	0.192	0.139

Table 9: Word Error Rates for Fongbe

Iban	Source A	Bible
Source A	0.192	0.249
Bible	0.440	0.036
Bible small	0.560	0.100
Source A+Bible	0.201	0.591
2phase Bible+A	0.156	0.068

Table 10: Word Error Rates for Iban