That doesn't sound right: Evaluating speech transcription quality in field linguistics corpora

Éric Le Ferrand[1], **Bo Jiang**[1], **Joshua Hartshorne**[2], **Emily Prud'hommeaux**[1]
[1]Department of Computer Science, Boston College, USA
[2]MGH Institute of Health Professions, USA

Abstract

Incorporating automatic speech recognition (ASR) into field linguistics workflows for language documentation has become increasingly common. While ASR performance has seen improvements in low-resource settings, obstacles remain when training models on data collected by documentary linguists. One notable challenge lies in the way that this data is curated. ASR datasets built from spontaneous speech are typically recorded in consistent settings and transcribed by native speakers following a set of well designed guidelines. In contrast, field linguists collect data in whatever format it is delivered by their language consultants and transcribe it as best they can given their language skills and the quality of the recording. This approach to data curation, while valuable for linguistic research, does not always align with the standards required for training robust ASR models. In this paper, we explore methods for identifying speech transcriptions in fieldwork data that may be unsuitable for training ASR models. We focus on two complimentary automated measures of transcription quality that can be used to identify transcripts with characteristics that are common in field data but could be detrimental to ASR training. We show that one of the metrics is highly effective at retrieving these types of transcriptions. Additionally, we find that filtering datasets using this metric of transcription quality reduces WER both in controlled experiments using simulated fieldwork with artificially corrupted data and in real fieldwork corpora.

1 Introduction

Automatic speech recognition (ASR) can support the creation of new linguistic resources for under-resourced and endangered languages, but such languages – which make up the vast majority of the world's 7000+ languages – vary considerably in their quantity of transcribed speech data. While some languages have well-curated speech datasets sourced from educational materials, mass media, or crowdsourcing efforts, many only have field recordings made by linguists as their primary source of speech data. Field linguists, whose work centers on describing languages and analyzing their linguistic properties, collect data primarily to support their academic research and the activities of the language community. Few linguists collect data with the goal of creating high-quality datasets for training speech technology models (Hanke, 2017; Le Ferrand, 2023). As a result, speech data from fieldwork may be only partially or unfaithfully transcribed due to issues of recording quality and the language skills of the linguist. Additionally, fieldwork transcripts often include ancillary information making it difficult to differentiate between transcription (word-level renderings of the speech) and annotation (glosses, translations, comments).

ASR models for widely-spoken languages are trained on enough professionally recorded and transcribed data that including a small number of inaccurately transcribed utterances is unlikely to significantly affect overall performance. For languages where data is scarce, however, even a small portion of low-quality data can severely degrade model performance. Detecting low quality transcripts can be done manually, but this process is tedious and time-consuming, underscoring the need for an automatic method to evaluate transcription quality.

In this paper, we explore two metrics for automatically assessing the transcription quality and accuracy of speech datasets: Phonetic Distance Match (PDM), a novel metric based on phoneme recognition, and the posterior probability of a Connectionist Temporal Classification (Graves and Graves, 2012) alignment (CTC). We evaluate the utility of these metrics for identifying poor transcriptions through experiments on clean datasets that we synthetically corrupt in ways that simulate common fieldwork data quality errors. We then demonstrate the ability of PDM in particular to

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 627–635
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

identify these types of errors. Finally we show that using our metrics to filter out these types of inaccurate transcripts can yield substantial improvements in ASR accuracy in both simulated and real-world fieldwork datasets.

2 Related work

Prior related work on filtering inaccurate transcripts has focused on leveraging ASR output itself, such as using the confidence score of an ASR model (Huang et al., 2013; Su and Xu, 2015; Koctúr et al., 2016) or using multiple ASR models to generate multiple predictions for the same speech utterance (Fiscus, 1997; Cucu et al., 2014; Li et al., 2016; Jouvet and Fohr, 2014). These methods, while useful in high-resource settings, assume the existence of an ASR model for the target language, which is not applicable for our work, where no strong ASR models exist. Using universal speech recognition models shows more promise for under-resourced languages. Models such as XLS-R (Conneau et al., 2021) and MMS (Pratap et al., 2024) have demonstrated promising ASR results in low-resource settings (Macaire et al., 2022; Guillaume et al., 2022; Tapo et al., 2024; Romero et al., 2024; Jimerson et al., 2023). However, since these models are trained on raw speech and lack textual information, they cannot provide direct feedback on individual segments. A more suitable alternative is a universal phoneme recognizer (Li et al., 2020), which can generate phone transcriptions for any language.

There is a robust history of prior work in evaluating the acoustic quality of audio using measures of speech intelligibility derived from output of ASR or proto-ASR systems (Holube and Kollmeier, 1996; Sakoe and Chiba, 1978; Spille et al., 2018; Arai et al., 2019). While this work is also relevant for filtering audio for ASR datasets, it is orthogonal to our own work, which focuses on identifying poor quality transcripts rather than poor quality audio.

3 Data

We apply our metrics (see Section 4.1) to two distinct classes of datasets. The CURATED class consists of five well-curated, high-quality speech datasets, ranging in size from 3.5h to 9h, which we synthetically corrupt to simulate common fieldwork transcription errors (see Section 4.2). The languages include Bunun (bnn), Saisiyat (xsy), and Seediq (trv), three Taiwanese indigenous languages extracted from FormosanBank (Mohamed et al., 2024). For each language, we use a subset of the ePark (Aboriginal Language Research and Development Foundation, 2023b) and ILRDF (Aboriginal Language Research and Development Foundation, 2023a) corpora, which consist of read speech recorded by native speakers (Hartshorne et al., 2024). We also included Mboshi (mdw), a Bantu language from Congo-Brazzaville, part of the LIG-Aikuma project[1], and Duoxu (ers), a critically endangered Sino-Tibetan language, included in the Pangloss collection (Michailovsky et al., 2014).

The second class consists of 2-hour fieldwork corpora from Pangloss (FIELDWORK) for Namakura (nmk), an Austronesian language spoken on Vanuatu, and Thulung Rai (tdh), a Sino-Tibetan language of Nepal. These consist exclusively of fieldwork recordings and include annotations and approximate transcripts. We use the FIELDWORK datasets to demonstrate the efficacy of our methods in a real-world fieldwork scenario.

All seven datasets[2] were partitioned into training (70%), validation (10%), and test (20%) sets. Dataset details are found in Table 1.

Several factors motivated our choice of these specific languages. First, the FormosanBank corpus contains an unusually large amount of high-quality data. These languages also posed an initial layer of complexity due to their orthographic conventions. For example, the glottal stop is typically represented with an apostrophe or straight single quote, and the voiceless alveolar affricate is denoted as *c*. Mboshi includes two non-ASCII characters—ɛ and ɯ—while Duoxu features systematic tone marking using superscript numerals (e.g., *ja*^{22}*nje*^{33} *xe*^{53}*nje*^{33} *tɕi*^{33} *o*), adding another dimension of orthographic variation.

4 Method

4.1 Transcript evaluation metrics

We consider two metrics for evaluating transcription quality[3]. First, we present **Phonetic Distance Match (PDM)**, a novel metric for evaluating orthographic transcriptions against their corresponding audio. PDM is calculated by transcribing an utterance recording using a phone-level transcription model and then measuring the edit distance between the resulting transcription and the manual reference transcription. Using Allosaurus (Li

[1] github.com/besacier/mboshi-french-parallel-corpus

[2] https://github.com/eleferrand/data_quality_eval

[3] https://github.com/eleferrand/data_quality_eval

et al., 2020) without fine-tuning, we automatically generate phone-level transcripts for each utterance in the corpus, which are then converted into their closest corresponding ASCII characters using the `unidecode` library.[4] The orthographic reference transcripts are also converted to ASCII to ensure a shared character set. Finally, we compute the normalized Levenshtein distance between the two transcriptions and subtract from 1 to generate a similarity metric ranging from 0 to 1. The scoring process is illustrated in detail in Appendix Fig. 4.

The rationale behind using an ASCII-ized version of IPA is as follows. We begin with the observation that many languages currently being documented are traditionally oral. Their orthographies are often introduced by outsiders who tend adopt the Latin alphabet, with minor modifications. Although exceptions exist (e.g., Ainu written in Japanese katakana or Inuktitut written using Indigenous Canadian syllabics), the Latin script remains the prevalent standard.

While we recognize that linguists and community members who use the Latin alphabet are free to use the characters as they wish in their writing systems, these newly devised orthographies are not arbitrary. They are frequently influenced by existing Latin-based writing systems and the International Phonetic Alphabet (IPA). For example, a voiced velar nasal is typically represented as ŋ or *ng*, and rarely as unrelated letters like *p* or *r*. Naturally, inconsistencies can occur—such as c representing /s/, /k/, or /ʃ/ in French, or /ts/ in Seediq, but overall, we expect the ASCII-ized forms to retain at least phonemic consistency.

There are two advantages to our approach. First, it does not require any prior knowledge about the language or its phonetic inventory, which might not be easily available for a poorly documented language. Second, it does not require additional effort or resources to create a rule-based or learned G2P transformation of the data. In short, the method can be applied to any language that uses at least a partially ASCII-based transcription system without requiring additional model training or in-depth research into the phonetic properties of the language.

The second metric is the **Connectionist Temporal Classification (CTC)** alignment posterior probability. We use a large wav2vec (Baevski et al., 2020) model[5] to extract a speech representation from each utterance, again without fine-tuning; we then apply CTC alignment (Graves et al., 2006) between the speech features and the manual transcription and output the alignment posterior probability. It is entirely independent of the PDM metric.

4.2 Synthetic dataset corruption

To simulate a dataset containing typical fieldwork transcription errors, we arbitrarily select 20% of each training set of the 5 CURATED datasets and introduce transcription errors using three different corruption methods: (1) **Deleted**: three random words are removed from the transcription; (2) **Cropped**: the final 50% of words in the transcription are removed; (3) **Swapped**: the transcription is randomly replaced with another from the training set. For each language, we create three corrupted datasets, each containing 20% of the utterances corrupted in one of these three ways. Each utterance/transcript pair in the three datasets is then scored with the two metrics described in Sec. 4.1. We then evaluate how accurate our metrics identified these corrupted utterances. Examples of corrupted utterances can be found in Table 2.

4.3 ASR model building

All experiments are conducted with XLSR-53 (Conneau et al., 2021), a multilingual model based on the wav2vec architecture. We train a CTC layer for 30 epochs, selecting the best model with the validation set. Decoding is performed using a trigram LM trained on the training set for each language and corruption setting. We follow the popular XLS-R tutorial[6] but do not freeze the feature extractor.

In our **simulated scenario**, we use the three corrupted versions of each CURATED dataset (cf. Sec. 4.2). For each corrupted dataset, as well as for the uncorrupted dataset, we train an ASR model to determine the impact of each corruption on WER. For each corrupted dataset, we then create three filtered datasets: one in which we filter out 20% of the utterances according to the strength of the PDM metric; one where we do the same according to the CTC metric. When training data is limited, removing utterances from the training set can negatively impact performance. To ensure a fair comparison, we also evaluate performance using a dataset where the same percentage of utterances is removed from the training data at random.

In our **real-world scenario**, we calculate the two

[4] https://pypi.org/project/Unidecode

[5] https://huggingface.co/facebook/wav2vec2-base-960h

[6] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

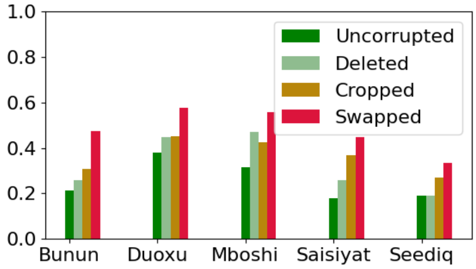


Figure 1: WER across corruption configurations.

metrics on the utterances of the two FIELDWORK datasets. For each dataset, we train ASR models on the unfiltered dataset and on filtered datasets, removing 5%, 10%, and 20% of the utterances using the more promising PDM metric and via random selection.

5 Results

5.1 Detecting corrupted transcripts

Figure 5 shows the full ROC curves and AUC values for all combinations of corruption type, dataset, and metric. We see that PDM achieves near-perfect AUC scores (0.89-0.98) for detecting utterances in the Swapped configuration, very high scores for Cropped (0.77-0.94), and strong scores for Deleted (0.64-0.85). In contrast, CTC is consistently and substantially less effective for all languages in all three corruption settings, with some AUC scores performing at chance in the Deleted and Cropped setting. Notably, Duoxu and Mboshi exhibit lower AUCs perhaps due to weak overlap in character set with the English wav2vec model used (cf. Table 1). The Deleted configuration appears to be the most challenging to detect for both metrics, but with PDM showing a clear advantage over CTC.

5.2 ASR evaluation: Simulated fieldwork

The baseline results for both the uncorrupted and corrupted CURATED datasets, shown in Figure 1, reveal a clear trend. The Deleted configuration causes the least degradation in WER. The Cropped configuration generally yields the second-worst results, except for Mboshi, where Deleted performs worse perhaps because of Mboshi's shorter average utterance length. Finally, the Swapped configuration consistently produces the weakest WER.

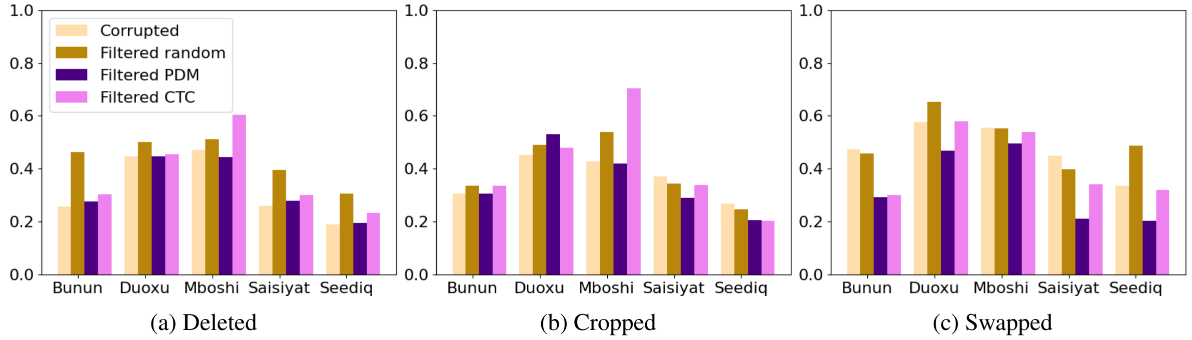
Figure 2 shows changes in WER in the corrupted datasets with and without filtering using the two metrics, PDM and CTC, as well as the random setting where 20% of the data is removed from a corrupted dataset at random. In the Deleted setting, PDM filtering has minimal impact, while CTC filtering generally degrades WER. In the Cropped setting, filtering with PDM improves WER except for Duoxu, while CTC filtering again generally degrades WER. In the Swapped setting, filtering with PDM systematically and often dramatically improves WER, while CTC filtering has little impact except for Bunun and Saisiyat. Overall, the superior performance of PDM filtering is quite consistent, yielding better results than CTC filtering in 14 out of 15 cases. The exception is Duoxu in the Cropped setting. As already noted, Duoxu's writing system contains many non-ASCII characters, which may limit the performance of the PDM metric in some cases.

In a few rare cases, ASR models trained on a corrupted dataset outperform those trained on data filtered using one of the two metrics. This typically occurs when the filtering metric lacks sufficient accuracy, as is observed in some languages with the Deleted and Cropped configurations (see Figure 5), leading to the unintended removal of clean data while allowing corrupted data to remain, ultimately degrading performance. The Deleted setting, which has a minor impact when utterances are relatively long, may also serve unintentionally as a form of corruption-based regularization.

5.3 ASR evaluation: Real-world fieldwork

Figure 3 shows the results of different thresholds of PDM filtering and random filtering on the two FIELDWORK real-world datasets. (We do not report results for CTC given the weak utility observed in the simulated fieldwork scenario both for corruption detection and as a filter.) For Thulung Rai, a 5% filtering threshold proved the most effective, resulting in a decrease of several points in WER, while higher thresholds and random filtering resulted in WER increases. With the Namakura dataset, WER consistently decreased as more data was filtered using the PDM score, suggesting that a significant portion of the corpus may contain transcription errors. Filtering randomly for Namakura yielded slight random variations in WER.

To better understand the utility of the PDM metric for identifying poor transcripts, we manually inspected the transcriptions of the lowest and highest 5% of utterances based on PDM scores for both corpora. In Thulung Rai, 61% of the lowest scoring utterances showed no issues, while 14% had mismatched transcriptions and 23% contained cropped transcriptions. In contrast, 93% of the top scoring

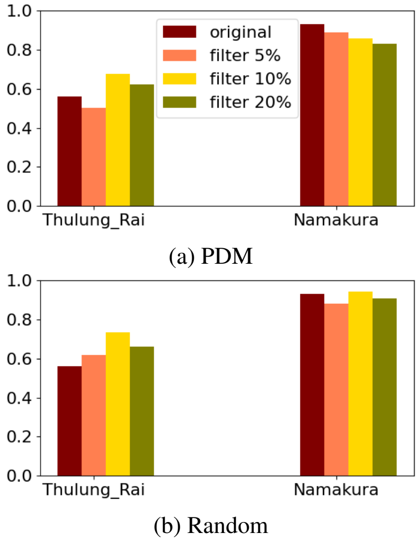


(a) Deleted

(b) Cropped

(c) Swapped

Figure 2: WER for corrupted and filtered CURATED datasets in the simulated fieldwork scenario.



(a) PDM

(b) Random

Figure 3: WER for unfiltered and filtered FIELDWORK datasets in the real-world fieldwork scenario.

utterances had correct transcriptions, with 6% missing a few words. For Namakura, only 11% of the lowest scoring utterances had accurate transcriptions, with 55% mismatched, 29% cropped, 3% with cropped audio, and 1% missing some words. Conversely, the highest scoring utterances had 97% correct transcriptions, with 1.5% cropped and another 1.5% missing words.

6 Conclusions and Future Work

This paper explores two metrics for identifying unsuitable and inaccurate speech transcriptions to improve ASR training from linguistic fieldwork data with the goal of supporting language documentation. We find that our novel PDM metric and, to a lesser extent, a CTC confidence metric are effective in identifying erroneous transcriptions in both simulated and real-world fieldwork datasets. Moreover, filtering data using the PDM metric consistently reduces WER in both simulated and real-world fieldwork scenarios. In our future work, we plan to investigate additional methods for identifying poor transcriptions and to explore the relationship between audio quality and transcription quality.

Limitations

Experimental results demonstrate that our PDM method is highly effective for languages with a limited number of non-ASCII characters. However, further experiments are needed to evaluate its performance on languages with a larger set of non-ASCII characters and non-Latin writing systems. The proposed metrics efficiently identify major errors, such as missing or mismatched transcripts, but are less likely to detect spelling mistakes or inconsistent transcription of specific speech sounds, which could also significantly impact WER. While these methods could be applied to high-resource languages like French or German, such languages may benefit more from approaches leveraging existing G2P models or pre-trained ASR systems trained specifically for these languages.

Ethics Statement

Researchers must always be respectful of language community concerns about data ownership when working with Indigenous language data. All of our data is gathered from public sources. In the case of the Formosan languages, the two organizations providing the data, the Indigenous Languages Research and Development Foundation and the ePark educational research organization, actively seek out collaborations with computational researchers. The other datasets are also made available on the Web by their creators specifically with the goal of furthering research in these languages, both linguistic and computational. We have permission from the creators and owners to redistribute the data in the form of ASR datasets.

Acknowledgments

The authors thank Yuyang Liu, Li-May Sung, and the Indigenous Languages Research and Development Foundation, especially Akiw and Lowking Nowbucyang, for generously providing data. 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

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

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

Kenichi Arai, Shoko Araki, Atsunori Ogawa, Keisuke Kinoshita, Tomohiro Nakatani, Katsuhiko Yamamoto, and Toshio Irino. 2019. Predicting speech intelligibility of enhanced speech using phone accuracy of DNN-based ASR system. In *Interspeech*, pages 4275–4279.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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

Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2014. Unsupervised acoustic model training using multiple seed asr systems. In *Spoken Language Technologies for Under-Resourced Languages*.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves and Alex Graves. 2012. Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks*, pages 61–93.

Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyên, Maxime Fily, Guillaume Jacques, and Alexis Michaud. 2022. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource setting. In *Proceedings of Interspeech*, pages 4905–4909.

Florian Hanke. 2017. *Computer-Supported Cooperative Language Documentation*. Ph.D. thesis, Ph. D. thesis, University of Melbourne.

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*.

Inga Holube and Birger Kollmeier. 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3):1703–1716.

Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu. 2013. Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration. In *Interspeech*, pages 2360–2364.

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.

Denis Jouvet and Dominique Fohr. 2014. About combining forward and backward-based decoders for selecting data for unsupervised training of acoustic models. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*.

Tomáš Koctúr, Ján Staš, and Jozef Juhár. 2016. Unsupervised acoustic corpora building based on variable confidence measure thresholding. In *2016 International Symposium ELMAR*, pages 31–34. IEEE.

Éric Le Ferrand. 2023. *Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities*. Ph.D. thesis, Charles Darwin University.

Sheng Li, Yuya Akita, and Tatsuya Kawahara. 2016. Data selection from multiple ASR systems' hypotheses for unsupervised acoustic model training. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5875–5879. IEEE.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics,* *Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

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*.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation*, 8:119–135.

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

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.

Monica Romero, Sandra Gómez-Canaval, and Ivan G Torre. 2024. Automatic speech recognition advancements for Indigenous languages of the Americas. *Applied Sciences*, 14(15):6497.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Constantin Spille, Stephan D Ewert, Birger Kollmeier, and Bernd T Meyer. 2018. Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, 48:51–66.

H Su and H Xu. 2015. Multi-softmax deep neural network for semi-supervised training. In *Proceedings of Interspeech*, pages 3239–3243.

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 *Proceedings of Interspeech 2024*, pages 5088–5092.

A Appendix

Table 1 shows the durations, token and type counts, and non-ASCII character proportions of each of the 7 datasets for reference purposes. As noted in the paper, we have released these corpora and their partitions for research purposes. We note that they are derived in their entirety from publicly available sources with licensing that permits redistribution in other formats.

Table 2 provides examples of the three types of corruption designed to mimic the kinds of errors observed in fieldwork transcripts.

Figure 2 provides a walk-through of the PDM calculation process with three example utterances.

Figure 5 plots all six ROC curves and reports AUC measures for using each of the two metrics, PDM and CTC, to identify corruptions for each CURATED dataset under of the three corruption settings.

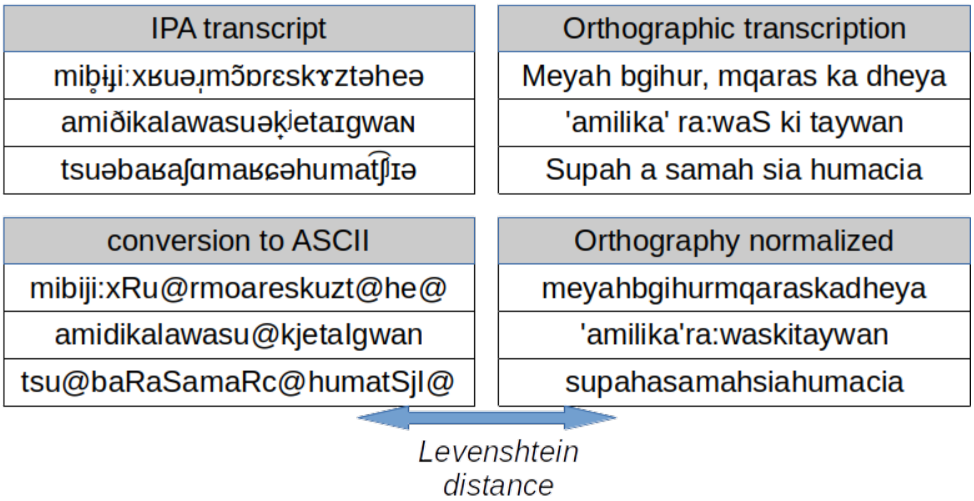
Table 3 shows the WER results presented in the paper in graphical format in Figure 2.

	Bunun	**Duoxu**	**Mboshi**	**Saisiyat**	**Seediq**	**Namakura**	**Thulung Rai**
Duration	8h34	7h57	3h28	8h14	8h53	1h53	2h18
Token	40166	61564	25671	39644	50123	18566	16296
Type	6846	2557	5621	4723	5608	1065	3965
Non-ASCII	0%	44%	23%	0%	2%	1%	8%

Table 1: Corpus size and token/type count for all datasets. We also provide the percentage of non-ASCII characters which may have an impact on the utility of the PDM metric.

Config.	**Original**	**Corrupted**
Deleted	wa adi pósá báabará wa kaá kobhá epɷrɷrɷ baá óyáála mwána anyɷɷ	wa pósá wa epɷrɷrɷ óyáála mwána
Cropped	maqasmav a abus malaitaz a savi to seediq msgelu sa seediq mneyah alang kiya	maqasmav a abus to seediq msgelu sa
Swapped	supah a samah sia humacia tai hari niqan rebuq watan dao su trebuq hii	anak anak sa ia maupacia minhanglas slii hini kanna nnapa namu bunga

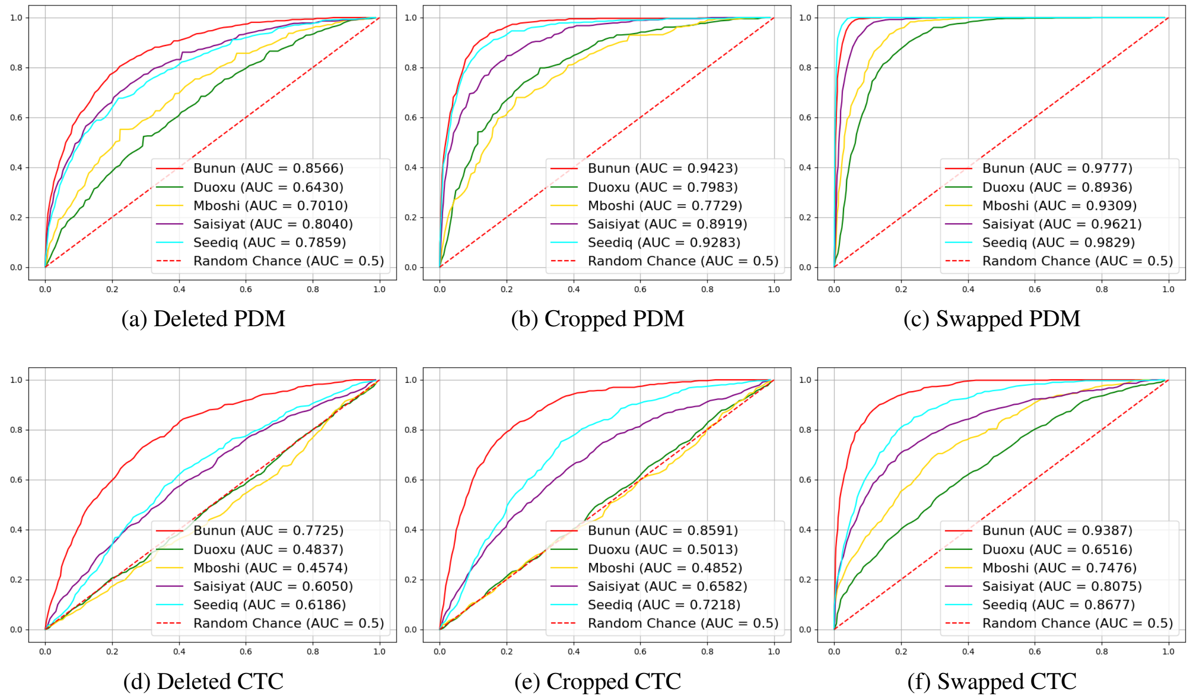
Table 2: Examples of input utterances and their corruptions from the three corruption configurations.



IPA transcript	Orthographic transcription
mib̥ɨɟiːxʁuəɹ̩mɔ̃ɒrɛskɤztəheə	Meyah bgihur, mqaras ka dheya
amiðikalawasuəḳʲetaɪgwaɴ	'amilika' ra:waS ki taywan
tsuəbaʁaʃɑmaʁɕəhumat͡ʃɪə	Supah a samah sia humacia

conversion to ASCII	Orthography normalized
mibiji:xRu@rmoareskuzt@he@	meyahbgihurmqaraskadheya
amidikalawasu@kjetaIgwan	'amilika'ra:waskitaywan
tsu@baRaSamaRc@humatSjI@	supahasamahsiahumacia

Figure 4: Demonstration of the PDM calculation method. In the upper left we see IPA transcripts generated from audio by Allosaurus. In the upper right we see the corresponding reference orthographic transcription for the three sample utterances. In the lower left are the phone-level transcripts converted to their ASCII equivalents often used to represent those IPA symbols (e.g., with Sampa). In the lower right, the reference orthographic transcriptions converted to ASCII, with spaces removed. We calculate normalized Levenshtein distance between the utterance in the lower panels and subtract from 1 to create the PDM metric.



(a) Deleted PDM

(b) Cropped PDM

(c) Swapped PDM

(d) Deleted CTC

(e) Cropped CTC

(f) Swapped CTC

Figure 5: ROC curves comparing performance of PDM and CTC for retrieving corrupted trasncriptions under the three corruption settings for all five of the CURATED datasets.

Corruption Setting	**Filtering Method**	**Bunun**	**Duoxu**	**Mboshi**	**Saisiyat**	**Seediq**
Deleted	*Unfiltered*	**0.2573**	**0.4464**	0.4717	**0.2600**	**0.1899**
	Random	0.4620	0.5008	0.5121	0.3943	0.3075
	PDM	0.2754	0.4478	**0.4446**	0.2779	0.1947
	CTC	0.3025	0.4562	0.6048	0.3002	0.2325
Cropped	*Unfiltered*	0.3065	**0.4514**	0.4268	0.3700	0.2692
	Random	0.3372	0.4911	0.5402	0.3428	0.2473
	PDM	**0.3062**	0.5311	**0.4198**	**0.2908**	0.2049
	CTC	0.3351	0.4806	0.7046	0.3387	**0.2034**
Swapped	*Unfiltered*	0.4743	0.5759	0.5562	0.4500	0.3356
	Random	0.4584	0.6538	0.5513	0.3994	0.4881
	PDM	**0.2940**	**0.4691**	**0.4951**	**0.2106**	**0.2036**
	CTC	0.3014	0.5800	0.5396	0.3421	0.3207

Table 3: WER for each combination of simulated corruption setting and filtering method for each of the five CURATED datasets. This same information is visualized in bar graph format in Figure 2. The lowest WER in for each language/corruption is boldfaced.