FRUSTRATINGLY EASY DATA AUGMENTATION FOR LOW-RESOURCE ASR

Katsumi Ibaraki

University of Notre Dame, USA
Dept. of Computer Science and Engineering
kibaraki@nd.edu

David Chiang

University of Notre Dame, USA
Dept. of Computer Science and Engineering
dchiang@nd.edu

arXiv:2509.15373v3 [cs.CL] 17 Jan 2026

ABSTRACT

This paper introduces three self-contained data augmentation methods for low-resource Automatic Speech Recognition (ASR). Our techniques first generate novel text—using gloss-based replacement, random replacement, or an LLM-based approach—and then apply Text-to-Speech (TTS) to produce synthetic audio. We apply these methods, which leverage only the original annotated data, to four languages with extremely limited resources (Vatlongos, Nashta, Shinekhen Buryat, and Kakabe). Fine-tuning a pre-trained Wav2Vec2-XLSR-53 model on a combination of the original audio and generated synthetic data yields significant performance gains, including a 14.3% absolute WER reduction for Nashta. The methods prove effective across all four low-resource languages and also show utility for high-resource languages like English, demonstrating their broad applicability.

***Index Terms*—** Automatic Speech Recognition (ASR), data augmentation, low-resource languages

1. INTRODUCTION

Despite substantial improvements in Automatic Speech Recognition (ASR), state-of-the-art models remain dependent on large-scale datasets; most progress has centered around high-resource languages [1]. Even research in low-resource settings often uses additional text data, such as dictionaries, for data processing or synthesis. However, these resources may be unavailable for endangered or less-studied languages, and new data collection is often infeasible due to a scarcity of speakers, the remote geographic locations where they reside, or other logistical barriers. This paper targets this gap by introducing simple data augmentation methods that expand a corpus using only its existing annotated data. We propose three techniques: (1) replacing words by words with the same gloss, (2) replacing words with random words, and (3) generating new sentences using a Large Language Model (LLM). We use these methods to synthesize new sentences, which are then converted into synthetic speech audio using a Text-to-Speech (TTS) model. Our replacement methods are an extension of semantic- and template-based replacement techniques [2, 3], adapted to work without external lexical databases, and allowing simultaneous, multi-word substitutions. We evaluate our methods on four low-resource languages—Vatlongos, Nashta, Shinekhen Buryat, and Kakabe—and on the high-resource English LibriSpeech [4], to assess their broader applicability.

In summary, we make three contributions. First, we present an approach that combines text and audio data augmentation to generate synthetic training data from a labeled corpus. Second, we develop ASR models for Vatlongos, Nashta, Shinekhen Buryat, and Kakabe, which, to our knowledge, are the first ASR systems for each of these languages. Third, we demonstrate that for ASR in low-resource settings, prioritizing phonemic and structural variety can be more effective for model training than preserving semantic coherence.

Our code is available at https://github.com/kibaraki/frustratingly-easy-asr-augmentation

2. RELATED WORK

Data augmentation is a well-established technique for improving text-based models. Common strategies range from simple operations—such as synonym replacement via WordNet [5] and random word insertion, swapping, or deletion [2]—to more sophisticated methods. These include using bidirectional LSTMs for contextual augmentation [6] and round-trip translation to paraphrase text for machine translation [7].

In ASR, these text augmentation techniques are combined with speech synthesis to expand the training corpora. For instance, Zevallos et al. [8], using delexicalization methods [3], substitute words within predefined semantic frames. While effective, the applicability of these methods is limited. First, the substitutions are typically restricted to a few common semantic categories within a sentence, which makes most of the sentence structure unchanged. Second, and more critically for our purposes, these techniques are dependent on external lexical or semantic resources to identify and supply replacements. More recently, LLMs have been used for synthetic data generation. Previous work has used LLMs for relatively constrained tasks, such as paraphrasing text to improve n-gram language models [9], or generating in-domain data for adaptation to new domains [10]. While more unconstrained, novel sentence generation has been explored in other contexts, its application specifically as a data augmentation technique for low-resource ASR, especially for languages that are truly unseen or typologically distant from a model's pretraining data, has received less attention.

3. METHOD

3.1. Synthetic Text Generation

We explore three different methods for synthetic text generation: gloss-based replacement, random replacement, and LLM-based generation. The two replacement methods are shown in Figure 1.

Gloss-based: Previous work [8, 3] has typically utilized datasets with consistent sentence structures and high semantic overlap, allowing for the extraction of frames like *distance* or *location*. In contrast, our data consists of more unstructured, freely spoken speech. This lack of regularity makes template-based augmentation impractical without extensive, handcrafted rules, and the varied contexts make it ineffective to focus only on frequent tags. Therefore, we

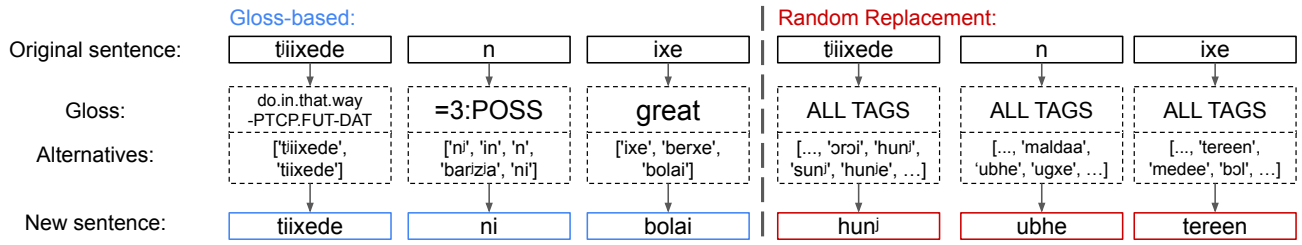


Fig. 1: Illustration of gloss-based replacement (left) and random replacement (right) applied to the beginning of the Shinekhen Buryat sentence, "Then the first queen told him." The gloss-based method replaces each word with an alternative from the set of all words sharing the same gloss in the training data. In contrast, the random replacement method ignores all linguistic information, substituting each word with a random selection from all words in the training data.

```
Given the following CSV, focus on columns [text,
clean_text, english, gloss] and generate {number
of sentences in train} sentences in a CSV with
all of the original columns, consisting of
only the new sentences; this is in {language},
{language description}; do not use Python code
to generate the sentences but rather use your
understanding of other languages as an LLM to
generate sentences; make sure that the text
and gloss generated match; this text will be
passed on to a TTS model to generate synthetic
audio, to use for additional training data for a
wav2vec2-based ASR model.
```

Fig. 2: LLM prompt for generating synthetic sentences.

adopt a more comprehensive strategy: we treat every word position as a candidate for substitution and use glosses (brief definitions and/or grammatical tags) instead of formal semantic frames, which would require nonexistent external resources. A dictionary is constructed from the training data by mapping each gloss to a list of all words that share it, based on the provided gloss and POS information. Treating each original sentence as a template for guiding gloss order for each sentence, we refer to the dictionary to randomly select a possible alternative, and continue the process for each position. Although the sentences maintain the same glosses as the original, this process does not ensure that the sentences are grammatically correct or are natural-sounding to a native speaker. While we acknowledge the limitations of this method, given the limited resources for these languages, this approximation is a pragmatic approach.

Random Replacement: As a more extreme variant of gloss-based replacement, we also introduce a random replacement method. In this approach, each original sentence serves as a structural template solely to define the length of the new sentence. For each word position, a replacement is randomly sampled from the entire vocabulary of the training data, completely disregarding any gloss or semantic information. While this technique ensures that all generated words are in-vocabulary, the resulting sentences are not expected to be grammatical or semantically coherent.

LLM-based: In addition to replacement-based techniques, we use Google's Gemini 2.5 Pro [11] to generate entirely new sentences. Our approach deliberately leverages the model's capacity for hallucination, a typically avoided behavior, to create novel words and syntactic structures not present in the original small training set. The model was prompted with the existing training data, including transcriptions, translations, and gloss/POS information, as detailed in Figure 2. This generative method was pursued after determining that other models, such as OpenAI's ChatGPT, were unsuitable, as they tended to produce code for rigid, template-based sentences using limited semantic categories, rather than the desired novel text.

3.2. Synthetic Audio Generation

The three types of synthetic text were converted into speech using Kokoro, an open-weight TTS model that can process both IPA and orthographic inputs.[1] For this synthesis, we used a standardized set of five voices across all languages rather than attempting to clone the voices of the original speakers. This decision was based on two factors. First, the limited audio per speaker was insufficient for fine-tuning individual, high-quality TTS models; using a uniform set of voices ensures experimental consistency. Second, preliminary experiments in voice conversion, where each original speaker's voice was used to generate all training sentences, did not yield any performance improvements.

3.3. Fine-tuning Wav2Vec2

We fine-tune the pre-trained Wav2Vec2-XLSR-53 model [12] using a Connectionist Temporal Classification (CTC) loss [13]. Hyperparameters were selected via preliminary experiments to prevent overfitting: we used a learning rate of 1.0×10^{-4} and set language-specific training epochs ranging from 10 to 50. To ensure a controlled experiment, we excluded other augmentations like SpecAugment [14] and maintained a 1:1 ratio of synthetic to original data, as initial tests showed larger proportions of synthetic data harmed performance.

4. EXPERIMENT SETUP AND RESULTS

4.1. Data

We focus on four low-resource languages and dialects, all with five hours or less of audio data. To our knowledge, all transcriptions were made manually, without the use of grapheme-to-phoneme tools. All audio clips were downsampled to 16 kHz for training.

For the data splits, we allowed speaker overlap, following Wei et al. [15] and Liu et al. [16], who demonstrated that this has a negligible impact in low-resource contexts. While the level of transcription detail varies across our data sources, we treat all IPA-based transcriptions as phonemic representations and refer to the units as *phonemes*. This aligns with the objective of our ASR models, which is to learn the contrastive sound system of each language.

[1] https://huggingface.co/hexgrad/Kokoro-82M

	Minutes	Speakers	Total Words	Total Unique	Train Words	Train Unique	Gloss	% Alt.	% Out
Vatlongos	286	60	66984	3644	23315	2823	422	10.0	55.4
Nashta	48	4	5473	1643	3982	886	637	29.5	95.3
Kakabe	99	34	17798	1830	12512	1525	259	24.3	20.5
Shinekhen Buryat	97	4	11338	3089	7895	2276	2242	11.4	60.0
English									
LibriSpeech-54	54	40	8738	2575	6164	2039	396	55.8	99.8
LibriSpeech-108	108	40	17449	4212	12236	3326	491	58.9	99.9
LibriSpeech-324	324	40	52576	8138	36842	6644	592	63.9	99.9
LibriSpeech-1207	1207	251	198291	17243	137411	14263	902	67.8	99.9

Table 1: Summary of corpus statistics for each language. The table shows the number of minutes of transcribed audio, speakers, total and unique words in the corpus, total and unique words in training, gloss tags in training, percentage of glosses with substitution alternatives in training, and the out-of-vocabulary rate for LLM-generated text relative to the training split. The LibriSpeech splits are marked with the number of minutes; the first three are from `test-clean` and 1207 is from `train-clean-100` as labeled in [4].

Vatlongos is a language of Vanuatu, an island nation in the South Pacific, with an estimated 3,000 speakers [17, 18]. For this study, we use a corpus collected by Ridge (Massey University) and hosted on the Pangloss Collection.[2] Each utterance is annotated with a text transcription (orthography and broad IPA conversion), English and Bislama translations, and gloss. An example is shown below:

(1) *Dilamun* *ba* *biteni* *mama* *nan*
3s.return.ind 3s.go.ind 3s.say.ind mother cl.gen-3s.poss
'He turned back and told his mother.'

Nashta is an under-documented South Slavic variety spoken in the Balkans that is related to literary Bulgarian and Macedonian and shows Greek influence [19]. The data comes from a corpus on the Pangloss Collection,[3] collected by Adamou (National Centre for Scientific Research in France), who established "Nashta" as a scholarly convention for this variety. The dataset is annotated with broad IPA transcriptions, French and English translations, and gloss. The following example is taken from this corpus:

(2) *i* *tam* *ˈsetne* *kaˈva* *ˈxodexa* *tam* *ˈnare*
and there then when go.IPFV there up
'And then, when they would go to the mountain'

Shinekhen Buryat is a Mongolic language variety spoken by approximately 6,000 people in Inner Mongolia, China [20]. It is a variant of the broader Buryat language, which has around 520,000 speakers across Russia, China, and Mongolia [21]. The data for this study was collected and transcribed by Yamakoshi (Tokyo University of Foreign Studies).[4] The corpus is annotated with a more narrow IPA transcription, English and Japanese translations, and gloss. An example from the corpus is provided below:

(3) *ii-g-eed* *zalg-aad* *ɔdɔɔ*
do.like.this-E-CVB.PFV continue-CVB.PFV now
ajan *xʊʊ-zʲe*
journey:INDF follow-CVB.IPFV
'Then he continued his journey.'

[2] `https://pangloss.cnrs.fr/corpus/Vatlongos?lang=en`

[3] `https://pangloss.cnrs.fr/corpus/Nashta?lang=en`

[4] `https://tufs.repo.nii.ac.jp/search?search_type=2&q=1729497608274`

Kakabe is spoken in Guinea, with approximately 50,000 speakers, and belongs to the Mande languages, but has influence from Pular [22]. We use the corpus collected by Vydrina (National Centre for Scientific Research in France), hosted on the Pangloss Collection.[5] The corpus contains a broad IPA transcription, English translations, and gloss. An example from the corpus is shown below:

(4) *a* *fɔ́ɔ* *káa,* *óma* *n'* *a* *fɔ́ɔ* *kélen* *kélen*
3SG say this.way 1PL.INCL OPT 3SG say one one
'Tell that we have to speak one by one.'

LibriSpeech is a corpus of read English speech [4]. To evaluate our methods on a high-resource language which the model has seen during pre-training, we use the `test-clean` and `train-clean-100` splits from this corpus. To emulate the documentation-style corpora, we use `spaCy` to add POS and dependency tags.

4.2. Evaluation

We evaluate model performance using three metrics: Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER). CER is computed for our orthography-based models, while PER, which is analogous to CER but operates on phonemes, is computed for our IPA-based models. For cross-linguistic analysis, both CER and PER are reported, though we acknowledge these metrics are not directly comparable. Additionally, we assess the statistical significance of improvements over the baseline for all metrics using paired bootstrap resampling [23, 24], with 10,000 samples and a significance level of 0.05.

4.3. Results

The results are displayed in Table 2. The improvements of the metrics mentioned below are based on the test set results.

For Vatlongos, all three data augmentation techniques improved performance over the baseline across all metrics, and we observed no differences between using IPA or orthography for sentence generation. The most effective method was random replacement, which achieved a 1.4% absolute (8.8% relative) reduction in PER and a 4.4% absolute (9.4% relative) reduction in WER. The gloss-based

[5] `https://pangloss.cnrs.fr/corpus/Kakabe?lang=en`

		Vatlongos				Nashta				Shinekhen Buryat				Kakabe			
		Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM
PER/CER	Val	16.0	14.8	**14.4**	15.2	26.4	22.3*	**22.0***	22.5*	13.9	13.2	**13.0***	13.4	23.7	23.2	**23.0**	23.4
	Test	16.3	15.3*	**14.9***	15.3	24.9	**20.0***	**20.0***	20.3*	14.9	**14.3**	14.9	15.0	22.4	22.0	**21.4***	21.6*
WER	Val	48.8	44.1*	**42.7***	45.1*	77.2	65.3*	**64.7***	64.8*	46.1	**44.4**	44.7	45.9	52.6	**51.1**	51.9	53.7
	Test	47.4	**43.0***	**43.0***	43.7*	75.4	**61.1***	63.2*	66.1*	48.1	**44.6**	44.9	47.7	50.5	49.7	49.3*	**49.1***

		English															
		LibriSpeech-54				LibriSpeech-108				LibriSpeech-324				LibriSpeech-1207			
		Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM	Base	Gloss	Rand.	LLM
PER/CER	Val	6.6	8.7	8.2	**4.8***	4.6	5.8	5.0	**3.5**	4.3	**2.7***	4.0*	2.8*	1.9	1.6	**1.5**	1.9
	Test	6.7	8.7	8.2	**4.6***	4.5	5.7	4.9	**3.6***	4.2	2.9*	4.0	**2.7***	1.8	1.6	**1.4***	1.8
WER	Val	24.9	32.7	30.9	**17.2***	17.4	22.3	18.8	**12.4***	16.7	**10.0***	15.4*	**10.0***	6.9	5.6	**5.3***	6.6
	Test	24.9	31.9	30.6	**16.3***	16.7	21.5	18.4	**12.6***	16.4	10.5*	15.4*	**10.3***	6.7	5.5	**5.1***	6.3

Table 2: The results for all models. All values in the table are in percentage (%). The improvements that are statistically significant from the baseline (at significance level of 0.05), using paired bootstrap resampling [23, 24], are indicated with an asterisk (*).

and LLM-based methods also yielded strong results, with the former achieving a WER reduction of 4.4% (9.4% relative). Across all methods, the relative improvement in WER was larger than in PER.

With Nashta, all methods yielded significant improvements, with gloss-based and random replacement performing best. The gloss-based method reduced PER by 4.9% (19.8% relative) and WER by 14.3% (18.9% relative). Notably, and in contrast to other languages, the relative improvements in PER for Nashta were comparable to those in WER, suggesting the augmentations provided balanced benefits at both the phoneme and word levels.

Of the low-resource languages tested, Shinekhen Buryat produced the most mixed results, with the effectiveness of the augmentation varying significantly by method. The gloss-based approach performed best, improving PER by 0.6% (4.1% relative) and WER by 3.5% (7.3% relative) over the baseline. Random replacement also provided an improvement; in contrast, the LLM-based method had a much smaller effect when compared to the other languages. For the two replacement methods, the relative gains in WER were more notable than in PER.

For Kakabe, we had statistically significant, though more modest, improvements. Random replacement was the most effective, achieving a PER reduction of 1.0% (4.6% relative) and a WER reduction of 1.2% (2.3% relative). The gloss-based replacement and LLM-based approach yielded similar results.

For LibriSpeech, the LLM-based augmentation provided consistent improvements, while the replacement methods underperformed at smaller data subsets. However, on larger subsets, both replacement methods became highly effective. Gloss-based replacement achieved a reduction of up to 1.3% (31.5% relative) in PER and 5.9% (36.1% relative) in WER for LibriSpeech-324, and random replacement achieved a reduction of up to 0.4% (22.5% relative) in PER and 1.5% (22.6% relative) in WER for LibriSpeech-1207.

5. DISCUSSION

The effectiveness of the three proposed augmentation methods varied significantly across different data resource conditions. In the four low-resource languages, the simple, self-contained replacement methods were the most reliable source of performance gains. The choice between the two methods appears language-dependent, with the random method excelling for Vatlongos, Nashta, and Kakabe, and the gloss-based method for Shinekhen Buryat. Contrary to expectations from prior work emphasizing semantic consistency, the fully random replacement method was marginally superior to the more structured gloss-based approach for several tests. This leads to a takeaway: in extremely data-scarce ASR settings, increasing the sheer variation of phonemic and structural patterns—even when constrained to the existing vocabulary—can be more beneficial for model training than preserving semantic coherence. In the high-resource LibriSpeech setting, both replacement methods were initially detrimental, only becoming beneficial after the dataset size was larger than any of the low-resource languages' datasets.

The LLM-based approach yielded more varied outcomes. It performed competitively for some languages, while underperforming the baseline or not improving for others. We speculate that its unique strength lies in its ability to hallucinate novel words not present in the training corpus, which may help the model generalize to unseen vocabulary. However, its unreliability makes it a higher-risk strategy. With LibriSpeech, it was the most consistent across different data sizes, but we note that because it was in English, the LLM did not need to synthesize in the same way as the other languages. Ultimately, simple and easily replicable methods like gloss-based and random replacement offer a reliable way to expand datasets and improve low-resource ASR.

6. CONCLUSION

This study evaluated three data augmentation techniques—gloss-based replacement, random replacement, and LLM-based generation—for low-resource ASR by generating synthetic data for four languages. All methods yielded significant performance gains, most notably for Nashta, where our best approach achieved absolute reductions of 4.9% in PER and 14.3% in WER. Meaningful WER reductions were also achieved for the other three languages. Furthermore, we demonstrated the versatility of these techniques by successfully applying them to English, confirming they are viable strategies for improving Wav2Vec2-based ASR models in data-scarce environments.

7. ACKNOWLEDGEMENTS

The authors have no relevant financial or nonfinancial interests to disclose.

8. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Pangloss Collection (Vatlongos: CC BY-NC-ND 3.0, Nashta: CC BY-NC 2.5, Kakabe: CC BY-NC-ND 3.0), Tokyo University of Foreign Studies (CC BY-SA 4.0), and Panayotov et al. [4] (CC BY 4.0). Ethical approval was not required as confirmed by the license attached with the open access data.

9. REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[2] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6382–6388.

[3] Y. Hou, Y. Liu, W. Che, and T. Liu, "Sequence-to-sequence data augmentation for dialogue language understanding," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1234–1245.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[5] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[6] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, June 2018, pp. 452–457.

[7] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, pp. 6256–6268, 2020.

[8] R. Zevallos, N. Bel, G. Cámbara, M. Farrús, and J. Luque, "Data augmentation for low-resource Quechua ASR improvement," in *Proceedings of Interspeech*, 2022, pp. 3518–3522.

[9] T. Nagano, G. Kurata, S. Thomas, H. J. Kuo, D. Bolanos, H. Jung, and G. Saon, "LLM based text generation for improved low-resource speech recognition models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[10] H. Su, T. Hu, H. S. Koppula, R. Vemulapalli, J. R. Chang, K. Yang, G. V. Mantena, and O. Tuzel, "Corpus synthesis for zero-shot ASR domain adaptation using large language models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12326–12330.

[11] Google, "Gemini," `https://gemini.google.com`, Aug. 2025.

[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech 2019*, 2019, pp. 2613–2617.

[15] X. Wei, C. Cucchiarini, R. van Hout, and H. Strik, "Automatic speech recognition and pronunciation error detection of Dutch non-native speech: Cumulating speech resources in a pluricentric language," *Speech Communication*, vol. 144, pp. 1–9, 2022.

[16] Z. Liu, J. Spence, and E. Prud'hommeaux, "Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 123–131.

[17] E. Ridge, "Language contexts: Vatlongos, southeast Ambrym (Vanuatu)," *Language Documentation and Description*, vol. 15, 2018.

[18] T. Crowley, "The language situation in Vanuatu," *Current Issues in Language Planning*, vol. 1, no. 1, pp. 47–132, 2000.

[19] E. Adamou, "Nashta grammatical sketch with examples linked to online corpus," *Slavische Mikrosprachen im absoluten Sprachkontakt*, 2022.

[20] Y. Yamakoshi, "Three folktales in Shinekhen Buryat," *Asian and African Languages and Linguistics*, vol. 6, pp. 109–136, 2011.

[21] B. Khabtagaeva, "The Shinekhen Buryat variety: Preliminary analysis," *International Journal of Eurasian Linguistics*, vol. 4, no. 2, pp. 217 – 236, 2023.

[22] A. Vydrina, *A corpus-based description of Kakabe, a Western Mande language: prosody in grammar*, Ph.D. thesis, Université Sorbonne Paris Cité, 2017.

[23] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1, pp. I–409.

[24] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 388–395.