Linguistically Informed Tokenization Improves ASR for Underresourced Languages

Massimo Daul
New York University
mmd9604@nyu.edu

Alessio Tosolini
McGill University
alessio.tosolini@mail.mcgill.ca

Claire Bowern
Yale University
claire.bowern@yale.edu

Abstract

Automatic speech recognition (ASR) is a useful tool for linguists aiming to perform a variety of language documentation tasks. However, modern ASR systems use data-hungry transformer architectures, rendering them generally unusable for underresourced languages. We fine-tune a wav2vec2 ASR model on Yanˆnhangu, an Indigenous Australian language, comparing the effects of phonemic and orthographic tokenization strategies on performance. In parallel, we explore ASR's viability as a tool in a language documentation pipeline. We find that a linguistically informed phonemic tokenization system substantially improves WER and CER compared to a baseline orthographic tokenization scheme. Finally, we show that hand-correcting the output of an ASR model is much faster than hand-transcribing audio from scratch, demonstrating that ASR can provide significant assistance for underresourced language documentation.

1 Introduction

Automatic Speech Recognition (ASR, also known as speech-to-text) is a natural language processing technology that converts spoken words to text. Most modern ASR systems, such as wav2vec2 (Baevski et al., 2020) and Whisper (Radford et al., 2022) are neural and transformer-based, working well for languages with large amounts of training data but falling short for those without. ASR is used for a broad range of human-computer interaction tasks, but there exists a big resource gap, where only a small number of languages have robust and freely available ASR models. Common Voice (Ardila et al., 2020), for example, covers only 2% of the world's languages. The geographic distribution of these languages is also unequal, with no Indigenous Australian languages represented in this corpus.

One reason for this asymmetry is that for low-resource languages, ASR training data comes from linguistic fieldwork where manual transcription and annotation are both time-consuming and require specialist knowledge. Few documentation projects have the resources to train and pay annotators (Chelliah, 2001), with transcription taking anywhere from 5 minutes to over an hour per minute of speech (Dwyer, 2006). That means that there is both much less training data for underresourced languages, and that the means to acquire such training data is extremely labor-intensive. Accurate ASR would vastly assist in this respect.

ASR models are trained to predict sequences of tokens: discrete textual units. In the context of ASR, tokenization refers to the process of segmenting transcribed speech into meaningful units, such as words, subwords, or characters. Different tokenization strategies may impact vocabulary size, error rates, and adaptability to various languages and domains, with the best tokenization strategies for high-resource languages not necessarily being the same as those for low-resource languages (Adlaon and Marcos, 2024; Bañeras-Roux et al., 2024). Additionally, linguistically informed tokenization strategies – i.e., ones where the tokenization occurs across phonological (Atuhurra et al., 2024; Liao and Shi, 2026) or morphological (Bayram et al., 2025; Hofmann et al., 2021) units – have been shown to improve model performance for some low-resource tasks. This paper investigates whether linguistically informed phonemic tokenization improves ASR accuracy for Yanˆnhangu, and evaluates its practical impact within a fieldwork documentation pipeline. We show that linguistically informed ASR improves error rates and substantially speeds up transcription. Furthermore, linguistically transparent tokenization aligns more closely with fieldwork pipelines by enabling local, interpretable computation, thereby supporting data sovereignty.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 31–37

2 Methodology

2.1 Data origin and preprocessing

The corpus for this test includes selected recordings of the Yan-nhangu language (Glottolog code YANN1237; ISO-639 JAY; Pama-Nyungan). Recordings were made from 5 fluent speakers of Yan-nhangu between 2004 and 2007. Since the Yan-nhangu data originated from different elicitation sessions, the recording quality is variable, though the recordings being made in the field allows the results from this experiment to be generalized to other field environments. The data was preprocessed to ensure consistency. For information regarding speaker demographics, recording and archiving information, and pre-processing see Appendix A.

2.2 Tokenization and Acoustic Models

Yan-nhangu is customarily written with a combination of Latin letters (e.g. *y*, *l*), accented letters (e.g. *ä* and *ṯ*), digraphs (e.g. *nh*, *th*), and characters not otherwise used in English standard segmental orthography (*ŋ*, *'*). There are 25 consonants and 6 vowels. For a complete consonant and vowel chart in IPA and orthography, see Appendix B. Note that for all figures in Section 3.2, capitalized tokens in the phonemic models represent apical consonants (e.g. "*N*" stands for n̪) or long vowels (e.g. "*A*" stands for aː) to emphasize the one-to-one mapping between Yan-nhangu phone and phonemic token. Additionally, whitespaces are represented by underscores.

Two identical models were trained on the same data with two contrasting tokenization methods. In the first class of models, the token was defined as the grapheme, splitting digraphs like *ny* /ɲ/ into *n* and *y*. The orthography is the same as the Yolŋu Matha orthography defined in (Zorc, 1996). In the second class of models, the token was defined as the phoneme, such that orthographic digraphs representing one phone like *ny* /ɲ/ are one token: *ɲ*. The analysis of Yan-nhangu phonology is based on the most detailed existing documentation of Yan-nhangu phonology (Baymarrwaŋa et al., 2006).

We use Wav2Vec2-BERT 2.0 model (Chung et al., 2021) accessed from HuggingFace, which has been pre-trained in a self-supervised manner on Facebook's multilingual corpus. Although this pretraining corpus spans over 140 languages, it does not include Indigenous Australian languages or typologically similar phonological systems, which suggests limited zero-shot transfer and motivates targeted fine-tuning in this setting. Our training dataset consists of up to 156 minutes of speech, and we optimize the model using Connectionist Temporal Classification (CTC) loss with an 80/20 train-validation split. Training is performed in Jupyter notebooks on one RTX-8000 GPU. [1] Each model is trained for 16 epochs using a linear learning rate scheduler, an initial learning rate of 1e-5, and early stopping to prevent overfitting. Training takes less than two hours per model.

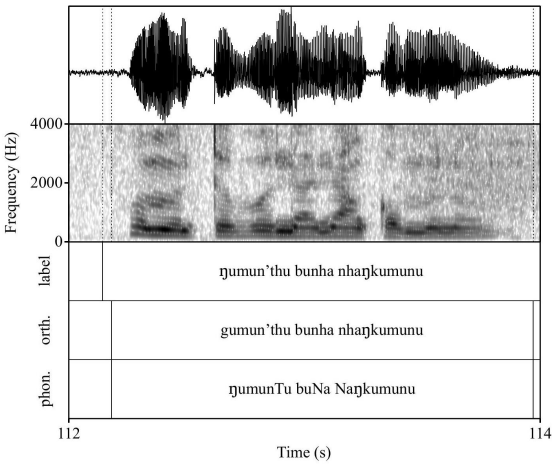


Figure 1: Waveform, spectrograph, and annotations for a sample testpoint.

2.3 Evaluation

Word Error Rate (WER) and Character Error Rate (CER) were used as evaluation metrics. We use CER when selecting top-performing models since post-alignment editing is common in language documentation workflows, meaning a lower CER reflects a greater speedup during manual correction compared to a lower WER. Models with orthographic and phonemic transcription schemes were trained and evaluated on 10, 30, 60, 90, 120, and 156 minutes of data. A qualitative analysis of the automatic transcription errors most common across models was also performed and reported on in Section 3.3. Following previous ASR error analysis research (Errattahi et al., 2018), we look at the Levenshtein distance between a held-out testing set and the best orthographic and phonemic ASR models' transcriptions. Finally, the last author manually corrected four minutes of automatically transcribed Yan-nhangu speech, such as

[1] The codebase is currently under active development as part of a broader open-source tool and will be released publicly following the associated corpus release.

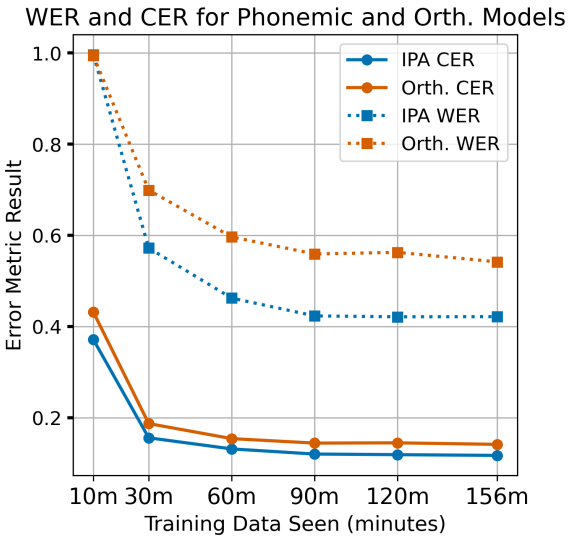


Figure 2: CER and WER metrics for phonemic and orthographic models across training set sizes.

Figure 1. Qualitative interpretations about the errors encountered were documented and the time saved between correcting automatic transcription and manually transcribing from scratch are discussed.

3 Results

3.1 WER and CER Across Models

Our results show that phonemic tokenization improves low-resource ASR performance. Figure 2 shows a consistent gap in WER and CER on held-out validation data, with the phonemic model outperforming the orthographic model across all training set sizes.

3.2 Levenshtein Distance Analysis of Errors

Analyzing the Levenshtein distance between manual annotations and the best orthographic and phonemic ASR outputs allows a more detailed investigation into model differences. The total number of errors made on the validation set are described in Table 1. Although the phonemic model shows lower overall Levenshtein distance from the manually transcribed label than the orthographic model, substitutions seem to be more frequent for the phonemic model.

The bar charts in Figure 3a and Figure 3d show the frequency of each deleted character. For both orthographic and phonemic models, spaces and short vowels are among the most commonly deleted characters. This is consistent with word boundaries being unclear during rapid speech, and

Model Type	Dels	Interts	Subs	Total
Phonemic	433	454	547	1434
Orthographic	624	592	438	1654

Table 1: Summary of Levenshtein distance between human annotated and ASR transcription by error type: deletions, intertions, and substitutions

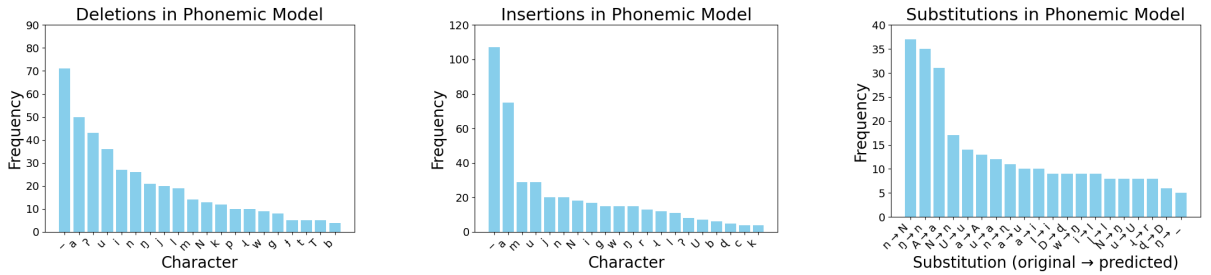
with vowel elision that occurs in Yan-nhangu, especially with *a*. Interestingly, the orthographic model frequently deletes tokens *n* and *h*, both of which are present in multiple digraphs. Across both models, sonorants (nasals, liquids, and vowels) are deleted much more frequently than stops.

Figure 3b and Figure 3e show the most frequently inserted characters, which mirror the most frequently deleted characters across both models. As such, whitespaces and *a* are the most frequent insertions, with *h* being the third most commonly inserted character for the orthographic model. Like with deletions, sonorants are inserted more frequently than stops.

Lastly, Figure 3c and Figure 3f show the most common token substitutions for the phonemic and orthographic models. This is the only setting in which we see a greater rate of errors for the phonemic model due to phonemic substitutions appearing as insertions or deletions for the orthographic model (e.g. $n \rightarrow N$, corresponding to $n \rightarrow nh$: in the orthographic model, this appears as an insertion). Another instance of ambiguous phones showing up as substitutions of phonologically similar phones include long vowels being substituted for their short counterparts, which occurs with an approximately equal and high frequency in both models. However, a stark asymmetry in substitution direction occurs with *n* and *ŋ*, where $ŋ \rightarrow n$ is the second most common substitution in the phonemic model, while $n \rightarrow ŋ$ is the second most common substitution in the orthographic model.

3.3 Further Comments on Errors

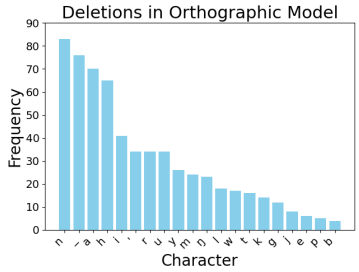
The last author, who is very familiar with Yan-nhangu, manually corrected 4 minutes of ASRed Yan-nhangu speech. It took 20 minutes to review each transcribed utterance to check, correct, and categorize the errors. At five minutes per minute of transcript, this is equal to the fastest unassisted transcription rate. Without using the ASR technology, it would take the last author approx. 15 minutes to transcribe 1 minute of Yan-nhangu, meaning that the introduction of ASR technology in



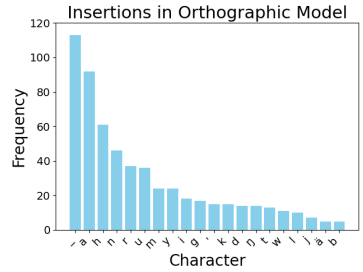
(a) Deletions for phonemic tokenization

(b) Insertions for phonemic tokenization

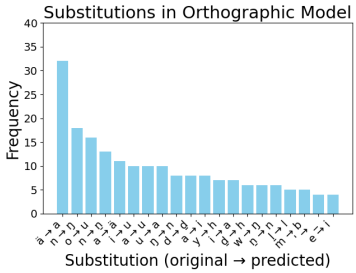
(c) Substitutions for phonemic tokenization



(d) Deletions for orthographic tokenization



(e) Insertions for orthographic tokenization



(f) Substitutions for orthographic tokenization

Figure 3: Counts for deletions, insertions, and substitutions for the best ASR models using phonemic and orthographic tokenization.

the transcription pipeline resulted in a three times speedup.

In the stretch of recording considered, errors fell into three (quantitative) categories. The first was grapheme substitutions, such as the word *d̲iltji* 'bush' was transcribed as *diltji*. The second set of "errors" involved items where ASR correctly rendered the material in the recording, but the transcription did not adhere to Yan-nhangu orthographic norms, such as word break errors or vowels in words not being transcribed due to them not being pronounced. Finally, there were 3 cases where, to the transcriber's ear, both the ASR word and the label word were possible representations of the recorded speech. The original transcripts were made in 2004–2008 and discussed with native speakers of Yan-nhangu, but there are several near homophone words that were, in context, also semantically plausible.

4 Discussion and Conclusions

This paper investigates whether phonologically informed tokenization improves ASR accuracy for the low-resource Australian language Yan-nhangu, providing the first comparative analysis of tokenization strategies for any Aboriginal language. Results show that phone-level tokenization improves WER and CER over a grapheme-level baseline for models trained on at least 30 minutes of data and improves CER across all models, suggesting that phonemic tokenization provides a more linguistically transparent representation of the language, reducing ambiguity and supporting more effective generalization in low-resource conditions. Models approach peak performance with approx. 90 minutes of training data, with the orthographic model showing more room for improvement. These findings align with prior research demonstrating the benefits of linguistically informed tokenization in low-resource NLP (Atuhurra et al., 2024).

We find that changing the tokenization scheme only results in meaningful differences in substitutions, insertions, and deletions if a speech sound is represented differently by the tokenizers. In Yan-nhangu, this occurs for single phonemes that are orthographic digraphs. We hypothesize that the substitution asymmetry involving *ŋ* and *n* arises because the orthographic model represents /n̪/ and /ɲ/ using the digraphs *nh* and *ny*, resuling in a single phoneme being encoded as two tokens. As a result, the probability of *n* becomes correlated with h or y. When the model encounters the velar nasal ŋ, the likelihoods of h and y (which cue apicals and palatals) decrease, narrowing down the places of articulation for the nasal leading to more frequent

substitutions of ŋ with *n*. This effect does not arise in the phonemic model, which lacks digraphs.

Speech transcription is essential but time-consuming in language documentation. While many factors affect transcription time, this experiment shows that correcting high-quality ASR transcriptions is about three times faster than manual transcription for someone familiar with the language. Based on the authors transcription rate, ASR-assisted transcription could provide a 10-hour speedup per hour of unannotated speech. This efficiency is crucial for developing resources for the world's most under-documented languages.

Limitations

This study is limited by its focus on a single language, Yan-nhangu. While this reflects realistic conditions in language documentation, the results may not directly generalize to languages with different phonological or orthographic properties. Similarly, by only testing on wav2vec2, we are limited in our ability to generalize these results to other model architectures. Additionally, the qualitative transcription speedup analysis is based on a single experienced annotator, and results may vary with regards to numerous factors which may lead to faster or slower transcription rates, including familiarity with the language, number of speakers, speech rate of participants, or complexity of the subject matter. Finally, our tokenization comparison is restricted to phonemic and orthographic representations, and does not explore alternative subword approaches (Bayram et al., 2025; Si et al., 2023), which remain an important direction for future work.

Ethical Considerations

While ASR-assisted transcription and annotation may increase the speed of language documentation workflows, its use raises important ethical considerations. Language documentation and fieldwork are often embedded within broader language revitalization efforts, where goals extend beyond corpus creation to include community engagement, skill development, and cultural preservation. In such contexts, the appropriateness of automated methods may depend on both the intent of the project and the nature of the materials being transcribed.

Prior work has shown that community members may prefer manual transcription even when ASR systems offer a faster alternative. For example, Prud'hommeaux et al. (2021) report that Indigenous participants preferred to transcribe from scratch, despite recognizing the efficiency of ASR-assisted transcription. This community emphasized the role of transcription in strengthening language proficiency through close engagement with the language. Participants also differentiated between content, remarking that ceremonial recordings should be transcribed manually, while less sensitive materials, such as childrens stories, may be more appropriate for ASR-assisted workflows.

In light of these considerations, linguists must work with language community members to identify the goals of the documentation project and how the methodology involved works to achieve them.

References

Kristine Mae M. Adlaon and Nelson Marcos. 2024. Finding the optimal byte-pair encoding merge operations for neural machine translation in a low-resource setting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14673–14682, Miami, Florida, USA. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Jesse Atuhurra, Hiroyuki Shindo, Hidetaka Kamigaito, and Taro Watanabe. 2024. Introducing syllable tokenization for low-resource languages: A case study with swahili. *Preprint*, arXiv:2406.15358.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Thibault Bañeras-Roux, Mickael Rouvier, Jane Wottawa, and Richard Dufour. 2024. A Comprehensive Analysis of Tokenization and Self-Supervised Learning in End-to-End Automatic Speech Recognition applied on French Language. In *32th European Signal Processing Conference (EUSIPCO)*, Lyon, France.

Laurie Baymarrwaŋa, Rita Gularrbanga, Laurie Milinditj, Rayba Nyaŋbal, Margaret Nyuŋunyuŋu, Allison Warrŋayun, and Claire Bowern. 2006. A learners guide to yan-nhaŋu.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümü, Sercan Karaka, Banu Diri, Sava Yldrm, and

Demircan Çelik. 2025. Tokens with meaning: A hybrid tokenization approach for nlp. *Preprint*, arXiv:2508.14292.

Shobhana L Chelliah. 2001. *The role of text collection and elicitation in linguistic fieldwork*, pages 152–165. Cambridge University Press.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Arienne M Dwyer. 2006. *Ethics and practicalities of cooperative fieldwork and analysis*, page Chapter 2. Mouton.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37. 1st International Conference on Natural Language and Speech Processing.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. 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 3594–3608, Online. Association for Computational Linguistics.

Disen Liao and Freda Shi. 2026. How tokenization limits phonological knowledge representation in language models and how to improve them. In *Tokenization Workshop*.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Kelly Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513. University of Hawaii at Mānoa.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Sub-character tokenization for chinese pretrained language models. *Preprint*, arXiv:2106.00400.

P Wittenburg, H Brugman, A Russel, A Klassman, and H Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*, pages 1556–1559. 00216.

R. David Paul Zorc. 1996. Yolngu-matha dictionary / r. david zorc.

A Additional Data Information

A.1 Speaker Demographics

All the data used in this experiment come from 5 women recognised by their community as authorities to speak about the language. At the time of recording they ranged in age from early 40s to early 70s. Four of the five were from Nhaŋu clans; the fifth was Warrawarra. The traditional lands of Yan-nhangu people are the Crocodile Islands of Northeast Arnhem Land (Northern Territory, Australia).

A.2 Recording and Archiving

Recordings were made with an Edirol R-01 solid state recorder and external microphone. They were recorded in a range of locations in Milingimbi Aboriginal Community and on Murruŋga Island. Language tasks included a range of structured and semi-structured elicitation tasks, including wordlist and sentence translation, descriptions of pictures and video clips, and discussion prompts around cultural concepts and practices.

Materials were transcribed in Elan (Wittenburg et al., 2006) by the last author and checked with speakers. Recordings, transcripts, and field notes have been deposited with the ELAR digital archive and the AIATSIS library. These materials are not publicly available due to usage agreements that respect Indigenous intellectual property; however, this work was conducted under agreements that permit the use of the recordings to support Yan-nhangu and other Indigenous language research.

A.3 Preprocessing

Incomplete transcriptions and audio containing English words were excluded, along with segments without transcriptions, leaving about 156 minutes of training data. Punctuation was removed, except for apostrophes, which denote glottal stops in Yan-nhangu orthography. The original transcriptions were all produced in Yan-nhangu orthography. Since the phonemic representation of Yan-nhangu words is predictable from the orthography, an automated script was used to generate phonemic transcriptions for all annotations used in training the phonemic models

p	t	ʈ (t̲)	t̪ (th)	c (tj)	k	ʔ (’)
b	d	ɖ (d̲)	d̪ (dh)	ɟ (dj)	g	
m	n	ɳ (n̲)	n̪ (nh)	ɲ (ny)	ŋ	
	l	ɭ (l̲)				
	r (rr)	ɹ (r)				
w				j (y)		

Table 2: Consonant inventory of Yan-nhangu

i	u	iː (e)	uː (o)
a		aː (ä)	

Table 3: Vowel inventory of Yan-nhangu

B Yan-nhangu Consonant and Vowel Inventory

Consonant and vowel charts are given below in Table 2 and Table 3 respectively in IPA, with orthography in parenthesis where different. Note that Yan-nhangu's orthography is the same as that used by the other Yolŋu languages of Arnhem Land.