Short-form verbal arts as a speech data resource in the field

Matthew Faytak[1], Tianle Yang[1], Pius W. Akumbu[2], Ivo Forghema Njuasi[3], Éric Le Ferrand[1]

[1]University at Buffalo, USA,
[2]LLACAN - CNRS, France
[3]University of Buea, Cameroon

Abstract

We propose a method for efficient field data collection of speech resource data which leverages short-form verbal arts, namely riddles and proverbs, which permit a predictable transcript to be assigned to naturalistic but conventionalized utterances. As a proof of concept, we describe a 5.25 hour corpus of proverbs and riddles collected for Kom, a low-resource language of Cameroon, and conduct ASR modeling experiments on the corpus. Results suggest that the proposed method yields high quality speech data, albeit with relatively low lexical diversity. We highlight the alignment of the collected data with community priorities for cultural education and preservation in the Cameroonian context.

1 Introduction

Recent advances in natural language processing, particularly the development of foundation models and transfer learning techniques, have made language technologies more accessible to a wider range of languages. The amount of data required to train state-of-the-art architectures has greatly decreased, but *some* annotated data remains necessary for model training. While high-resource languages benefit from abundant, diverse online data, many low-resource languages studied by field linguists have no available online data sets. As a result, field collection of data resources remains essential to the development of automatic speech recognition (ASR) systems, which are increasingly used to facilitate further linguistic data collection (Seifart et al., 2018; Michaud et al., 2018).

Naturalistic, spontaneous speech data is commonly collected in the field, given the difficulty of implementing controlled tasks and engaging speakers. However, these generate less predictable speech content which can only be deciphered with extensive work from expert transcribers (Himmelmann, 2018). In this paper, we introduce a novel method for efficient field collection of speech resource materials based around short-form verbal arts: namely, proverbs and riddles, which are not only easily obtained and efficiently transcribed, but also better aligned with community interests and priorities. We present as proof of concept a speech corpus created for Kom (Grassfields Bantu, Cameroon; ISO 639-3: bkm) using this approach.

In this paper, we first review the literature on challenges related to field data collection and ASR for field linguistics. We then describe in detail our use-case language, Kom, our data collection method, and aspects of the short-form verbal arts materials at issue. Finally, we present ASR experiments conducted on the collected corpus, which suggest generally good quality of the collected data but relatively low word character diversity.

2 Related work

Datasets collected for ASR in well-resourced languages often rely on speech or text data generated by institutions or media such as parliamentary discussions, audiobooks, or radio broadcasts (Wang et al., 2021; Panayotov et al., 2015; Kocabiyikoglu et al., 2018; Gelas et al., 2012). For low-resource languages, the range of available speech data sources is drastically reduced and is often limited to religious texts (Black, 2019; Zanon-Boito et al., 2020; Pratap et al., 2024), models trained on which may not extend well to other domains (Le Ferrand et al., 2025). While training models in low-resource contexts has become easier with the advent of transformers-based architectures – foundation models fine-tuned on minimal data have been reported to have good performances across domains (Havard et al., 2025; Billings and McDonnell, 2025; Geng et al., 2025) – collecting new datasets for fine-tuning is still a challenge for many languages and a popular research topic (Taguchi et al., 2024; Ngue Um et al.,

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 38–45
March 26, 2026. ©2026 Association for Computational Linguistics

2025).

Linguistic fieldwork is often the only source of materials for speech resource development (Michailovsky et al., 2014; Paschen et al., 2020; Tapo et al., 2024). Corpora of naturalistic, spontaneous speech are typically the major targets of this work. This reflects the difficulty of controlling the field recording environment, as well as the fact that speakers may be difficult to engage in repetitive or highly controlled tasks (Bowern, 2015; Le Ferrand et al., 2022). It also reflects the priority of naturalistic speech events in language documentation (Lüpke, 2010; Woodbury, 2011). However, development of annotated speech corpora is greatly facilitated by predictable transcripts, which spontaneous speech lacks. The resulting *transcription bottleneck* requires time-consuming manual transcription before further work with the resource can be done (Himmelmann, 2018; Seifart et al., 2018).

At issue in this paper are short-form verbal arts such as proverbs and riddles, specifically in the African context. In our view, they present a happy medium of naturalistic speech with consistent, manageably-sized transcripts. They also offer a crucial additional benefit: work centered on culturally significant verbal arts is more engaging for many participating speakers, compared to the arbitrary lists of text sentences often used for speech resource development (Gutkin et al., 2020; Gelas et al., 2012; Godard et al., 2018). Below, we consider the affordances of proverbs and riddles for efficient collection of a high-quality speech corpus in a field context, as well as their benefits for engaged speaker communities.

3 Background

3.1 Use-case: Kom

Kom (Itaŋikom) is a Grassfields Bantu language of Cameroon spoken by about 300,000 people. An official orthography is used in childhood education and small-press publications (Chia and Kimbi, 1992; Chuo, 2022). However, Kom totally lacks speech technology resources to our knowledge. The inventory of segments and surface tones is given in Table 1. The segmental inventory is notable for possessing front rounded vowels and *fricative vowels* /^{z}ɨ, ^{v}ɨ/, which are associated with frication or affrication of preceding consonants (Connell, 2007; Faytak, 2017). Kom also has a complex tonal inventory, with at least eight surface tone contours and considerable postlexical changes to lexical tone patterns (Hyman, 2005).

Vowels			
^{z}ɨ ⟨zɨ, sɨ⟩			^{v}ɨ ⟨vɨ, fɨ⟩
i	y ⟨ue⟩	ɨ	u
e	ø ⟨oe⟩		o
ɛ ⟨ae⟩			
		a	

Consonants			
b	t d	tʃ ⟨ch⟩ dʒ ⟨j⟩	k ⟨k, ’⟩ g
f v	s z		
m	n	ɲ ⟨ny⟩	ŋ
w	l	j ⟨y⟩	ɰ ⟨gh⟩

Surface tones	
⟨V⟩	H, HM, M, MH
⟨V̀⟩	L, LM
⟨V̂⟩	HL, ML

Table 1: Kom phonological inventory (Shultz, 1993; Hyman, 2005; Faytak, 2017); graphemes given in ⟨. . . ⟩.

3.2 Proverbs and riddles as genres

Here, we highlight aspects of proverbs and riddles as the most conventionalized short-form verbal arts, with a particular focus on Kom as our use-case language. Table 2 shows a range of Kom proverbs and riddles.

Proverbs are fixed sayings, transmitted orally within the speech community, which have educational, critical, or advisory functions depending on the situation, both generally (Anchimbe, 2011; Etta and Mogu, 2012; Yankah, 1989) and in Kom specifically (Nkwi, 1987; Njwe, 2015). Proverb use demonstrates a speaker's linguistic competence and higher-order thinking skills and is associated with wisdom and advanced age (Nkwi, 1987; Finnegan, 2012) . Proverbs generally convey moral lessons and emphasize pro-social behavior. They usually indirectly hint at their deeper meanings through figurative language, irony, and exaggeration. Because they are seen as indirect and conveying collective cultural wisdom rather than personal opinion, proverbs may be used to manage interpersonal conflict or negotiations (Finnegan, 2012; Fonkem, 2014).

A riddle is a verbal puzzle consisting of an short figurative description which guessers must identify with a scenario or object. Riddles are traditionally used in the more informal context of evening entertainment (Okpewho, 1992; Jick and Ngam, 2016; Akumbu et al., 2025). In Kom

Orthography	Literal translation
Awu à mò’ a nɨ n-kulɨ wi ibu’.	One hand does not tie a bundle.
Ghelɨ nɨ n-kôŋ sà chà’ kɨ ibi zɨ̀ a yi n-chem.	People like to pick up kola that has dropped.
Wà tɨm àvɨ a kɨa, a kfɨ kɨ iti.	If you hit your foot, may only the stone break.
Chɨ̀sɨ nɨ̀ n-ku wi ɨlvâ ɨ yum	Charms don’t catch an empty belly.
Afo kì a fòyn làlì sɨ achɨ, a kɨ du’i.	A thing that sits when the chief stands up. Answer: fɨnjâenjàe (a fly)
Ghɨ se’ ìghoŋ, a yɨ woynda.	They go to war, the children win. Answers: mɨlvɨ (soldier ants), ayôyn (speargrass)

Table 2: Representative Kom proverbs (top) and riddles (bottom) drawn from the collected corpus.

and other nearby societies, if the riddler outsmarts all guessers, they may demand symbolic payment to reveal the answer, in the form of titles of local chiefs (Jick and Ngam, 2016; Akumbu et al., 2025). Unlike proverbs, riddles usually do not convey moral lessons, but are seen as a sort of beneficial cognitive exercise.

3.3 Affordances of proverbs and riddles

Because proverbs and riddles are ubiquitous – the “palm oil with which words are eaten” (Achebe, 1959/1994, 14) – they are broadly known to speakers, and consistent transcripts can be generated for them. Because verbal arts are an oral tradition, they can be recalled by speakers without literacy in the target language. In our experience, short-form verbal arts are very effective at engaging participants in data collection. They also align better with the priorities of partner communities for cultural preservation than arbitrarily chosen materials, as proverbs are a repository of a community’s worldview and epistemology. Recordings of short-form verbal arts may also have applications in childhood formal education, potentially filling gaps in availability of pedagogical material in local languages (Echu, 2004; Chiatoh, 2013) and contributing to decolonial practice in the classroom (Wolff, 2016; Akumbu et al., 2025).

4 Corpus collection and characteristics

For the verbal arts corpus, eighteen Kom speakers (7F, 11M) participated in recording near Douala, Cameroon, with a Zoom H4n digital recorder and Shure SM10A head-mounted cardioid dynamic microphones. Data were recorded as 16-bit mono WAV at a sampling rate of 44.1 kHz. For the out-of-domain evaluation described in Section 5.2, two speakers (1F, 1M) participated in similar recording sessions to collect spontaneous-speech narratives using different equipment: a Zoom H6Essential digital recorder with Shure WH20 head-mounted cardioid dynamic microphones, recording as 32-bit mono WAV at a rate of 44.1 kHz. The male participant was also a participant in the collection of the verbal arts corpus. The recording setting was a quiet but reverberant room in a concrete building without sound treatment. Because the data were originally collected for phonetic research on connected speech, the microphone hardware was chosen for its directional response and rejection of background noise. Speakers were recorded alone or in pairs, with the directional microphones ensuring isolation of one speaker per audio channel. Due to logistical limitations, not all proverbs or riddles are recorded for all speakers, and the total amount of material per speaker varies according to their availability.

Verbal arts data were collected in interaction with the first author; for full details on data collection see (Faytak et al., 2026). Speakers were prompted to re-speak proverbs read aloud from published sources by the first author (Loh, 1997; Lo-ah, 2018; Njwe, 2015). Speakers frequently volunteered conceptually related proverbs not attested in published sources; all riddles were volunteered, since riddles do not appear in published sources to our knowledge. Transcripts in Kom orthography were drawn from published resources or generated in consultation with participants if the proverb or riddle was volunteered. Due to the high degree of conventionalization of both proverbs and riddles, variation among speakers in the lexical or structural characteristics of a given item was quite low, and once a transcript was established it typically applied with few or no modifications to the versions provided by other speakers.

Once a proverb or riddle was successfully recalled, participants repeated it four to five times.

Repetition of this sort is common in phonetics research, as it increases the number of tokens per type and improves the statistical power of later analysis (Maddieson, 2001; Ladefoged, 2003; Bowern, 2015). Because speakers repeated the same utterance multiple times, the resulting corpus has reduced lexical diversity, which partly explains the low WER described in Section 5.

The resulting publicly available corpus[1] contains 315 minutes of Kom speech, with 55,092 word tokens covering 782 word types. This yields a very low token-type ratio of 0.0142. The out-of-vocabulary (OOV) rate on a random 80/20 split of the full corpus is 0.0243, and even across speakers (leave-one-out), the average OOV rate is 0.07%±0.04. This suggests that lexical variation is constrained, as may be expected given the repetitive data collection method. The token-type ratio remains low even if repetition is taken into account: taking *unique* utterances as the basis for calculations yields 14,469 word tokens and raises the token-type ratio only to 0.054.

5 Modeling

5.1 Experimental setup

As a proof of concept for the method, we trained two standard ASR architectures on the corpus under two evaluation strategies. Transcripts were preprocessed to remove tone diacritics due to inconsistent application across the corpus. The first evaluation setup was a regular random split where the utterances were shuffled and 80% were used for the training and the remaining 20% for testing. The second evaluation is a cross validation where the data collected from a given speaker is left out for evaluation for each speaker in the dataset.

The two ASR architectures used are XLSR, the multilingual version of wav2vec (Conneau et al., 2021; Baevski et al., 2020), and MMS (Pratap et al., 2024). For both architectures we used hyperparameters provided by the main HuggingFace tutorial with some modifications[2][3]. Models were trained for 20 epochs with a batch of size 16; acoustic features were kept unfrozen and CTC loss set to zero infinity. Since the classic MMS framework does not allow this, we evaluated only the XLSR models using trigram language models trained on the training sets of each model.

5.2 Results

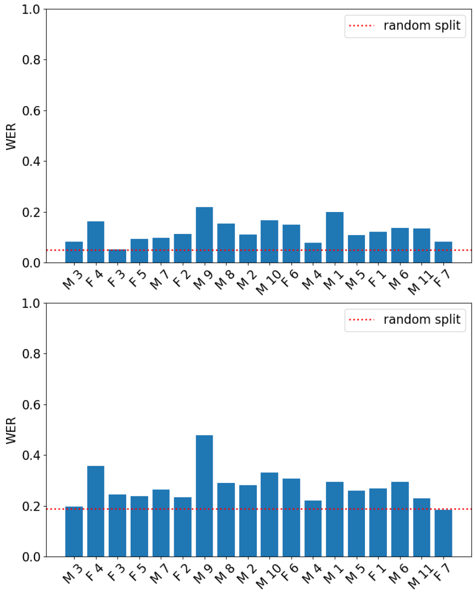

Figure 1: Word Error Rate by speaker for the cross validation (XLSR - top, MMS - bottom) compared with random split evaluation.

Word Error Rates for XLSR and MMS can be found in Figure 1. While WER is higher for MMS than XLSR, likely due to the former lacking a language model during decoding, all scores are below 0.2 for XLSR and 0.5 for MMS, which is within the range or below what is usually expected for a field-collected dataset (Jimerson et al., 2023). This confirms that the orthography of the corpus is consistent and the speech consistently intelligible.

To better inform future modeling decisions—particularly around use of the standard orthography for transcription of training data and around tokenization using characters rather than graphemes—we also conducted an analysis of the character recognition, using the Needleman-Wunsch algorithm to align the predictions to the gold standard at the character level. For both architectures, errors involving ⟨h⟩ are in the top three character recognition errors (Table 3). Because ⟨h⟩ appears in two digraphs ⟨ch⟩ and ⟨gh⟩ and is also written in some interjections (e.g. *mmhmm*, *eehee*), it may be hard to interpret due to its association with several phones' acoustic features.

To assess the quality of the data collected as ASR training data, we evaluated our baseline XLSR model on an additional 10 minutes of spontaneous speech. This additional data set consists of two unprompted monologues, one by a female speaker and one by a male speaker, concerning the

[1] https://huggingface.co/datasets/PhonLab-Buffalo/Kom
[2] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2
[3] https://huggingface.co/blog/mms_adapters

XLSR	CER
h	0.05495
<space>	0.05253
y	0.03628

MMS	CER
<space>	0.10946
h	0.09041
e	0.08239

Table 3: Top 3 CER per character by architecture.

recent history of their family and the traditional political structure of the Kom chiefdom, respectively. To obtain the gold standard transcriptions, raw ASR outputs were used as a starting point for corrections completed by the fourth author (a first-language speaker of Kom).

Using the baseline XLSR model, we obtained a WER of 63.9 and a CER of 28.6. The overall low level of performance is expected given low lexical diversity in the training data and the general lack of overlap between the domains at issue. In particular, the spontaneous speech data contained a number of English loanwords and proper nouns from both English and Kom; these categories of words are systematically absent from proverbs and riddles, reflecting the latter's concern with timeless generalities and normative values. However, the WER and CER obtained are surprisingly low for out-of-domain testing considering the low lexical diversity, and well within the range of results one should expect for this kind of dataset (Le Ferrand et al., 2025; Liang and Levow, 2025; Geng et al., 2025). The WER is also low relative to the size of the training data, in a similar range to rates obtained from models trained on Bible recordings (Le Ferrand et al., 2025), which typically provide five to ten times as much data as the present corpus (Meyer et al., 2022; Black, 2019).

6 Discussion and future work

The method presented here leverages short-form verbal arts to collect high-quality data sets in the field, improve transcriptional efficiency and consistency, and better align field data collection with speaker-community goals in cultural education and preservation. As a source of data for ASR model development, the data collected using this method appear to be very consistent for in-domain testing and offer a solid baseline for transcription of speech in other domains.

The proposed method shows no clear *degradation* of performance relative to other approaches to gathering ASR training data in low-resource contexts. Due to the many factors which affect ASR performance, we hesitate to claim an *improvement* at this stage, with such determinations requiring careful evaluation in future work. For instance, better-than-expected performance in OOD evaluation on spontaneous speech may be due to closer-than-expected vocabulary overlap with proverbs and riddles, due in part to the latter being embedded within everyday speech to a greater extent than domains involving a distinct performance style (e.g. traditional narratives, Bible recordings).

While the current low lexical diversity of the dataset limits its performance, materials collected using the proposed method offer a starting point for future model development, especially if augmented with additional training data. As such, future work aims to increase the lexical and domain diversity of the collected data. For instance, spontaneous explanations of the deeper significance of proverbs and riddles can be elicited, which may help to collect material complementary in style and more diverse in lexical content compared to the verbal arts items themselves. Refinements to transcription and tokenization may also yield incremental improvements in model performance: for instance, future models may tokenize by grapheme rather than character, avoiding issues related to recognition of ⟨h⟩ in the digraphs ⟨ch, gh⟩. Removal of tone diacritics for model training also likely reduced somewhat the number of lexical types in the corpus, and their reintroduction in future corpora may increase lexical diversity.

Limitations

Low WER and CER, as well as high recognition scores, suggest the present corpus has high recording quality and transcriptional consistency. However, we are aware that WER is also likely reduced due to low word character diversity, as may be expected given the repetition of the corpus materials. We also acknowledge that verbal arts, particularly proverbs, are not necessarily used by all speakers in a community, a limitation shared by other specialized genres such as traditional narratives. Performance in out-of-domain testing is rather low: WER and CER obtained in out-of-domain testing are fairly high, as may be expected, and the corpus needs to be augmented with more diverse data in order to build a robust ASR system for Kom.

We view these shortcomings as acceptable in light of the ease of use of short-form verbal arts materials and the efficiency with which they can be used to collect a large amount of data in a range of field situations. The data which results is easy to consistently transcribe, and also has additional cultural significance for participating communities. The models presented here should be viewed as a starting point for further speech resource development, which is more viable if efficiency gains from imperfect ASR can be used to speed the annotation of additional training data.

Acknowledgments

We thank the members and executive of the Kom Cultural and Development Organization (Douala branch), Ikom Christopher Achuo, Pierpaolo Di Carlo, and Jeff Good for their logistical support. This work was supported by United States National Science Foundation award #2438916 to Matthew Faytak and Pius W. Akumbu.

References

Chinua Achebe. 1959/1994. Things Fall Apart.

Pius W Akumbu, Roland Kießling, Constantine Kouankem, Justine G Nzweundji, and Cornelius W Wuchu. 2025. Integrating traditional stories in formal education in the Cameroonian Grassfields. *Journal of the Cameroon Academy of Sciences*, 21(3):205–228.

Eric A. Anchimbe. 2011. *Language policy and identity construction: The dynamics of Cameroonian multilingualism*. Peter Lang.

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. Curran Associates, Inc.

Blaine Billings and Bradley McDonnell. 2025. Connecting automated speech recognition to transcription practices. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 128–132, Honolulu. Association for Computational Linguistics.

Alan W Black. 2019. CMU Wilderness multilingual speech dataset. In *Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, pages 5971–5975. IEEE.

Claire Bowern. 2015. *Linguistic fieldwork: A practical guide*. Springer.

Emmanuel N. Chia and Joseph C. Kimbi. 1992. *Guide to the Kom alphabet*. SIL Cameroon.

Blasius A Chiatoh. 2013. Cameroonian languages in education: enabling or disenabling policies and practices. In *Language policy in Africa: perspectives for Cameroon*, pages 32–51. Miraclaire Academic Publications.

Godfrey Kain Chuo. 2022. *Achievements and Challenges of the Kom Multilingual Education Longitudinal Experience and the Impact on Cameroons Educational System.* SIL Cameroon.

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

Bruce Connell. 2007. Mambila fricative vowels and Bantu spirantisation. *Africana Linguistica*, 13(1):7–31.

George Echu. 2004. The language question in Cameroon. *Linguistik Online*, 18(1).

Emmanuel Efem Etta and Francis Ibe Mogu. 2012. The relevance of proverbs in African epistemology. *Lwati: A Journal of Contemporary Research*, 9(1).

Matthew Faytak. 2017. Sonority in some languages of the Cameroon Grassfields. In Martin J. Ball and Nicole Müller, editors, *Challenging Sonority: Cross-Linguistic Evidence*, pages 77–97. Equinox.

Matthew Faytak, Ivo Forghema Njuasi, Nicholas Mori, and Angelique Griffith. 2026. A speech corpus of Kom verbal arts and its applications. In Vicki Carstens, Katherine Russell, Olawale Akingbade, Deborah Morton, and Michael Diercks, editors, *Pamoja tena 'Together again': African linguistics after COVID*, pages 441–465. Language Science Press.

Ruth Finnegan. 2012. *Oral literature in Africa*. Open Book Publishers.

Achankeng Fonkem. 2014. Bekem in Peacemaking in Nweh Society. *Indigenous Conflict Management Strategies in West Africa: Beyond Right and Wrong*, page 307.

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*, Cape Town.

Mengzhe Geng, Patrick Littell, PENÁĆ, Aidan Pine, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOŦEN Language Documentation Efforts with Automatic Speech Recognition. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39.

P. Godard, G Adda, Martine Adda-Decker, J Benjumea, Laurent Besacier, J Cooper-Leavitt, G-N

Kouarata, L Lamel, H Maynard, M. Müller, A Rialland, S. Stüker, F. Yvon, and M Zanon-Boito. 2018. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki.

Alexander Gutkin, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kọ́lá Túbọ̀sún. 2020. Developing an Open-Source Corpus of Yoruba Speech. In *Interspeech 2020*, pages 404–408.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Speech Technologies with Fieldwork Recordings: the Case of Haitian Creole. In *Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 40.

Nikolaus P. Himmelmann. 2018. Meeting the transcription challenge. In Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors, *Reflections on Language Documentation 20 Years after Himmelmann 1998*, pages 39–55. University of Hawai'i Press.

Larry M Hyman. 2005. Initial vowel and prefix tone in Kom: Related to the Bantu Augment. In Koen Bostoen and Jacky Maniacky, editors, *Studies in African comparative linguistics with special focus on Bantu and Mande: Essays in honour of Y. Bastin and C. Grégoire*, pages 313–341. Rüdiger Köppe Verlag Cologne.

Henry K. Jick and Gilead N. Ngam. 2016. Generic varieties and performance principles in Kom oral literature. *International Journal of Liberal Arts and Social Science*, 4(5):14–25.

Robert Jimerson, Zoey Liu, and Emily Prud'Hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proc of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proc of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Peter Ladefoged. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Blackwell.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an Australian aboriginal community. In *Proc of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4988–4998.

Éric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. Faithful transcription: Leveraging Bible recordings to improve ASR for endangered languages. In *Proc 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*, Mumbai. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In *Proc of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37.

F. Lo-ah. 2018. *The Wisdom of Fon Vincent Yuh II of Kom*. Bookman Communications.

Pius Loh. 1997. *Itaŋikom i tɨmlɨni-i*. SIL Cameroon.

Friederike Lüpke. 2010. Research methods in language documentation. *Language Documentation and Description*, 7:55–104.

Ian Maddieson. 2001. Phonetic fieldwork. In *Linguistic Fieldwork*. Cambridge University Press.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Interspeech 2022*, pages 2383–2387.

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

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit. *Language Documentation & Conservation*, 12.

Emmanuel Ngue Um, Francis Tyers, Eliette-Caroline Emilie Ngo Tjomb, Florus Landry Dibengue, Blaise-Mathieu Banoum Manguele, Blaise Abbo Djoulde, Mathilde Nyambe, Brice Martial Atangana Eloundou, Jeff Sterling Ngami Kamagoua, José Mpouda Avom, Zacharie Nyobe, Emmanuel Giovanni Eloundou Eyenga, and André Likwai. 2025. Speech technologies datasets for african under-served languages. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 82–90.

Eyovi Njwe. 2015. "The palm oil with which words are eaten": Proverbs from Cameroons endangered indigenous languages. In Elizabeth C. Zsiga, One Tlale Boyer, and Ruth Kramer, editors, *Languages in Africa: Multilingualism, language policy, and education*, pages 118–126. Georgetown University Press.

Paul Nchoji Nkwi. 1987. *Traditional Government and Social Change: a study of the political institutions among the Kom of the Cameroon Grassfields*. Fribourg University Press.

Isidore Okpewho. 1992. *African oral literature: Backgrounds, character, and continuity*, volume 710. Indiana University Press.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proc 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

George Shultz. 1993. *Notes on the phonology of the Kom language*. SIL Cameroon.

Chihiro Taguchi, Jefferson Saransig, Dayana Velásquez, and David Chiang. 2024. Killkan: The Automatic Speech Recognition Dataset for Kichwa with Morphosyntactic Information". In *Proc of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9753–9763, Torino. ELRA and ICCL.

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 *Interspeech 2024*, pages 5088–5092.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 993–1003.

H. Ekkehard Wolff. 2016. *Language and development in Africa: Perceptions, ideologies and challenges*. Cambridge University Press.

Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.

Kwesi Yankah. 1989. Proverbs: The Aesthetics of Traditional Communication. *Research in African Literatures*, 20(3):325–346.

Marcely Zanon-Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In *Proc of the twelfth language resources and evaluation conference*, pages 6486–6493.

A Appendix

	XLSR		MMS	
Spkr.	WER	CER	WER	CER
M3	0.0823	0.0371	0.1979	0.0817
F4	0.1617	0.0632	0.3562	0.1110
F3	0.0519	0.0208	0.2448	0.1022
F5	0.0930	0.0408	0.2391	0.0762
M7	0.0975	0.0396	0.2650	0.0767
F2	0.1118	0.0479	0.2344	0.0871
M9	0.2181	0.0893	0.4781	0.1485
M8	0.1541	0.0632	0.2900	0.0895
M2	0.1112	0.0460	0.2807	0.1083
M10	0.1675	0.0849	0.3318	0.1265
F6	0.1494	0.0593	0.3083	0.1066
M4	0.0779	0.0351	0.2202	0.0924
M1	0.1993	0.0902	0.2947	0.0891
M5	0.1075	0.0462	0.2592	0.0901
F1	0.1220	0.0475	0.2688	0.1014
M6	0.1357	0.0664	0.2942	0.1049
M11	0.1341	0.0581	0.2299	0.0815
F7	0.0823	0.0333	0.1845	0.0605
Rand.	0.0600	0.0265	0.1880	0.0623

Table 4: Word and Character Error Rate by speaker and random split, for XLSR (left) and MMS (right).