Building a Functional Machine Translation Corpus for Kpelle

Kweku Andoh Yamoah[1] **Jackson Weako**[2] **Emmanuel J. Dorley**[1]
[1]University of Florida
[2]Liberia Language Institute
kyamoah@ufl.edu, weakojackson@gmail.com, edorley@ufl.edu

Abstract

In this paper, we introduce the first publicly available English-Kpelle dataset for machine translation, comprising over 2,000 sentence pairs drawn from everyday communication, religious texts, and educational materials. By fine-tuning Meta's No Language Left Behind (NLLB) model on two versions of the dataset, we achieved BLEU scores of up to 30 in the Kpelle-to-English direction, demonstrating the benefits of data augmentation. Our findings align with NLLB-200 benchmarks on other African languages, underscoring Kpelle's potential for competitive performance despite its low-resource status. Beyond machine translation, this dataset enables broader NLP tasks, including speech recognition and language modeling. We conclude with a roadmap for future dataset expansion, emphasizing orthographic consistency, community-driven validation, and interdisciplinary collaboration to advance inclusive language technology development for Kpelle and other low-resourced Mande languages.

1 Introduction

Several notable initiatives have sought to address the challenges of low-resource languages, particularly in Africa. Collaborative projects like Masakhane (Nekoto et al., 2020; Orife et al., 2020) have created and publicly released several machine translation datasets and baseline models for African languages (Nekoto et al., 2020; Orife et al., 2020; Nakatumba-Nabende et al., 2024). The Lacuna Fund has also played a vital role in accelerating the creation of openly accessible text and speech datasets for various African languages (Nakatumba-Nabende et al., 2024; Asamoah Owusu et al., 2022; Vydrin et al., 2022; Asmelash Teka Hadgu et al., 2022; Wanjawa et al., 2024; Adelani et al., 2022). Additionally, there is Meta's "No Language Left Behind" (NLLB) project aimed to develop high-quality machine translation systems for over 200 languages, including many low-resource languages in Africa (Team et al., 2022). Despite these efforts, languages such as the Kpelle language have not been explored, leaving the language marginalized in natural language processing (NLP) research.

Kpelle is a language primarily spoken in Liberia and Guinea, with over one million speakers across these two countries (Vydrin, 2018). It is classified as a macro-language due to distinct variants—Liberian Kpelle and Guinean Kpelle—that, while closely related, constitute separate linguistic entities (Vydrin, 2018). Belonging to the Southwestern subgroup of the broader Mande language family, Kpelle is part of a larger linguistic family that includes approximately 70 languages spoken by at least 25 million native speakers and an additional 30 million second-language speakers throughout West Africa (Konoshenko, 2008; Vydrin, 2018). Within Liberia specifically, Kpelle represents the largest indigenous language, spoken by approximately 20% of the population (Vydrin, 2018).

Although Kpelle boasts a considerable number of speakers, it remains largely absent from digital platforms, including AI tools. Kpelle is a low-resourced language, which means the language lacks sufficient digital resources to support the development of NLP applications. Therefore, by extension, Kpelle faces the same challenges that are unique to low-resourced languages. These challenges include data scarcity (Kusampudi et al., 2021; Maillard et al., 2023; Nakatumba-Nabende et al., 2024; Nguyen et al., 2022), data quality(data limited to specific domains like religious texts) (Nakatumba-Nabende et al., 2024; Maillard et al., 2023; Kusampudi et al., 2021; Team et al., 2022), multilingualism, and dialectical variations(difficulty determining boundaries within dialects) (Konoshenko, 2024).

To address this significant gap, we present the first-ever dataset for Kpelle. This dataset is de-

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 52–63

signed for machine translation and language learning of Kpelle and English and vice versa. Our work aims to lay the foundations for intensive research for Kpelle and other low resource Liberian languages, enabling the development of NLP applications and solutions that can enhance the way speakers of the language interact with everyday technologies. This paper begins with an introduction highlighting our work's foundations and motivations. The continuing sections present the related work for machine translation for African languages. We then present the history of the Kpelle language, examining its unique linguistic features. Following that, we discuss the dataset creation process and the corpus benchmarking using the NLLB model and the results obtained. Our contributions are as follows:[1] *(a) Created a bilingual English-Kpelle corpus that has 3234 translation pairs. (b) The methodological data collection, cleaning, and alignment approach offers a replicable framework for other researchers working with low-resource languages. (c) Benchmarked the dataset on NLLB achieving a BLEU of $\approx$ 30 for kpe_Latn $\rightarrow$ eng_Latn translation and a BLEU of $\approx$ 24 eng_Latn $\rightarrow$ kpe_Latn translation.*

2 Related Work

2.1 Review of Efforts in Low-Resource Language Datasets

The development of robust NLP tools for low-resource languages is limited by data scarcity, creating significant challenges for tasks like machine translation. Addressing this challenge has prompted various initiatives to expand language coverage and improve translation quality. Community-led projects like Masakhane have played a pivotal role in building datasets and models for African languages through a collaborative approach involving researchers and native speakers (Nakatumba-Nabende et al., 2024; Akinfaderin, 2020). The Lacuna Fund has further supported these efforts by funding the creation of open-source text and speech resources for African languages (Akinfaderin, 2020; Nakatumba-Nabende et al., 2024; Asamoah Owusu et al., 2022; Vydrin et al., 2022; Asmelash Teka Hadgu et al., 2022; Wanjawa et al., 2024; Adelani et al., 2022). Meta's ambitious "No Language Left Behind" (NLLB) project has made significant progress in building machine translation systems for over 200 languages, including many that are under-resourced (Team et al., 2022). The NLLB Team et al. (2022) used data mining to transform vast monolingual datasets into new training data for low-resource languages and employed new modeling approaches, like the Sparsely Gated Mixture of Experts, to improve translation quality (Team et al., 2022). However, NLLB (Team et al., 2022), like many other initiatives, primarily focuses on languages with established written standards, leaving languages with limited or no written traditions largely unaddressed .

Beyond large-scale projects, creating specialized corpora has proven vital in addressing the data diversity and domain adaptation needs of specific languages and regions (Agyei et al., 2024; Maillard et al., 2023). The Twi-2-ENG corpus from (Agyei et al., 2024) is a recent example, providing a comprehensive resource for the Twi language, encompassing a wide range of genres relevant to Ghanaian Twi-speaking communities. This corpus aims to support NLP applications like machine translation and linguistic research by offering a searchable platform for accurate translations and a deeper understanding of Twi linguistics (Agyei et al., 2024; George et al., 2024; Williams et al., 2018). Another example is the LORELEI program, initiated by DARPA, which targets research and development of language technologies that aim to reduce the dependency on manually transcribed and translated corpora (Nguyen et al., 2022; Agyei et al., 2024; Goyal et al., 2021). This program has facilitated the collection of language samples and data for several African languages, including Hausa, Zulu, Yoruba, Twi, Somali, Swahili, and Wolof, contributing to the growth of language resources for these languages (Agyei et al., 2024; Goyal et al., 2021; Team et al., 2022).

2.2 Prior Work on the Mande Language Family

Existing NLP research on the Mande languages primarily focuses on individual languages, with limited cross-linguistic studies or comprehensive datasets representing the broader family (Vydrin, 2018). A few studies have investigated specific linguistic phenomena, such as the origin of the S-O-V-X word order (Vydrin, 2018), motion events in Bambara (Vydrin, 2018), and the evolution of tonal systems (Konoshenko, 2008; Vydrin, 2018). Efforts in language documentation and corpus creation for Mande languages have also been undertaken (George et al., 2024; Nakatumba-Nabende

[1] Dataset is made available at https://huggingface.co/datasets/IARG-UF/English-Kpelle-Corpus

et al., 2024; Akinfaderin, 2020; Team et al., 2022). For instance, a grammatical sketch of Beng, a Southern Mande language, has been developed (Paperno, 2014). Additionally, research on the Kakabe language, a Western Mande language, has focused on prosody in grammar (Vydrina, 2017). However, these efforts typically focus on individual languages or specific linguistic phenomena, and thus do not provide comprehensive resources or datasets necessary for cross-lingual NLP applications across the broader Mande language family.

2.3 Gap Filled by the Kpelle Dataset

The Kpelle Dataset aims to address a critical gap in the current research by providing the first, publicly available bilingual dataset for the Kpelle language. Despite being one of the most widely spoken languages in Liberia and Guinea, Kpelle remains severely underrepresented in NLP research, lacking any existing publicly available datasets. This absence stems from several factors, including Kpelle's status as a low-resource language with limited digital presence, the complexities arising from its dialectal variations across Guinea and Liberia (Konoshenko, 2008; Vydrin, 2018), and the lack of standardized orthography (Konoshenko, 2024). The dataset from this work will provide a much-needed resource for developing and evaluating NLP tools for Kpelle, enabling advancements in tasks like machine translation, language modeling, and speech recognition. By making this dataset publicly available, the project contributes to the broader goal of promoting language diversity and inclusion for African Languages.

3 Overview of Kpelle

As previously mentioned, Kpelle belongs to the Southwestern Mande branch of the larger Mande language family. Figure 1 illustrates how Kpelle fits within this broader linguistic context, demonstrating its relationship to other languages spoken throughout Liberia.

Kpelle boasts of a rich oral tradition, with storytelling, proverbs, and songs playing a pivotal role in preserving the history and cultural values of the people (Thach, 1981). Oral tradition has been key in maintaining the language across generations, especially since written text is limited (Thach, 1981). Also, Kpelle faces challenges in representation and expansive linguistic research due to its low-resources status.

Further, external influences have impacted the Kpelle language. In the 19th and 20th centuries, interactions with European colonizers and neighboring ethnic groups introduced new vocabulary into the language (Thach, 1981). However, Kpelle has kept its core linguistic structure and continues to thrive as a means of communication and cultural identity for its speakers (Thach, 1981).

3.1 Linguistic Features

In this paper, we focus on **Liberian Kpelle** which exhibits distinct linguistic features that set it apart within the Mande Language family.

3.1.1 Phonetics

Kpelle uses a sound system with a rich array of consonants and vowels (Thach, 1981; Vydrin, 2018; Konoshenko, 2024; Thach et al., 1981). Notably, it includes labiovelar stops such as /gb/ and /kp/, which are said simultaneously at the velar and bilabial places of articulation and represent single consonant sounds (Thach, 1981; Thach et al., 1981; Vydrin, 2018). These sounds are relatively rare in global languages and contribute to Kpelle's unique phonological profile. The vowel system in Kpelle has seven oral vowels and their nasal counterparts, making for a complex vocalic inventory (Vydrin, 2018). Dialectical variations influence pronunciation, particularly with the /s/ sound (Thach, 1981). In some regions, the /s/ can resemble the English /s/; in others, it may sound like /ʃ/ (as in "ship") or /h/ (Thach, 1981). These forms of variations can pose difficulties for language learners.

3.1.2 Syntax

Kpelle follows a Subject-Verb-Object(SVO) sentence structure, which aligns with the syntactic patterns of many languages in the world, including English (Thach, 1981; Vydrin, 2018; Konoshenko, 2008). This syntactic structure facilitates the translation of Kpelle to English to some extent. Kpelle also distinguishes between dependent and independent nouns, akin to the idea of inalienable and alienable possession seen in other languages (Thach, 1981; Vydrin, 2018). For example, body parts and kinship terms are treated differently grammatically compared to other nouns, affecting possessive constructions (Thach, 1981; Vydrin, 2018).

Modifiers in Kpelle usually follow the nouns they describe (Thach, 1981), and the language employs postpositions rather than prepositions (Vydrin, 2018). Verb serialization is also a feature

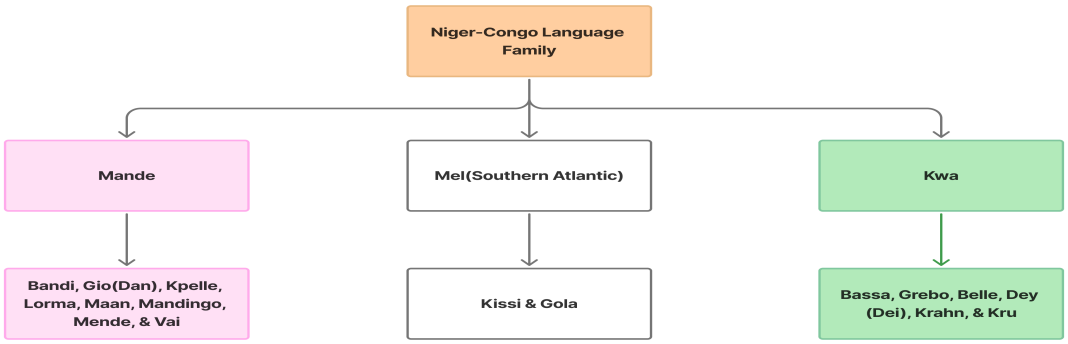


Figure 1: Overview of Liberian language family under the Niger-Congo Branch.

in Kpelle (Heine and Reh, 1984), where multiple verbs are used sequentially to convey complex actions or events without conjunctions.

3.1.3 Grammar

Kpelle grammar has a complex system of pronouns that reflect distinctions in person, number, and sometimes gender (Thach, 1981; Vydrin, 2018). The verb system marks tenses, aspect, and mood through affixes and particles (Thach, 1981; Thach et al., 1981). For example, there are specific markers for past, present, and future tenses and for completed and ongoing actions (Thach et al., 1981).

Noun classes in Kpelle are less prominent than in some other African languages but do exist and can affect agreement within the sentence (Vydrin, 2018). Kpelle employs emphatic particles like *'bé'* to convey emphasis or focus within a sentence (Thach, 1981). Since tone and stress are primarily used to convey lexical and grammatical meaning (Thach, 1981; Thach et al., 1981; Vydrin, 2018)- these particles play an important role in adding nuance and emphasis without altering the tonal structure.

3.1.4 Tonality

Liberian Kpelle is a tonal language, meaning that the pitch at which a syllable is said can change the word's meaning entirely (Thach, 1981; Thach et al., 1981; Vydrin, 2018; Konoshenko, 2024, 2008). Kpelle features three tone levels: high, mid, and low (Thach, 1981; Vydrin, 2018; Konoshenko, 2008). Tones can be level (staying the same throughout the syllable) or contour (changing pitch within the syllable) (Thach, 1981; Thach et al., 1981; Konoshenko, 2008). This tonal system is essential for distinguishing words that are otherwise identical phonetically (Konoshenko, 2008). For example, (Konoshenko, 2008) presents that "simple words in Kpelle form several groupings according to the tonal patterns which are assigned to these words lexically," and the groupings can be binned into categories known as *tonal classes* (Konoshenko, 2008). Also, a single syllable pronounced with a high tone might mean one thing (*lá, meaning mouth*); in a mid-tone, that same syllable communicates (*la, meaning it*), while the same syllable with a low tone means something entirely different (*là, meaning if*) (Thach, 1981).

Tone also plays a grammatical role in Kpelle, affecting verb tenses and aspects (Konoshenko, 2008). Tonal patterns indicate whether an action is completed, ongoing, or habitual. This reliance on tone adds a layer of complexity to Kpelle learning and computational processing since accurate tonal representation is critical, especially for this work. Table 1 presents the tones seen in Kpelle with examples.

Table 1: Tonal Levels in Kpelle adapted from (Thach, 1981; Weako, 2024)

Tonal Level	Mark	Kpelle Example	English Version
High	ˊ	zóo	native doctor
Mid	no mark/¯	tuna	rain
Low	ˋ	nyɔ̀ɔ	be afraid
High-Low	ˆ	sâa	today
Mid-High-Low	ˆ	tisô	sneeze
Low-High	ˇ	kǒ	to plant
Nasal	˜	sã	to dance

3.1.5 Writing System

Historically, Kpelle has been primarily an oral language, but people have worked to develop writ-

ing systems that promote literacy and documentation. An example is the Kpelle syllabary created by Chief Gbili in the 1930s, an indigenous script designed to represent the sounds of Kpelle (African 671, 2019). However, few people use this script today (African 671, 2019).

More commonly, Kpelle is written using Latin-based orthography (Vydrin, 2018). This system has been influenced by various scholars and linguists, such as William E. Welmers, who worked on developing practical orthographies for African languages in the mid-20th century (Konoshenko, 2008). The Latin-based orthography often has diacritical marks to show tonal variations (Konoshenko, 2024; Thach, 1981); moreover, the lack of standardization leads to inconsistencies in written materials (Konoshenko, 2024; Thach, 1981).

The Kpelle dictionary by (Leidenfrost and McKay, 2005) incorporates tonal markings and provides valuable resources for language learners and researchers (Thach, 1981; Konoshenko, 2008). Materials from the Kpelle Literacy Center in Totota also use the Latin script to promote written literacy among native speakers of Kpelle (Thach, 1981). The absence of a universally accepted orthography remains challenging, considering the variations between Liberian and Guinean Kpelle (Thach, 1981; Konoshenko, 2008).

4 Dataset Creation

Creating the English-Kpelle dataset involved planning and execution to ensure the data's relevance, accuracy, and cultural appropriateness. Our primary goal was to compile a corpus facilitating effective communication for individuals who may not speak Kpelle, particularly in everyday social interactions and essential services. This section outlines the data collection sources and methods, preprocessing steps, and translation alignment processes used in building the dataset.

4.1 Data Collection

4.1.1 Sources

The sources used in building the dataset covered a combination of practical and culturally relevant scenarios:

Travel and Tourism Phrases. We identified common phrases and questions frequently asked by tourists and travelers when they visit a new location. Usually, due to their unfamiliar disposition to the place, we focused on phrases that covered greetings, inquiries about locations, costs, weather conditions, and other essential interactions. The phrases were sourced from the following respected travel and language teaching website: *Business Insider's Travel Language Phrases*(Abadi, 2018), *EF Education First's Essential Phrases*(B, 2018), *Online Teachers UK's English for Tourism and Travel* (Writer, 2017), *Go Overseas' Language Phrases Before Travelling* (Perez, 2022), *Accessible Travel Phrasebook by Premiki* (Limited, 2018), and *Wikivoyage's Afrikaans Phrasebook* (Wikivoyage, 2005).

Religious Texts. Religious literature, like the Bible, often contains a wealth of translated material that can be valuable for language datasets. We added a few excerpts from publicly available religious texts that have been translated into Kpelle.

Educational Material. Significant portions of the dataset were sourced from the book *A Learner Directed Approach to Kpelle by Sharon V. Thach* (Thach, 1981), *English-Kpelle Dictionary, with a Grammar Sketch and English-Kpelle Finder List* (Leidenfrost and McKay, 2005), *We Have Come To Learn Kpelle* (Ricks, 2009). These resources had bilingual content, including matching English-Kpelle sentence pairs, standalone English paragraphs, and standalone Kpelle paragraphs.

4.1.2 Methods

Data Extraction. We gathered a list of essential phrases and sentences relevant to everyday communication from the travel and tourism websites. These phrases were selected based on their frequency of use and utility in facilitating introductory interaction.

Translation. For English or Kpelle paragraphs that did not have the corresponding translation, we engaged a native Kpelle speaker with linguistic expertise to provide accurate translations.

Segmentation of Paragraphs. In cases where the source material provided paragraphs rather than individual sentences, we segmented the text into sentence pairs. This approach increased the granularity of the dataset, making it suitable for machine translation tasks.

Expert Verification. All translated sentences were reviewed by Kpelle language experts to verify the accuracy of the translations, the correctness of tone and grammar, and the appropriateness of context.

4.2 Data Preprocessing

4.2.1 Cleaning

The raw data collected contained inconsistencies such as typographical errors, informal language, and irrelevant content. We performed a thorough cleaning process to remove these anomalies. This included spell-checking, correcting grammatical errors, and eliminating duplicate entries. Special attention was given to resolving translation inconsistencies, especially where multiple translations existed for a single English phrase. The most accurate and contextually appropriate translation was selected based on expert advice.

4.2.2 Normalization

Given Kpelle's lack of a universally accepted writing system, we adopted the Latin-based orthography commonly used in educational materials and literacy programs. Diacritical marks were standardized to represent tonal variations accurately. All text data was encoded using UTF-8 Unicode to ensure compatibility across different platforms and tools. This was essential for preserving special characters and tonal markers unique to Kpelle. To maintain consistency, all text was converted to a standard case format, except where capitalization was necessary for proper nouns and the beginning of sentences.

4.2.3 Segmentation

The text was segmented into individual sentences using punctuation cues and linguistic rules specific to Kpelle. This process was manually verified due to the potential for misinterpretation by automated tokenizers not tailored to Kpelle. Within sentences, words were tokenized based on whitespace and morphological patterns. This facilitated subsequent processing tasks such as alignment and statistical analysis. Kpelle often uses contractions and compound words. These were carefully identified and treated according to linguistic guidelines to ensure accurate tokenization.

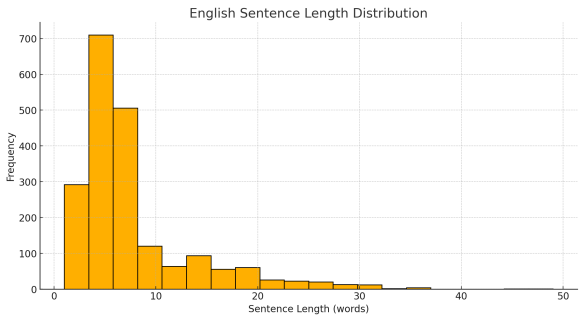
5 Dataset Statistics and Analysis

5.1 Quantitative Overview

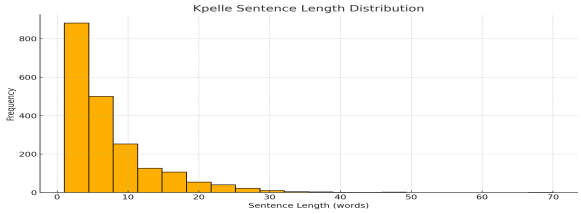
The dataset has 3234[2] entries corresponding to unique Kpelle-English translation pairs. Typically, each entry has one Kpelle sentence paired with its English equivalent; however, some entries contain sentences under a single translation unit (e.g., compound or complex sentences kept intact to preserve context). In total, the dataset contains 30,021 words (14,790 in Kpelle and 15,231 in English) and 4,369 sentences (2,202 in Kpelle and 2,167 in English). The longest sentences contain 70 Kpelle words and 49 English words, with the shortest being a single word in either language. Moreover, there are 4,702 unique Kpelle words and 3,579 unique English words, resulting in an overall vocabulary of 8,281 entries. These statistics make this the largest publicly available bilingual English–Kpelle resource to date.

5.2 Sentence Length

After our distribution analysis, we observed that most of the English sentences ranged from 3 to 15 words, with an average length of around 8 words per sentence. The Kpelle sentences vary more due to certain functional words' presence (or absence) and the possibility of encoding multiple concepts in a single phrase. However, the average Kpelle sentence length approximates 7 words, with most sentences falling between 3 and 12.



(a) English sentence length distribution



(b) Kpelle sentence length distribution

Figure 2: Sentence length distributions for English (top) and Kpelle (bottom), illustrating the corpus's inherent variability.

The wide range of sentence lengths reflects the dataset's inclusion of both simple and more complex utterances. Short, single-word sentences often correspond to exclamations, greetings, or short prompts, while longer sentences derive from reli-

[2]This count refers specifically to Version 2 of our dataset, which extends the initial 1,518 sentence pairs to 2,005 and increases word entries from 1,181 to 1,229.

gious or educational materials that contain embedded clauses and descriptive text.

5.3 Vocabulary Frequency

In terms of vocabulary, the top ten most frequent english words were *man (116), good (93), town (93), want (83), go (75), going (65), one (61), house (59), baby (54), went (53)*. Similarly, the top ten most frequent Kpelle words were *su (177), pâi (143), la (123), kaa (123), kɛ̀ (112), mɛ (108)ni, pɔ̂ri (104), li (101), kɛ (99), kɛ̂i (82)*.

Even though we remove common stop words, frequent English words indicate a high presence of articles, pronouns, and commonly used verbs, mirroring everyday conversational usage. On the Kpelle side, repeated use of function words like ***a, da,*** and ***e*** underscores similar syntactic necessities. These observations led to an English Hapax Legomena (words that appear once) of 1732 and a Kpelle Hapax Legomena of 2714.

A high number of hapax legomena suggests a rich and diverse vocabulary, but it also indicates that many words appear in the dataset with minimal frequency. This sparsity could pose challenges for certain NLP models, as low-frequency words often result in less robust embeddings and higher rates of out-of-vocabulary (OOV) tokens.

5.4 Domain Coverage

We conducted a keyword-based classification across common categories to understand the dataset's topical breadth. Table 2 shows that **Daily Conversation (664)** and **Household (214)** predominate, while underrepresented categories were **Religion (27)**, **Health (21)**, and **Education (14)**. It is worth noting that around 30% of the dataset remains unclassified, reflecting idiomatic expressions and content not easily mapped to predefined categories. However, the broad coverage of the dataset, given the number of entries, ensures the dataset can serve a variety of use cases.

5.5 Observations and Challenges

Even though we adopt a standardized Latin-based script, Kpelle orthography's dynamic and evolving nature continues to introduce spelling and tone-marking variations throughout the dataset. These inconsistencies highlight the broader challenges of documenting a language with limited written traditions and underscore the importance of ongoing refinement in orthographic conventions. Additionally, the low representation of domains such as Religion, Health, and Education highlights future avenues for data collection to achieve more balanced coverage. The distribution of topics also shows that key domains, such as Religion, Health, and Education, remain underrepresented, emphasizing potential areas for future data collection and corpus expansion to achieve more balanced coverage.

Domain	Number of Sentences
Daily Conversation	664
Household	214
Business & Finance	142
Family	93
Time & Events	91
Nature & Environment	80
General Purpose	59
Religion	27
Health	21
Travel & Tourism	19
Education	14
Unclassified	581

Table 2: Distribution of Sentences by Domain

6 Experiments and Benchmarking

This section presents our machine translation experiments and benchmarking using the NLLB model by (Team et al., 2022). We describe our baseline models, outline the fine-tuning process, report quantitative results using standard evaluation metrics, and provide an analysis comparing our outcomes with previously reported NLLB-200 performance in other African languages. Figure 3 visually summarize this process.

6.1 Baseline Models and Experimental Setup

Given its strong performance across low-resource African languages, we leveraged Meta's NLLB model as a baseline. Our experiments focus on two Kpelle dataset versions:**Version 1 (V1)** contains 1,667 Kpelle and 1,638 English sentences (3,852 and 2,952 unique words). **Version 2 (V2)** benefits from data augmentation efforts, yielding 2,202 Kpelle and 2,167 English sentences (4,702 and 3,579 unique words). We aimed to assess how expanding the corpus (from 1,518 to 2,005 translation pairs) affects translation quality in both English → Kpelle (eng_Latn → kpe_Latn) and Kpelle → English (kpe_Latn → eng_Latn). We split each dataset into two sets, train and test, according to a 9:1 ratio and hold out the test set. Then, these sets were fine-tuned for 10k, 30k, and 60k steps on top of NLLB using Adafactor as the optimizer

(a) Fine-tuning NLLB-200 with Kpelle. (b) Translation example using the fine-tuned model.

Figure 3: NLLB-200 fine-tuning with Kpelle: (a) Model adaptation for bidirectional translation, and (b) a sample translation.

with a batch size of 8, constrained by the memory requirements of the Quadro RTX 6000 GPU, training times ranged between 30 minutes to 6 hours, dependent on the number of steps and the version of the dataset. We trained a Kpelle-specific tokenizer (a SentencePiece model (Kudo and Richardson, 2018)) on data from Penedo et al. (2024) to handle out-of-vocabulary tokens and then enriched the standard NLLB tokenizer with any missing tokens, ensuring compatibility with the model's subword vocabulary. Finally, we used sacreBLEU (Post, 2018) to measure BLEU, 1–4-gram precision, brevity penalty (BP), hypothesis/reference lengths, and chrF2++ to evaluate the fine-tuned model.

6.2 Results

	eng_Latn → kpe_Latn							
	Steps	BLEU	chrF2++	Precision (1-4 grams)				BP
				1-gram	2-gram	3-gram	4-gram	
NLLB	10k	**24.09**	38.24	49.3	28.7	20.5	16.8	0.913
	30k	**24.46**	38.20	50.2	29.1	20.6	16.8	0.918
	60k	**24.00**	38.19	50.1	28.2	19.5	16.1	0.930
NLLB V2	10k	19.80	**38.26**	49.6	25.4	15.5	10.0	0.942
	30k	19.97	**38.42**	49.1	24.6	15.2	10.2	0.961
	60k	20.79	**38.83**	51.5	26.9	16.9	11.4	0.915
	kpe_Latn → eng_Latn							
	Steps	BLEU	chrF2++	Precision (1-4 grams)				BP
				1-gram	2-gram	3-gram	4-gram	
NLLB	10k	23.16	38.29	42.5	24.7	18.6	14.7	1.000
	30k	24.31	39.60	44.1	26.6	19.4	15.3	1.000
	60k	23.65	39.41	43.1	25.2	18.9	15.2	1.000
NLLB V2	10k	**26.39**	**40.22**	50.0	30.6	20.9	15.2	0.999
	30k	**30.03**	**44.00**	52.4	34.0	24.7	18.4	1.000
	60k	**30.28**	**44.28**	53.4	34.5	24.8	18.3	1.000

Table 3: NLLB performance when fine-tuned on two versions of the English-Kpelle dataset (V1 and V2) at 10k, 30k, and 60k steps. Metrics (BLEU, chrF2++, 1–4-gram precision, and BP) are reported for both eng_Latn→kpe_Latn and kpe_Latn→eng_Latn. Bold scores denote the best performance.

Table 3 summarizes the results for NLLB fine-tuned on V1 and V2 of the Kpelle dataset across 10k, 30k, and 60k training steps.

We observe that moving from **V1** (1,518 entries) to **V2** (2,005 entries) improved BLEU scores in some scenarios, particularly for kpe_Latn → eng_Latn translation at higher step counts (e.g., 30k, 60k). This outcome aligns with the broader expectation that additional in-domain data can boost model performance in low-resource settings. Further, we also observe that increasing the fine-tuning steps from 10k to 30k and 60k generally yielded incremental gains for both versions. However, the improvements were again more pronounced when translating from Kpelle to English. In contrast, eng_Latn → kpe_Latn translation showed modest gains, suggesting that further optimization may be necessary to achieve comparable results in translation quality for Kpelle.

6.3 Analysis and Comparison with NLLB-200 Benchmarks

Reports by Team et al. (2022) highlight NLLB-200's performance across multiple African languages (e.g., Hausa, Igbo, Swahili, Yoruba). As shown in Table 4, M2M-100, MMTAfrica, and NLLB-200 yield varying BLEU and chrF2++ scores for these languages. Given the differences in language structure, dataset sizes, and domain coverage, cross-lingual comparisons should be made cautiously. However, **the scores we observe for Kpelle (BLEU in the range of 20–30 depending on the direction and training steps) are generally consistent with NLLB-200's range for other African languages.**.

	eng_Latn–xx			**xx–eng_Latn**		
	MMTAfrica	M2M-100*	NLLB-200	MMTAfrica	M2M-100*	NLLB-200
hau_Latn	-/-	4.0/-	**33.6/53.5**	-/-	16.3/-	**38.5/57.3**
ibo_Latn	21.4/37.2	19.9/-	**25.8/41.4**	15.4/38.9	12.0/-	**35.5/54.4**
lug_Latn	-/-	7.6/-	**16.8/39.8**	-/-	7.7/-	**27.4/46.7**
luo_Latn	-/-	13.7/-	**18.0/38.5**	-/-	11.8/-	**24.5/43.7**
swh_Latn	40.1/53.1	27.1/-	37.9/**58.6**	28.4/56.1	25.8/-	**48.1/66.1**
wol_Latn	-/-	8.2/-	**11.5/29.7**	-/-	7.5/-	**22.4/41.2**
xho_Latn	27.1/44.9	-/-	**29.5/48.6**	21.7/48.6	-/-	**41.9/59.9**
yor_Latn	12.0/28.3	13.4/-	**13.8/25.5**	9.0/30.6	9.3/-	**26.6/46.3**
zul_Latn	-/-	19.2/-	**36.3/53.3**	-/-	19.2/-	**43.4/61.5**

Table 4: BLEU/chrF2++ performance on selected African languages (eng_Latn ↔ xx) for MMTAfrica, M2M-100*, and NLLB-200 from (Team et al., 2022).

Our kpe_Latn → eng_Latn best BLEU of **30.28** at 60k steps surpasses NLLB-200's lower-bound performances (22.4 BLEU on Wolof), mid-range(24.5 BLEU on Luo, 26.6 BLEU on Youruba, 27.4 BLEU on Luganada) results, though it remains below the model's high performance (48.1 BLEU on Swahili). The eng_Latn → kpe_Latn

translation lags slightly behind, reaching approximately **24.46** BLEU with V1 at 30k steps. This result is comparable and higher to NLLB-200's results ($\approx$ 25.8 BLEU in some languages) but lower than its highest observed values (37.9 BLEU in Swahili). Kpelle translations have the potential to reach NLLB-200's highest performance levels with further data augmentation and fine-tuning. However, language-specific nuances, such as Kpelle's orthographic variations, limited standardization, and relatively small corpus size, currently limit model performance.

7 Conclusion

This paper introduced the first publicly available English-Kpelle dataset for machine translation. Our corpus has over 2,000 translation pairs from diverse domains, such as daily conversation, household activities, and religious texts. We demonstrated the dataset's usability by fine-tuning Meta's NLLB model on two corpus versions. Our experiments revealed that data augmentation significantly benefits translation performance, particularly in the Kpelle-to-English direction at higher fine-tuning steps. These findings highlight the importance of domain-specific data expansion in enhancing translation quality for low-resource languages. Moreover, comparative analysis against reported NLLB-200 results highlights the potential for Kpelle NLP systems to achieve competitive performance levels, given continued data curation and iterative fine-tuning.

8 Limitations

1. **Dataset Expansion and Domain Coverage:** While we have made progress in building a representative English-Kpelle dataset, some gaps remain. Future efforts could focus on collecting domain-specific materials from underrepresented categories such as nature, environment, and specialized technical fields to enhance the domain coverage of the dataset further. Adding more varied dialectal data is also essential to capture the linguistic richness of Kpelle more comprehensively.

2. **Broader NLP Applications:** Beyond machine translation, the dataset can be a foundation for other NLP tasks, including speech recognition, language modeling, and sentiment analysis. We intend to explore these avenues, building on the groundwork established here to develop robust and context-aware Kpelle language tools.

3. **Limited Cross Model Evaluation:** Our current evaluation relies exclusively on fine-tuning Meta's NLLB model. While NLLB provides a strong baseline for low-resource translation, this restricts our understanding of how the dataset performs across diverse architectures. As future work, we plan to benchmark an expanded version of the dataset on additional models, including M2M-100 and BLOOMZ, to better assess transferability and generalization. We also intend to incorporate complementary evaluation metrics, such as TER and METEOR, to provide a more comprehensive analysis of model performance.

4. **Lack of Qualitative Error Analysis:** The current scope of this work sought to present the first English-Kpelle dataset and understand Kpelle's potential by benchmarking on a strong baseline like Meta's NLLB model. Given this, we failed to conduct a qualitative error analysis on the translation generated for the held-out test set. In future work, we plan to introduce human evaluation loops where native Kpelle speakers assess translation quality and identify systematic errors. This feedback will guide targeted model improvements and support a more fine-grained understanding of the dataset's linguistic challenges.

8.1 Call to Action

We invite researchers, linguists, and language technology enthusiasts to collaborate in expanding and refining this dataset. By contributing additional Kpelle text resources, validating translations, or developing novel NLP techniques, the research community can help bridge the digital divide faced by low-resource languages. We hope the work presented here will spark renewed interest in Kpelle and other underrepresented Mande languages, ultimately driving innovation and inclusivity in multilingual NLP.

9 Acknowledgments

We acknowledge the contributions of Mr. Aaron D. Y. Pope, Cuttington University, and Mr. Better Jallah, University of Liberia, who served us our Kpelle translation experts, providing Kpelle

translation pairs for gathered English sentences and words.

References

Mark Abadi. 2018. I've been to 25 countries and i can tell you there are only 11 phrases you need to get by anywhere.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. *arXiv preprint*. ArXiv:2205.02022.

University of Wisconsin-Madison Students in African 671. 2019. Kpelle- history and brief intro. Lesson on syllabary and alphabet. Publisher: Pressbooks.

Emmanuel Agyei, Xiaoling Zhang, Stephen Bannerman, Ama Bonuah Quaye, Sophyani Banaamwini Yussi, and Victor Kwaku Agbesi. 2024. Low resource twi-english parallel corpus for machine translation in multiple domains (twi-2-eng). *Deleted Journal*, 27.

Adewale Akinfaderin. 2020. HausaMT v1.0: Towards English–Hausa neural machine translation. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 144–147, Seattle, USA. Association for Computational Linguistics.

D. Asamoah Owusu, A. Korsah, B. Quartey, S. Nwolley Jnr., D. Sampah, D. Adjepon-Yamoah, and L. Omane Boateng. 2022. Github - ashesi-org/financial-inclusion-speech-dataset: A speech dataset to support financial inclusion created by ashesi university and nokwary technologies with funding from lacuna fund. https://github.com/Ashesi-Org/Financial-Inclusion-Speech-Dataset. [online] GitHub.

Asmelash Teka Hadgu, Gebrekirstos G. Gebremeskel, and Abel Aregawi. 2022. Machine Translation Benchmark Dataset for Languages in the Horn of Africa. Original-date: 2021-12-05T14:04:38Z.

Sara B. 2018. 13 important phrases to know in your second language.

Gideon George, Olubayo Adekanmbi, and Anthony Soronnadi. 2024. TangaleNLP: Building po tangle to english parallel corpora and machine translation of the tangle (tangale) language. In *5th Workshop on African Natural Language Processing*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

B. Heine and M. Reh. 1984. *Grammaticalization and Reanalysis in African Languages*. H. Buske.

Maria Konoshenko. 2024. Quotatives in guinean and liberian kpelle: A study of parallel bible corpora and non-biblical texts. *Acta Linguistica Petropolitana*, 19(3):558–583.

Maria Yu Konoshenko. 2008. Tonal systems in three dialects of the kpelle language. *Mandenkan*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Siva Subrahamanyam Varma Kusampudi, Anudeep Chaluvadi, and Radhika Mamidi. 2021. Corpus creation and language identification in low-resource code-mixed Telugu-English text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 744–752, Held Online. INCOMA Ltd.

Theodore E. Leidenfrost and John S. McKay. 2005. *Kpelle-English Dictionary, with a Grammar Sketch and English-Kpelle Finder List*. Language-Literacy-Literature and Bible Translation Center- Lutheran Church in Liberia, Totota.

Lonely Planet Global Limited. 2018. 35 languages covered accessible travel phrasebook.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukiibi, Eric Peter Wairagala, Chodrine Mutebi, Tobius Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, and Andrew Katumba. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: Experiences and lessons. *Applied AI Letters*, 5(2):e92.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Vinh Van Nguyen, Ha Nguyen, Huong Thanh Le, Thai Phuong Nguyen, Tan Van Bui, Luan Nghia Pham, Anh Tuan Phan, Cong Hoang-Minh Nguyen, Viet Hong Tran, and Anh Huu Tran. 2022. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5494–5502, Marseille, France. European Language Resources Association.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane – machine translation for africa. *Preprint*, arXiv:2003.11529.

Denis Paperno. 2014. Sample texts in beng. *Mandenkan*, pages 106–111.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages.

Olivia Christine Perez. 2022. Helpful language phrases to learn before you travel | go overseas.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Paul Kanmu Ricks. 2009. *Kwaa Pa Kpelee-Woo Maa Kori(We Have Come to Learn Kpelle)*, first edition. Cuttington University.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Sharon V Thach. 1981. *A Learner Directed Approach to Kpelle. A Handbook on Communication and Culture with Dialogs, Texts, Cultural Notes, Exercises, Drills and Instructions [microform] / Sharon V. Thach and Others*. Distributed by ERIC Clearinghouse. Also contributed by Michigan State Univ., East Lansing. African Studies Center. Accessed: 20 February 2025 via National Library of Australia.

S.V. Thach, D.J. Dwyer, and Michigan State University. African Studies Center. 1981. *Kpelle, a Reference Handbook of Phonetics, Grammar, Lexicon and Learning Procedures*. [Prepared] for the United States Peace Corps at the African Studies Center of Michigan State University.

Valentin Vydrin. 2018. Mande languages.

Valentin Vydrin, Jean-Jacques Meric, Kirill Maslinsky, Andrij Rovenchak, Allashera Auguste Tapo, Sebastien Diarra, Christopher Homan, Marco Zampieri, and Michael Leventhal. 2022. Machine learning dataset development for manding languages. urlhttps://github.com/robotsmali-ai/datasets.

Alexandra Vydrina. 2017. *A corpus-based description of Kakabe, a Western Mande language: prosody in grammar*. Theses, Institut National des Langues et Civilisations Orientales.

Barack Wanjawa, Lilian D. A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2024. Kencorpus: Kenyan Languages Corpus.

Jackson Weako. 2024. Libtralo kpelle keyboard help. *Keyman.com*.

Contributors Wikivoyage. 2005. West germanic language, spoken in south africa and namibia.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Liam G.-Staff Writer. 2017. English for tourism: Essential uk travel phrases with examples.