Does Generative AI speak Nigerian-Pidgin?: Issues about Representativeness and Bias for Multilingualism in LLMs

David Ifeoluwa Adelani[*]
Mila - Quebec AI Institute
McGill University, Canada CIFAR AI Chair
david.adelani@mcgill.ca

A. Seza Doğruöz[*]
LT3, IDLab, Universiteit Gent
as.dogruoz@ugent.be

Iyanuoluwa Shode
Bloomberg
ishode@bloomberg.net

Anuoluwapo Aremu
University of Trento
Lelapa AI

Abstract

Nigeria is a multilingual country with 500+ languages. Naija is a Nigerian Pidgin spoken by approximately 120M speakers and it is a mixed language (e.g., English, Portuguese, Yoruba, Hausa and Igbo). Although it has mainly been a spoken language until recently, there are some online platforms (e.g., Wikipedia), publishing in written Naija as well. West African Pidgin English (WAPE) is also spoken in Nigeria and it is used by BBC to broadcast news on the internet to a wider audience not only in Nigeria but also in other West African countries (e.g., Cameroon and Ghana). Through statistical analyses and Machine Translation experiments, our paper shows that these two pidgin varieties do not represent each other (i.e., there are linguistic differences in word order and vocabulary) and Generative AI operates only based on WAPE. In other words, Naija is underrepresented in Generative AI, and it is hard to teach LLMs with few examples. In addition to the statistical analyses, we also provide historical information on both pidgins as well as insights from the interviews conducted with volunteer Wikipedia contributors in Naija.

1 Introduction

Between 16th-19th centuries, there were contacts between Europeans and non-Europeans outside Europe. In West Africa, contacts between English and West African languages led to simplified and mixed languages combining linguistic features from several languages. These new forms of languages were lingua francas (i.e., common or bridge languages) that served for a mutual understanding between speakers of different languages for various purposes (e.g., trade, plantation agriculture, mining) (Mufwene, 2024). The terms "pidgin" and "creole" are used to refer to these languages. Although there is a lack of agreement about the precise definitions and coverage of these terms, pidgin roughly refers to the "speech-forms which do not have native speakers, and are therefore primarily used as a means of communication among people who do not share a common language" (Muysken et al., 1995). Creoles, on the other hand, are assumed to be extended pidgins which are more established and have native speakers especially in urban environments (Muysken et al., 1995).

Nigeria is a multilingual country in West Africa hosting over 500 different languages spoken by approximately 220 million people across 371 ethnic tribes (Eberhard et al., 2019). It is the sixth most populous country in the world and Africa's most populous country. English is the official language and acquired mostly through formal education in Nigeria (Agbo and Plag, 2020). The three major tribes in Nigeria with their respective languages include Hausa (spoken by 63M speakers), Igbo (27M speakers), and Yorùbá (42M speakers). Nigerian Pidgin (Naija) is a mix of English with local languages (e.g., Portuguese, Yorùbá, Igbo, Hausa) (Balogun, 2013a), (Oyebola and Ugwuanyi, 2023). Naija is widely spoken (approx. 120M speakers) as a first and second language (Adelani, 2022) around the Southern part of Nigeria (e.g., Lagos and Niger-Delta) with origins going back to the English-Creole Atlantic Krio language family. It is also adopted as the unifying and unofficial language for communication across ethnically diverse groups. According to some researchers (e.g., Muysken et al. (1995)), Naija has evolved into a creole over time and has now native speakers as well. However, it is still referred to as a pidgin among the locals.

Although Naija may have words that sound similar to English, their meanings may vary and there

[*] Equal contribution

Findings of the Association for Computational Linguistics: NAACL 2025, pages 1571–1583
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

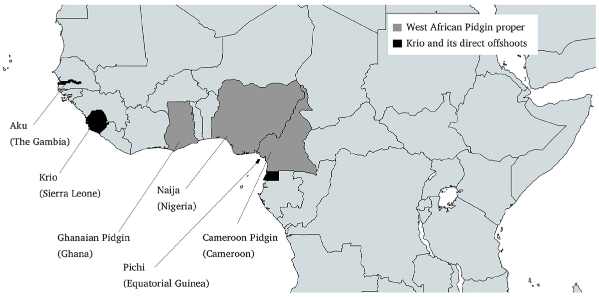
▨ West African Pidgin proper
■ Krio and its direct offshoots
Aku (The Gambia)
Krio (Sierra Leone)
Naija (Nigeria)
Ghanaian Pidgin (Ghana)
Cameroon Pidgin (Cameroon)
Pichi (Equatorial Guinea)

Figure 1: **WAPE Locations**: West African countries whose official language is English (The Gambia, Sierra Leone, Liberia, Ghana and Nigeria), Cameroon (North-Western and South-Western Anglophone region), and Krios immigrants to Equatorial Guinea. Map is obtained from Yakpo (2024).

is no standardized orthography for Naija (Marchal et al., 2021; Akande and Salami, 2021; Lin et al., 2024).

Until recently, **West African Pidgin English (WAPE)** has been mainly a spoken language with many local varieties (e.g., Nigerian Pidgin, Ghanaian Pidgin English, Cameroonian Pidgin English). Despite the large number of speakers across West Africa countries, WAPE remained as a spoken language until 2017 when the British Broadcasting Company (BBC) launched a news website (West African Pidgin English). It aims to target the diversity of WAPE speakers across different countries. [1] Since 2022, Naija is also accepted as one of the languages on Wikipedia [2]. Although they are mutually understandable, there are linguistic and social differences between the two written varieties. For example, (WAPE on BBC website) resembles English in terms of word order and vocabulary (see example (1)) with a simplified grammar, lacking auxiliary "were"). However, Naija on Wikipedia has a different word order and vocabulary choice (e.g., ”moto” instead of a "car" and ”wund" instead of ”injured" or ”wounded").

Example (1)

WAPE Two pesin in di car dey injured.

Naija Na wund di two pesin get for di moto.

English Translation: Two persons in the car were injured.

From the sociolinguistic perspective, WAPE is favored mostly by educated Nigerians (who also speak, read and write in English) whereas Naija is used mostly in everyday life and it is more accessible for a larger audience (Akande and Salami, 2012). In terms of the resources, there is much more data on the Internet about WAPE than Naija (160K sentences vs. 25K sentences) (Ogueji et al., 2021).

There is a growing need for more research for multilingual and low resource languages (Doğruöz and Sitaram, 2022; Doğruöz et al., 2021) in Generative AI systems. This need is even enhanced for pidgin and/or creole languages due to their high numbers of speakers but lack of data (Lent et al., 2022, 2023). However, there are also unresolved issues about to what extent the available data on the Internet represent the language spoken in real life (Doğruöz et al., 2023). It is crucially important that the different genres of the same pidgin and creole languages are also represented in these systems to be inclusive and accessible for all speakers/users with diverse backgrounds.

In our paper, we address these issues for WAPE and Naija in the Nigerian context with the following contributions. We introduce WARRI [3] as a new MT evaluation data set including written WAPE on the BBC website and written Naija on Wikipedia. We also had interviews with the Naija Wikipedia contributors to understand their motives and regulations about written Naija. Our paper is the first to systematically analyze the similarities and differences between WAPE and Naija in Nigeria. Through a Machine Translation (MT) experiment, we find that Generative AI models (e.g., GPT-4o and LLAMA 3.1 8B (Touvron et al., 2023)) are biased towards WAPE and they do not include Naija despite large numbers of speakers. Further analysis shows that LLMs are hard to teach with few examples (e.g., 5-shots) to generate text in Naija. For reproducibility purposes, we release the Warri dataset and our evaluation code on GitHub under CC-BY-4.0 license.[4]

2 Related Research

Available research on representativeness originates from corpus linguistics where it is important to include samples from different textual sources to have a balanced and representative (smaller) corpus reflecting the variation in the (larger) corpora

[1] https://www.bbc.com/news/world-africa-40975399

[2] https://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Nigerian_Pidgin

[3] Warri is the name of a city in Delta State in the South-South region of Nigeria where Naija is widely spoken

[4] https://github.com/McGill-NLP/Naija-representation-in-LLMs

(Biber, 1993). Similarly, Crowdy (1993) states the significance of representative sampling corpora to minimize bias and maximize the credibility and consistency of the linguistic analyses. Therefore, representative sampling encompasses a broad spectrum of language usage across various contexts, genres, and demographic factors.

Generative AI systems depend on the availability of large data sets on the Internet. However, this assumption does not consider the representativeness of the variation in the available data sets which is especially difficult to obtain for multilingual and low resource languages (Doğruöz et al., 2023). While developing language technologies for multilingual and low resource languages, it is crucially important to be aware of the linguistic variation (e.g., WAPE and Naija) within these languages and aim for representing the variation in a balanced way to prevent potential bias.

To investigate the potential bias in the Nigerian context, the first step is to establish to what extent WAPE and Naija are similar or different from each other linguistically.

3 WARRI MT benchmark dataset for WAPE and Naija

To establish the WARRI data set, we used a portion of WAPE BBC news data, previously used in MasakhaNER dataset (Adelani et al., 2021b). It is a Named Entity Recognition (NER) dataset with available untokenized texts. We downloaded the Naija Wikipedia data from the Hugging Face.[5] After the data collection, we created a parallel data set in English by recruiting two bilingual speakers. They translated about 505 sentences from WAPE BBC data and Naija Wikipedia data into English. In this way, we maintained high-quality datasets by preventing the translators from mixing the features of the two pidgins into one. However, this also introduced a new obstacle (i.e., comparison of two test sets from slightly different domains (news vs. Wikipedia)).

To handle the domain related obstacle, we created a **multi-way parallel dataset** for Wikipedia domain. First, we asked a bilingual speaker to translate the Naija Wikipedia sentences into English. Then, we asked a professional translator (a different person), to translate the English sentences into WAPE following the BBC style of writing.

Table 1 provides the details of our new WARRI dataset, containing **single-way parallel sentences** (translated from WAPE BBC to English by two native speakers), and **multi-way parallel sentences** (where the English sentences have parallel translations in both WAPE BBC and Naija Wikipedia). Our test set composed of 500 sentences and the remaining five sentences were for few-shot/in-context learning for LLMs.

Other datasets in Naija There are also other parallel translation datasets (i.e., Naija-English). We also perform an analysis and evaluation on them, and compare them to our WARRI dataset. Table 2 provides one example each per dataset.

(a) Bible: We found two Naija Bibles online. The first one was translated by Wycliffe Bible Translators, and is part of the freely available eBible corpus (Akerman et al., 2023). Naija Wikipedia contributors also agree that this Bible conforms with Naija rather than WAPE. The other Bible translation was created by the Mercy Christian Ministry International (MCMI), [6] which is written to be closely similar to an African languages in Nigeria including the use of underdot diacritics as they are used in the Igbo language (e.g., "Wọd" for "Word"). Table 2 shows the two Naija Bible styles but there is not an established standard. We focus on our analysis on Wycliffe Bible (Naija 1) since the MCMI Bible (Naija 2) does not have the complete Bible online. We divided the Wycliffe Bible data into 31,051/1,500/1,500 TRAIN/DEV/TEST split.

(b) JW300: Similar to the Bible, JW300 (Agić and Vulić, 2019) is based on religious texts, bible studies and missionary reports of Jehovah Witness ministry in various languages. JW300 covers 343 languages including Naija. We divided the data into 23,322/ 1,500/ 1,500 TRAIN/DEV/TEST split.

(c) UD-Pidgin: This is based on the Universal Dependecy (UD) project for Naija (Caron et al., 2019). The data is based on the transcript of a conversationbetween two Naija speakers. [7] We divided the data into 6,241/ 1,500/ 1,500 TRAIN/DEV/TEST split.

[5] We make use of the 20231101 version, https://huggingface.co/datasets/wikimedia/wikipedia

[6] https://nigerianpidgin-bible.yolasite.com/

[7] https://github.com/UniversalDependencies/UD_Naija-NSC

Dataset	Creole	Domain	Average length (Pidgin)	TRAIN	DEV	TEST
Bible	PCM	religious	25.2	31,051	1,500	1,500
JW300	PCM	religious	17.2	23,322	1,500	1,500
UD	PCM	spoken	10.6	6241	1,500	1,500
MAFAND	PCM	news	25.0	4,790	1,484	1, 564
WARRI (single-way)	WAPE	news (BBC)	20.8	5	-	500
WARRI (multi-way)	WAPE & PCM	Wiki	21.3	5	-	500

Table 1: **WARRI and other datasets:** WARRI is only used for evaluation in zero or few-shot (e.g. 5) setting. WARRI (multi-way) have the same sentences in both WAPE and Naija (PCM is the ISO 639-3 code) pidgins unlike WARRI (single-way). We label each dataset based on the specified pidgin assigned by the creators of the dataset.

Lang.	Sample Sentences (English & Pidgin)
Bible	
English	And the Word became flesh, and dwelt among us
Naija 1	Den di Word kon shange to pesin and e stay with us for dis world
Naija 2	Di Wọd kọm bikọm human bin an Im liv wit ọs
JW300	
English	What can we do to make wise use of our freedom?
Naija	Wetin go help us use our freedom well?
MAFAND	
English	Each group is supposed to submit its needs
Naija	Each group suppose bring di things wey dem need kom
UD	
English	And I love the job with all my heart
Naija	And I love di job as in wit all my heart
BBC	
English	It is great - nothing is better than proving people wrong
WAPE	E dey great - nothing better pass make you prove pipo wrong
Wikipedia	
English	He married one wife with 7 children.
Naija	Na one wife im mari an dem don bon 7 pikin.

Table 2: Example of different styles of Pidgin used in different corpora

(d) MAFAND: This is based on the news domain. The news articles were obtained from English Daily Trust newspaper (published in Nigeria), and translated to Naija (Adelani et al., 2022). We make use of the same split as the MAFAND corpus with 4,790/1,484/1,564. Unlike the other datasets, it can be considered as "general domain" similar to Wikipedia.

Aside from parallel corpora, large amounts of WAPE unlabelled texts have been collected in literature from BBC (Ogueji et al., 2021) to train language models such as AfriBERTa. The AfriBERTa corpus has more than 160,000 sentences. Other sources of data for Naija are often smaller (e.g., Naija tweets (Muhammad et al., 2022)). However, we primarily focus on the parallel data sources for our analyses.

4 Experimental setup

We conduct three types of experiments to find out if WAPE and Naija are similar to each other: (1) Statistical analyses of the texts obtained from different datasets to measure their similarity to English and to each other. (2) Cross-corpus zero-shot transfer results when an MT model is trained on one dataset and evaluated on another. We expect domains that are similar should have a higher performance (Adelani et al., 2021a; Lee et al., 2022). Similarly, we expect transfer results to be higher if the pidgins are similar in terms of writing. (3) Prompting an LLM to find out whether WAPE or Naija is represented in Generative AI. We compare the results to the evaluation of WARRI MT benchmark dataset when trained on MAFAND.

4.1 Statistical analysis of the texts

First, we compute the lexical similarity between the English portion of each dataset and Pidgin by measuring **Jaccard similarity** (in percentage) for each corpus unigram, bigram, and trigram tokens. Secondly, we compute the **Levenshtein distance** (Levenshtein, 1965) which is an edit distance between each English test sentences and their translations to WAPE and Naija. Finally, we make use of three additional text generation metrics to measure their similarity to English: BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017) and BERTScore (Zhang et al., 2020). BLEU and ChrF++ are n-gram matching metrics, while **BLEU** focuses on word-level matching, **ChrF++** helps with evaluating character-level differences which are more common for Pidgin. Therefore, it is more reliable. **BERTScore** is an embedding-based metric that measures the semantic relationship between the sentence embeddings of two sentences. Therefore, it has better correlations with the human judgments.

					single-way	multi-way	
Metric	**Bible**	**JW300**	**UD**	**MAFAND**	**WAPE)**	**Wiki (WAPE)**	**Wiki (Naija)**
Jaccard Similarity ([0,1] range)							
Unigram ↑	0.133	0.295	0.537	0.554	0.712	0.802	0.517
Bigram ↑	0.025	0.086	0.149	0.178	0.289	0.371	0.167
Trigram ↑	0.002	0.025	0.055	0.076	0.151	0.207	0.084
Levenshtein distance ↓	88.5	56.5	30.3	58.0	26.6	21.6	53.6
BLEU ↑	0.8	11.2	16.8	20.9	36.1	46.8	23.4
ChrF++ ↑	20.5	30.4	43.7	54.7	65.2	73.4	51.3
BERTScore ↑	72.4	79.6	82.6	82.4	87.4	90.5	79.8

Table 3: **Lexical overlap and Levenshtein distance on WARRI benchmark**. Lexical overlap is measured by Jaccard similarity between English and WAPE (BBC) and Naija (Wikipedia). For multi-way WARRI corpus, the source of the data is from Wiki. The WAPE translation is denoted as Wiki (WAPE) but the original text is in Naija.

4.2 Cross-corpus zero-shot transfer results

We evaluate the performance of training an MT model on a source corpus and evaluate the performance on a target corpus. The source corpus are MAFAND, Bible, JW300, and UD, while the target corpus can be one of source corpora, and the WARRI dataset i.e. single-way and multi-way test sets. We perform an evaluation based on ChrF++ due to its reliability about capturing the character-level differences between Pidgin and English. Following Adelani et al. (2022), we leveraged a pretrained model to train an MT model by fine-tuning M2M-100 (418M) on each source data, and evaluated on the remaining test sets of our datasets.

4.3 Prompting of LLMs

We prompted GPT-4O [8] and LLAMA 3.1 8B & 70B (Dubey et al., 2024) to generate translations in either Pidgin or English in both zero-shot or few-shots settings (with one or five examples). A sample prompt is provided in Appendix A. The prompting result is compared with the supervised training of MT models on the MAFAND dataset which is also in the general domain.

5 Experimental Results

5.1 Statistical analysis results

In Table 3, by computing a lexical similarity between the n-gram tokens of each genre, we show that WAPE in both news (BBC) and Wikipedia domains consistently have a *higher* Jaccard similarity score with its parallel English corpus for all n-grams, compared to other datasets with the Naija label. For example, the unigram similarity score for WAPE was around $0.712 - 0.802$ while the others are much lower between 0.133 (Bible) and 0.554 (MAFAND). UD, MAFAND and WARRI WIKI data sets have similar Jaccard similarities.

Furthermore, Levenshtein distance provides an additional evidence of a difference between WAPE and Naija. It takes more than twice edit-distance to transform the English sentences to Naija (WIKI) than to WAPE (BBC) and WAPE (WIKI). Naija (WIKI) requires more edits in characters, which shows that it is farther from English compared to the WAPE. In other words, these two pidgins are quite different than each other linguistically. Similarly, we find longer Levenshtein distance for other datasets: JW300 MAFAND, and BIBLE with 56.5, 58.0 and 88.5 respectively. On the otherhand, UD dataset has a shorter Levenshtein distance compared to others which we attribute to the shorter utterances of the dataset (see Table 1).

Finally, our experiments on text generation metrics (e.g., BLEU, ChrF++ and BERTScore) show that WAPE (BBC) is more similar to English than any of the other Pidgin datasets we evaluated. We find BLEU to be less reliable for this evaluation, achieving only 0.8 for the BIBLE while ChrF achieve relatively higher scores. We attribute this result to several character-level differences between the Bible Pidgin and the English. In general, we find *higher* scores for both WAPE (BBC) and WAPE (WIKI) ($65.2 - 73.4$ ChrF++) than Naija (WIKI) (51.3). BERTScore evaluation also confirmed this finding by reaching to a score of 90.5 for WAPE and 79.8 for Naija.

5.2 Machine translation evaluation

While statistical analysis already proves the linguistic difference between WAPE and Naija, evaluation on a task provides additional perspectives (i.e., Would a model trained on WAPE, perform well on Naija, and vice versa? What is the transfer performance of a model trained on one pidgin to

[8] GPT-4O pre-training data is up to December 2023.

	Single-way (news)			Multi-way parallel (Wiki)					
	WAPE (BBC)			WAPE (Wiki)			Naija (Wiki)		
Evaluation Task	BLEU	ChrF++	BERTScore	BLEU	ChrF++	BERTScore	BLEU	ChrF++	BERTScore
`wape/pcm` → `en`									
0-shot: MAFAND → WARRI	57.6	76.3	**94.5**	**68.6**	83.4	96.2	35.0	59.1	86.7
0-shot: LLAMA 3.1 8B	51.5	74.8	92.1	58.5	79.5	93.6	37.8	64.5	89.0
0-shot: LLAMA 3.1 70B	56.1	76.4	93.5	63.6	81.6	94.3	42.7	67.7	90.7
0-shot: GPT-4O	**59.3**	**78.8**	94.4	65.5	**83.6**	**96.3**	**43.5**	**68.9**	**92.2**
`en` → `wape/pcm`									
0-shot: MAFAND → WARRI	54.7	75.2	91.6	61.0	79.5	92.9	26.5	51.8	83.6
0-shot: LLAMA 3.1 8B	41.3	66.8	88.4	45.0	68.5	89.9	22.9	48.0	81.9
1-shot: LLAMA 3.1 8B	41.0	67.9	87.9	44.0	72.1	87.7	26.1	50.2	82.3
5-shot: LLAMA 3.1 8B	49.2	72.0	90.2	53.4	76.4	91.2	26.4	50.7	83.2
0-shot: LLAMA 3.1 70B	46.1	69.8	89.7	43.8	67.8	89.5	25.4	50.9	83.2
1-shot: LLAMA 3.1 70B	50.6	73.1	90.8	56.0	76.9	92.0	29.0	53.4	84.2
5-shot: LLAMA 3.1 70B	58.1	77.2	92.1	61.5	80.3	93.2	28.0	53.1	84.7
0-shot: GPT-4O	51.8	72.3	91.4	53.8	74.8	92.0	26.7	51.7	83.1
1-shot: GPT-4O	58.7	76.9	92.6	57.7	79.7	92.8	29.6	54.3	84.6
5-shot: GPT-4O	**63.5**	**79.6**	**93.2**	**64.9**	**83.1**	**93.8**	**30.0**	**54.7**	**85.1**

Table 4: **Evaluation on WARRI dataset: single-way and multi-way parallel (same sentences translated to both pidgins) test sets**: We compared the performance of MT to different genres using GPT-4-Turbo and adapted M2M-100 (418M) from MAFAND training set.

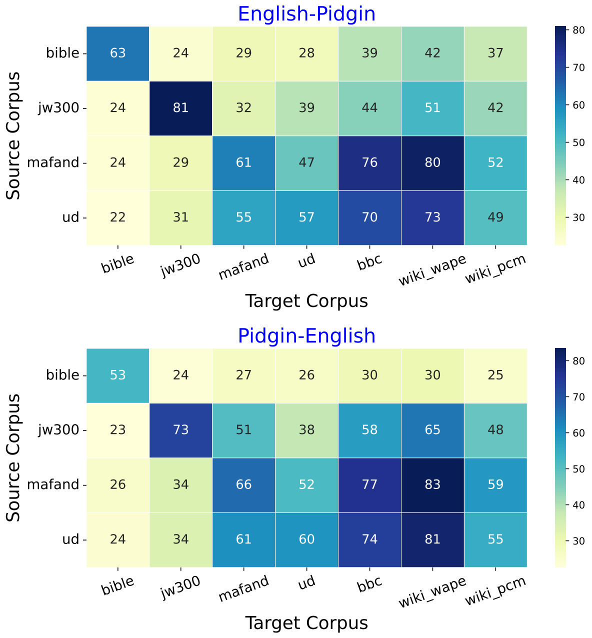

Figure 2: **Cross-corpus transfer results**: Evaluation based on ChrF++

another for text generation tasks?).

Figure 2 shows our results on MT. The transfer performance depends on both the similarity of the domains (e.g. religious vs. news) and similarity of the pidgin style of writing. For example, religious datasets (e.g., BIBLE and JW300) generally transfer poorly to other domains. Similarly, MAFAND and UD in the news and spoken conversation domains also do not transfer well to religious domains.

Both MAFAND and UD datasets often have a higher zero-shot transfer result to WAPE (BBC and WIKI) than to Naija (WIKI). In the English-Pidgin direction, MAFAND achieved between 76 – 80 ChrF++ on WAPE while achieving only 52 ChrF++ on Naija (PCM). Surprisingly, we find MAFAND transfering better to WAPE than to its own test set showing the simplicity of generating WAPE compared to Naija. We have a similar observation when transfering from UD. Our evaluation on the Pidgin-English also confirms this hypothesis that translating from WAPE to English is easier for MT systems than Naija.

5.3 LLM performance on WARRI Benchmark

In this section, we focus on finding out which pidgin is represented in the current LLMs, and whether they support several pidgin variants which are accessible for different communities of Naija speakers. We evaluated the performance of LLM in translating from and into WAPE and Naija.

MAFAND MT model and LLMs represent WAPE more Table 4 shows the result of evaluation of the WARRI MT results. In the direction of **wape/pcm→en**, adapting MAFAND MT model to WAPE gave an impressive result in both single-way (76.3 ChrF++) and multi-way parallel (83.4 ChrF++) scenarios. However, the performance on Naija (WIKI) is much worse (−24.3 drop in ChrF++). This shows that the fine-tuning corpus most likely represents the WAPE. Similar observation was found in GPT-4O and LLAMA 3.1 8B & 70B evaluation, although the performance of the latter was worse especially on Naija. Similarly, for the **en→wape/pcm**, in zero-shot setting, MAFAND MT model gave the best performance over GPT-4O and LLAMA 3.1 8B on WAPE in zero-shot setting, and competitive performance on the Wikipedia genre (51.8 ChrF++) compared to GPT-4O (51.7 ChrF++).

Can we teach LLMs different genres with only a few examples? Our result (Table 4) of prompting GPT-4O, LLAMA 3.1 8B and LLAMA 3.1 70B shows that providing one or five examples is effective for extra performance boost to generate generating Pidgin sentences.[9] For GPT-4O, the performance improved over zero-shot result by +4.9 ChrF++ when the LLM is prompted with one example translation of WAPE, and +8.3 ChrF++ when prompted with five examples, on the multi-way test set. However, the boost in performance is very small when Naija (WIKI) examples are provided. It is only +2.6 and +3.0 when one example and five examples are provided during the prompting of GPT-4O. This shows that GPT-4O is more biased toward the WAPE than Naija and it is difficult to teach the LLM with few examples. The reason for this performance difference is because the WAPE (BBC) is the largest unlabelled data available on the web (Ogueji et al., 2021). Other sources that are more representative of Naija, are often in smaller quantity (e.g., Wikipedia). We observe a similar trend for the LlaMa models where LLAMA 3.1 70B attained up to 80.3 ChrF++ with 5-shots (−2.7 points when compared to GPT-4O), while LLAMA 3.1 8B achieved 76.4.

We provide a qualitative example in Table 5, where we show that with one or five examples, the LLAMA 3.1 70B LLM slightly changes its writing style to be more similar to Naija but sometimes the model combine the vocabulary of WAPE and Naija which leads to misunderstandings. For example, in the 5-shot translation of "...was expected of them to do in their different areas", LLAMA 3.1 70B translated it to be **"suppose do for dia different areas"** which is more similar to the WAPE translation of the same sentence. However, in Naija the words like "suppose", "different" and "area" are spelled differently (e.g., **"sopos du for dia difren aria"**). On the other hand, GPT-4O produced an (almost) accurate translation into Naija except the use of "suppose" rather than "sopos" and "becos" (a WAPE word) instead of "bikos". This implies that the model is able to learn in-context. However, it is still biased towards WAPE without few-shot examples. With more examples, we may be able to teach the model Naija with supervised fine-tuning of the instruction data containing Naija-English parallel sentences. Qualitative examples for WAPE show that the LLMs are able to generate sentences correctly in zero-shot setting without additionally few shot examples which confirms our hypothesis that the LLMs are biased towards WAPE.

6 Qualitative interviews with Naija Wikipedia contributors

To validate our study, we interviewed two native speakers of Naija who contribute to the writing and editing of Naija Wikipedia articles. Some Wikipedian contributors have online public profiles with links to their email addresses and social media accounts (e.g, Twitter or Linkedln). We sent emails to two Naija contributors with the online public profiles and conducted interviews ($\sim$ 1 hour each) with each of them. [10]

Our first observation is that the Naija Wikipedia contributors are not linguists or language experts but they are volunteers without a formal linguistic training. They have a passion for Naija and make an effort toward establishing a writing system which is very similar to the way it is spoken in their community (within Nigeria), rather than targeting a wider West African audience like BBC. They make efforts to create a standardized way of writing during the Naija Wikipedia incubator program. Through these efforts, Naija is included as a separate language on Wikipedia. The volunteers were part of the Wikipedia incubator program from the start, and they are part of the editors team of Naija Wikipedia. This team mentors new contributors about how to write the Naija reflecting the patterns how Naija is spoken (sometimes with a few adjustments to make it readable since original spoken Naija form could be different than English (e.g. "moto" instead of "car")).

The volunteers also mentioned that they consult the available literature (e.g., Ofulue and Esizimetor (2010)) about Naija and follow the recommended rules by Naija linguists (Balogun, 2013b; Aghoghovwia et al., 2010) before starting to contribute to Naija Wikipedia.

In terms of content of the Naija entries on Wikipedia, the contributors focus on the biogra-

[9] The exact 1-shot and 5-shots examples are provided in the Appendix D

[10] We provided honorarium of $11 to each interviewee.

Method	Translation
English	Rimi said he gave them the money because they did all that was expected of them to do in their different areas.
Reference (Naija)	Rimi sey im giv dem di moni bikos dem "folo do evritin wey dem sopos du for dia difren aria.
Reference (WAPE)	Rimi tok say im give dem di money becos dem do evritin wey dem suppose do for dia different areas.
English-Naija translation	
0-shot: GPT-4o	Rimi talk say e give dem the money because dem do everything wey dem suppose do for their different areas.
1-shot: GPT-4o	Rimi tok say im give dem di moni because dem do all di tin wey dem suppose do for dia different area.
5-shot: GPT-4o	Rimi tok se im dash dem di moni becos dem do all di tins wey dem suppose do for dia diffren areas.
0-shot: LLaMA 3.1 70B	Rimi talk say him give dem money because dem do all wetin dem suppose do for dia different area.
1-shot: LLaMA 3.1 70B	Rimi tok say im give dem moni because dem do all wet dem suppose do for dia different areas.
5-shot: LLaMA 3.1 70B	Rimi talk say im dash dem di moni bikos dem do wetin dem suppose do for dia different areas.
English-WAPE translation	
0-shot: GPT-4o	Rimi talk say e give dem di money because dem do all wey dem suppose do for their different areas.
1-shot: GPT-4o	Rimi talk say e give dem the money because dem do all wey dem expect make dem do for their different areas.
5-shot: GPT-4o	Rimi talk say e give dem di money because dem do all wey dem expect dem to do for their different areas.
0-shot: LLaMA 3.1 70B	Rimi tok say him give dem money because dem do all wetin dem suppose do for dia different areas.
1-shot: LLaMA 3.1 70B	Rimi talk say im give dem moni because dem do all wetin dem suppose do for dia different areas.
5-shot: LLaMA 3.1 70B	Rimi talk say im give dem di money because dem do all wet dem expect dem to do for dem different areas.

Table 5: **Qualitative analysis on Predicted translations on WARRI dataset:** multi-way parallel output. Words/phrases expressed in Naija are in violet color, WAPE words are in cyan, while English words that ought to be translated are in red.

phies of notable people (e.g., musicians and actors) in Nigeria. They are not allowed to contribute to sensitive topics (e.g., health) except when it is a direct translation from an high-resource language (e.g., English). To achieve this, they prefer words that come from local Nigerian languages (e.g., Hausa, Igbo and Yoruba), which many Nigerians are familiar with rather than words that are commonly understandable across West Africa. In general, the interviews with the Naija Wikipedia contributors confirm our results that they follow some convention distinguishing them from the WAPE writing convention.

7 Conclusion

Different versions of Pidgins mixed with English and local languages are used in West Africa but not all of them have standardized writing systems. Since 2017, BBC broadcasts (on Internet) in WAPE target the West African countries with the goal of reaching a wider audience across countries in a standardized writing style.

Nigeria is a multilingual country with both richness and challenges that come along with the linguistic diversity. Although the official language is English, Naija is a lingua franca that brings speakers of different Nigerian languages together regardless of their linguistic, social or educational backgrounds. Since 2022, it is also a written language on Wikipedia.

Although both pidgin varieties are used in Nigeria, we prove that WAPE and Naija are different from each other linguistically and current Generative AI models are built upon WAPE only. This is probably due to more availability of data on the Internet for the WAPE rather than Naija.

Lack of data on low resource languages is a key challenge for current AI systems. In our paper, we show that the situation is much more challenging for linguistically rich areas (e.g., West Africa). More specifically, pidgin varieties with the most data on the Internet gets represented on AI systems and the others may not be visible. This could potentially lead to a bias towards favoring language preferences of certain speakers/users instead of being more inclusive toward the users/speakers of other pidgins. Although our analysis focuses on Naija spoken in Nigeria, we hope to extend our analysis to other English-based pidgins in West Africa (e.g., Ghananian Pidgin, Cameronian Pidgin, and Krio in the future) as well.

8 Limitation

There are few limitations of our work (1) Our evaluation dataset is small, although we argue that 500 may be good enough as a test set for MT. However, we only have a maximum of 5 sentences we could use for the few-shot learning or in-context learning. Moreover, with additional sentences (e.g. 2.5K-5K parallel sentences as recommended in (Adelani et al., 2022)), we may be able to adapt M2M-100 model to produce better generation of the Wikipedia genre. (2) Our analysis is limited to one task which is machine translation, we hope to extend this analysis to other tasks in the future as well.

Acknowledgment

We are grateful to Sergiu Nisioi and John Pavlopoulos for their time and feedback on an early draft of the paper. We thank OpenAI for granting the API credits (through their Researcher Access API program to Masakhane) enabling the evaluation of GPT-4 LLMs. We thank the volunteer Wikipedia contributors for Naija for their insights and explanations, anonymous reviewers, area chairs and senior area chairs of NAACL25 for their feedback and recommendations on our paper.

References

David Adelani, Jesujoba 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 Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani. 2022. Natural language processing for african languages.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Ogechi Florence Agbo and Ingo Plag. 2020. The relationship of nigerian english and nigerian pidgin in nigeria: Evidence from copula constructions in ice-nigeria. *Journal of Language Contact*, 13(2):351–388.

Philip O. Aghoghovwia, C. Ailende Ativie, Rose Aziza, Harrie Bazunu, Bernard Caron, Francis Egbokhare, O. Ndidiamaka Ejechi, David Oshorenoya Esizimetor, Rudolf Gaudio, Rita Mebitaghan, Macaulay Mowarin, Christine Ofulue, Albert Okelome, Sony Okpeadua Okpeadua, and Mabel I. Osakwe. 2010. Minutes from meeting of naija languej akedemi. In *IFRA Nigeria*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Akinmade T. Akande and Oladipo Salami, editors. 2021. *Current Trends in Nigerian Pidgin English*. De Gruyter Mouton, Berlin, Boston.

Akinmade Timothy Akande and L. Salami. 2012. Use and attitudes towards nigerian pidgin english among nigerian university students.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Tae Young Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *ArXiv*, abs/2304.09919.

Temitope Abiodun Balogun. 2013a. In defense of nigerian pidgin. *Journal of languages and culture*, 4(5):90–98.

Temitope Abiodun Balogun. 2013b. In defense of nigerian pidgin. *Journal of Languages and Culture*, 4:90–98.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8:243–257.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24, Paris, France. Association for Computational Linguistics.

Steve Crowdy. 1993. Spoken corpus design. *Literary and Linguistic Computing*, 8:259–265.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline C. Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2019. Ethnologue: Languages of the world(22nd edn.). dallas, tx: Sil international. *Online version: http://www. ethnologue. com [08.08. 2020]*.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Richard Fekete, Esther Ploeger, Li Zhou, Hans Erik Heje, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loic Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Sogaard, and Johannes Bjerva. 2023. Creoleval: Multilingual multitask benchmarks for creoles. *ArXiv*, abs/2310.19567.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves nlp performance for nigerian pidgin. *Preprint*, arXiv:2404.18264.

Marian Marchal, Merel Scholman, and Vera Demberg. 2021. Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 84–94, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

S. Sangol Mufwene. 2024. Pidgin. *Encyclopedia Britannica*.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and

Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Pieter Muysken, Norval Smith, et al. 1995. The study of pidgin and creole languages. *Pidgins and creoles: An introduction*, 1:14.

Christine I. Ofulue and David O. Esizimetor. 2010. Guide to standard naijá orthography. an nla harmonized writing system for naijá publications. In *IFRA Nigeria*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Folajimi Oyebola and Kingsley Ugwuanyi. 2023. Attitudes of nigerians towards bbc pidgin: A preliminary study. *Language Matters*, 54(1):78–101.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Kofi Yakpo. 2024. West african pidgin: World language against the grain. *Africa Spectrum*, 59(2):180–203.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompt Template

B MAFAND training

We fine-tune MAFAND dataset on M2M-100 (418M) using the same hyparameters stated in Adelani et al. (2022) i.e. number of training epochs of 10, batch size of 32, source and target maximum sequence length of 200, and beam size of 10.

C Licence of WARRI

We plan to release it publicly under the CC-4.0-NC due to BBC portion of the dataset that cannot be for commercial use. However, WARRI (multi-way) has a licence of CC-4.0 international.

D Five shot examples provided

	Prompt
Task Description	You are a helpful assistant who is an expert in translating English sentences to Pidgin using two varieties: West African Pidgin English and Nigerian-Pidgin, I would provide you with five examples of the different varieties, your task is to follow the style of the writing of the specified variety when translating the sentences.
Example	**Example 1:** **English**: Innocent Ujah Idibia was born on 18 September 1975, that is well known as 2baba, a Nigerian singer, songwriter, producer, philantropist. **West African Pidgin English**: Innocent Ujah Idibia wey dem born for 18 September 1975, wey dem know as 2baba, be a Nigerian singer, songwriter, producer, philantropist. **Nigerian-Pidgin**: Innocent Ujah Idibia (dem bon am for 18 September 1975), wey pipul no wel wel as 2baba, na Naija singa, songraita an podusa an im sabi dash pipul moni an gift wel wel.
Example	**Example 2:** **English**: He was born in Jos, Nigeria **West African Pidgin English**: Dem born am for Jos, Nigeria **Nigerian-Pidgin**: Dem bon am for Jos for inside Naija.
Example	**Example 3:** **English**: He is from the Idoma ethnic group **West African Pidgin English**: Im be from di Idoma ethnic group **Naija**: Im na Idoma pesin.
Example	**Example 4:** **English**: Idoma is in the southern part of Nigeria **West African Pidgin English**: Na southern part of Nigeria Idoma dey **Nigerian-Pidgin**: Idoma dey for di south side for Naija.
Example	**Example 5:** **English**: Before July 2014, he used 2face Idibia as his stage name **West African Pidgin English**: Before July 2014, i dey use 2face Idibia as im stage name **Nigerian-Pidgin**: Bifor July 2014 na 2face Idibia bi di nem wey im dey yuz for stej.
Prompt	’Translate this sentence to Nigerian Pidgin
Input	Alexander Abolore Adegbola Akande was born on 17 January 1980, well known as 9ice, a Nigerian singer, dancer, and songwriter.
Output:	Alexander Abolore Adegbola Akande (dem bon am for 17 January 1980), wey pipul sabi well well as 9ice, na Naija singa, dansa, an songraita.

Table 6: **Prompt template used for MT**. An example prediction by GPT-4o