Prompting Towards Alleviating Code-Switched Data Scarcity in Under-Resourced Languages with GPT as a Pivot

Michelle Terblanche, Kayode Olaleye, Vukosi Marivate
Data Science for Social Impact
Dept. of Computer Science
University of Pretoria
South Africa
michelle.terblanche@gmail.com,
kayode.olaleye@cs.up.ac.za,
vukosi.marivate@cs.up.ac.za

Abstract

Many multilingual communities, including numerous in Africa, frequently engage in code-switching during conversations. This behaviour stresses the need for natural language processing technologies adept at processing code-switched text. However, data scarcity, particularly in African languages, poses a significant challenge, as many are low-resourced and under-represented. In this study, we prompted GPT 3.5 to generate Afrikaans–English and Yoruba–English code-switched sentences, enhancing diversity using topic-keyword pairs, linguistic guidelines, and few-shot examples. Our findings indicate that the quality of generated sentences for languages using non-Latin scripts, like Yoruba, is considerably lower when compared with the high Afrikaans–English success rate. There is therefore a notable opportunity to refine prompting guidelines to yield sentences suitable for the fine-tuning of language models. We propose a framework for augmenting the diversity of synthetically generated code-switched data using GPT and propose leveraging this technology to mitigate data scarcity in low-resourced languages, underscoring the essential role of native speakers in this process.

Keywords: code-switch, LLM, few-shot, prompting

1. Introduction

Multilingual communities, exemplified well by various African countries, often engage in code-switching, where two or more languages are used within a single discourse (Poplack, 2001a). This language practice highlights the need to develop more advanced natural language processing (NLP) technologies that can smoothly process and produce code-switched sentences. This will move the needle towards equitable representation of the world's under-resourced languages, ensuring that everyone has equal access to these technologies (Solorio, 2021).

There are numerous challenges in code-switching research. The main three are highlighted by Doğruöz et al. (2021) as follows: i) data, which is related to quantity, quality and availability; ii) evaluation, which refers to benchmarks and metrics; and iii) challenges related to end-to-end applications, particularly the ability to process and produce code-switched data.

The focus of this paper is on the first challenge regarding ***data***. While code-switching frequently occurs in written forms, due to the ubiquitous use of social media platforms, leveraging this data in NLP applications for code-switching presents many challenges. These platforms, with their extensive and diverse linguistic expressions, can be invaluable in gathering code-switched data. Yet, the practical utility of such data is hindered by various factors, including the informal, inconsistent nature of online language (Çetinoğlu et al., 2016). It is common to use acronyms, emojis and make spelling mistakes which affect quality and usability of such data (Srivastava et al., 2019). Furthermore the diversity of such data is limited to a specific type of language use (Winata et al., 2022).

To address the shortage of available data, efforts have been made to create synthetic code-switched data using different methods: from using parallel corpora with linguistic constraints on where a switch can occur (Pratapa et al., 2018; Rizvi et al., 2021) to employing transformer-based models to generate diverse sentences that adhere to lexical and syntactic rules (Riktika et al., 2022). A more recent study evaluated prompting of large language models (LLMs) to generate code-switched data for South East Asian languages (Yong et al., 2023). They explored a few prompting templates with a limited number of topics in a zero-shot manner and cautioned against the use of synthetically generated data without involving native speakers of the language.

In this paper, we build on the work of (Yong et al., 2023) to address the question about *GPT's* ability to generate code-switched data. Our work overlaps in that we also use an LLM, *OpenAI's GPT*, and various topics in the prompts. We increase the number of topics and provide topic-related keywords in an

effort to increase diversity and reduce the model's propensity to default to certain words. Our goal is not to evaluate various prompting templates, however, we add linguistic guidelines in the prompts to further increase diversity. We propose this as an approach towards language agnostic prompting. We also test the performance of GPT 3.5 with few-shot in-context examples. We specifically consider whether *GPT* can support the generation of larger code-switched datasets and to what extent.

Our contributions are as follows: (i) we provide a framework to increase the diversity of synthetically generated code-switched data by prompting *OpenAI's GPT*; and (ii) we position GPT as a pivot to address code-switched data scarcity in low-resource languages while emphasising the need for native speakers in the loop.

Increasing data availability is at the center of developing language models that serve multilingual communities. Our work is a step towards closing the gap in low-resourced and under-represented languages.

2. Related Work

2.1. Code-Switching Research

Various types of code-switching have been identified but the type that attracts the most academic research is intra-sentential code-switching which can occur anywhere within a sentence boundary (Poplack, 1980) and as a result, adds complexity in evaluation (Poplack, 2001b). Another complex type is intra-word code-switching where the stem of one language is bound to another language (Çetinoğlu et al., 2016; Van der Westhuizen and Niesler, 2018).

Over and above the issue of data diversity (Winata et al., 2022), one of the major challenges in code-switching studies is related to data availability (Doğruöz et al., 2021). A survey by (Winata et al., 2022) showed that up until October 2022, a relatively small amount of papers (ACL Anthology, 2023 and ISCA Proceedings, 2023) focused on code-switching research in African languages with very few publicly available datasets. Eleven publications mention South African languages. The non-English South African languages referenced are isiZulu, isiXhosa, Setswana, Sesotho and Afrikaans. Only one proceeding includes Afrikaans code-switching (Niesler and De Wet, 2008) with no published dataset. A paper by Van der Westhuizen and Niesler (2018) introduced the first corpus on isiZulu, isiXhosa, Setswana, Sesotho curated from transcribed soap opera speech data and eight of the papers makes use of this dataset and is mainly focused on automatic speech recognition (ASR) systems.

Code-switching in Kiswahili–English is studied in two papers but no datasets were made available (Otundo and Grice, 2022; Piergallini et al., 2016). In addition to a survey by Winata et al. (2022), one other paper was found that addresses Sepedi–English code-switching. Modipa et al. (2013) develop a corpus from a set of radio broadcasts to evaluate the implication of code-switching in ASR systems. This dataset is publicly available. This brief review of the state of code-switching research in an African context motivates our work to develop methods for addressing data scarcity.

A predominant approach to mitigating data availability issues involves augmenting existing datasets through the generation of synthetic code-switched data. Some of the methods to augment the earlier mentioned South African speech corpus include the use of word embeddings to synthesise code-switched bigrams to find similar words in the sparse training data (Westhuizen and Niesler, 2017). Biswas et al. (2018) evaluated adding out-of-domain monolingual data and synthesised code-switched data using an LSTM to augment the dataset.

For non-African languages, Rizvi et al. (2021) developed a toolkit that generates multiple code-switched sentences using either the Equivalence Constraint or the Matrix Language Frame. The limitations are that it relies on a good sentence aligner and parser and parallel translated sentences as input. The notion is that this approach should work on any language pair. Winata et al. (2019) implemented a sequence-to-sequence model for English-Mandarin code-switched data. Although the model does not require external knowledge regarding word alignments, it still relies on an existing English–Mandarin code-switched dataset and parallel corpora. The work of (Liu et al., 2020) introduced an attention-informed zero-shot adaptation method that relies on a limited number of parallel word pairs. The languages covered are German, Italian, Spanish and Thai, the latter two for natural language understanding. The shortcoming of the above-mentioned approaches is the diversity of data. Most existing code-switched datasets were collected from social media platforms such as Twitter and therefore limits the type of code-switching (Doğruöz et al., 2021).

To this issue, Riktika et al. (2022) developed an encoder-decoder translation model for controlled code-switched generation. It uses monolingual Hindi and a publicly available Hindi–English code-switched dataset as input to generate data that is faithful to syntactic and lexical attributes.

Yong et al. (2023) proposed an approach that is independent of existing code-switched datasets or parallel corpora through prompting of LLMs. Their objective was to test whether multilingual LLMs

can generate code-switched text through prompting. They evaluated a variety of prompt templates and found that those explicitly defining code-switching gave the highest success rate. However, they also highlighted the sentences often contained word-choice errors and semantic inaccuracies which was more prevalent in the languages that don't use the English alphabet and Latin script. They limited the scope to five topics and did not include diversity as a measure. Their findings were that GPT's capability to generate code-switched data is superior to other LLMs, however, using this method without humans-in-the-loop is not advised.

Jha et al. (2023) elaborated on LLMs such as GPT being prone to hallucinations where it provides factually inaccurate or contextually inappropriate responses. A solution to address this is to ensure carefully curated prompts. Furthermore, to avoid encoded biases, Bender et al. (2021) emphasises the need to also evaluate appropriateness in relation to a particular social context.

With the rapid adoption of LLMs in everyday life, these are a low-cost alternative to alleviate data scarcity in low-resourced and under-represented languages by synthetically generating text. In this paper we expand on the work of Yong et al. (2023) and position GPT as a pivot in generating code-switched data rather than a self-sufficient solution.

3. Code-Switched Text Generation via GPT-3.5 Prompting

Our prompt-based approach to code-switched (CS) text generation is heavily inspired by the work of Yong et al. (2023), who collected synthetic CS data by prompting LLMs with requests along languages and topics. Their focus was on code-switching English with South-East Asian languages. In our case, we focus on two under-explored and under-resourced code-switching scenarios: Afrikaans–English and Yoruba–English. Although Afrikaans and English are typologically dissimilar (van Dulm, 2007), they are both West Germanic languages and generating CS text should be easier. Yoruba is a tonal language and even more dissimilar to English which could provide challenges when creating synthetic CS data. We extend the limited topics covered in Yong et al. (2023) and present GPT-3.5 not as an autonomous solution to CS data scarcity, but as a potential tool for supporting CS data curation efforts for under-resourced African languages. We specifically use GPT-3.5, firstly as a baseline to compare with the findings from Yong et al. (2023) and secondly, due to the unavailability of the GPT-4 API at the time of our experiments[1].

[1]The API for GPT 4 was made available after we finished the majority of the experiments, https://openai.com/blog/gpt-4-api-general-availability

3.1. Prompting for Afrikaans–English CS Sentences

Building on the prompt template from Yong et al. (2023), which uses topics as guidelines, our approach extends this by (i) incorporating specific code-switching words related to each topic within the prompt and (ii) evaluating the effect of prompt complexity from basic (Section 3.1.1) to more comprehensive prompts (Section 3.1.2).

We curate a non-exhaustive list of common conversation topics and associate typical English words from native speakers of Afrikaans and from available online platforms. We cover 22 topics with a total of 355 keywords. For this paper we generate one sentence per keyword for the various prompts. We also develop a general list of words used in code-switching that is not directly linked to a specific topic consisting of 138 words. ∼90% of the keywords are nouns, verbs and adjectives which is in line with the notion that switching is more likely to occur on these open word classes as opposed to close word classes (such as pronouns and conjunctions) (Kodali et al., 2022).

3.1.1. Topic-Keyword Basic Prompting

In the six different prompting templates of Yong et al. (2023), one prompt specifically requests a native speaker to give a mixed sentence. This is an indirect way to impose a matrix language (ML). We explicitly include the use of a matrix language in our prompts (Jake et al., 2002). This is to ensure that we adequately represent the low-resourced language. However, we recognise that grammatical constraints on CS is an open research question with varying definitions of acceptability that evolves over time (Bhat et al., 2016).

The following shows the basic prompt we used (Prompt 1.1) and a few examples to highlight the behaviour of GPT-3.5 (English translation in *Italics*).

Prompt 1.1: Generate an Afrikaans-English code-switch sentence with Afrikaans as the matrix language. Typical words used in code-switching are: **general**[2]. The topic is *[insert topic]* and must contain the word *[insert keyword]*.

Topic: education and training; **keyword**: skills

Example 1: Ek_{af} moet_{af} my_{af} skills_{en} verbeter_{af} om_{af} 'n_{af} beter_{af} $\text{werksgeleentheid}_{af}$ te_{af} kry_{af}.
I must improve my skills to get a better job opportunity.

Topic: general conversation; **keyword**: try

[2]List of general words provided

Example 2: Ek$_{af}$ sal$_{af}$ probeer$_{af}$ to$_{en}$ finish$_{en}$ my$_{af}$ assignment$_{en}$ op$_{af}$ tyd$_{af}$.
I will try to finish my assignment on time.

The matrix language is Afrikaans in Example 1 and English in Example 2. We see from these examples that GPT 3.5 does not necessarily follow the prompt with regards to the matrix language.

We do not evaluate word-level language identification therefore we do not explicitly measure adherence to the matrix language prompt in this paper.

The results of the generated sentences therefore indicate that GPT 3.5 is capable of generating some coherent sentences and can be corrected where the grammatical structure follows English. Section 4.3 gives a more detailed analysis of code-switch acceptability.

A key observation from using this basic prompt for generating Afrikaans–English sentences is that sentences are one-dimensional with ∼80% of sentences starting with a singular personal pronoun: **'Ek'** (*English: 'I'*) (Section 4.2.1). This creates the opportunity to explore ways of adding diversity to the type of sentence through the use of basic linguistic guidelines (such as specifying pronouns) which is discussed in the following section.

3.1.2. Linguistic-Based Prompting

Since the word lists contain nouns, verbs and adjectives related to specific topics, content diversity in the sentences is addressed. These are also words that are most typically code-switched (Kodali et al., 2022). To add further diversity in the type of sentence, we add basic linguistic guidelines in the form of varying pronouns (personal, impersonal, interrogative etc.), tenses (past, present and future that alters the verb) and using negative particles. The inclusion of negative particles is randomly initialised and not in each prompt. We also impose a rule that conjunctions must be in the matrix language since conjunctions are part of closed word classes and should less likely be switched.

Prompt 2.1 is an example of a prompt using linguistic guidelines following with an example of the generated sentence (English translation in *Italics*). In Example 3 the prompts are adhered to, however, the conjunctions 'but' and 'and' are in English therefore note adhering to the guideline. Our preliminary observation is that the prompting approach can support the generation of CS sentences that are diverse. The effect of varying pronouns on sentence diversity is further evaluated in Section 4.1. Word order structure mimics that of natural speech and can be corrected where needed. We give additional examples and an evaluation of the quality of the sentences in Section 4.3.

Prompt 2.1: Generate an Afrikaans-English code-switch sentence with Afrikaans as the matrix language. Typical words used in code-switching are: **general**. The topic is *[insert topic]* and must contain the word *[keyword]*. Start the sentence with *[insert pronoun]* using the *[insert tense]*. *A conjunction must be Afrikaans. [Use a negative particle]*.

Topic: physical health and fitness; **keyword**: race; **Pronoun**: impersonal; **Tense**: past; **Use a negative particle**: No

Example 3: Dit$_{af}$ was$_{af}$ super$_{en}$ lekker$_{af}$ om$_{af}$ die$_{af}$ race$_{en}$ te$_{af}$ hardloop$_{af}$, but$_{en}$ ek$_{af}$ ignore$_{en}$ die$_{af}$ consequences$_{en}$ and$_{en}$ het$_{af}$ te$_{af}$ veel$_{af}$ geëet$_{af}$ afterwards$_{en}$.
It was super nice to run the race, but I ignore the consequences and ate too much afterwards.

3.1.3. Few-Shot Prompting

In the work from (Yong et al., 2023) they did not evaluate the effect of few-shot examples. We therefore evaluate two additional prompts: Prompt 1.2 and Prompt 2.2 where we add five examples of code-switched sentences to Prompts 1.1 and 2.1 respectively. These are general examples and not in the context of the topic.

3.2. Prompting for Yoruba–English CS Sentences

In this section we apply the same methodology (Section 3.1) used to generate Afrikaans–English CS sentences to generate Yoruba–English CS sentences and provide brief observations. We develop similar topic keyword lists for Yoruba with most words overlapping with those developed for Afrikaans–English. In future work we will focus on developing common lists that cover a more diverse set of languages. The following are a few examples of the generated Yoruba–English sentences:

Topic: information technology; **keyword**: spreadsheet; **Pronoun**: indefinite; **Tense**: future; **Use negative particle**: Yes

Example 1: Mo$_{yo}$ ni$_{yo}$ ko$_{yo}$ relax$_{en}$, infact$_{en}$ mo$_{yo}$ gba$_{yo}$ surprise$_{en}$ pe$_{yo}$ spreadsheet$_{en}$ je$_{yo}$ Yoruba$_{yo}$ word$_{en}$.
I said you should relax, infact I accept the surprise that spreadsheet is a Yoruba word.

Topic: social media; **keyword**: cope; **Pronoun**: indefinite; **Tense**: present; **Use negative particle**: Yes

Example 2: Kòyo síyo èèyànyo tóyo yànyo ònàyo níyo wáhálàyo, viewen yìíyo niyo awọnyo ẹ̀dáyo tíyo wọ̀nyo ṣeyo làtiyo copeyo.
There is no person that chooses problems as a path, this view is what the creatures XXX did to cope

Examples 1 and 2 both follow the prompt guidelines with respect to the matrix language and tense. Example 1, however, uses a personal pronoun instead of an indefinite pronoun with Example 2 using the correct pronoun. XXX in Example 2 indicates a phrase that cannot be translated.

We observe that the prompting approach can also support the generation of Yoruba–English sentences that are diverse.

We provide observations on the coherence and naturalness of synthetic sentences in Section 4.4.

4. Evaluation of Generated Data

In this section, we evaluate our work in three parts: (i) we evaluate the diversity of the generated sentences, (ii) we comment on GPT 3.5's adherence to the prompts provided, and (iii) we evaluate the quality of the sentences generated through a combination of statistical analysis and human evaluation of the sentences. We use the four prompt guidelines as discussed in Section 3. For this paper we Romanised the Yoruba–English sentences for easier evaluation, however, we will include this in future work.

4.1. Data Diversity

4.1.1. Content Diversity

In Figure 1a (from Prompt 1.1) we see a large amount of general words being used compared with the number of sentences. We also note that the top three keywords (*amazing, acknowledge, anyway*) is the same as the top three keywords in the alphabetised list. In Prompt 2.1 we provide a randomised general word list to GPT 3.5 and in Figure 1b we observe a more even distribution of general words as a result. This indicates GPT 3.5's sensitivity to prompts and the context provided.

4.1.2. Linguistic Diversity

Since Prompts 2.1 and 2.2 asked "start the sentence with...", all sentences were evaluated accordingly. We used a list of common Afrikaans and Yoruba pronouns to evaluate this prompt.

From Figure 3 we observe an increase in diversity of the types of sentences with regards to the distribution of pronouns (Prompts 2.1/2.2). For Afrikaans–English, more than 90% of the sentences start with one of the specified pronouns. We also see an increase in the diversity of Yoruba–English sentences, however, there are still ∼35% of sentences starting with words other than the requested pronouns. It is not well understood why GPT 3.5 ignored these prompts. In the absence of linguistic guidelines in the prompt, we note that by adding few-shot examples, we lack diversity (Prompts 1.2 and 2.2).

Similarly to pronouns, we use Afrikaans and Yoruba keywords that indicate past and future tense, negation (negative sentiment) and conjunctions to evaluate the effect of adding these guidelines to the prompts. In Table 1 we highlight the impact of these factors on distribution in sentences using Prompts 1.1 and 2.1 (prompts without example sentences).

	Afrikaans		**Yoruba**	
Prompt	**1.1**	**2.1**	**1.1**	**2.1**
Past Tense	42%	34%	17%	23%
Future Tense	55%	39%	10%	12%
Negation	26%	39%	15%	27%
Conjunction*	14	4	1	2

*The ratio of Afrikaans/Yoruba to English.

Table 1: Distribution of tenses and negation and ratio of conjunctions.

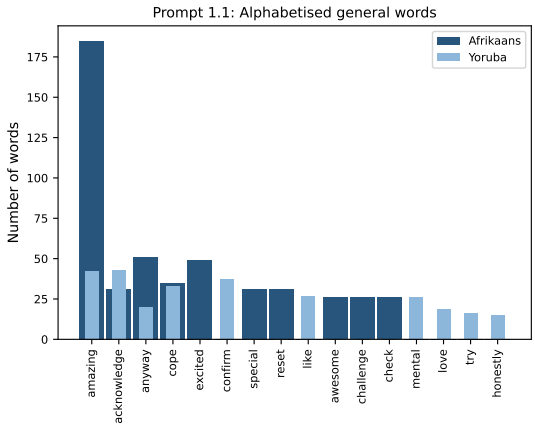
We see in Table 1 that for Afrikaans–English, both the distribution of tenses (equal distribution between past and future) and the presence of negation improved. However, it is only negation that improved for Yoruba–English. We further elaborate on this observation in Section 4.2.1. The ratio of Afrikaans:English conjunctions decreased showing the guideline is not efficient. For Yoruba:English conjunctions we observe a slight improvement.

The above statistical evaluation of diversity shows that adding various linguistic guidelines to the prompts improves diversity. However, this does not consider whether a prompt is adhered to. In the next section, we evaluate GPT 3.5's ability to execute prompts.

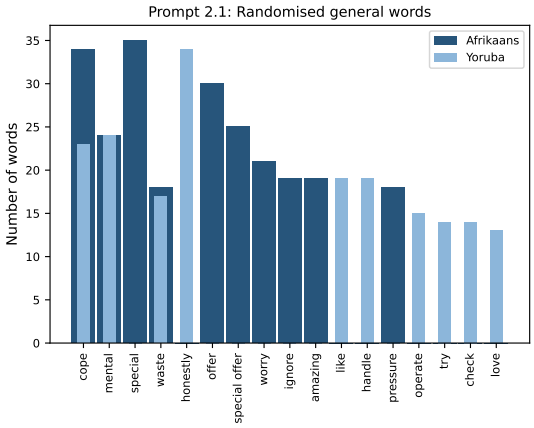
4.2. Prompt Adherence

In Section 3.1 we already observed that GPT 3.5 does not always adhere to using the specified matrix language and since we do not consider word-level language identification in this paper, we exclude this when determining adherence.

We apply a simple approach to calculate prompt adherence. We express the number of prompts adhered to as a percentage of the total prompts given. In Prompt 1.1, the only prompt given is the topic keyword hence a total of one prompt (the same for Prompt 1.2). In Prompt 2.1, there are five prompts given: topic keyword, pronoun, tense,



(a) Distribution of general words (alphabetised).



(b) Distribution of general words (randomised).

Figure 1: Distribution of top 10 general CS words across all topics.

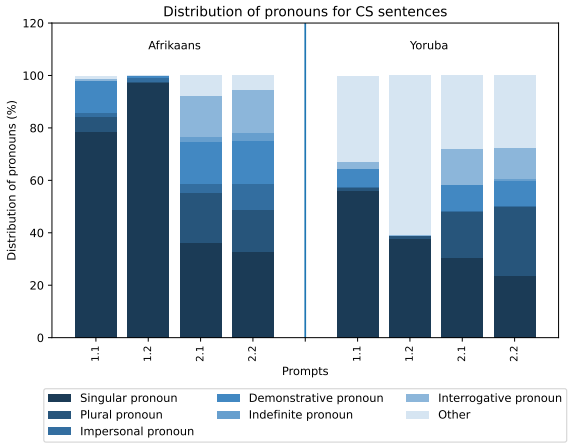


Figure 2: Distribution of pronouns.

Figure 3: Distribution of the use of pronouns at the beginning of a generated sentence.

negative particle and conjunction. The average prompt adherence across the sentences is then used to represent overall prompt adherence.

4.2.1. Statistical Evaluation of Prompt Adherence

In this section we present the prompt adherence for the four prompt guidelines. Keywords for pronouns, tenses, negative particles and conjunctions as per Section 4.2.1. Table 2 shows the overall prompt adherence.

Prompt	1.1	1.2	2.1	2.2
Afrikaans	83%	90%	74%	78%
Yoruba	83%	92%	53%	58%

Table 2: Overall prompt adherence.

From Table 2 we see that the adherence to prompts for Yoruba–English is much lower than for Afrikaans–English in the linguistically guided prompts (Prompts 2.1 and 2.2).

In Afrikaans there are a few specific keywords such as **'nie', 'nooit', 'nee'** *(English: not, never, no)* that indicate negation. Similarly for tenses, words like **'was', 'gister', 'wil', 'more'** *(English: was, yesterday, will, tomorrow)* can be used for past and future tense. However, the Yoruba language is more complex and keywords like the above-mentioned are not adequate to identify negation and tenses, hence the lower prompt adherence.

In the next section (Section 4.3) we use manual annotation of sentences for tenses and negation to re-evaluate prompt adherence.

4.2.2. Manual Evaluation of Prompt Adherence

For manual evaluation of generated sentences, we sample 100 sentences each from the four prompt methods.

We manually annotate the sentences of Prompts 2.1 and 2.2 with tense (past or future) and negation (whether the sentence expresses some negative sentiment). In future work, external annotators will also be used.

In Table 3 we show the impact on the calculated prompt adherence (using Prompt 2.1) for the statistical (1) and manual (2) evaluation of the 100 sentences. The prompt adherence for Yoruba–English increased to 66% from 59% with a significant increase in the adherences to tenses.Afrikaans–English prompt adherence remains constant. The adherence to negation reduced slightly for both languages. This confirms the earlier comment that it is statistically more difficult to calculate prompt adherence for Yoruba–English without a human in the loop.

Prompt	Afrikaans (1)	Afrikaans (2)	Yoruba (1)	Yoruba (2)
Tense	79%	84%	41%	72%
Negation	47%	41%	40%	36%
Total	72%	72%	59%	66%

Table 3: Comparing prompt adherence for both a statistical and manual annotation perspective.

We conclude that there is potential in using GPT 3.5 as a supporting tool to generate diverse sentences with linguistically guided prompts. In the following sections we provide an overview of the quality of generated sentences to further determine the role that GPT 3.5 can play in addressing code-switched data availability.

4.3. Code-Switch Acceptability

The final part of our analysis looks at the quality of generated sentences. As mentioned in Section 4.3, we sampled 100 sentences from each of the four prompt methods. For this part of the analysis, we rated the acceptability of a code-switch sentence according to: i) Yes, ii) Yes, with minimal changes or iii) No. We adopt the constraint-free approach of MacSwan (2000).

The results of the manual annotation are shown in Figure 4. We observe that the acceptability of Afrikaans–English sentences far outweighs that of Yoruba–English. We also see that adding few-shot examples increases acceptability (Prompts 1.2 and 2.2). Although we observe an increase in diversity through linguistic guidelines, the quality of sentences are sub-optimal. Subsequent work will focus on how correctable sentences can be used for improved prompting and/or fine tuning of language models. However, with further analysis and improvement, there is potential to use GPT 3.5 to support synthetic data generation.

4.4. Language Specific Observations

4.4.1. Afrikaans–English

In order to quantify the acceptability observed from internal evaluation, we randomly select 5 Afrikaans-English sentences from the dataset used for manual evaluation (Section 4.3). Table 4 gives the sentences with translations and comments.

In our general overview we find that the typical mistakes made are as a result of following English grammar structure. However, for many sentences this does not affect the meaning and can be corrected.

The results from the various experiments therefore indicate that using GPT 3.5 (and it's followers)

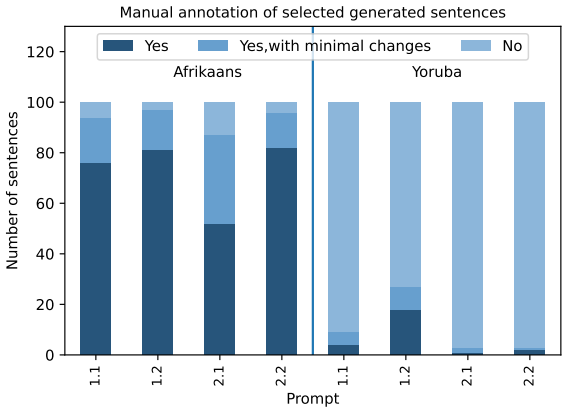


Figure 4: Evaluation of manual annotation of sentences.

can be considered as a method to generate large-scale data in Afrikaans-English code-switching.

4.4.2. Yoruba–English

Similarly to quantifying the Afrikaans-English sentences, we give 5 randomly selected Yoruba-English sentences in Table 5 from the dataset used for manual evaluation (Section 4.3).

It is hypothesised that the exposure of GPT 3.5 to the Yoruba language is to a much lesser extent than Afrikaans yielding a substantial amount of unacceptable sentences. Furthermore, as was postulated by Yong et al. (2023), languages using the English alphabet and Latin script perform better on LLMs. Further analysis is required to improve prompting and quality of sentences.

5. Conclusions and Future Work

In this paper we extended on the of Yong et al. (2023) where they used prompting of LLMs (including GPT 3.5) to generate code-switch sentences. Our approach evaluates three dimensions: (i) diversity, through a wider range of topics, keywords, linguistic guidelines and few-shot examples; (ii) prompt adherence, to understand the ability of GPT 3.5 to follow these prompts; and (iii) quality, to determine the use of GPT 3.5 as a supporting tool to address code-switched data scarcity. We evaluated two typologically diverse language pairs: Afrikaans–English and Yoruba–English.

Our main findings are: (i) using topics, keywords and general context words increases coverage; (ii) linguistic-based guidelines increases diversity in the types of sentences, (iii) few-shot prompting increases the quality of sentences but is limited in diversity of the types of sentences;(iv) quality of sentences are much lower for languages that use non-Latin script (such as Yoruba); and (v) evalu-

	Sentence	Accept	Comments
1	Ek is so excited om my nuwe partner te ontmoet. *(I am so excited to meet my new partner.)*	✓	-
2	Ons moet takeaways hê for dinner, maar ek wil nie weer McDonald's eet nie. *(We must have takeaways for dinner, but I don't want to eat McDonald's again.)*	✓	The use of English 'for' instead of Afrikaans 'vir' is less typical but can be accepted
3	Ons het 'n nuwe app gedownload om die fotos te organise. *(We downloaded a new app to organise the photos.)*	✓	'gedownload' is an example of intra-word code-switching
4	Ek moet 'n nuwe uitdaging in my loopbaan aanpak. *(I have to tackle a new challenge in my career.)*	✗	No code-switching, only Afrikaans
5	Daai kursus was 'n disaster, ons het reset van die begin af. *(That course was a disaster, we reset from the beginning.)*	✗	Unclear about the intended meaning with the use of 'reset', however, can be corrected in context

Table 4: Generated Afrikaans–English sentences, translations and comments on acceptability.

	Sentence	Accept	Comments
1	o ma install software yii ni computer mi. *(You will install this software in my computer.)*	✓	The model is not clear about the right orthography for the Yoruba words in the sentence and used the word "ni" instead of 'sorii' which translates to 'on' in Yoruba
2	60 million naira yen fe po die fun mi. I need to buy orange juice for the party. *(That 60 million naira seems to be a bit too much for me. I need to buy orange juice for the party.)*	✓	This is an inter-sentential code-switched sentence. However, this can be accepted by just dropping the second sentence
3	Mo n gbadun ojo meta ti n si se fun mi ni lockdown ni ojo kan, but honestly, e wa wo mi, I don tire. The pressure don too much, and I just dey try survive. *(I am enjoying the three days XXX during lockdown in one day, but honestly, come and see me, I am tired.)*	✗	These sentences make no sense. Contains the Nigerian version of Pidgin-English mixed with Yoruba and English. The 'XXX' indicates phrases that cannot be translated
4	eniyan miran naa maa click si awon idile mi lati ba wa. *(That other person will click to my family to come with.)*	✗	This sentence makes no sense
5	o ma jabo ile-ise yi niwaju wireless connectivity yi. *(You will XXX this company in front of this wireless connectivity.)*	✗	The English translation for the word 'jabo' cannot be inferred without knowing the diacritics. The sentence makes no sense

Table 5: Generated Yoruba–English sentences, translations and comments on acceptability.

ating quality of data requires a human-in-the-loop. We provide a framework for linguistically-guided prompting and we conclude that *OpenAI's GPT* exhibits the ability to support synthetic code-switched data generation and can be invaluable to address the issue of data availability.

In future work we will address the following: i) include external annotation to cross-validate the quality of generated sentences; (ii) improve on the prompting guidelines to increase quality; (iii) use correctable sentences to improve the performance of the latest generation of *OpenAI's GPT* to support large-scale generation; and (iv) expand to more African languages in an effort to develop a language agnostic approach to synthetically generate data.

6. Ethical Considerations

Data Generation Research in code-switching is not only focused on the grammatical aspects of this phenomenon but also the socio-pragmatic characteristics in discourse (Nel, 2012). Large language models such as *OpenAI's GPT* are influenced by social views and inherit encoded biases (Bender et al., 2021). Our work propose the use of GPT to support efforts in synthetically generated code-switched data to increase the prevalence of under-resourced languages. We therefore carefully considered the method in which GPT was prompted to eliminate the introduction of bias. We use general topics and keywords with the goal to generate a diverse range of acceptable sentences.

Human Evaluation The generated sentences were internally evaluated by native speakers of Afrikaans and Yoruba. We ensure the data is respectful to culture and social norms. We will continue to include humans-in-the-loop to ensure faithful data generation.

7. Acknowledgements

We thank JP Morgan and ABSA for their financial support, and OpenAI for providing API credits.

8. Bibliographical References

ACL Anthology. 2023. Welcome to the ACL Anthology. Accessed: 2023-10-08.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Gayatri Bhat, Monojit Choudhury, and Kalika Bali. 2016. Grammatical constraints on intra-sentential code-switching: From theories to working models. *arXiv e-prints*. ArXiv ID: 1612.04538, [Online] Available: http://arxiv.org/abs/1612.04538.

Astik Biswas, Ewald van der Westhuizen, Thomas Niesler, and Febe de Wet. 2018. Improving ASR for code-switched speech in under-resourced languages using out-of-domain data. *6th Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2018*, pages 122–126.

Justine Calma. 2023. Twitter just closed the book on academic research. Accessed: 2023-10-06.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of Computational Processing of Code-Switching. *EMNLP 2016 - 2nd Workshop on Computational Approaches to Code Switching, CS 2016 - Proceedings of the Workshop*, (1980):1–11.

Matt Crabtree. 2023. What is prompt engineering? a detailed guide. Accessed: 2023-10-06.

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 ofthe 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1654–1666.

European Parliament. 2016. General Data Protection Regulation. *Regulation (EU) 2016/679*. Online. [Available]: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504 [Accessed: 4 May 2023].

ISCA Proceedings. 2023. Welcome to the ISCA archive. Accessed: 2023-10-08.

Janice L Jake, Carol Myers-Scotton, and Steven Gross. 2002. Making a minimalist approach to codeswitching work: Adding the matrix language. *Bilingualism: language and cognition*, 5(1):69–91.

Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D. Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152.

Prashant Kodali, Anmol Goel, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022. SyMCoM-syntactic measure of code mixing a study of english-hindi code-mixing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 472–480.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Koena Ronny Mabokela, Madimetja Jonas Manamela, and Mabu Manaileng. 2014. Modeling code-switching speech on under-resourced languages for language identification. In *Spoken Language Technologies for Under-Resourced Languages*.

Jeff MacSwan. 2000. The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: language and cognition*, 3(1):37–54.

Thipe I. Modipa, Febe De Wet, and Marelie H. Davel. 2013. Implications of sepedi/english code switching for asr systems. *Pattern recognition association of South Africa (PRASA)*.

Joanine H. Nel. 2012. *Grammatical and sociopragmatic aspects of conversational code switching by Afrikaans-English bilingual children*. MA in Linguistics for the Language Professions, University of Stellenbosch.

Thomas Niesler and Febe De Wet. 2008. Accent identification in the presence of code-mixing. In *Odyssey*, page 27.

Billian Khalayi Otundo and Martine Grice. 2022. Intonation in advice-giving in kenyan english and kiswahili. *Proceedings of Speech Prosody 2022*, pages 150–154.

Mario Piergallini, Rouzbeh Shirvani, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of the second workshop on computational approaches to code switching*, pages 21–29.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of codeswitching. *Linguistics*, 18(7-8):581–618.

Shana Poplack. 2001a. Code-switching (linguistic). In *International Encyclopedia of the Social and Behavioral Sciences*, pages 2062–2065. Elsevier Science Ltd.

Shana Poplack. 2001b. Code switching: Linguistic. *International Encyclopedia of the Social and Behavioral Sciences*, pages 2062–2065.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1543–1553.

S. Mondal Riktika, S. Pathak, P. Jyothi, and A. Raghuveer. 2022. CoCoa: An encoder-decoder model for controllable code-switched generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2479.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*, pages 205–211.

Sebastian Ruder. 2022. Acl 2022 highlights. [Online]. Available: https://www.ruder.io/acl2022/#next-big-ideas [Accessed: 5 May 2023].

Thamar Solorio. 2021. Moving the Needle in NLP Technology for the Processing of Code-Switching Language. [Online]. Available: http://solorio.uh.edu/wp-content/uploads/2021/08/Solorio-NAACL-2021.pdf [Accessed: 5 May 2023].

South Africa. 2013. Protection of Personal Information Act, No. 4 of 2013. *Government Gazette*, 581(37067). Online. [Available]: https://www.gov.za/documents/protection-personal-information-act [Accessed: 4 May 2023].

Ankit Srivastava, Vijendra Singh, and Gurdeep Singh Drall. 2019. Sentiment analysis of twitter data. *International Journal of Healthcare Information Systems and Informatics*, 14:1–16.

Ewald Van der Westhuizen and Thomas Niesler. 2018. A first South African corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ondene van Dulm. 2007. *The grammar of English-Afrikaans code switching*. PhD Dissertation, Radboud Universiteit Nijmegen.

Ewald Van Der Westhuizen and Thomas Niesler. 2017. Synthesising isizulu-english code-switch bigrams using word embeddings. In *Proceedings of Interspeech 2017*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. ArXiv ID: 2302.11382, [Online] Available: https://arxiv.org/abs/2302.11382.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. *arXiv e-prints*. ArXiv ID: 2212.09660, [Online] Available: http://arxiv.org/abs/2212.09660.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. ArXiv ID: 1909.08582, [Online] Available: https://arxiv.org/abs/1909.08582.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023.

Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. *arXiv e-prints*. ArXiv ID: 2303.13592, [Online] Available: https://arxiv.org/abs/2303.13592.