AFRICAN LANGUAGES IN THE DEVELOPMENT OF CORPUS LINGUISTICS AND OPENAI/CHATGPT: A CASE STUDY OF THE YORUBA LANGUAGE

Anthonia Adunola Abe
abeadunola@gmail.com
Departments of Linguistics, Adekunle Ajasin University, Akungba-Akoko, Ondo State

Abstract
This study, African Languages in the Development of Corpus Linguistics and Open AI/Chat GPT: A Case Study of the Yoruba Language, looked into the flaws of Chat GPT when it comes to inputting African languages into the model. Chat GPT is largely trained in European languages corpora, but to a limited extent, in African languages corpora. The research uses the Yoruba language as a case study. Yoruba corpora data were gathered and analyzed for the benefit of understanding the language structure adequately. The differences between African and European languages were also looked into, which results in one of the flaws of Chat GPT, in giving the right response in African languages. Most African languages, such as Yoruba, Igbo, Zulu and so on are tonal languages while most European languages are non-tonal. It was evident that Chat GPT has challenges in differentiating words with the same segments but different tones, and it was also discovered that Chat GPT was giving wrong data for the Yoruba language. Many African languages are considered low-resource, meaning there is limited digital content available in those languages. This scarcity in training data may impact models' ability, such as Chat GPT, to understand and generate content accurately. This study, therefore, laid the foundation for other researchers who may be interested in working on the corpus of African languages.
Keywords: African linguistics, Corpus linguistics, Chat GPT/AI.

Background of the Study

According to **UNESCO**, "Africa alone is home to a third of the world's languages, with an estimated **1,500 to 3,000 languages.**" African languages exhibit a fascinating array of features that set them apart from languages spoken in other parts of the world. These features contribute to the rich linguistic diversity found across the African continent (*Heine & Nurse, African Languages: An Introduction*, Cambridge University Press, 2000).

Corpus linguistics approaches the study of language in use through corpora (singular: corpus). A corpus is a large, principled collection of naturally occurring examples of languages stored electronically. In short, corpus linguistics serves to answer two fundamental research questions:

1. What particular patterns are associated with lexical or grammatical features?
2. How do these patterns differ within varieties and registers?

According to Sinclair (1991), who is one of the most influential scholars of modern-day corpus linguistics, he detected that a word in and of itself does not carry meaning, but that meaning is often made through several words in sequence. This idea of Sinclair forms the backbone of corpus linguistics.

ChatGPT was launched on November 30, 2022, by San Francisco-based OpenAI. It is a natural language processing tool driven by AI technology that allows you to have human-like conversations. The language model can answer questions and assist you with tasks, such as composing emails, essays, and code. Chat GPT runs on a language model architecture created by OpenAI called the Generative Pre-Trained Transformer (GPT). Generative AI models of this type are trained on vast amounts of information from the internet, including websites, books, news articles, and more. The language model was fine-tuned using supervised learning as well as reinforcement learning. The use of reinforcement learning from human feedback (RLHF) is what makes ChatGPT distinctively unique *(OpenAI, 2022)*.

Chat GPT is trained on a large corpus of text data, which is a fundamental aspect of corpus linguistics. Chat GPT is a product of corpus linguistics, as it relies heavily on large text data sets to generate language. Corpus linguistics can be used to analyse the output of ChatGPT and other language model, to understand better the linguistic patterns and structures that underlie their language and look into the role of corpus linguistics in developing question capabilities.

Observing the difference between African languages and European languages, we can deduce that most African languages are tonal languages. That is, the variation in pitch or tone on words with similar segments can convey different meanings. However, most European languages are non-tonal languages.

There is no doubt that most linguistic corpora in ChatGPT are from and in European languages. That is, a linguistic corpus has been generated on many European languages, and different aspects (like grammatical category and some other aspects) have been looked into.

The tonal feature of African languages gives a clear cut difference between African languages and European languages, generating a linguistic corpus of African languages like Yoruba language with the tonal feature into ChatGPT/Open.AI will help to show the dichotomy between words that have similar segments like igbá (calabash) and ìgbà (time).

Also, with this tonal feature, researchers will be able to generate linguistic corpora into ChatGPT as well as the grammatical category of African languages, especially the Yoruba language.

Therefore, this study will look into the dichotomies between African languages and European languages and how African languages, especially the Yoruba language corpus as a case study, can be inserted into ChatGPT, as well as problems that may occur in the process of inputting the Yoruba corpus into ChatGPT.

Statement of the Problem

It has been an obvious fact that ChatGPT works according to the corpus that is inputted into it. Most of the corpus data in ChatGPT is in European languages, and not enough corpus in African languages, which results in the malfunction of ChatGPT when asked or required to give feedback in African languages. Also, it is discovered that the African language corpus, especially the Yoruba language corpus, has not been extensively looked into or worked on. This research study aims to look into the gathering of the African language corpus (Yoruba language corpus) and the issues that may arise when African languages are installed into ChatGPT.

Research Questions

This research addressed the following questions:

1. What are the differences between African languages and European languages?

2. What are the likely issues that may arise in the process of inputting a large corpus of African languages, most especially the Yoruba language, into ChatGPT?

3. What is the best possible way to gather an African language corpus using the Yoruba language corpus as a core study?

Methodology

This study adopts corpus linguistics as its primary research methodology. Corpus linguistics, in itself is a well-established method, a rapidly evolving approach that applies statistical and computational analysis to large collections of written or spoken texts (corpora) in order to investigate linguistic patterns and phenomena. Central to this approach is the use of a large, principled, and naturally occurring body of texts as the basis for systematic linguistic analysis.

For the purpose of this study, data were gathered with the assistance of five Yoruba native speakers, who served as informants in the compilation of the corpora. A native speaker is defined as an individual who has acquired a particular language as their first language. Such individuals are considered to have full communicative competence in the language, making them reliable sources for accurate linguistic data, particularly for usage, pronunciation, and contextual meanings.

Other sources included Yoruba-language films, most notably Jagunjagun, which contain extensive and authentic use of Yoruba expressions and speech patterns. Online platforms exclusively featuring Yoruba texts, such as Yorùbá Yé Mi, were also consulted. Further materials included selected Yoruba literary texts (e.g., Ìgbà Lonígbàá kà, Eégún Aláré), Yoruba-language newspapers, and the Yoruba Bible.

The digital tools employed for data collection, processing, and analysis include Microsoft Word, CamScanner, AntConc, and the speech analysis software Praat, which was used for phonetic annotation and acoustic analysis.

Presentation of Data and Discussion of Results

This section deals with data presentation and analysis of the Yoruba corpora gathered for the purpose of this research. Focus is laid on the differences of Yoruba and European languages, the analysis of the corpora with the use of Praat (which is used to annotate audio data to text) and Antconc (which is used to annotate the textual data).

Differences between European and African Languages

Most European languages are non-tonal, which means the differences in pitch do not affect the meaning of words. Most African languages, including Yoruba, Igbo, Zulu and so on are tonal, that is, the variation in pitch affects the meaning of words. Looking at the gathered data (Yoruba corpus), we can see this obvious fact.

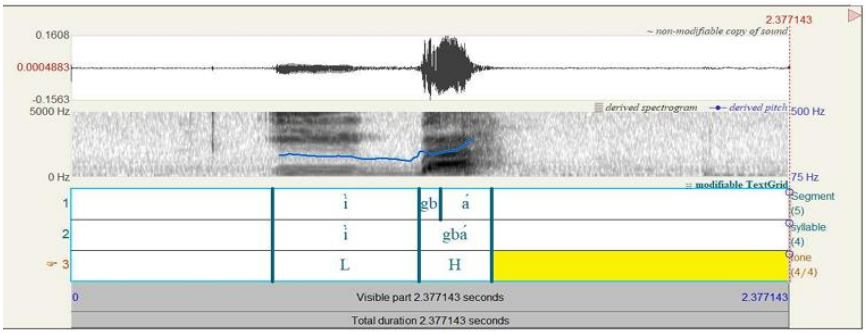
Pitch Variation

European languages do not attest tonal or pitch variation but African languages (Yoruba as a case study) attest tonal variation. Tones in Yoruba are lexically significant, that is, they make differences in meaning of words that have the same segments. For example;

A	B	C
igbá “calabash”	owó “money”	eni “person”
ìgbà “time”	òwò “business”	ení “mat”

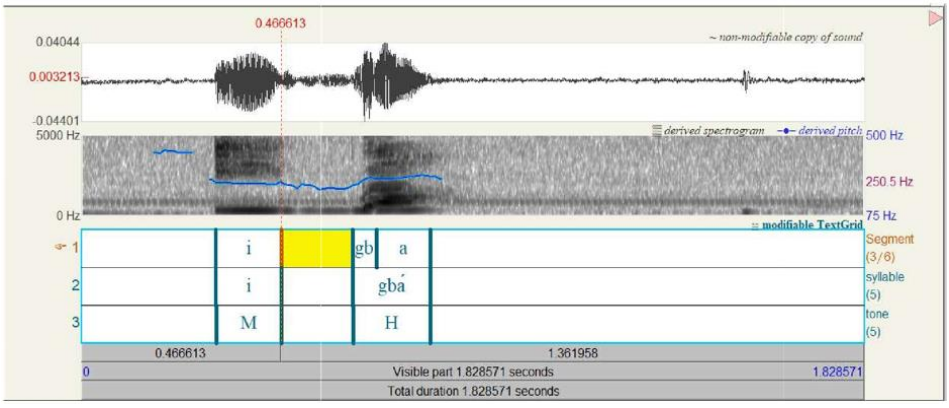
It can be observed in the examples given above that the tones on each same segment bring about a change in meaning.

Pitch variations in the Yoruba language from the audio data, which was annotated to text using Praat;



The chart above shows the annotation of the word ìgbá, which means “Garden egg”. We can see here that the word consists of a low and a high pitch, which operate on different frequencies as represented by the wave lines. The first notable wave line was shown as not too obvious or thick, which represents the low tone and the second wave line, which was extremely thick and obvious, represents the high tone.

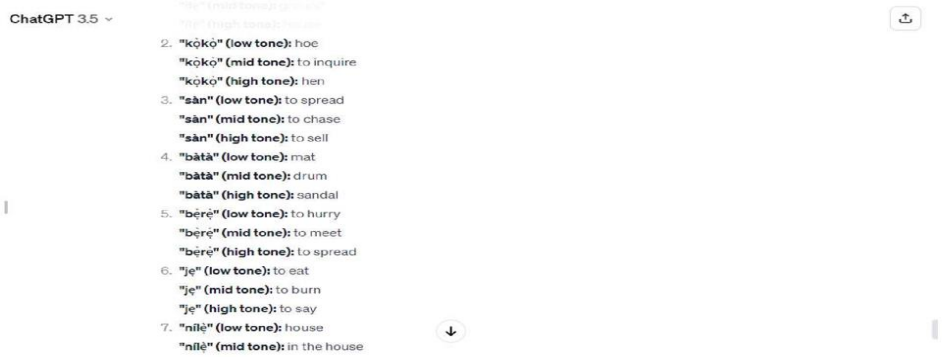
Let us consider another word in the Yoruba corpora data gathered, which has exactly the same segment as the above word but has a different tone:



The chart above shows the annotation of the word igbá, which means “Calabash”. It has a mid-tone and a high tone, which specifically differentiate it from the word ìgbá “garden egg”. The segments are exactly alike, but the differences in pitch institute a difference in meaning.

We can see here from the stated examples derived from the Yoruba corpus gathered that Yoruba language, which is an African language, is obviously a tonal language, that is, a difference in pitch results in a change in meaning.

In Zulu language, which is a southern Bantu language of the Ngunui branch spoken in South Africa, possesses words such as:
/úmfúndisi/ which means “Priest”
/úmfundísi/ which means “Teacher”
Like almost all other Bantu and African languages, Zulu is tonal, which implies change in pitch can result to change in meaning, as seen in the examples given above.
Igbo language is the principal native language cluster of the Igbo people, an ancient ethnicity in the southeastern part of Nigeria. Igbo language is tonal (the meaning of a word can be altered depending on the tone used when pronouncing it). For example;
Ákwá “cry”
Àkwá “egg”
Àkwà “cloth”
We can see from the examples above that the differences in pitch or tone result in a change in meaning in the Igbo language.
Most European languages are non-tonal. In the English language, changes in pitch do not alter a change in meaning; instead, stress and intonation patterns are used for emphasis, emotion, or to convey grammatical features. For example;
TA-ble (NOUN): The TA-ble is outside.
ta-BLE (VERB): She ta-BLES the matter unexpectedly.
CON-vert (NOUN): He is a CON-vert to Christianity.
con-VERT (VERB): How do you con-VERT feet into meter?
We can see here that the stress on each word does not result in a change in meaning, though their classes were changed.
We can also see in other European languages such as German, Spanish, French, Italian and so on, which are all non-tonal languages that do not rely on pitch like tonal languages to alter meaning, but rather they rely on stress, intonation and context to convey meaning and distinctions. For example, in the Italian language, the difference between piano, that is “softly” and piano, which also means “floor”, depends on the pronunciation and context to determine the intended meaning. Likewise, in Spanish, the word Canto, which means both “song” and “I sing”, the two meanings are differentiated based on the context and sentence structure. These languages are non-tonal; therefore, they do not depend on pitch to alter meaning.
The tonal and non-tonal differences between African and European languages, respectively, bring about ChatGPT's flaws. ChatGPT is widely trained on European language corpora. Although we could observe that ChatGPT is also trained in African language corpora, but not to a great extent. This is observed here by the image screenshot of ChatGPT's response to the question: Give a list of words in the Yoruba language that are of the same segments but different pitch. The image screenshot can be seen below:



The first example in the image above refers Ìgbà (a low tone and a low tone) as calabash, which is completely wrong. According to our analysis and annotation of the audio data derived from a Yoruba native speaker using Praat, Ìgbà with a two-tone on each syllable, respectively, refers to "time" and not “calabash”. The second example refers Igbà (a mid-tone and a low tone) as time, time is referred to as Ìgbà (a low tone and a low tone) in Yoruba. The third example is completely out of line, the forest is referred to as “Igbó” in Yoruba and not Ìgba (low-mid) as identified in the image above. Drum is referred to as “ìlù” and not ígbá (High-High) as seen above.
It is definitely an obvious fact that ChatGPT has challenges or issues when it comes to tonal languages such as the Yoruba language, most especially in the aspect of differentiating words with the same segments but different tones or pitch. Using Chat gpt as a tool of learning tonal languages or the Yoruba language for L2 learners will not be a best choice to make, as it will provide a lot of wrong data about the language to the L2 learners.

Consonant Clusters

A consonant cluster is when two or more consonant sounds appear in a word without intervening vowels. African languages such as Yoruba, Igbo, Xhosa and Ibibio disallow consonant clusters while European languages such as English, German, and Swedish allow consonants cluster. For example: In the Yoruba language, we have words such as pàtàkì, pátápátá, omi, ìlù, irun, e.t.c. We can also observe that whenever the Yoruba language borrows words from other languages, if the loan words contain a consonant cluster, the cluster is broken with the use of vowels *(Akinlabi & Liberman, 2000)*. For example, the following words in the English language are borrowed into the Yoruba language which resulting in a change in their structure:

ENGLISH	YORUBA
Bread	búrẹ́dì
Slate	Síléétì
Radio	Rédíò
Teacher	Tísà
Grammar	Gírímà
Card	Káàdì
Pastor	Pásítò

We can also observe that Yoruba language does not allow a consonant at the word final position, which is strongly allowed in English languages and other European languages such as Dutch (e.g Dag “bye”, Totziens “see you later” and Jaar “year”) and Swedish (e.g Hej “hello”, Nej “no” and Natt “Night”).

Discussion of Findings

In our analysis, we have been able to discover the dichotomy between European and African languages, which led to the major flaws of ChatGPT in the aspect of giving the right response in African languages. It was discovered that most African languages are tonal languages, while most European languages are mostly non-tonal. African language such as Yoruba, Igbo, Xhosa and so on, disallows consonant clusters and consonants occurring at the word final position and European languages such as English, Swedish, Dutch and so on, allow consonant clusters and consonants occurring at word final position.

ChatGPT has challenges, especially in the aspect of differentiating words with the same segments but with different pitch or tones. Using Yoruba as a case study, Yoruba language has lexically significant tones, i.e tones have the capability of changing the meaning of words. It was discovered that ChatGPT could not differentiate words with the same segments but different tones; it would rather give the same tones for each of them or allocate an entirely wrong tone for each of them. It was also discovered that ChatGPT was given the wrong data for the Yoruba language. For example, ChatGPT refers kòkò as "hoe", “hen” and “to inquire”. This is complete wrong data as none of the words mentioned refer to as “kòkò” in the Yoruba language.

During the process of this research, it was discovered that the African languages corpus has not really been looked into or worked on; most available corpora are in European languages, especially English. The software used for our corpus analysis (Antconc) has pre-built corpora, but none of them are in African languages. This shows that there is limited availability of African language corpora. This research work has been able to successfully gather Yoruba language corpora, which can also be useful to other researchers.

Conclusion

ChatGPT is trained on a diverse range of internet text, but it may not be exposed to a balanced representation of African languages or dialects. This bias in training data can lead to limited understanding and proficiency in generating contextually relevant responses in African languages. African languages are incredibly diverse, with thousands of languages spoken across the continent. The lack of sufficient training data for each language can result in a lack of expertise in generating accurate responses. Many African languages are considered low-resource, meaning there is limited digital content available in those languages. This scarcity in training data for low-resource languages may impact the model's ability to understand and generate content accurately. As discussed earlier, African languages often have complex linguistic structures, tones, and grammar rules that might differ significantly from widely spoken languages like English. If the model is not adequately exposed to these linguistic complexities during training, it may struggle to handle them effectively.

This research encourages other linguistic researchers to work on various African language corpora. The research provides and works on Yoruba language corpora, which leads to the availability of training data for models like ChatGPT for the development and improvement of these models. These corpora are not only useful for models such as ChatGPT, but they are also of benefit in language teaching and learning, lexicography, stylistic analysis, discourse analysis and so on. Therefore, this research has laid a foundation for other researchers who may be interested in working on the corpus of African languages.

References

Akinlabi, A., & Liberman, M. (2000). The phonology of Yoruba. In H. E. Wolff & O. Gensler (Eds.), *Proceedings of the 2nd World Congress of African Linguistics* (pp. 1–18). Köln: Rüdiger Köppe Verlag.

Heine, B., & Nurse, D. (Eds.). (2000). *African languages: An introduction*. Cambridge University Press

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

UNESCO. (n.d.). *UNESCO and the promotion of languages in Africa: Cultural diversity and multilingualism*. UNESCO IICBA. https://www.iicba.unesco.org/en/unesco-and-promotion-languages-africa-cultural-diversity-and-multilingualism

Sinclair, J. (1991) Corpus, Concordance and Collocations. Oxford: Oxford University Press.