A Stochastic Collocation Algorithm Method for Processing the Yoruba Language Using the Data Context Approach Based on Text, Lexicon, and Grammar

Enikuomehin A. Oluwatoyin[1*] **and Adewumi O. Opeyemi**[2]

[1]Lagos State University, Ojo, Lagos, Nigeria.
[2]Lagos State Ministry of Education, Lagos, Nigeria.

E-mail: toyinenikuomehin@gmail.com*

ABSTRACT

In this paper, we show an initial attempt to generate a self-extractive text processor for the Yoruba language. The Yoruba language is a language spoken by about 60 million persons across America, Europe, and majorly West Africa. This is implemented with the use of a holder codenamed "YOTEX". YoTEx is a Yoruba language text repository which simply learns from the English Language corpus with much emphasis on the agglutinative tendencies of the Yoruba language.

In the building of the data repository, the development of the system considered parameters for existing relations as available in other textual corpora like the WordNet English corpus, which is used in this work as a case study. We used stochastic collocation algorithm to show relationship within entities. The choice of the algorithm is based on the tonal orientation of the language. Hidden Markov model was extended in line with the aim of carrying out deep text analysis. The developed system performs well against known benchmarks in the formulation of an appropriate tagging, part of speech, stemming, chunking etc. system for the Yoruba textual terms. The resulting YoTex will improve the "codinazation" of the Yoruba Language in particular and the other agglutinative language in general. Such will enhance the computer processing efficacies of the Yoruba language. This work presents a novel approach of testing some known language models on a Yoruba lexical corpus.

(Keywords: Yoruba, lexicon, part of speech, collocation algorithm, WordNet, hidden Markov model)

INTRODUCTION

YoTex stands as an abbreviation for Yoruba Text Lexicon developed as an extractive repository from the generalized English syntactic forms. The systems logically correlates terms in the Yoruba text automatically using a combination of itemized linguistic information. The aim is to assist researchers in having access to the linguistic framework of the language. Such will enable researcher and language linguist have access to viable useful information while reducing lingucides. The framework for the development of this platform is based on the large set of experimental results achieved in many language modelling concepts including the use of bench mark analysis such Part of speech tagging, Syntactic analysis among others, will help researchers in non-English domain to understand the differences and relationships alike.

YoTex is an output of several language-logic based frameworks incorporating several types of data. The system can assist in improving the accuracy of the computer application for language modeling. If given appropriate data, the system can generate corresponding grammatical and phonological representation of the data. This has been a method for improving techniques in text to speech research. YoTex is not part of the fundamental language listing for the benchmark IR systems but it has the potentiality of being an effective language for testing in the IR domain. The thought is to propose the best classification method for a large set of document based on the language of discussion. If the repository is conjoined, then the retrieval processes will, have a large subset to contend with.

Yoruba is a Niger-Congo language (sub classification: Kwa > Yoruboid) spoken natively by nearly 20 million people, the vast majority of them in southwestern Nigeria. There are also approximately a half million Yoruba speakers in Benin, as well as speakers in Togo and Ghana and among the emigrant populations in the United States and the United Kingdom. In addition, roughly two million people in Nigeria speak Yoruba as a second language (reference: Yoruba Global Database).

The Yoruba language is spoken in many countries across continents with varied cultural background. Earlier work (1)(2), has shown that the language is syntactically similar to others like---- because of its agglutinative nature. Of note is the tonal orientation of the language (2). Similar tonal languages have been studied, however, the Yoruba language whose base is more extensive in terms of geographical distribution has not received the needed attention from the computational research domain. This has led to dearth in findings and improvement, including generation of new word context as an extraction of the Yoruba word (3).

In the case study, Nigeria Yoruba is predominantly the language in the south west with about 5 million native speakers. The importance of studying this language includes the understanding of the divergent nature of the language across several multicultural societies and the aim to use computational techniques to model. Yoruba language contains some distinct differentials from word formation theories in English, which make this study as crucial as demanded.

The pluralization problem (4) seems to be the first set of constraints where many English language analyzers seems to fail. Root words are not derived by stemming words as done in English terms because these words may exist as another term with the likelihood of having a separate meaning if treated differently. Table 1 shows examples of such terms.

Leveraging on the large number of speakers, appropriate research can be carried out due to high availability of the heterogeneous data. The data set is suitably classified as heterogeneous because of the high level of diversity in the language. That is, two Yoruba speakers may not understand what each other say at the facial level of it until tonal variations are added.

Table 1: Terms and their Pluralized Form Existing as a Different Word.

Persons	Singular	Plural
First Person	mo	A
Second Person	o	ẹ
Third Person	ó	wọ́n

The Data Set

For this work, we rely on the global lexical database which has been ascribed to being the largest Yoruba text corpus. This database consists of the Yoruba terms as used by the Yorubas of Nigeria, Trinidad, and Tobago. The corpus consists of over 450,000 Yoruba words. The choice of this database is encouraged by the context to which the database was built. The corpus consists of detailed lexicographic entries such as the part of speech of the Yoruba term, its English translation and the meaning, among others. Morphological transformation is carried out for each data item, that is, the Yoruba term (5).

The main interest of the work in the long run is to develop a system that can handle the Yoruba spoken word. Earlier attempts have shown that such applications are applicable in English however, this seems to be the first step in the development of a prosody based platform for the Yoruba word. This, we predict, may yield to a general application that can transform many other tonal languages. The limitation of the dataset is that it has not covered the large diversity that exists in the language based on its tone but the collection seems to be usable as a first step to enforce computerization.

EXPERIMENTAL SET UP

The experiment is set to test the performance of *ad hoc* algorithms on an agglutinative tonal language and compare with a classical outcome for a conventional generalized language in context of use for appropriate retrieval of relevant documents. The outcome is essential framework for building text transformation model. We use the NLTK based on python.

NLTK is one of the earliest most reliable platform for building human usable languages majorly based on python. The toolkit stands as backup for front end services such as WordNet among almost 70 other corporals which enables text processing such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

As a first instance, we test the lemmatization process of these native terms in the context of well formulated English based *ad hoc* models. We set to test a sample of about 40,000 of the terms and collections in the database, these terms are structured to include a spelling corrector, a morphological analysis for possible lemmatization.

Stemming and Lemmatizing

Stemming is the processing of transforming a word into its base form. In the information retrieval domain, stemming remains a key concept in generating usable result in any retrieval model. The process of stemming is well pronounced in English language but the stemming in Yoruba language may not be as straight forward as reported in literature, this is because the extensions in word, including pluralization, in most cases changes the term and not just and an extension of the base form. “Iwe” which means book in singular form is “awon iwe” in its pluralized form as against books in English (6). These are well understood in the lemmatization process in English which is the benchmark, a lemma is used in its lexicon so as the simple base terms are used in context.

In grammar, terms like organize, organizes, organization, organizing, democracy, democratization, etc., are all extension of a base form. These are words that are crucial in search domains. In Natural Language Processing, these related terms are fractioned as a concept of a generalized form. Stemming as well as lemmatization are some methods used in the reduction of inflectional form. Same approach is used for the process of reduction of derivational forms.

However, the two words differ in their flavor. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known the lemma (7).

In a well-structured lexical process, the lexicon is built with the capability of identifying the lemma of any term so as to enable stemmers to carry out the transformation to base words. The built lexicon in the Yoruba language should be able to process the Yoruba terms with its complexity.

Grammatically, Yoruba is a Subject-Verb-Object (SVO) language orientation. If, as example, we consider the word “awon omo”, the step is to consider the base form which is omo as the awon simple shows its pluralized nature. The existing challenge is that many systems cannot classify whether awon is the pluarizing term or a word with another meaning. The Yoruba language and other tonal languages are most appropriate in building language engines. In this case, awon, may exist as more than a plularizing term. It may also exit as another word in its own context, as if pronounced in another tone, it may mean trap.

The complexity can only be solved by building engines that do not only consider the term but the sentence as an entity. The dataset was abstracted on tested to identify close performance between terms in the database and other conventional languages, as example, the used NLK toolkit returns the following for the English expression: Balogun was a great warrior in the ancient oyo town. The following was retrieved:

Word	lemma	Char begin	Char end	POS	Normalized NER	Speaker	Sentiment	ID
Balogun	Balogun	0	7	NNP		PERO		1
Was	Be	8	11	VBD		PERO		2
A	A	12	13	DT		PERO		3
Great	Great	14	19	JJ		PERO		4
Warrior	Warrior	20	32	NN		PERO		5
Then	Then	28	38	WRB		PERO		6

Parse Tree

(ROOT (S (NP (NNP Balogun)) (VP (VBD was) (NP (NP (DT a) (JJ great) (NN warrior)) (UCP (ADJP (WHADVP (WRB when)) (JJ alive)))))))

Uncollapsed dependencies

root (ROOT-0 , warrior-5)

nsubj (warrior-5 , Balogun-1)

cop (warrior-5 , was-2)

det (warrior-5 , a-3)

amod (warrior-5 , great-4)

advmod (alive-7 , when-6)

amod (warrior-5 , alive-7)

Enhanced dependencies

root (ROOT-0 , warrior-5)

nsubj (warrior-5 , Balogun-1)

cop (warrior-5 , was-2)

det (warrior-5 , a-3)

amod (warrior-5 , great-4)

advmod (alive-7 , when-6)

amod (warrior-5 , alive-7)

Conference resolution graph

(ROOT (NP (NP (NNP Balogun)) (NP (NN je) (NN jajagun) (NN nla)) (NP (NN nigba) (NN aye) (NN re))))

<u>Uncollapsed Dependencies</u>

- root (ROOT-0 , Balogun-1)
- compound (nla-4 , je-2)
- compound (nla-4 , jajagun-3)
- dep (Balogun-1 , nla-4)
- compound (re-7 , nigba-5)
- compound (re-7 , aye-6)

dep (Balogun-1 , re-7)

<u>Enhanced Dependencies</u>

- root (ROOT-0 , Balogun-1)
- compound (nla-4 , je-2)
- compound (nla-4 , jajagun-3)
- dep (Balogun-1 , nla-4)
- compound (re-7 , nigba-5)
- compound (re-7 , aye-6)

dep (Balogun-1 , re-7)

We deduce that there is a gross difference between the deep NLP lemmatization process of English words and their corresponding Yoruba term. This awareness will help NLP in building an appropriate collocation system of the Yoruba verb.

Part of Speech Tagging and Parsing

In natural language processing domain, part of speech tagging which is also known as word category disambiguation is the process of assigning a part of speech to a word. It is major in the syntactic analysis of natural language word. Given a corpus, word existence in such corpus can be associated with a particular part of speech. Just as in the English words, the part of speech in Yoruba language can be distinctly categorized using its context of existence. Accordingly, the transformation of the same word due to context of use is also a huge possibility. In the Yoruba domain, four parts of speech are prominently used in theory, Yoruba adjective, noun, pronoun, and verb.

There are many algorithms that have been developed over the years for the purpose of performing efficient tagging. TnT tagger, has been ascribed to being the most efficient algorithm for tagging part of speech however, there are doubts if these algorithms that can perform tagging on more than a million token, can be effective on non-English word. TnT, the short form of Trigrams'n'Tags, is a very efficient statistical part-of-speech tagger that is trainable on different languages and virtually any tagset. Tagging is a crucial stage in achieving and appropriate parse of terms. The step seems to be co-dependent but each step is implemented individually. Tagging is about grammatical classification while parsing is about restricting the class to form a sentence or phrase. These are structural analysis of natural languages.

In the Yoruba language, tagging and parsing are collectively exhaustive in which makes language engineers to always consider the level in the design of any related experimental framework. In this research, we are not developing a new corpus as that is not the intention of the work but rather use an existing corpus, the YORUBA WEB CORPUS which has about 2.8 million Yoruba words and the global Yoruba database version 1.0 with over 450,000 words. A major challenge with this corpus is that there are no efficient algorithms and models to test its efficiency. The differential computational difference between the Yoruba and English languages are obvious form the tagging and parsing state, more when the syntactic rules are clearer in English than in many tonal languages. The correctness is dependent on the meaning extracted from the stage, in this job, the corresponding parser for the contextual tagging will also be validated.

DISCUSSION

Text, Lexicon and Grammar

The processes of developing a processor for English has majorly been dependent on general usage and knowledge. As it is, the structure has remained a baseline for the development of other language models. Of course, due to the varied differences, this same assumptions and general knowledge cannot be applied to the Yoruba language, this is also shown in Tables 1 and 2. Being an agglutinative language, the morphological structure is quite different and therefore requires deeper analysis as word stem and affixes may not a change in morpheme with or without word fusion. Formally, Agglutinating language is a language which has a morphological system in which words as a rule are polymorphic and where each morpheme corresponds to a single lexical meaning. It had pronouns, but mostly suffixes instead of prepositions.)

We suppose this consideration was taken care of in the building of the global Yoruba lexical database which has appropriate translation with terms containing the same letters but with different tonal carriage Yotex.

Normally, for a text extraction process of this nature to be developed, several factors must be considered and of importance is the dataset for testing the algorithm. For its robustness, we agree that the selected data set may be sufficient to test and implement the desired framework. Like many other language model, the basis for the transformation is English language and thus the gap between Yoruba as a standalone language and the English context must be reduced. To achieve this, we aim to build the template over a broad spectrum of agglutinative language with or without tonal bias. Language identification is crucial in becoming the first step in this approach, many algorithms have been implemented in this area for identifying several non-English language inclusive of the Yoruba language. We adjust the algorithm to suite the setting and provision of our database. This setting has been done in NLTK which is used in his work. We implement a transliteration process to mine the Yoruba English pair as available in the corpus. The following algorithm was used to measure the translational probability to verify its acceptance ratio.

Algorithm 1

```
Match (w1, w2, translation probability (w1,w2))

1. Find the language of both the words by using the 1st character of each and checking in the character list.

2. Calculate Soundex equivalent of w1 and w2 using Soundex algorithm.

3. Check if both the soundex codes are equal.

4. If yes, return both as transliteration pairs.

5. Else, check the LCS between the soundex codes of w1 and w2.

6. If the distance is found to be 1,

7. Check if the translation probability for w1 to w2 is more than 0.5.

8. If Yes, return “both are transliteration pairs”.

9. Else “both are not transliteration pairs”
```

The experimental implementation shows that the algorithm yields expected result which enables us to ignore the tonal classifications. Since the terms are now being efficiently classified, we can appropriately implement the tagging system. The correct POS has been identified and used to find a ratio to the total available tags in the collection.

A Tag dictionary is a word dictionary, which contains specified POS tags for the tokens. The accuracy was improved by re appropriating the tags to token process. In English corpus, as the annotation increases the language model becomes easier to implement. Same is observed in this regard.

SUMMARY

In this paper, we have shown that the advancement achieved so far in the Natural Language research domain is well suited for testing and analyzing many natural languages however a lot of these tools seems not sufficient to perform well on non-English corpora as shown in the Yoruba language corpus. The fundamental processes such as POS, stemitization and lemmatization did not perform well in the development of YOTEX. Thus, the need for manual transformation and plug-ins. The process of transliteration was used as a basis for effective use of these tools which were developed primarily using the English syntactic structures.

CONCLUSION

In this paper, we developed a framework to which a functional Yoruba text processor can be developed. In this context, we consider the appropriateness of all fundamental text processing template such as POS, lemmatization etc. We have shown clearly that well pronounced templates which has been very functional and useful in English language domain has failed in the analysis of agglutinative tonal language such as the Yoruba such as the Yoruba language.

Consider Tables 1 and 2, the generative structure of the terms in English seems to be correct in conventional use however the failure becomes prominent in Table 2 as action words which are identified as verbs are categorized as nominal noun. The paper recommends that purpose built algorithm and concepts should be built in the analysis and transformation of non-English tonal language with the aim of assisting the computerization of the language.

REFERENCES

1. Ward, I.C. 1956. *An Introduction to the Yoruba Language*. W. Heffer: Cambridge, UK.

2. Ekundayo, S.A. and F.N. Akinnaso. 1983. “Yoruba Serial Verb String Commutability Constraints”. *Lingua*. 60(2-3):115-133.

3. Tek, S., L. Mesite, D. Fein, and L. Naigles. 2014. “Longitudinal Analyses of Expressive Language Development Reveal Two Distinct Language Profiles among Young Children with Autism

Spectrum Disorders”. *Journal of Autism and Developmental Disorders*. 44(1):75-89.

4. Kubota, R. 2014. “The Multi/Plural Turn, Postcolonial Theory, and Neoliberal Multiculturalism: Complicities and Implications for Applied Linguistics”. *Applied Linguistics*, amu045.

5. Linguistic Data Consortium. 2017. “Global Yoruba Lexical Database 1.0”. https://catalog.ldc.upenn.edu/LDC2008L03.

6. Hucks, T. E. 2012. *Yoruba Traditions and African American Religious Nationalism*. UNM Press: Albuquerque, NM.

7. Stanford Natural Lanugage Program. 2017. “Stemming and Lemmatization”. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

SUGGESTED CITATION

Enikuomehin, O.A. and O.O. Adewumi. 2018. “A Stochastic Collocation Algorithm Method for Processing the Yoruba Language Using the Data Context Approach Based on Text, Lexicon, and Grammar. *Pacific Journal of Science and Technology*. 19(1):175-181.

Pacific Journal of Science and Technology