The Design of 'Gègé ÀkọTọ́' – An Orthography Model for Yorùbá Language Use

Àìná Akíndélé Àkànjí
Department of Nigerian and Foreign Languages
Ọlabisi Ọnabanjọ University
Agọ-Iwoye, Ogun State
larexacademie@yahoo.com

Abstract

'Gègé àkọtọ́' is the name adopted for the Yorùbá orthography editing model designed by this study. Writing a good orthography in any human language is very important because any little error can affect and twist the whole meaning of a perfect work. As developments increase in our society, many more are responding to the need of making regular use of Yorùbá as ***our new media*** *in their day to day business, newspaper articles, radio and television broadcasting. But it is observed that so many of these articles have not conformed to the adopted orthographic standard. Many users do not ensure conformity to the standard orthography.*

The modern trend now is to employ and train computers to perform or support human tasks. This paper presents a conceptual approach to address the problem of employing machine to support writing standard Yorùbá orthography. The Yorùbá orthography rules were captured and stored into a library file that has four modules namely: the lexicon, the

parser, the orthography discriminator and the output translator. The algorithm and flowchart for development and implementation was carried out. PyQt4 suite is used for the GUI designs and Python programming language was used for the entire source code. The user input the write up in the input section of the user's interface and the model will re-write and edit it into the current Yorùbá orthography standard. Bearing the fact in mind that human language continues to change and scholars have varied opinions on some of the principles behind the current Yorùbá orthography, the model is designed such that the library files could be edited and optimised though locked up in database.

Key words: Gègé àkọtọ́, Yorùbá orthography, Orthography Model, Yorùbá language.

Introduction

Gègé àkọtọ́ is the name adopted for the Yorùbá orthography systemdesigned and implemented in this study. Orthography is defined by Oxford Advanced Learner's Dictionary (2010 edition) as "The system of spelling in a language". It is a conventionally adopted symbol for representing or writing speech in a particular language. In other words, intentions and information of the mind is expressed through speaking, these messages are represented in writing symbols, but the writing standard which all users must conform to is the orthography. Like the language use, writing a good orthography is a question of competence and performance. Just as specified in the

Chomsky Universal Grammar (UG), each speaker will have to learn mastery of the rules of writing good orthography. However, the new trend in the study of human language is to employ information and communication technology to aid human performance, so that machine can assist human in his tasks. This study therefore is a response to the clarion call that researchers should start beaming their searchlight on modern techniques to develop our language for national development in turn.

Background to the Study

Writing good Yorùbá orthography involves learning and mastery of the agreed conventions. Hence, there is tendency for errors in applying those conventions. Sometimes too, memorising the specific rules guiding Yorùbá orthography can fail the writer. Yet a standard write-up, article or publication demands the writer to apply the conventions pretty well. Then there is need for tools that can assist human to adhere to these standards. The advent of employing computers to aid human tasks then becomes a necessity in writing good Yorùbá orthography. This study captures the laid down principles involved in writing good Yorùbá orthography. The model designed will achieve a great feat in the following areas:

i. Building capacities in the Yorùbá language use for those who are not in Yorùbá studies, but who are always in

need of using Yorùbá language in their day to day affairs. We are in the era of indigenous language use re-orientation. Many more are responding positively to this re-orientation and they adventure into writing their messages in Yorùbá for political, religious, social, trade and mass media. This model will aid this class of people in producing standard Yorùbá documents and write ups.

ii. The Yorùbá specialist will find this model useful as an editing tool. All that is needed is to input the text and allow the model scan the text line by line and output the standard document. If the user is interested in querying the basis for any alterations done to the original text, Gègé àkọtọ́ will output the particular rule that guided his re-arrangement of writing.

iii. This system can be further developed into an ontology model that will allow machines of different application, effective in portability, share ability and interoperability which are the main goals of ontology developed environments.

Statement of the Problem

Clarion calls have gone out for adoption of our indigenous languages for developmental purposes both nationally and internationally. Some of these calls were recorded in Bamgbose (2006), Owolabi (2006, 2007), Awobuluyi (2011) and many people are responding to these calls.

Newspapers are springing up in Yorùbá language medium, radio and television programs are being broadcast in Yorùbá language. Good developments as this is, however, many writers do not bother so much in ensuring that their write ups and articles are in perfect Yorùbá orthography. Few ones are not even aware there is a standard to be conformed to. The Joint Consultative Committee of 1974 has prescribed a standard to follow and scholars have provided one review or the other to these guidelines. A comprehensive history of the development of Yorùbá orthography can be explored from Arohunmolase (1987). The two Yorùbá metalanguage volumes edited by Bamgbose and Awobuluyi can be of good assistance to train oneself in mastery Yorùbá metalanguage.

Interestingly, few individuals have responded to the clarion call of employing computational methods to analyse Yorùbá language. Research in this area includes: Automated Speech Recognition (ASR) for Yorùbá Tone Systems in order to build capacity in human language (Adegbola 2006, 2008, and 2009), Machine Translations (Hassan 2009, Odoje 2010, 2013, 2017, Eludiora (2012), Ontology developments (Hassan, Odejobi, Ogunjobi and Adejuwon 2013), (Aina 2016), Speech synthesis (Abimbola, 2014). Akinade (n.d.) developed 'Tákàdá'- a text processing tool for typing Yorùbá language to conform to the correct Yorùbá orthography. However, this studyseeks to achieve another feat by

exploring linguistics rules behind the standard Yoruba orthography in building a model that assists users in writing Yorùbá orthography.

Aims and Objective of the Study

This paper aims at constructing an orthography system that will enhance the effective use of Yorùbá language. It has the following as its objectives:

- Isolate the current standard Yorùbá orthography rules based on Joint Consultative Committee on education (JCCE)1974's recommendation.
- Design an interactive user's Interface where the user can input the document and retrieve a re-written version corrected according to the captured standard.
- Help those who are not familiar with the basics behind the standard orthography learn and master the rule.
- Contribute to the efforts to develop Yorùbá language as the new media for national development.

Data for the Study

The data used for building the model is the standard Yorùbá orthography as recommended by the Joint Consultative Committee on Education of 1974. .

Scope of the Study

The model designed is limited to the 1974 JCCE recommendations. It is a clear fact that various scholars had reviewed the JCCE recommendation based on different linguistic perspectives. (Arohunmolase 1987, Bamgbose 1990, Adekunle 2008). Also, various conferences and workshops have been held on Yorùbá orthography and each one had come up with their recommendations. We attempted to adopt and compared the moderation to a fair extent since all of these recommendations are similar. The design of the model's library file is basically on JCCE 1974's recommendation. The system is designed in such a way that it can be extended to take any further recommendations whenever it is desired.

Literature Review

There are numerous books on Yorùbá orthography. Some of them are foundation materials for learners of Yorùbá language and some are designed for pedagogical use in secondary and tertiary institutions. However, as said earlier only few are selected since our model will be based on that of 1974's J.C.C.E approved list. Arohunmolase (1987) is a developmental history book of Yorùbá language from 1800 to 1985. The book reviews the efforts of earlier missionaries, slave traders, religious preachers toward reducing Yorùbá into writing. It examines various scholars

associations and their impact in improving what the earlier forerunners have done. It also reveals the various committees inaugurated to work on Yorùbá orthography right from 1875, 1965, 1966, 1969/70 and eventually of 1974 and the orthography standard they have adopted. The book also presented various criticisms levelled against the earlier orthography standard and the solution proffered to these faults. No account of linguistic rules that lead to the derivation of orthography was given in the book. A reader who is interested in these details is encouraged to consult the said literature.

Adebiyi (2000) carried out a summary of the previous literature on Yorùbá orthography. It is a book designed for Colleges of Education students as such it is a highlight of the rules adopted for Àkọtọ́ Yorùbá. Adekunle 2008 explores the various linguistic rules as the basis at which the Yorùbá orthography was based. Adopting Bamgbose (1990) and Chomsky (1972)'s view, the paper explains that the theory of language structure, acquisition and use, forms the basis to follow in describing standard orthography for any natural language. He divided the discussion into the linguistic levels of phonology, morphology, syntax and semantic. Yet none of these materials have considered the possibility of storing the rules of proper orthography in any memory database, hence an added feat this study achieves.

Data Presentation

The linguistic level of phonology, morphology syntax and semantics is at the background of the standard Yoruba orthography and this is followed in presenting the data and its analysis. Our aim is to capture a fairly large lexicon incorporated into the library files. Therefore, we shall present this data one after the other as follows:

Linguistics and the Standard Orthography

Bamgbose (1965) stated that two phonetic symbols should not represent one phoneme. Only the distinct phoneme sound perceived should be represented by the phonetic symbol therefore the following words in Set B follow instead of Set A

(1)	**Set A**	**Set B**
	Àìyà	Àyà
	Ẹiyẹ	Ẹyẹ
	Aiyé	Ayé
	Ẹ̀ìyà	Ẹ̀yà
	Ng ó lọ	N ó lọ
	Nwọn	Wọn
	Ẹ̀nyin	Ẹ̀yin
	Nyín	Yín

For the purpose of structural organisation of our database, we have to take the following procedure to arrange the above words in alphabetical order.

i. Arrange the words in alphabetical order
ii. Write the possible meaningful variances based on tone variations of the same phonetically similar structures.
iii. Pair the words into syllable VCV structures.
iv. Assign precedence to the three distinct Yorùbá tone taking low tone (̀) as the first, followed by the mid (N[1]) and high (́).

So for the set of data in (1a) we have two (2) variants

2ai	Àìyà	Àyà	(chest)
ii	Aiya	Aya	(wife)

To the best of our knowledge for now, no other variants could be formed from the above patterns that give any semantically meaningful noun in Yorùbá language.

Having exhausted that we take the next:

2b.

i.	Aiyé	Ayé	(world)
ii.	Ẹ̀ìya	Èya[2]	(race)
iii.	Ẹiyá	Ẹyá	(a kind of mammal)
iv.	Ẹiyẹ	Ẹyẹ	(bird)
v.	Ẹ̀ìyẹ	Ẹ̀yẹ	honour)

[1] Please note that N stands for neutral, because the mid tone is not phonetically marked in Yorùbá so it is Neuter.

[2] This indicates that you can either assign mid or low tone for the same expression to refer to the same noun.

It is counter intuitive to use symbol in representing the lengthening of syllable in some Yorùbá expression. Such lengthening should be done by providing the sound and not symbols as shown in (3) below:

3.

i.	Ãgo	Aago	(watch/clock)
ii.	Alãfia	Àlàáfíà	(peace)
iii.	Ãnu	Àánú	(mercy)
iv.	Alãnu	Aláàánú	(merciful)
v.	Olõto	Olóòótọ́	(truthful person)
vi.	Lãrin	Láàárín	(between)
vii.	Õto	Òótọ́	(truth)
viii.	Mewã́	Mẹ́wàá	(ten)
ix.	Mésã̀n	Mẹ́sàn-án[3]	(nine)
x.	Yĩ	Yìí	(this)

The adopted standard in set B complies with the phonological rule that insists that every distinctive sound provided should be written rather than one phonemic sound representing two. That is map one to one representation from grapheme to morpheme.

Nasality Assimilation Rule

The nasality assimilation rule proposed by Schane (1973:50) states that oral vowels contiguous to nasal

[3] See Aina 2012 for the reason for this occurrence.

vowels must assimilate to the nasality feature through progressive assimilation so that it becomes nasal too.

In the following set A and B, A does not attest to this rule and the standard Yorùbá orthography takes into account those phonological underpinnings.

Set A	**Set B**
Dín-i	Dín-in
Fún u	Fún-un
Gan-a	Gan-an
Mẹ́sàn-á	Mẹ́sàn-án
Pọn ọ́	Pọn-ọ́n
Rán a	rán-an

Another phonological rule establishes that two distinct phonemes should not refer to one grapheme symbol. The old orthography in set A does not take this into consideration whereas the set B conforms to this rule as in:

Set A	**Set B**
Ìddó	Ìdó
Ọ̀ttà	Ọ̀tà
Òshogbo	Òṣogbo
Ògbómọ̀shọ́	Ògbómọ̀ṣọ̀
Shàǹgótọ́lá	Ṣàngótọ́lá́
Ilésha	Iléṣà
Shàgámù	Ṣàgámù

Phonology rules also states that the graphemic representation should correspond exactly as pronounced. Consider the two sets below:

Set A	**Set B**
Ènìà	Ènìyàn
Ná	náà
Yíò	yóò

Phonology rules also states how the following words are realised in the final output:

Sọ + òótọ́ ⟶ sòótọ́
Ṣe + àárẹ̀ ⟶ Ṣàárẹ̀

Morphology studies internal structure of words. In analysing the standard Yorùbá orthography, some rules of morphology takes precedence in adopting the standard. Some of the rules are discussed as follows:

(i) A word can stand as root morpheme, and each morpheme must stand separately on its own to indicate word class. Such examples include:

Wípé	Wí pé
Ńlọ́	Ń lọ
Sọpé	Sọ pé
Nígbàgbogbo	Nígbà gbogbo
Nítoríná	Nítorí náà
Nítorípé	Nítorí pé

(ii) In all languages of the world, no affix can stand as a word category, this why Yorùbá does not attest a nominal prefix to stand alone except being fused to a root morpheme. This rule is the basis for adopting the following orthography:

Àì + gbọ̀ràn	⟶	Àìgbọ̀ràn
À + ti lọ	⟶	Àtilọ
Ṣe + àì + gbọ̀ràn	⟶	Ṣàìgbọ̀ràn
Ṣe + àì + sùn	⟶	Ṣàìsùn

(iii) The SY requires hyphenation in the following set of words:

Ẹja-n-bákàn
Àbù-ù-bùtán
Àjẹ-gbági
Àìbalè-àyà
À-sọ-jásan

(iv) The following standard expressions emanated as a result of syntactic rules:

Nítorínà	Nítorí náà
Nítoríti	Nítorí tí
Gbogbogbò	Gbogbo gbòò
Nígbàgbogbo	Nígbà gbogbo
Wipe	Wí pé
Mẹ́nuba	Mẹ́nu bà
Ríi	Rí i
Pèe	Pè e

Ṣáa	Ṣá a
Ṣeéṣe	Ṣe é ṣe
Bàá	Bà á

Materials and Method

Python programming language is the backbone of Natural language toolkit (NLTK). Python can run in the operating system such as windows, vista, Linux and so on. But commonly in our locality here, windows and vista are in common use. So these applications operate on this easy-to-get operating system. The GUI is designed using the PyQt4 of the python language. The python interpreter, Python shell and the NLTK can be downloaded from http://python.org/

Gègé àkọtọ́'s Library Design

The crux of this model lies in the library file designs. It is the database for the system. A good structure of the library is needed and it is partitioned according to the linguistic rules which guided the formation of Yorùbá orthography. The major sets of the lexicon for the standard orthography were extracted from different literature but majorly from Arohumolase (1987) and the library file stores this different data set outlined in the previous section.

Conceptualisation of the model

The study focus on a system that can scan through and rewrite an unedited text or corpora of Yorùbá language, checking those words that are wrongly written in a non-standard form. The medium of the system is only text-to-text medium. The diagram below illustrates the conceptualisation paradigms.

The Conceptual chart of Gègé Àkọtọ́

The System Architecture

The entire system architecture is designed in four modules for easy partitioning and comprehension. These are the lexicon i.e. the standard orthography, the Parser, the Orthography discriminator and the output translator.

i. **The Lexicon Module:** This consists of the extracted Standard Orthography. The operating system notepad was used as the processor of the lexicon at the first stage but it is implemented in phython shell 2.7.1. Version. To solve the problem of tone and its peculiarity in Yorùbá language, the items in the lexicon is firstly processed in MSword. The MSword has been configurated using the "combining Diacritical marks of ASCII code digit of the symbol interface in Microsoft word processor. It is then copied and pasted to notepad software which accepts the input Yorùbá tone lexicon untampered.

ii. **The Parser Module**

This module accepts the input string from the lexicon module. It scans through the input string and identifies the structure of every token in the strings and check if it tallies with the items as stored in library files. It flags an error message and compiles the total items which are not in conformity to the standard input and parses it to the discriminator module. Some designs issues arise as a result of the peculiarities in Yorùbá language some of which are listed below:

a. The first phonological rule in the previous section requires that two distinct phonetic symbol (vowels in this case) should not represent one phoneme … e.g. aiya – aya

We have to declare all the oral vowels in Yorùbá and specify that if any word is preceded by these oral vowel onset except 'u', the onset syllable must not take the 'i', otherwise an error message should be flagged immediately. The algorithm for this specific rule is presented below as an example.

Algorithm 1: IMPLEMENTATION OF SYO VCV SYLLABLE PATTERN

Comment: For each of the oral vowels declared i.e. v:{ a, e,ɛ, I, O,Ɔ, u [oral v]

Comment: Is there any 'i' contiguous to initial V of VCV pattern?
if not (read the next token of VCV) syllable structure)
then identify all the token initials as elements of V
else note [occurrence of the contiguous V]
Comment: for each of the i contiguous V noted for V. [oral V]

do
  do
    Comment: Delete not minding position
    Delete in every case of content V
    Comment: Deletion of contiguous V.
    If not (implement contiguous to C of CVC syllable structure then display the implemented CVC syllable part
    else return an error message
    then exit.

The same declaration was made for the consonant initial in 1 (Set A) in previous sections and the algorithms for it is also developed.

iii. **The Orthography Discriminator Module**: This module accepts the scanned error strings from the parser module and arranges them in tokens discriminatively. After discriminating the input tokens, it then re-arranges them

according to the specification in the library file. This is the semi-final stage and it sends it to the output translator.

iv. **The Output Translator** links other modules. It is designed to link the lexicon, parser and discriminator modules so as to make the output translation easier between the database, GUI and the parser. This module is the interactive interface where the user interacts with Gègé-Àkọtọ́. The user type the words of which its orthography is to be checked and it gives the output as OK or flag an error message then represent it the way it should be. Atext, sentence or phrase can be loaded for the same operation, only the user will instruct the system of his intention. It has provision for reset, save the output or print the output. PyQt4 was used for the coding of the GUI and whole source code for the modules is written in python. The diagram below shows the system design architecture:

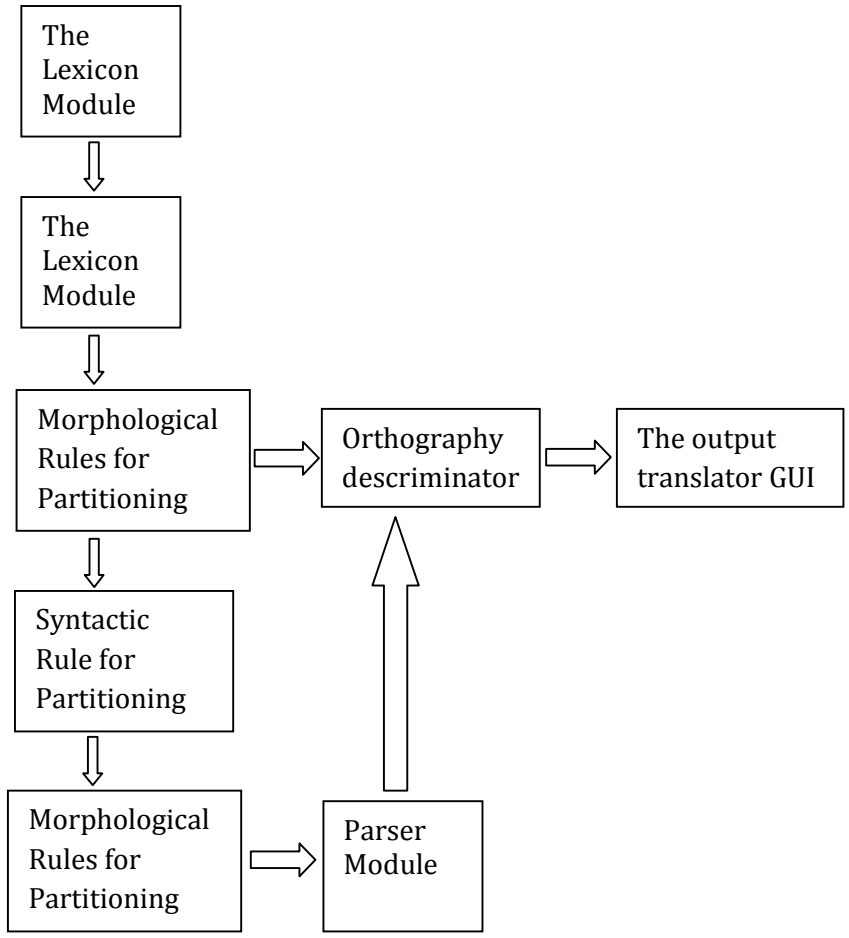

Figure 1: Gègé Àkọtọ́ System Architecture

Implementation of the Model

Implementation of the standard Yorùbá orthography data of course involves various degree of complexity because linguistic rules behind these various items differ. Some of

these design issues or complexities were mentioned in the previous section about the data in (2b).

The implementation of linguistic rules for writing the standard orthography in (3a) demands that we declare that the vowels carrying vowel lengthening diacritics should be copied into two syllables rather than displaying the lengthening symbol. But the problems however is that some lengthened syllables are more than two e.g.

Oloto	o –ló-ò-ó-tọ́
Alanu	a-lá-à-á-nú

How will the system determine the number of lengthened syllable so as to implement it? The solution is to include phoneme to grapheme recognition devices i.e. speech to text synthesis whereby the user will pronounce the word and the machine synthesizes the number of syllabic units to determine the orthography. However, the scope of this work cannot implement this solution.

What we do is to apply orthography learning method, whereby the data are gathered from available corpus and stored into the database. This is the option implemented in this study.

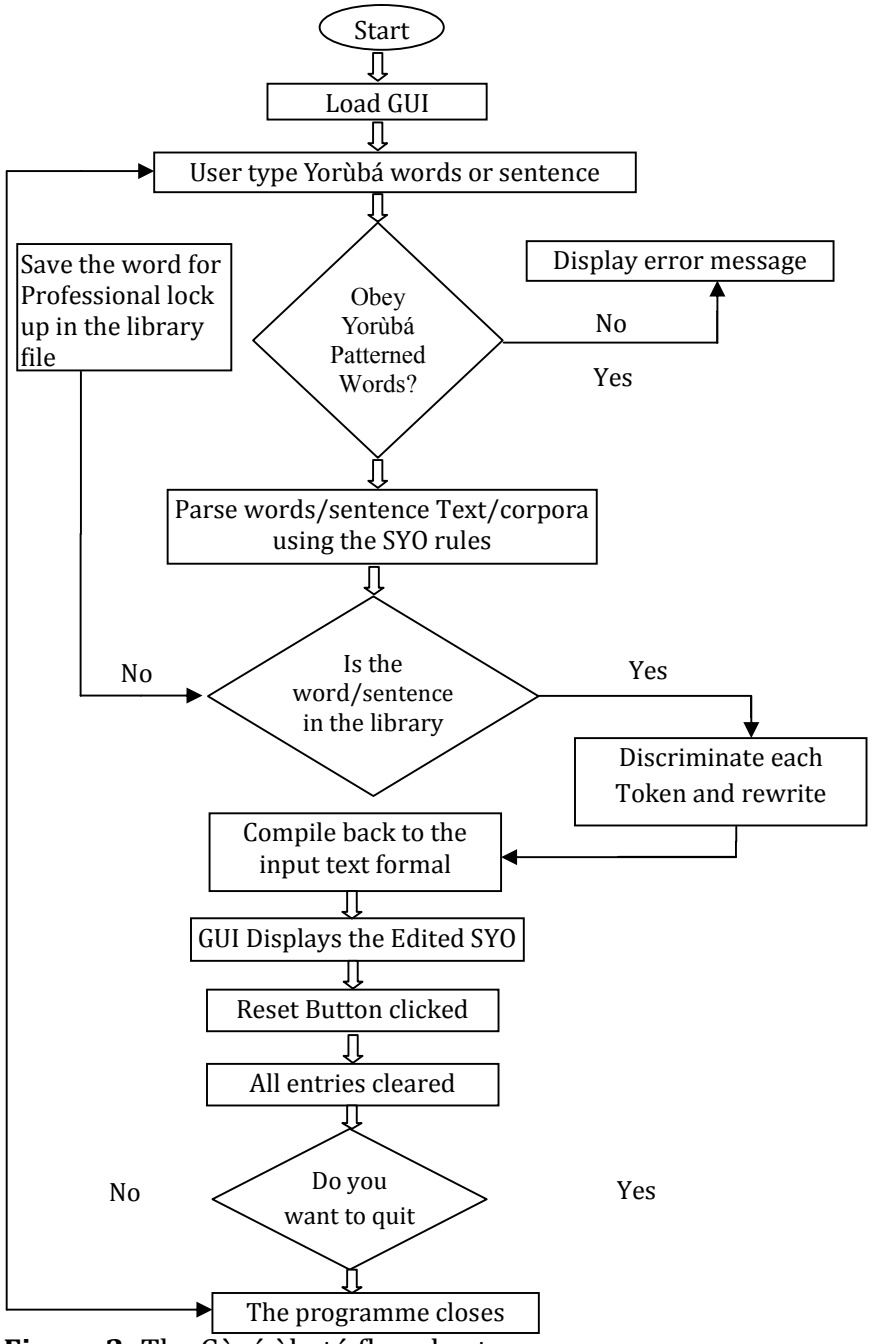

Figure 2: The Gègé àkọtọ́ flowchart

Conclusion

This study has exposed us to the linguistic rules involved in writing Standard Yoruba orthography. Various materials that were used to build a model that implements the orthography has been examined and machine learning approach of storing the data corpus into the library files having four modules of implementation has been considered. The system will interact with a user to re-write a standard Yoruba orthography. All these show that machine or computer can be employed and trained effectively to assist and support various linguistics tasks.

References

Adebiyi A. (2000). Ìdàgbàsókè àti Àkọtọ́ Èdè Yorùbá nínú Tunji Ọ̀pádọ̀tun (Olóòtú), *Ẹ̀kọ́ Èdè Yorùbá fún Ilé-Ẹ̀kọ́ Olùkọ́ni Àgbà.* Federal College of Education, Òṣíẹ̀lẹ̀ Abẹ́òkúta. Pgs 20-32.

Adegbola, T. (2006). Globalization: Colonizing the Space Flows. *Globalization and the Future of African Languages.* Egbokhare F. and Kolawole C. (eds). Ibadan: Ibadan Cultural Studies Group.

Adegbola, T. (2009). "Indigenizing Human Language Technology for National Development". A Lecture delivered as First ARCIS Distinguished Guest Lecture, U.S.A.

Adegbola, T. (2008). Building Capacities in Human Language Technology for African Languages. EACL

2009 Workshop on Language Technologies for African Languages – AfLaT 53-58. Athens: Association for Computational Linguistics

Adekunle A.M. (2008). Ipa Àwọn Ẹ̀ka Gírámà Nínú Yorùbá Àjùmọ̀lò. Pépà Sẹminá fún Àwon Akẹ́kọ̀ọ́ Yunifásítì Ọlábísí Ọnàbánjọ, Àgọ́- Ìwòyè.

Aina, A. (Forthcoming). "Development and Preservation of Yorùbá Cultural Identity through Ontology Development- An Introduction" *Culture and Identity among the Yorùbá People.*

Akinade (n.d.). 'Tákàdá'- A Text Processing Tool for Typing Yorùbá Language. Retrieved online Mar.12, 2014 from http/www.ajocict.

Arohunmolase, O. (1987). *Àgbéyèwò Ìdàgbàsókè àti Àkọtọ́ Èdè Yorùbá,* Ibadan. *Onibon*-oje Press (Nig.) Ltd.

Bamgbose A. (1990). *Fonọ́lọ́ji àti Gírámà Yorùbá*: Ibadan: UPL.

Bamgbose, A. (1965). *Yorùbá Orthography.* NERDC

Chomsky, N. (1972). *Language and Mind* Enlarged (ed.) Harcourt Brace Jovanovich, NY.

Eludiora, S.I. (2012). Development of an English to Yorùbá machine translation. Ph.D. Thesis Dept. of Computer Science and Engineering. O.A.U. Ile-Ife. xi+206pp.

Hassan, J. A., Odejobi. O.A., Ogunfolakan, B.A. and Adejuwon, A, (2013). Ontology Engineering in Yorùbá cultural Heritage Domain. *African Journal of Comp. & ICT*s IEEE. 6.5:181-198. Retrieved online Oct. 20, 2015 from http/www.ajocict.

Iyanda, A.R. (2014). Design and Implementation of a Grapheme-to-Phoneme Conversion System for Yorùbá Text-to-Speech Synthesis. Ph.D. Thesis. Dept. of Computer science. Obafemi Awolowo University. Ile Ife, Osun State, Nigeria. X+186pp.

Odoje, C.O. (2010). English-Yorùbá Rule Based Syntax in Machine Translation. M. A. Project, Department of Linguistics and African Languages. University of Ibadan, Ibadan.

Odoje, C.O. (2013). "Language Inequality: Machine Translation as the Bridging Bridge for African Languages" in *Àgọ́-Ìwòyè Journal of Languages and Literatures,* Vol. 4, Pg 22-49.

Odoje, C.O. (2017). Linguistic Rule in Statistical Machine translation. Ph.D. Thesis. Dept. of Linguistics and African Languages, University of Ibadan. xvi+218pp.

Owolabi, K. (2006). Nigeria's Native Language Modernisation in Specialised Domains for National Development. A Linguist's Approach", University of Ibadan, Inaugural Lecture, Ibadan. Universal Akada Books (Nig) Ltd.

Owolabi, K. (2007). Ó Tó Gẹ́ẹ́, Ọmọ Odùduwà: Ogun Ìṣàmúlò Èdè Yorùbá Ní Ibikíbi. Ní Ipòkípò àti Ní Àyèkáyè Di Jíjà Wàyí. Ìdánilékọ̀ọ́ Ní Ibi Àyájó Èdè Àbínibí. Mọ́kọ́lá Ìbàdàn.

Oxford Advanced Learner's Dictionary (2010 edition)

Schane, S. (1973) *Generative Phonology* U. S. A. University of California.