Development of English to Yorùbá Machine Translation System for Yorùbá Verbs' Tone Changing

Safiriyu I. Eludiora
Dept. of Computer Sci. & Engrg., Obafemi Awolowo University, Nigeria

Abayomi O. Agbeyangi
Dept. of Computer Engrg., Moshood Abiola Polytechnic, Abk., Nigeria

O.I. Fatusin
Dept. of Computer Sci. & Engrg., Obafemi Awolowo University, Nigeria

ABSTRACT

In translating English sentences (text) to Yorùbá sentences (text), some Yorùbá verbs change tone from the bilingual dictionary low-tone to mid -tone when they are translated to Yorùbá. They are called tone change verbs. These tone change verbs do pose some challenges in English to Yorùbá machine translation. Most of the time it changes the meaning of the sentence. These changes usually depend on the positions of the nouns and pronouns in the sentence . The verbs in this category were collected from different Yorùbá sentences that contain tone change verbs. Some re-write rules were designed for the two languages . The re -write rules were tested using JFLAP. Apart from re-write rules, there are other grammatical rules considered and the rules affected the Yorùbá translations. The software was designed using unified modelling language (UML). The Rule-based approach was used for the translation. Python programming language was used for the software development. The python has natural language tool kits that are used for the sentence parsing. The system accept English sentence then discover the pattern for the sentence. The system was implemented and tested for twenty tone change verbs within the home domain. The two languages are subject verb object (SVO) sentence structure with some differences. The results show that the MT system can translate these tone change verbs. The system is efficient in its response time. IFEMT2 system was developed to handle the translation of tone changing verbs. IFEMT2 was used to compare with the Google translator. IFEMT2 system performs better than the Google translator. The Google translator could not adequately translate these tone change verbs.

Keywords

Tone-change, *Yorùbá*, Linguistics, rules, translation.

1. INTRODUCTION

Yorùbá is one of the major languages spoken in Africa. Other languages in this category include: Fulfude, Hausa, Lingala, Swahili and Zulu. *Yorùbá* has a speaker population of about 25 million (South West Nigeria only) according to 2006 population census conducted by the National Population Commission of Nigeria [1]. *Yorùbá* language has many dialects but all speakers can communicate effectively using the standard *Yorùbá* (SY) which is the language of education, mass media, and everyday communication [2].

Yorùbá is a tonal language with three phonological contrastive tones: High (H), Mid (M) and Low (L). Phonetically, however, there are two additional allotones or tone variances namely, rising (R) and falling (F) [3] and [4]. The *Yorùbá* alphabet has twenty-five letters comprising eighteen consonants and seven vowels. There are five nasalised vowels in the language and two pure syllabic nasal vowels [3] and [5]. *Yorùbá* has a well-established orthography which has been in use for over ten decades (around 1843). *Yorùbá* is relatively well studied when compared with other African languages and there are literatures on the grammar of the language.

The present work is one of the works that have examined the machine translation systems in the context of text to text translation technology. In translating from English to *Yorùbá*, some verbs differ in tonal pattern when translated to *Yorùbá* compare to other monosyllabic verbs when used with object nouns in *Yorùbá* sentence. These types of verbs are called Tone change monosyllabic verbs which are the main focus of the research work reported in this paper. Example of such tone change verbs in a sentence is 'Hit' which implies 'nà' in *Yorùbá*.

The context free grammar (CFG) is the most common way of modelling constituency.

CFG = Context-Free-Grammar => Phrase Structure Grammar => BNF = Backus-Naur Form.

The idea of basing a grammar on constituent structure dates back to Wilhelm Wundt (1890), but not formalized until 1956 by Chomsky, and independently formalized by Backus.

Grammar (G) is a mechanism to describe the language ; it is used to describe how language may be formed . The automats structure of the English and Yorùbá sentences with respect to the context free grammar can be described as a four tuple grammar given as:

G = {T, N, S, P}, Where:

- T is set of terminals (lexicon)
- N is set of non- terminals for natural language processing (NLP), we usually distinguish out a set P subset of N of pre terminals which always rewrite as terminals.
- S is start symbol (one of the non-terminals)
- P is rules or productions of the form X => ß, where X is a non-terminal and ß is a sequence of terminals and non-terminals.

2. MT APPROACHES

Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora [6], Statistical machine translation (SMT) is characterised by the use of machine learning methods, for example, Hidden Markov Model for POS tagging. In less than two decades, SMT has come to dominate academic MT research and has gained a share of the commercial MT market [7].

Rule-based Machine Translation (RBMT) also known as `Knowledge-based Machine Translation', `Classical Approach' of MT is a general term that denotes machine translation systems based on linguistic information about source and target languages. Basically the linguistic information can be retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological and syntactic regularities of each language [8] and [9]

Having input sentences (in source language), a RBMT system generates the output sentences (in target language) on the basis of morphological, syntactic and semantic analyses of both the source and the target languages. This involves a concrete translation task [8] and [9]. We used rule-based approach for the research work that is being reported here.

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies [10]. Several MT companies (Asia Online, LinguaSys, Systran, PangeaMT and UPV) are claiming to have a hybrid approach using both rules and statistics. The approaches differ in a number of ways:

1. Rules post-processed by statistics: Translations are performed using a rule-based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.

2. Statistics guided by rules: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalisation. This approach has a lot of power, flexibility and control when translating.

Hybrid machine translation is a combination of both statistical and rule-based translation methodologies [11].

Split-verbs concept in Yorùbá language was investigated by [12]. These verbs are view in number but they affect the final translation.

The research done by [13] provided foundation for the English to Yorùbá machine translat ion system. The study was able to highlights some of the challenging issues.

3. SYSTEM FRAMEWORK DESIGN

In this section, the linguistics rules for the tone changing Yorùbá verbs and re-write rules for the system.

3.1 The Linguistics Rules for the Tone Changing Yorùbá Verbs

In this subsection, the linguistics rules were used to design the system.

Rule 1: When verbs require noun as object , tone change affects low tonal verbs. This tone must change to mid tone in Yorùbá language Bamgbose (1990). For example:

1. English sentence: He knows the house.

1(a). Yorùbá sentence: Ó mọ̀ ilé náà. [L] Wrong

1(b) Yorùbá sentence: Ó mọ ilé náà. [M] Correct

2. English sentence: He bought meat.

2(a) Yorùbá sentence: Ó rà ẹran. [L] Wrong

2(b) Yorùbá sentence: Ó ra ẹran. [M] Correct

Rule 2: Low tonal verbs maintain their form when pronouns come as their objects. For example:

3. English sentence: Ade hit it.

3(a) *Yorùbá* sentence: *Adé nà á.*

4. English sentence: He hit it.

4(b) *Yorùbá* sentence: *Ó nà á.*

The "*nà*" does not change tone.

Rule 3(a): tone change does not affect mid tone of transitive verbs in *Yorùbá* language. For example:

English sentence: He resembles his Elder.

Yorùbá sentence: *Ó jọ ẹ̀gbọ́n rẹ̀.*

Rule 3(b): tone change does not affect high tone of intransitive verbs in *Yorùbá* language. For example:

English sentence: She plucked the leaves.

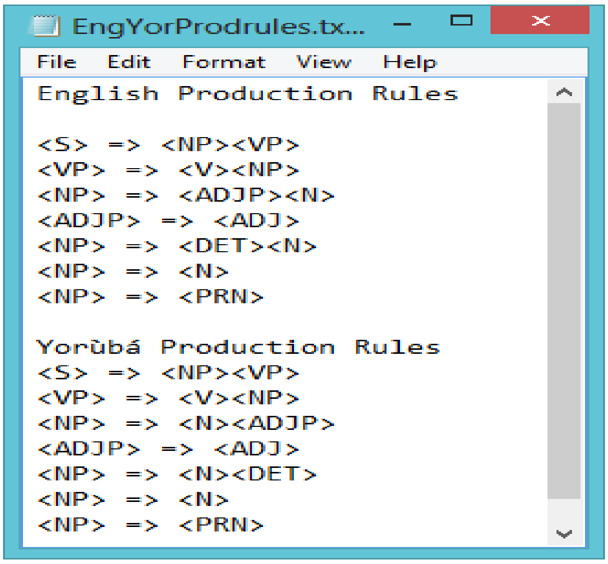
Yorùbá sentence: *Ó já ewé náà.*

3.2 English and Yorùbá Re-write Rules

These re-write rules were designed by considering all the possible ways of translate those sentences that contain the tone changing verbs. Fig. 1 shows the re-write rules. Figs. 2 and 3 show the realisation of the re-write rules using state diagrams for the two languages. Figs. 4, 5 and 6 show the JFLAP test results for some of the sentences used. The re-write rules *were* used for the tests. The three sentences have different rules. The three sentences used for the rules tests are shown in Table I.

Table 1. Sentences used for the rules' test in JFLAP

English	Yorùbá
He hit the dog	*Ó na ajá náà*
He hit it	*Ó nà á*
John hit the dog	*John na ajá náà*


```
EngYorProdrules.tx...
File  Edit  Format  View  Help
English Production Rules

<S> => <NP><VP>
<VP> => <V><NP>
<NP> => <ADJP><N>
<ADJP> => <ADJ>
<NP> => <DET><N>
<NP> => <N>
<NP> => <PRN>

Yorùbá Production Rules
<S> => <NP><VP>
<VP> => <V><NP>
<NP> => <N><ADJP>
<ADJP> => <ADJ>
<NP> => <N><DET>
<NP> => <N>
<NP> => <PRN>
```

Fig 1: Re-write rules of the two languages

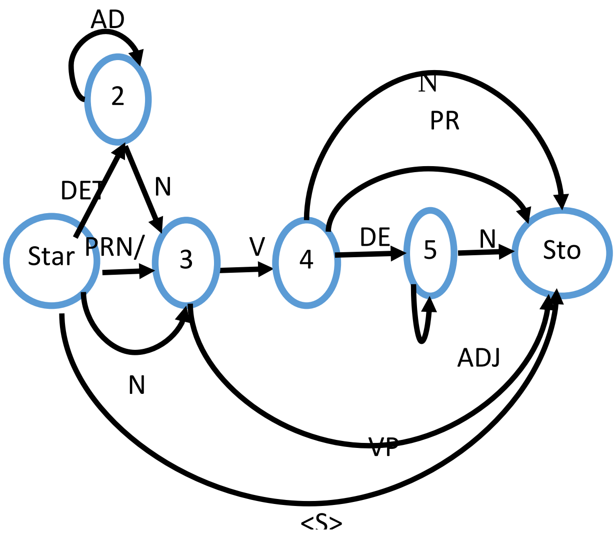


Fig 2: State diagram of the English Re-write rules

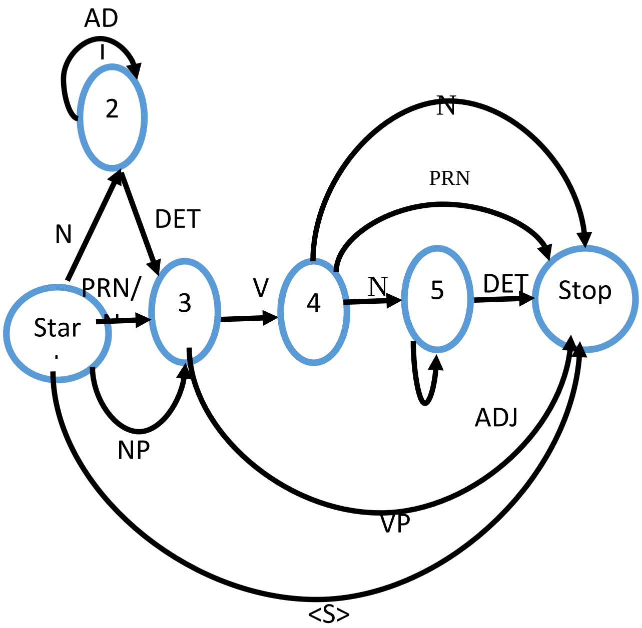


Fig 3: State diagram of the Yorùbá Re-write rules

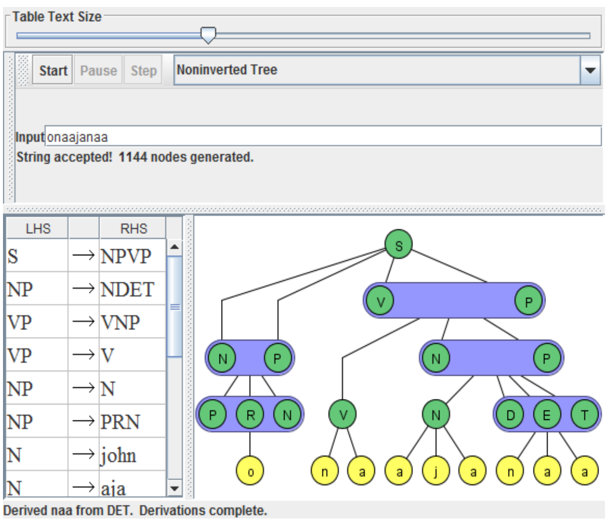


Fig 4: JFLAP sentence sample test I

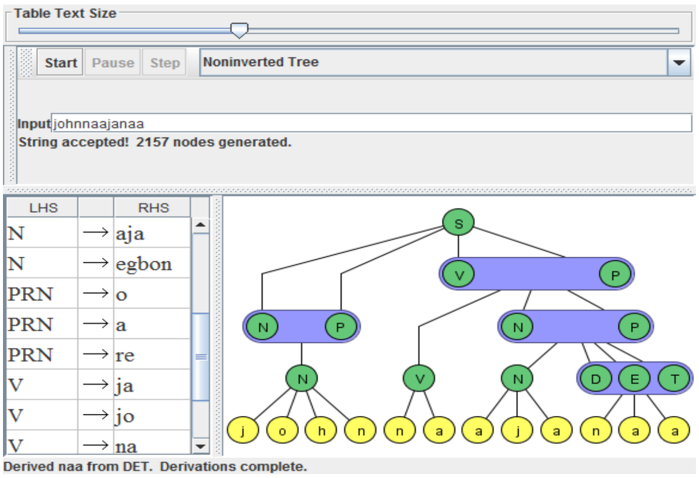


Fig 5: JFLAP sentence sample test II

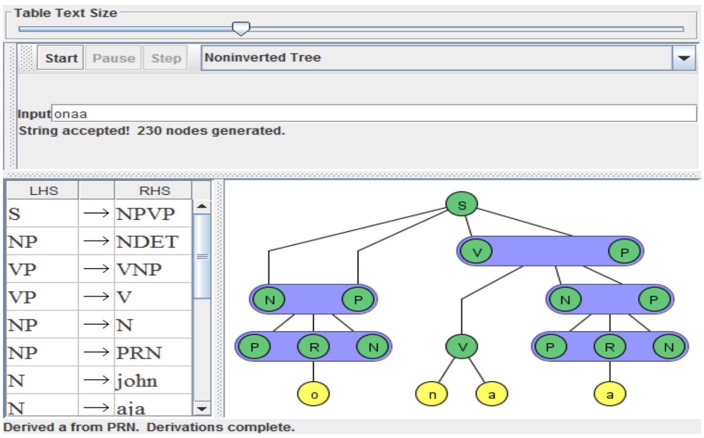
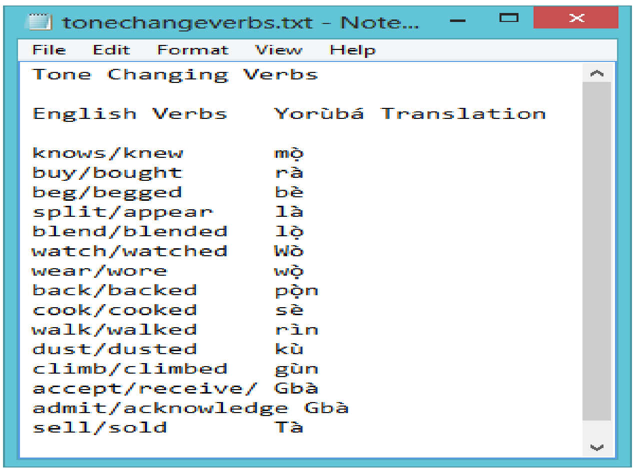


Fig 6: JFLAP sentence sample test III

4. DATABASE DESIGN

The database was designed by categorizing all the parts of speech into their different grammatical functions. This approach made it easy to access the database and update the content. Figs. 7 – 10 describe different parts of speech that were represented.



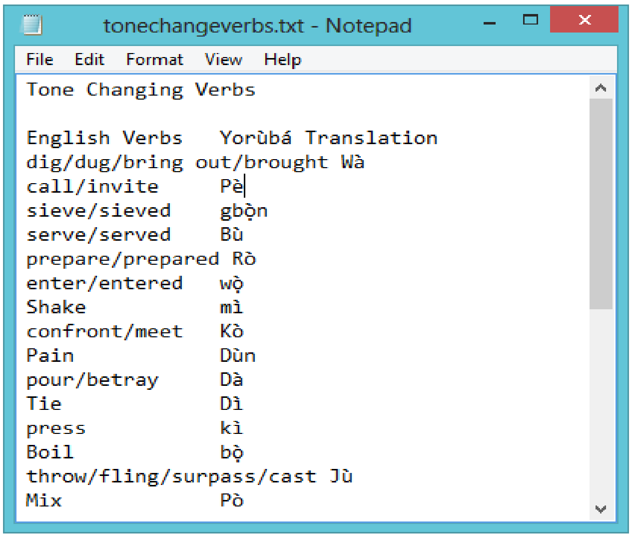
tonechangeverbs.txt - Note...

File Edit Format View Help

Tone Changing Verbs

English Verbs	Yorùbá Translation
knows/knew	mọ̀
buy/bought	rà
beg/begged	bẹ̀
split/appear	là
blend/blended	lọ̀
watch/watched	Wò
wear/wore	wọ̀
back/backed	pọ̀n
cook/cooked	sè
walk/walked	rìn
dust/dusted	kù
climb/climbed	gùn
accept/receive/	Gbà
admit/acknowledge	Gbà
sell/sold	Tà

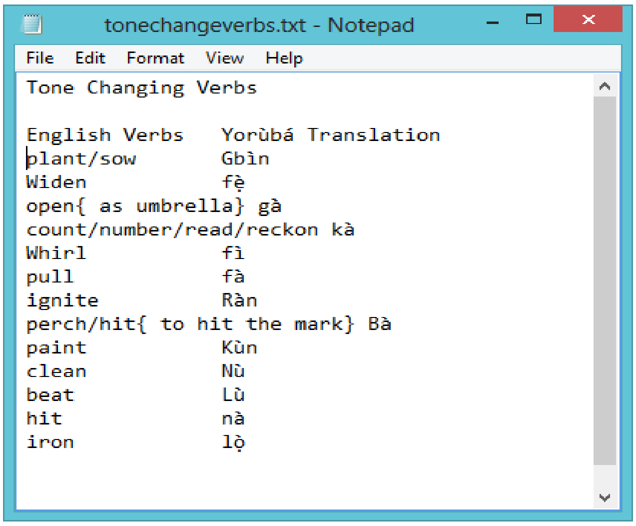
Fig 7a: Sample of Tone changing verbs


tonechangeverbs.txt - Notepad

Tone Changing Verbs

English Verbs	Yorùbá Translation
dig/dug/bring out/brought	Wà
call/invite	Pè
sieve/sieved	gbọ̀n
serve/served	Bù
prepare/prepared	Rò
enter/entered	wọ̀
Shake	mì
confront/meet	Kò
Pain	Dùn
pour/betray	Dà
Tie	Dì
press	kì
Boil	bọ̀
throw/fling/surpass/cast	Jù
Mix	Pò

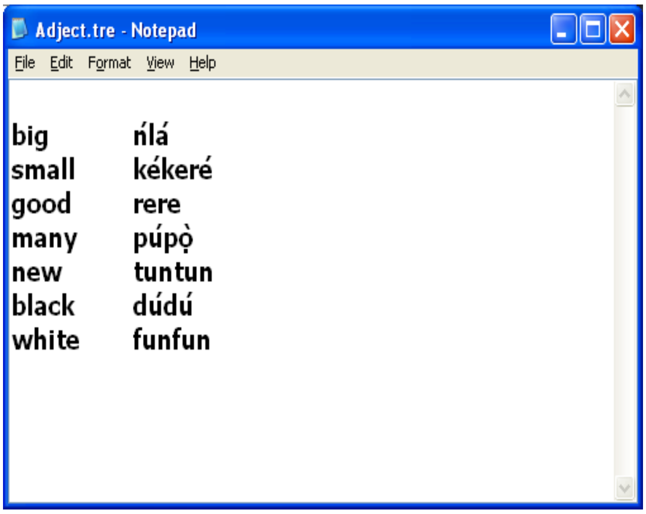
Fig 7b: Sample of Tone changing verbs


tonechangeverbs.txt - Notepad

Tone Changing Verbs

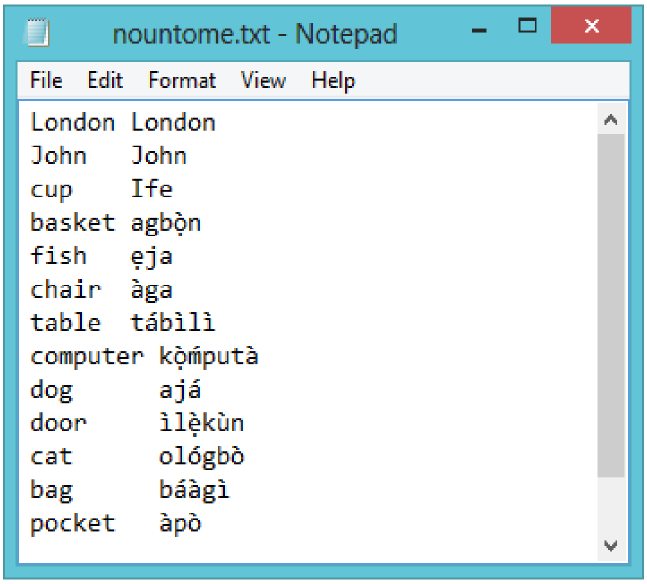
English Verbs	Yorùbá Translation
plant/sow	Gbìn
Widen	fẹ̀
open{ as umbrella}	gà
count/number/read/reckon	kà
Whirl	fì
pull	fà
ignite	Ràn
perch/hit{ to hit the mark}	Bà
paint	Kùn
clean	Nù
beat	Lù
hit	nà
iron	lọ̀

Fig 7c: Sample of Tone changing verbs


Adject.tre - Notepad

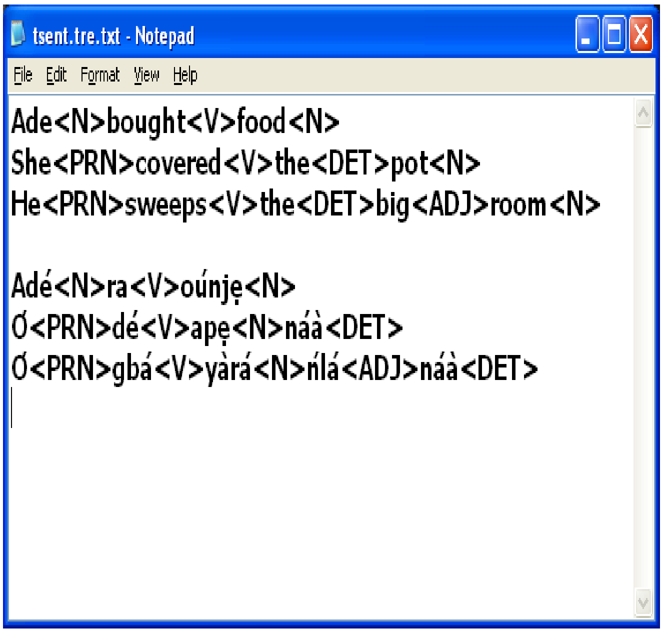
big	ńlá
small	kékeré
good	rere
many	púpọ̀
new	tuntun
black	dúdú
white	funfun

Fig 8: Lists of Adjectives


nountome.txt - Notepad

London	London
John	John
cup	Ife
basket	agbọ̀n
fish	ẹja
chair	àga
table	tábìlì
computer	kọ̀mpútà
dog	ajá
door	ìlẹ̀kùn
cat	ológbò
bag	báàgì
pocket	àpò

Fig 9: Lists of Nouns


tsent.tre.txt - Notepad

Ade<N>bought<V>food<N>
She<PRN>covered<V>the<DET>pot<N>
He<PRN>sweeps<V>the<DET>big<ADJ>room<N>

Adé<N>ra<V>oúnjẹ<N>
Ó<PRN>dé<V>apẹ<N>náà<DET>
Ó<PRN>gbá<V>yàrá<N>ńlá<ADJ>náà<DET>

Fig 10: Sample of tagged sentences

5. SOFTWARE DESIGN

Fig. 11 is the sequence diagram used for the design of the software. Essentially there are four modules. The system graphical user interface (GUI), translator, database, and parser. The GUI is meant to display what the user enter as inputs in the first plane, the second plane is the word for word translation and the third plane is the *Yorùbá* translation. The translator is the module that coordinates all the modules. The database module store the data used for the translation. The parser test all the rules used as it affects each sentence. Fig. 12 explains the steps that can be taken while using the system in line with what operations of fig. 12 specified.

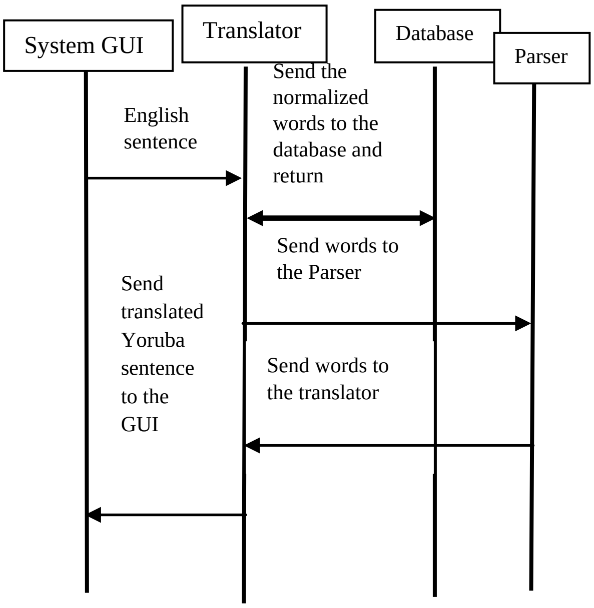


Fig 11: Sequence diagram

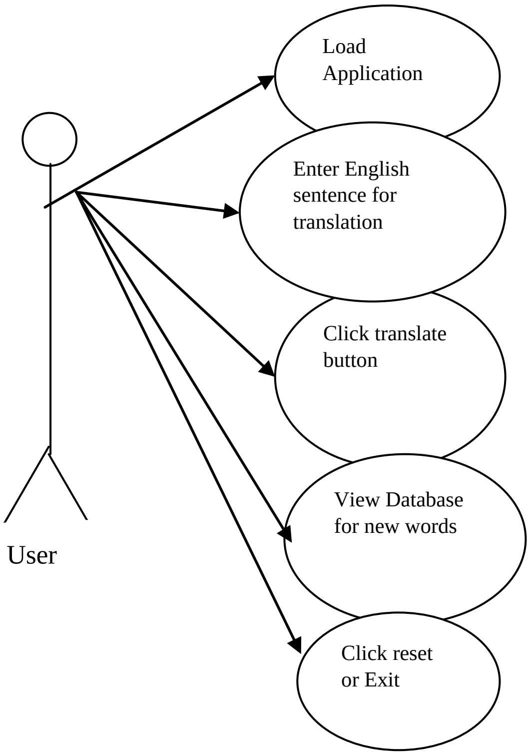


Fig 12: Use case diagram

6. SYSTEM AND SOFTWARE IMPLEMENTATION

Fig. 13 shows the GUI of the software. The user can type the sentence in the first plane and press the translate button. The word for word and Yoruba translated sentences being displayed in Fig. 14.

Fig 13: Sample I (input/output)

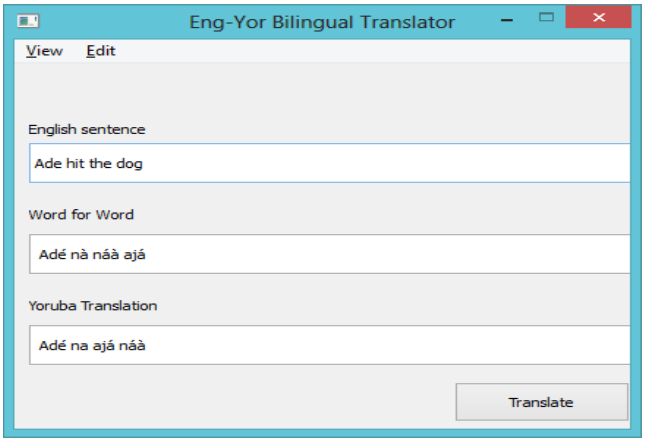


Fig 14: Sample II (input/output)

7. RESULTS AND DISCUSSION

Table II and III show the results of the comparison between the Google translator and IFEMT2. It is obvious that IFEMT2 do better in the translation of those sentences. The Yorùbá Language professional can judge the two translators. The words *lu* and *na* are the same in this context. Fig. 15 Shows Google translated sentences samples.

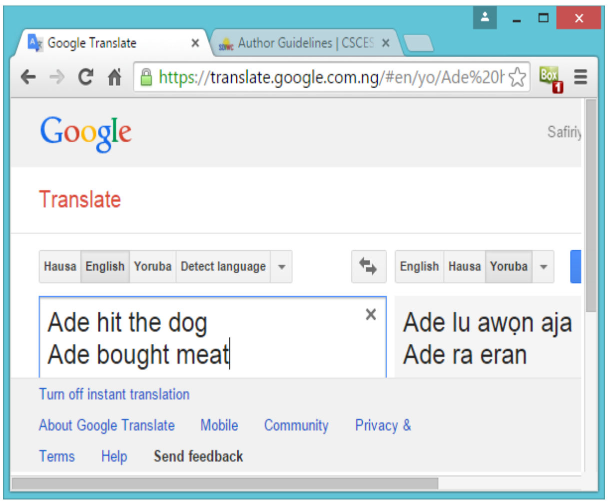


Fig 15: Google translated sentences

Table 2. Sentences used for the translation (Google Translator)

English Sentences	Google Translations
Ade hit the dog.	Ade lu awọn aja.
Ade hit dog.	Ade lu aja.
He hit it.	O si lu o.
Ola bought the yam.	Ola rà ni iṣu.
Ola bought yam.	Ola rà iṣu.
Lola knew the house.	Lọla mọ ile.

Table 3. Sentences used for the translation (IFEMT2)

Ade hit the dog.	Adé na ajá náà
Ade hit dog.	Adé na ajá
He hit it.	Ó nà á
Ola bought the yam.	Ọlá ra iṣu náà
Ola bought yam.	Ọlá ra iṣu
Lola knew the house.	Lọlá mọ ilé náà

8. CONCLUSION

There are many concepts in Yorùbá Language that general grammar rules cannot work for. Theory and the system developed can conveniently address the problem. This work will be integrated with English to Yorùbá Machine Translation System which is on-going project. The issue of serial verbs will be addressed within the context of machine translation.

9. REFERENCES

[1] NPC (2006). National Population Commission, 2006 Census. url: www.population.gov.ng, (Accessed: 25/06/2012).

[2] L .O. Adewole, The Categorial Status and the Functions of the Yoruba Auxiliary Verbs with Some Structural Analysis in GPSG. PhD thesis, University of Edinburgh, Edinburgh, 1988.

[3] A. Bamgbose, A Grammar of Yoruba. Cambridge University Press, London, 1966.

[4] D. Cornell, and D. R. Ladd, (1990). "Aspects of Pitch Realisation in Yoruba". Phonology 1990.

[5] L. O. Adewole, The Yoruba Language: Published Works and Doctoral Dissertations. Helmut Buske, Hamburg, 1987.

[6] Weaver, Translation, in Machine Translation of Languages. MIT Press, Cambridge, 1955.

[7] A. Lopez, "Hierarchical Phrase-Based Translation with Suffix Arrays", In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CNLL, Association for ComputationalLinguistics.url: http://www.aclweb.org/anthology/D/D07/D07-1104, (Accessed: 23/06/2011), 2007, pp., 976-985.

[8] F. Bond, "Introduction to Machine Translation". url: www.cs.mu.oz.au/research/it/nlp06/materials/Bond/mt-intro.pdf, (Accessed: 02/03/2011), 2006.

[9] M. Osborne, MT History and Rule-Based Systems. School of Informatics, University of Edinburgh, 2012.

[10] A. Boretz, AppTek Launches Hybrid Machine Translation Software. Magazine, Speech Technology. url: http://www.speechtechmag.com/Articles/News/News-eature/AppTek-Launches-Hybrid-Machine-Translation-Software-52871.aspx (Accessed: 04/08/2011), 2009.

[11] E. M. YeaseenurRahmanTahin, E. M. (2010). "Bootstrapping of a Rule-Based English, Bangla Machine Translation System Using Work Done for a Sister Language", Master's thesis, BRAC University, School of Engineering and Computer Science, 2010.

[12] S. I. Eludiora, T. Okunola and O. A. Odejobi, Computational Modeling of Yorùbá Split-Verbs for English to Yorùbá Machine Translation System , published in International Journal of Advanced Research in Computer Science and Applications, Volume 3, Issue 4, April, 2015, pp. 1-12.

[13] S. I. Eludiora, Development of English to Yorùbá Machine Translation System, unpublished Ph.D Thesis, Obafemi Awolowo University, Ile-Ife, Nigeria, 2014, pp 1- 241.