Development of a Verb Group Machine Translation System

Safiriyu Eludiora,
Gabriel Elufidodo
Department of Computer
Science & Engineering,
Obafemi Awolowo University,
Ile-Ife, Nigeria

ABSTRACT
The study reported in this paper considered the translation of English language verbs' group to Yorùbá language verbs' group. The study considered the verb group issue among different issues that affect English to Yorùbá machine translation (EYMT) system. The EYMT is a project that started some years back. The EYMT project was experimented and then raised a lot of issues that raise questions. The Yorùbá language extinction is of concern to the speakers and researchers. The total dominance of English language over Yorùbá language in almost all human endeavours is a major challenge. The linguistic rules and the automata theory are considered for the elicitation of the theoretical framework. The re-write rules were designed for the two languages. The Unified Modelling Language (UML) were used to design the system software, and python programming language was used for the system implementation. The evaluation was carried out using mean opinion score approach. The expert average was 100 percent and that of the experimental subject respondents was 81 percent while that of the developed system was 95 percent.

Keywords
Verb group, machine translation, Yorùbá language, re-write rules, acronyms

1. INTRODUCTION
Translation has always been understood to refer to a written transfer of a message or

meaning from one language to another. It refers to the process and result of transferring a text from a source language into a target language [1]. Translation can also be described as the transfer of the meaning of a text from the source language (SL) to the target language (TL). This implies that translation does not mean direct substitution of the word(s) from the source language to the target language but the translated word must convey the same meaning in the target language (TL) with meaning in the source language (SL) [2]. This means that the machine translator developer must understand the grammar of the two languages in order to convey the meaning of the translation.

"Translation is a linguistic process between languages and any theory of translation must derive from a theory of language" [3]. The system is a uni-direction in which the source language is translated into the target language.

There are two types of translation which are:

The full translation and the partial translation. In full translation, the entire text is translated from the source language to the target language text, example is given below:

1. Ade goes to market

1a *Adé lọ sí ọjà,*

2. She is coming –

2a *Ó ń bọ̀*

In partial translation, only few words from the source language are translated to the target language. The need for translation from the English language to a Yorùbá language is becoming paramount. The development of a Machine translation system has helped in reducing the problem of language barrier.

They are three major MT approaches: Data-driven, Hybrid, and Rule-based approaches. Literatures provide information about the strengths and weaknesses of each approach [4].

1.1 Yorùbá Language and Culture
Yorùbá language is one of the official languages spoken in Nigeria with over 30 million speakers in the south-western part of the country [5]. After a thorough research, it has been discovered that there is insufficient parallel English – Yorùbá corpus, hence English – Yorùbá statistical machine translator is not common (probably Yorùbá Google translator). There are basically three indigenous languages in Nigeria, they are the Hausa language spoken IN the northern part of Nigeria, the Igbo is spoken by the Eastern part of the country and the Yorùbá which is spoken in the south-western part of Nigeria [6].

The English language is the official language use in communication in Nigeria and it becomes the language of debate and record in spite of the use of major indigenous Nigerian languages [7].

The Yorùbá language (target language) is a tonal language spoken by people of the south- western part of Nigeria, which covers states like Ọ̀yọ́, Ọ̀sun, Ògùn, Òndo, Èkìtì, Lagos, Kogi and Kwara.

1.2 Translation of Verb Group
The verb group is the morphological unit which realizes the verb element in the sentence. The term "verb" refers to some classes of words with certain morph syntactic characteristics, one of which is their ability to function as elements of the verb group. It is formed as a result of a combination of two or more verbs which follows some rules in their combination. Verb group is formed in various ways. The pattern of formation is given below and the way it is translated into the target language (Yorùbá).

An auxiliary verb such as will, could, shall, and ought to etc combine with the lexical verb to form a verb group. ''The rule that the formation follows is that the modal auxiliary verb must come before the lexical verb''. Sometimes, preposition

"to" always follow the verb group formed. Below is an example that shows the formation of verb group consisting of a modal auxiliary verb and a lexical verb.

From the illustration given above, it is clear that in the sentence formation, it follows the **subject + verb group + object formation**. Table 1 shows some examples.

Table 1. Examples of English and Yorùbá verb group

English	Yorùbá
Ade will go to school	*Adé maa lọ sí ọjà*
He can jump the chair	*Ó lè fo àga*

1.3 Rules for verbs' group use in the Target Language

Rule 1: Noun can start the sentences in the verb group.

1. Ade will go to the market

1a. Adé *yóò lọ* sí ọjà

Rule 2: Pronoun cannot start sentence that begins with the third person singular Ó (She/He/It) yóò in the VBG.

3. She/he/it will/would/ go

2a. *Ó yóò lọ* (not correct), but *Ó maa lọ* (is correct)

Rule 3: Mo/I cannot be used with *yóò* but can be used with *maa*

4. I will/would go

3a. Mo yóò lọ (not correct), but *Mo maa lọ* (is correct)

Section 1 introduces the study, section 2 discusses the literature review. System design is described in section 3. Section 4 addresses software implementation and section 5 discusses results and discussion.

2. LITERATUR REVIEW

Translation processes for translating English ambiguous verbs are proposed by [8]. A machine translation system was developed for this purpose. Context-free grammar and phrase structure grammar were used. The rule-based approach was used for the translation process. The re-write rules were designed for the translation of the source language to the target language. The MT system was implemented and tested. For example, Ade saw the saw, *Adé rí ayùn náà* [8].

Ref [9] experiment the concept of *Yorùbá* verbs' tone changing. For instance, Ade entered the house, *Adé wọ ilé.* In this case, the dictionary meaning of enter in *Yorùbá* is *wọ̀*. This verb takes low tone, but in the sentence above it takes mid-tone. The authors designed different re-write rules that can address possible different *Yorùbá* verbs that share these characteristics. The machine translator was designed, implemented and tested. The system was tested with some sentences.

Ref [10] research on split verbs as one of the issues of English to *Yorùbá* machine translation system. The context-free grammars and phrase structure grammar were used for the modelling. Authors used rule-based approach and design re-write rules for the translation process. The re-write rules were meant for split-verbs' sentences only. The machine translator can translate split verbs sentences. For instance, Tolu cheated Taiwo, *Tolú rẹ́ Táíwò jẹ*.

Ref [11] propose the alternatives for the use of He/she/it => *Ó* of the third personal plural of English to *Yorùbá* machine translation system. *Yorùbá* language is not gender sensitive, authors observed the problem that does arise when the identity of the doer/speaker cannot be identified in the target language. Authors proposed different representations for he/she/it. *Kùnrin* was proposed for he, *Bìnrin* was proposed for her, and *ǹkan* was proposed for it.

Ref [12] propose a rule-based approach for English to *Yorùbá* Machine Translation System. There are three approaches to machine translation process. The authors reviewed these approaches and considered rule-based approaches for the translation process. According to Authors, there is limited corpus that is available for *Yorùbá* language this informs the rule-based approach.

3. SYSTEM THEORETICAL FRAMEWORK AND DESIGN

System theoretical framework, design, and database designs are considered in this section.

3.1 Theoretical Framework

In this section, the theoretical framework of the system was addressed. English and Yorùbá are languages that have similar sentence structure such as subject-verb-object (SVO) pattern (Eludiora, 2014). The English verbs are inflectional and Yorùbá verbs are non-inflectional. However, there are some syntactic similarities and differences.

The lists of Yorùbá acronyms used is shown in Table 2. The essence of the acronyms is to provide the equivalence of the phrases used in English in Yorùbá language.

3.2 System Design

The design architecture of the system is based on the architecture of a window-based application where it provides a link between the interface and the database. The system design considered all the principles and rules guiding the translation from the source language to the target language.

The design procedure is that the users are allowed to enter a text in the source language which is the English Language, the texts are broken into token (lexemes). The token is then patterned according to the re-write rules. The re-write rules are designed and developed using the automata theory provisions. The lexemes are fetched from the database. The outputs of the system are then displayed through the Graphical User Interface (GUI).

Table 2. Lists of Acronyms

English	**Yorùbá**
Sentence	Gbólóhùn (GB)
Noun Phrase (NP)	Àpólà ọ̀rọ̀ Orúkọ̀ (APỌO)
Prepositional Phrase (PP)	Àpólà ọ̀rọ̀ Atọ́kùn (APATK)
Verb Phrase (VP)	Àpólà ọ̀rọ̀ ìṣe (APỌI)
Adjectival Phrase (ADJP)	Àpólà Òrò Àpónlé (APỌA)
Prepositional (PRE)	ọ̀rọ̀ Atọ́kùn (ATK)

Noun (N)	Òrò Orúkọ (ỌO)
Pronoun (PRN)	Arópò Òrò Orúkọ (AỌO)
Adjective (ADJ)	Òrò Àpónlé (ỌA)
Determinant (DET)	Asàpéjúwe Ìlò ọ̀rọ̀ orúkọ (AIỌO)
Auxiliary (AUX)	Bẹ́ẹ́ ọ̀rọ̀ ìṣe (BỌI)
Verb (V)	ọ̀rọ̀ ìṣe (ỌI)
Verb Group (VBG)	àpapọ̀ ọ̀rọ̀ ìṣe (AỌI)

3.2.1 Grammar and Production Rule

A production rule is used to specify how a grammar transforms one string to another defining a language associated with the grammar. The designed rules provide the pattern in which the translation process is followed. Figures 1 and 2 explain possible transition models of the two languages. It is used to translate various sentences and an example is given in figure 3 and 4. The NLTK was used to test the re-write rules as shown in figures 5 and 6.

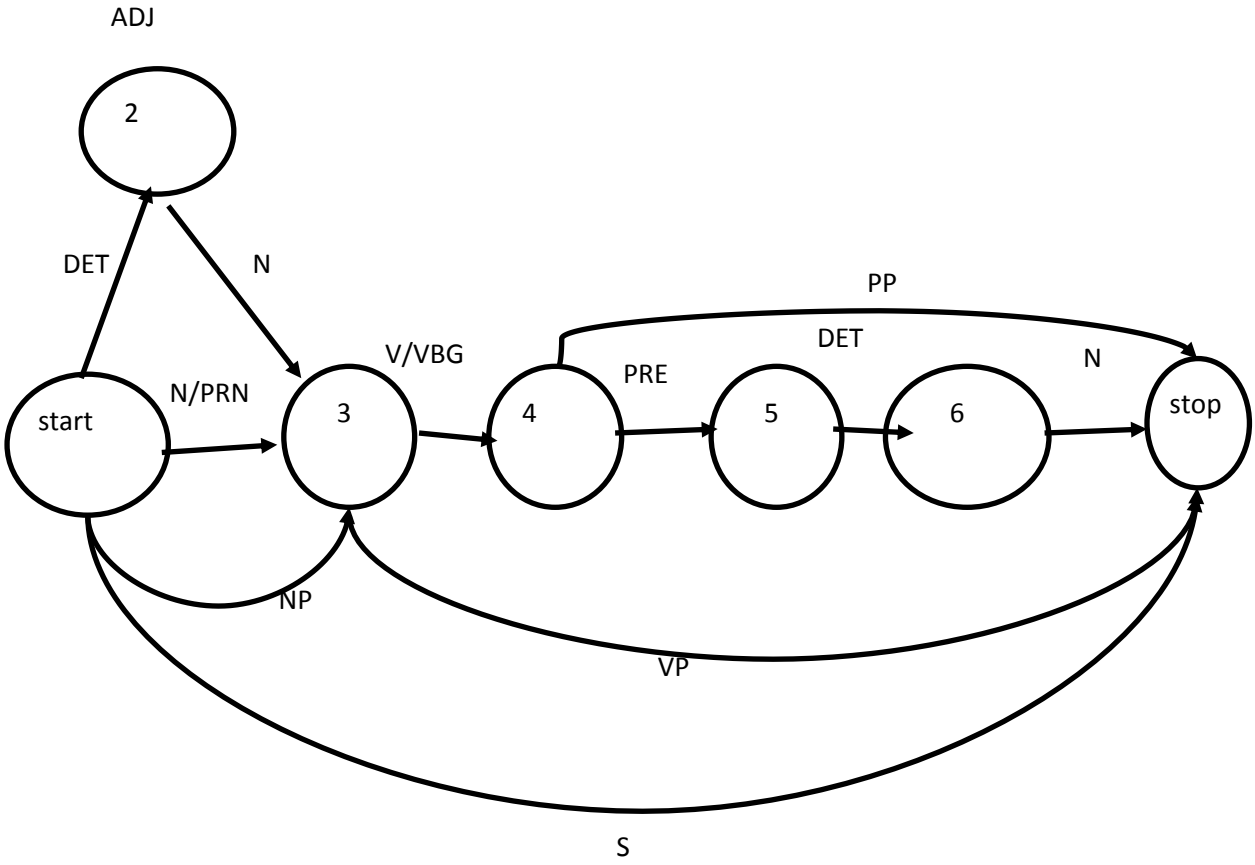


Fig 1: Transition model for English Verbs' group Sentence structure

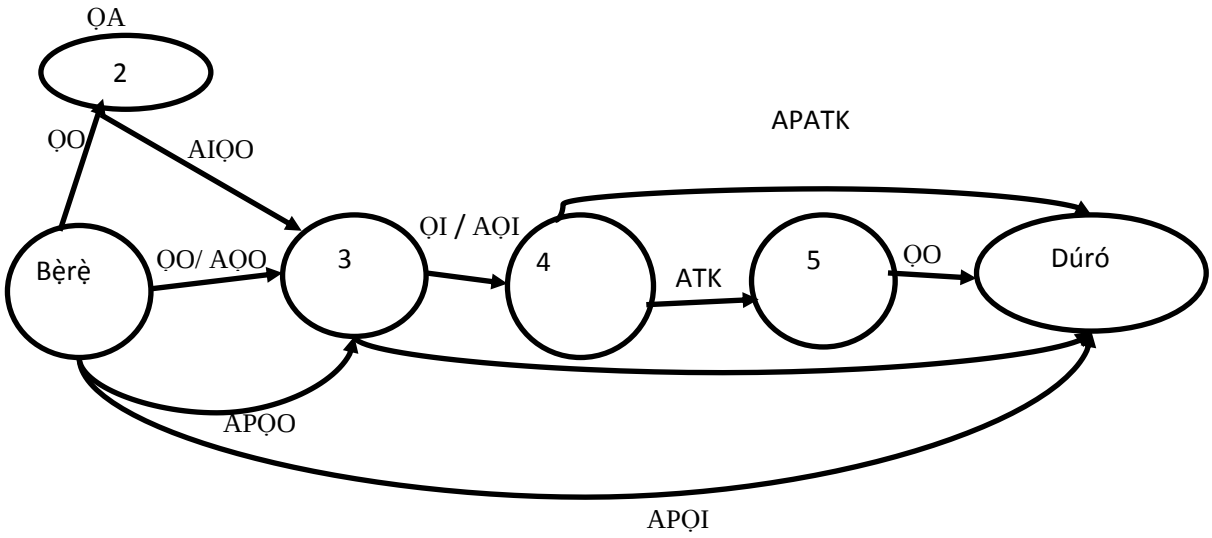


Fig 2: Transition model for Yorùbá Verbs' group Sentence structure

Ade can kill the Rat

S → NP VP
NP → DET NP
NP → ADJP NP
NP → N PRN
VP → V NP
VP → VBG PP
PP → PRE NP
N → Ade | Rat
VBG → can kill
DET → the

Fig 3: Translation rules for the English Sentence

Adé lè pa eku náà

Gbólóhùn (S) → APỌO APỌI
APỌO → APỌO AIỌO
APỌO → ỌO | AỌO
AỌI → BỌI ỌI
APỌI → ỌI
ỌO → Adé | Eku
BỌI → lè
ỌI → pa
AIỌO → náà

Fig 4: Translation rules for the Yorùbá Sentence

The mode of translation is based on the grammar designed for both English language and Yorùbá language. The parse trees in figures 5 and 6 explain the formation of English sentence and the word for word Yorùbá sentence and the real Yorùbá sentence which are generated using natural language tool kit (NLTK).

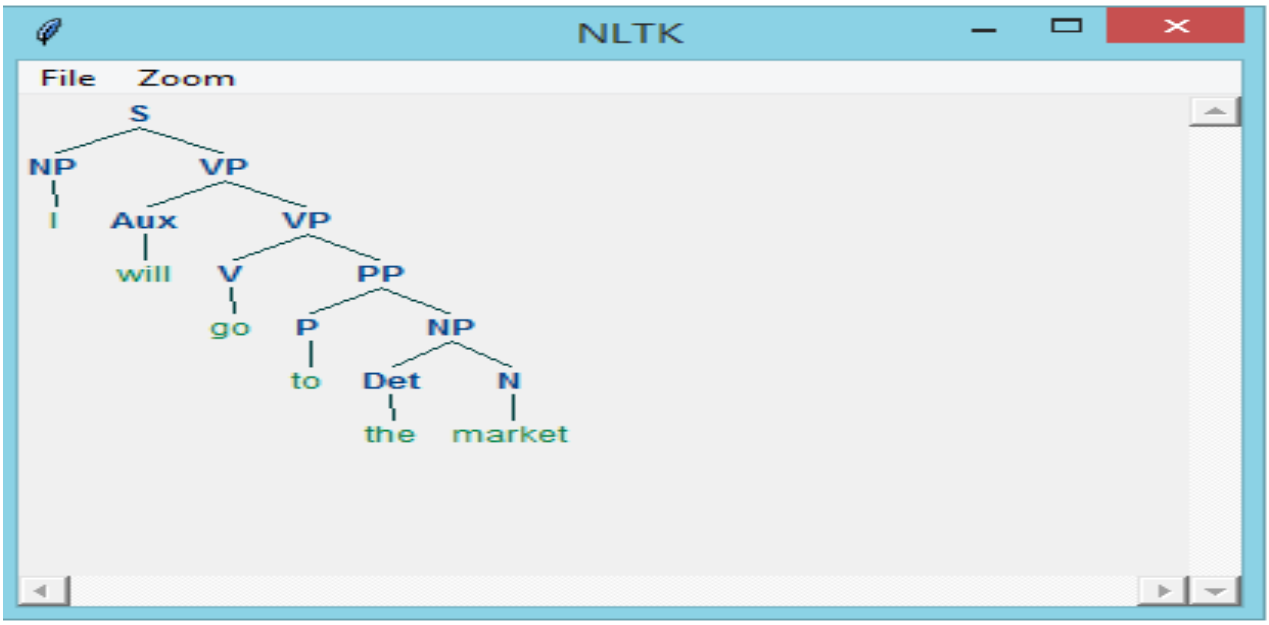


Fig 5: NLTK Parse tree: English verbs' group

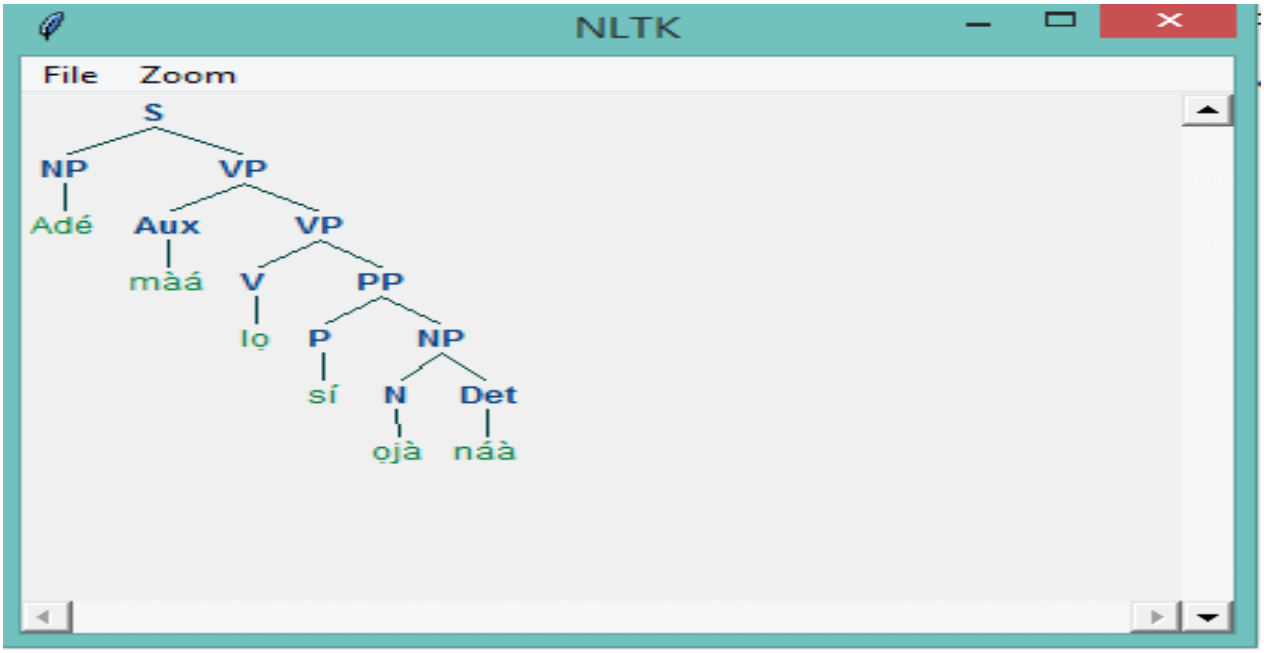


Fig 6: NLTK Parse tree: Yorùbá verbs' group

3.4 Database Design

In the process of acquiring data for the development of this system, various digital resources (Yorùbá corpus) were consulted. The Yorùbá language has a dearth of digital resources that can be used to develop the system library. The Internet, online and offline books, magazines and newspapers were consulted. The system was developed using home domain terminologies. People are familiar with these home domain lexical items (lexemes).

The database design for the work was designed using dictionary format. The English words are arranged with their respective equivalent Yorùbá words. These words are arranged according to their parts of speech. Samples of the database is shown in figures 7, 8, 9 and 10. Table 3 shows the verbs' group.

```
 1  Ade adé
 2  accident ìjàmbá
 3  ademola adémólá
 4  ajibola ajíbólá
 5  arm apá
 6  arms àwọn-apá
 7  awolowo awólówò
 8  ayo ayò
 9  baby ọmọ
10  bag báágì
11  bathroom ilé-ìwẹ̀
12  bayo báyò
13  bed ibùsùn
14  bed ibùsùn
15  beds àwon-ibusun
16  benjamin benjamin
17  benson benson
18  bicycle kèké
19  bicycles àwon-keke
20  biodun bíódún
21  biola bíólá
22  biology bàólógì
23  birds àwon-eye
24  birthday ojò-ìbí
25  blacksmith alágbèdẹ
26  bodies àwon-ara
27  body ara
```

Fig 7: Sample of Nouns' Database

```
1  accommodate|accommodates|accomodated gbà
2  ache|ached|aches ro
3  bathe|bathes wè
4  beat|beats|beat nà
5  break|breaks|broke fó
6  burn|burns|burnt jó
7  buy|buys|bought ra
8  call|calls|called pe
9  came|come|comes dé
10 catch|catches|caught mú
11 collect|collects|collected gbà
12 cook|cooked|cooks se
13 cried|cry|crys sokún
14 cut|cuts gé
15 dance|danced|dances jó
16 died kú
17 dig|digs|dug gbé
18 drive|drives|drove wà
19 eat|eats|ate je
20 faint|faints|fainted dákú
21 fetch|fetches|fetched pon
22 fight|fights|fought jà
23 find|finds|found ri
24 give|gives|gave fún
25 go|went|goes lo
26 graduate|graduates|graduated jáde
27 had|has|have ní
```

Fig 8: Sample of the database for the Verbs

```
1   average tóníwọ̀n
2   beautiful arewà
3   big nlá
4   brilliant tóló́yẹ
5   cheap olówó-póókú
6   cunning alágàbàgebè
7   delicious aládídùn
8   desolate ìkòsílè
9   dirty dídòtí
10  elegant tójáfáfá
11  experienced tónírìrí
12  fast tóyára
13  fat sísanra
14  flat pẹlẹbẹ
15  fresh òòjó
16  gentle oníwàpèlé
17  good dídára
18  handsome arewà
19  happy aláyọ̀
20  heavy tówúwo
21  hot gbígbóná
22  hungry ẹlébi
23  large títóbi
24  lazy ọ̀lẹ
25  light tófúyẹ́
26  little kékeré
27  many púpọ̀
```

Fig 9: Sample of the database for adjectives

```
1   a|an kan
2   her|his rẹ
3   my mi
4   our wa
5   that yen
6   the naa
7   their won
8   these wònyí
9   this èyí
10  those wònyen
11  your re
12
```

Fig 10: Sample of the database for pronouns

Table 3: Verb group

Modal Auxiliary verb	**Lexical verb**	**English Verb group**	**Yorùbá verbs group**
Will / would | yóò	Go | lọ	Will/would go	*Yóò lọ*
May/might | maa	Go | lọ	may/might go	*Màa lọ*
Can / could | le	Go | lọ	Can/could go	*lè lọ*
Shall / should | maa	Sing | kọrin	Shall/should go	*màá kọrin*
Ought to | yẹ	Go | lọ	Ought to go	*yẹ kó lọ*

3.5 System Software Design and Implementation

The software was designed in line with the translation process discussed in section three. Figure 11 shows the system sequence diagram. The software design is divided into different modules; the Graphical User Interface (GUI) which is designed using UML and implemented using python programming language. The GUI has three planes, the first plane is where user enters the sentences. The second plane display input sentences word for word. The third plane displays the translated Yorùbá sentences. After the sentences have been typed, the translator module of the code begins to execute. The sentence is broken into lexemes, it then tagged into different parts of speech.

The sequence diagram is used to depict the interaction between the object in a sequential order. The translator module will accept input sentence from the GUI module break it down and send it to the library or database module to confirm that the lexemes are in the database. However, if the lexemes are not in the library error message will be generated requesting the user to enter the correct lexemes. The final translated sentence is then displayed by the GUI. This is illustrated in figure 12. Python programming language was used in the software coding and the interface of the machine is designed using Tkinter. The lexemes are manually tagged and each word is categorised according to its parts of speech. The parser module was using the Natural Language Tool Kits (NLTKs).

The translation process is based on the phrase grammar rules built in the source code which implements the re-write rules. The machine translation system has the capability to translate sentences that contain a combination of two verbs which is referred to as “verb group” from the English Language to Yorùbá language in its textual form. These verb groups are usually the combination of one auxiliary and main verb.

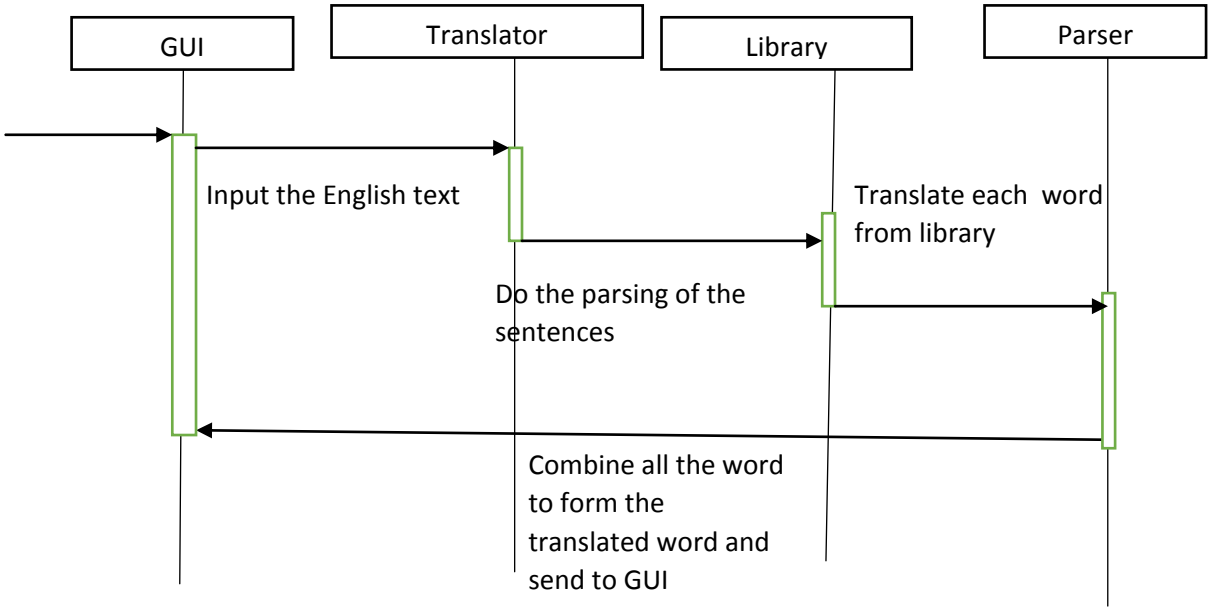


Fig 11 the system sequence diagram

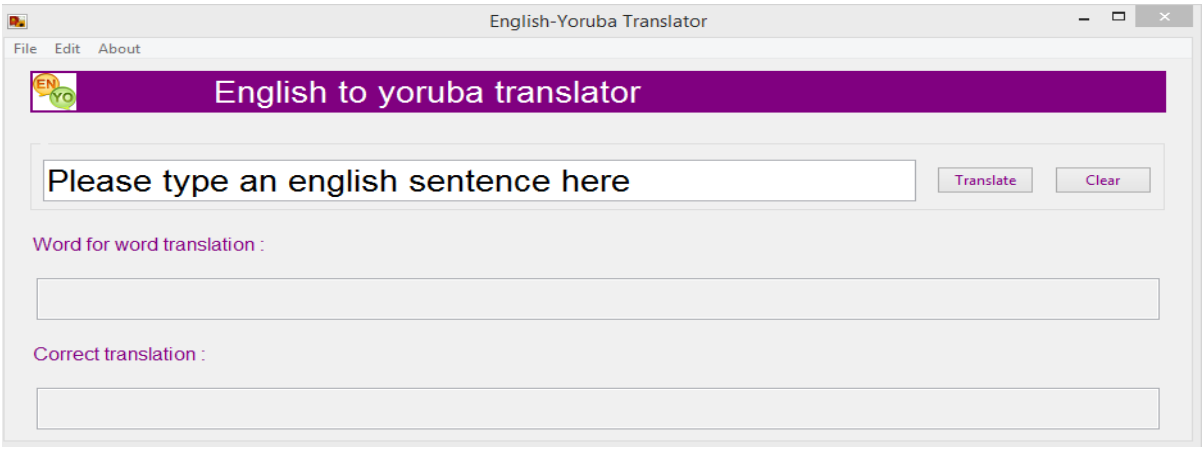


Fig 12: The Graphical User Interface of the System

3.5.1 System Output

The use case diagram shown in figure 13 describes how the user can use the system. The user loads the system and types the sentences he wants to translate. The system translates the sentence if there are no errors.

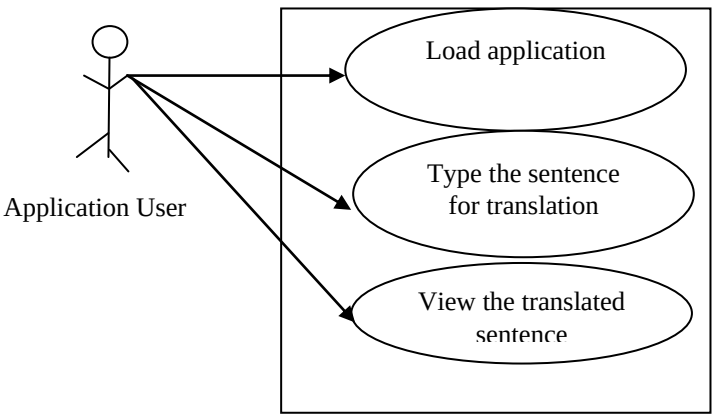


Fig 13: Use case diagram

3.5.2 System Sample Outputs

Figures 14 and 15 shows some of the sample outputs generated by the system.

Fig 14: The system translates "the boy will go to the market"

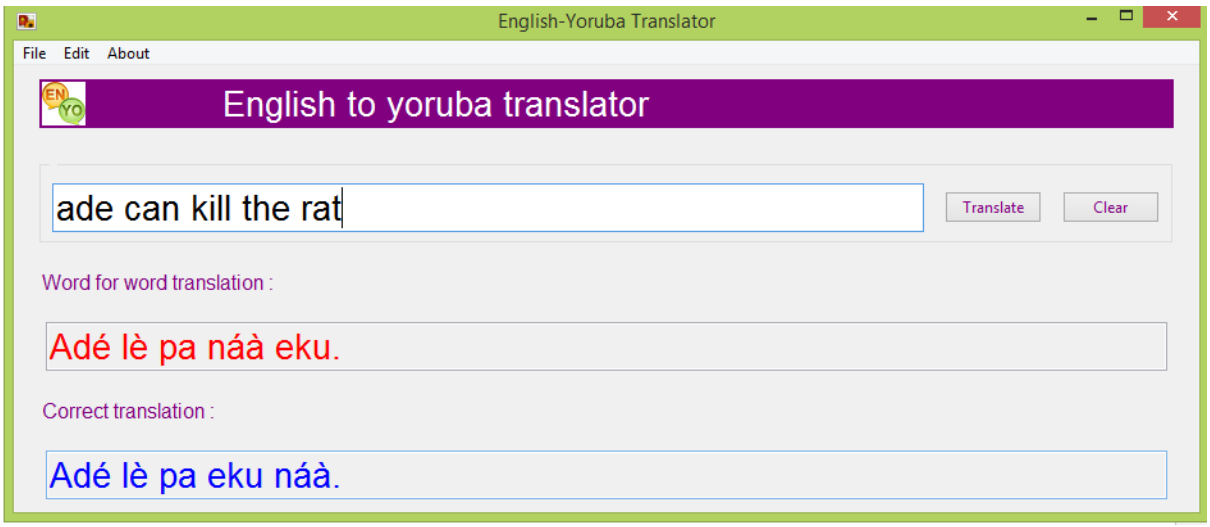


Fig15: The system translates *"ade can kill the rat"*

4 SYSTEM EVALUATION

In this section, system evaluation is discussed. The mean opinion score approach was used and questionnaires were designed and distributed. The experimental subject respondents submitted their view about the system.

4.1 The Mean Opinion Score

The Mean opinion score (MOS) is a subjective measurement of people's opinion. The Expert i.e. the professional translator translates the sentences from English language to Yorùbá language. The purpose of this evaluation is to compare the translations from machine and experimental subject respondents with the Expert translations.

4.2 Questionnaire Design

The questionnaire designed has simple sentences that consist of verb group which is designed to test the experimental subject respondent on the ability to translate simple sentences. The questionnaire has nine (9) simple sentences. The sentences in the questionnaire are meant to test the respondent translations' accuracy considering Yorùbá language orthography and the syntax of the language which is described in term of tone marks and diacritics (dotted vowels and consonant).

4.2.1 Questionnaire Administration

The questionnaires were administered in Obafemi Awolowo University Ile-Ife, Osun state, Nigeria. The environment was chosen because there are literate Yorùbá speakers and the

questionnaires were distributed among the Yorùbá speakers from the Yorùbá ethnic group.

4.3 Result and Discussion

The system was evaluated to determine the performance of the developed system. This shows the strength and weakness of the developed system in terms of the system accuracy using word orthography (tone marking and under dotting) accuracy. The reason is that majority of the experimental respondents got the translation, but many do not know how to tone mark words. This was carried out by comparing the Expert translated sentences with the one translated by the machine and the experimental respondents using the mean opinion score (MOS) technique to perform the evaluations. The result of the evaluation shown in figure 16. This is evaluation based on the word orthography which includes the tone mark and the under dots correctness.

It was observed from the graph plotted that the machine scores are higher than the average score of the experimental subject respondents and the expert has the highest.

Six sentences were used for the graph plotted. The results are shown in Table 4

Table 4. Evaluation Results Analysis

Sentences	Expert	Respondents Ave	Machine
1.	100	78	100
2.	100	78	90
3.	100	80	100
4.	100	90	100
5.	100	80	82
6.	100	85	100
Average	100	81.83	95.33

The expert average was 100 percent and that of the experimental subject respondents was 81 percent while that of the developed machine translator was 95 percent. The graph depicts that the machine correctness is close to that of the Expert and more accurate than that of the average experimental subject respondents.

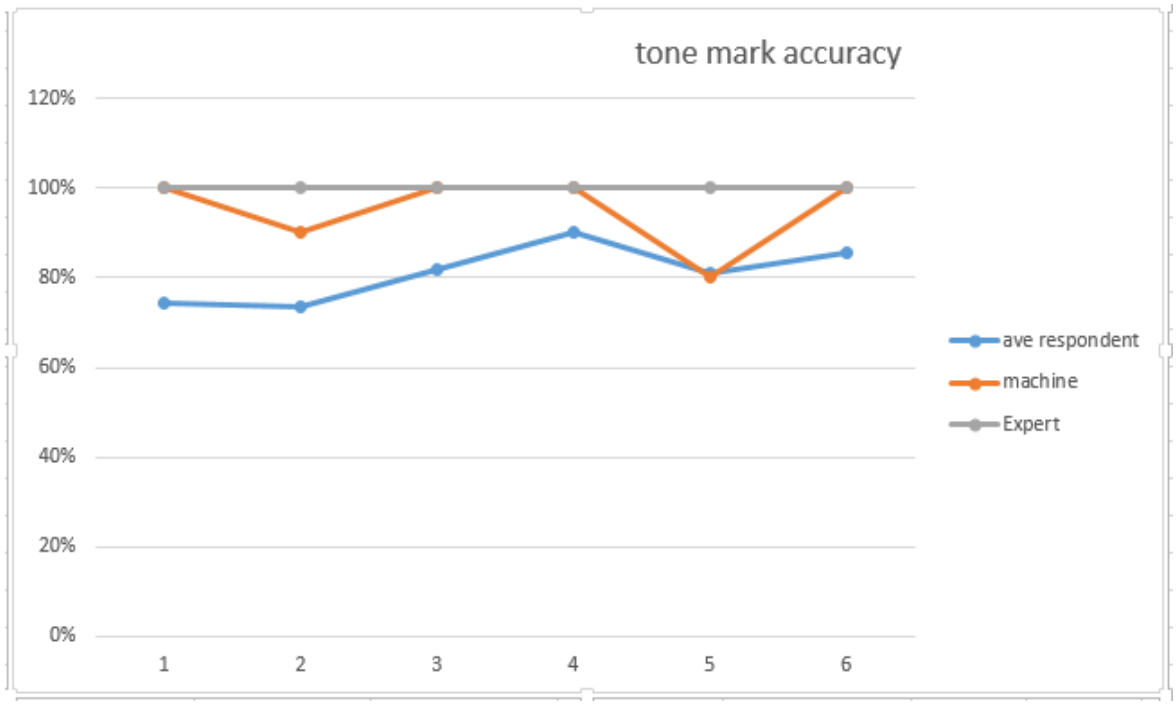


Fig 16: Translated sentence orthography accuracy

5 CONCLUSION

This developed English to Yorùbá verb group machine translator will contribute to the body of knowledge in the area of machine translation. This study was carried out to experiment the properties of verb group in the English to Yorùbá machine translation system perspective. The results gotten reflect that the people are not good at writing Yorùbá language again. This study also reflect the difference between speaking and writing, most especially in a tonal language like Yorùbá. The tone marks give the meaning of what one is writing. The study will be integrated with the on-going English to Yorùbá machine translation system project.

6 REFERENCES

[1] Bussmann, H. (1996). Routledge Dictionary of Language and Linguistics. Routledge, London.

[2] Koblomoje A. (2008) "Problem of Translation Yorùbá and English in Focus" B.Sc Thesis Department of Linguistics and Nigeria Languages, Faculty of Arts, university of Ado-Ekiti, Ekiti State, Nigeria.

[3] Catford, J. C. (1965) A Linguistics theory of translation, Oxford University Press, London.

[4] Eludiora S.I (2014), "Development of English to Yorùbá machine translation system" Unpublished thesis: Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife.

[5] National population commission, 2006 census URL: www.population.gov.ng (accessed: 25/06/2012).

[6] Yusuf, O. (2007) "Basic Linguistic for Nigeria Language Teachers" Published by Linguistic Association of Nigeria, Language and languages pg. 332.

[7] O. D. Ninan and Odetunji A. O (2013). Theoretical Issues in the Computational Modelling of Yorùbá Narratives. Workshop on Computational Models of Narrative 2013. Turkey

[8] Eludiora, S. I., Agbeyangi, A. O. and Ojediran, D. I. (2015a) Word Sense Disambiguation in English to Yorùbá Machine Translation System, Journal of Multidisciplinary Engineering Science and Technology, Berlin, Germany, vol 2, issue 7, 1814-1819.

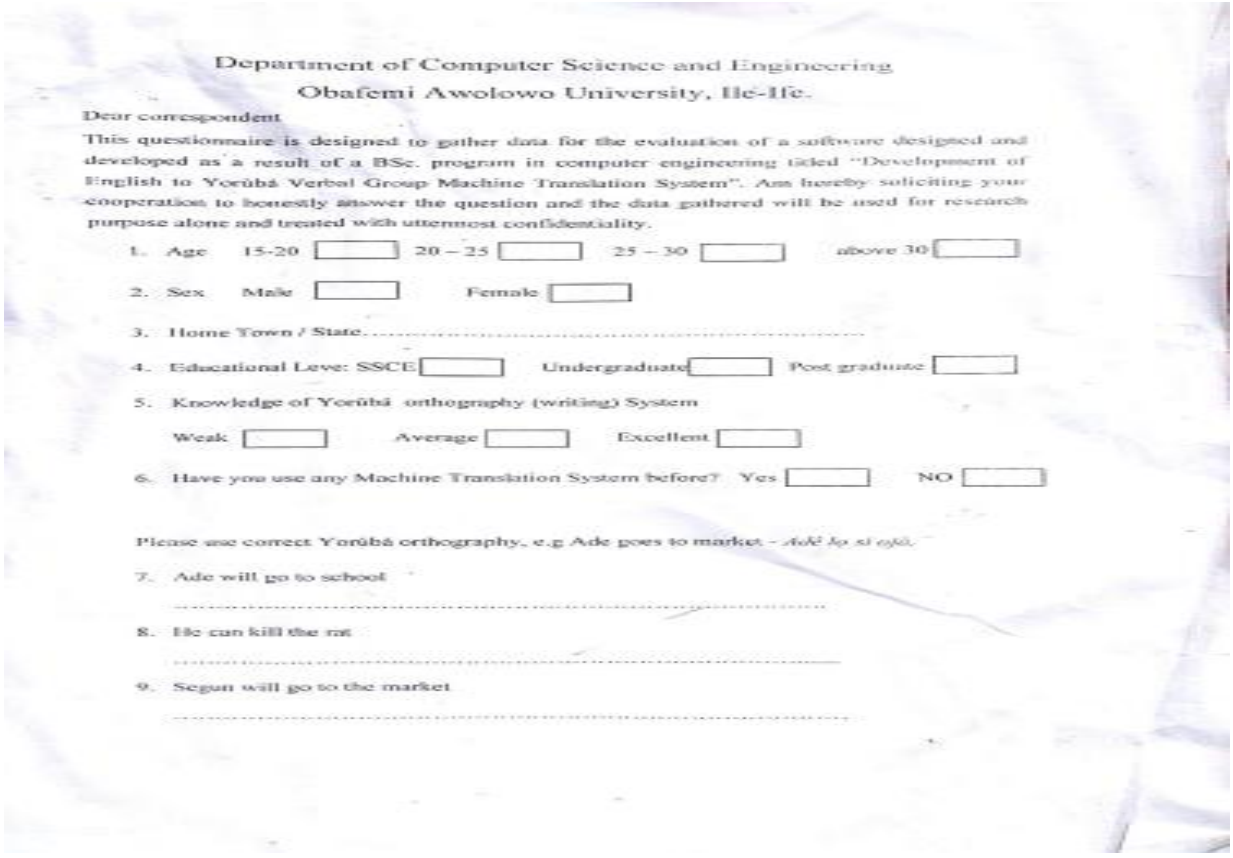
[9] Eludiora, S. I., Agbeyangi, A. O. and Fatunsin , A. (2015b) Development of an English to Yorùbá Machine Translation System for Yorùbá Verbs' Tone Changing , International Journal Computer Application, USA, vol 129, number 10, 12-17.

[10] Eludiora, S. I., Okunola, M. A and Odejobi, O.A. (2015c): Computational Modelling of Yorùbá Split-Verbs for English to Yorùbá Machine Translation System, *International Journal of Advanced Research in Computer Science and Applications,* Bangalore, vol (3), issue no (4), 1-12.

[11] Eludiora, S. I., Awoniyi, A. and Azeez, I. O. (2015d) Computational Modelling of Personal Pronouns for English to Yorùbá Machine Translation System, a paper presented at IEEE and The Science and Information Organisation (SAI) Intelligent Systems Conference 2015 (IntelliSys 2015) held in London, United Kingdom, November 10-11, 2015. 733-741

[12] Agbeyangi, A. O., Eludiora, S. I. and Adenekan, D. I. (2015) English to Yorùbá Machine Translation System using rule-based approach, Journal of Multidisciplinary Engineering Science and Technology, Berlin, Germany, vol 2, issue 8, 2275-2280.

7 APPENDIX



Department of Computer Science and Engineering
Obafemi Awolowo University, Ile-Ife.

Dear correspondent

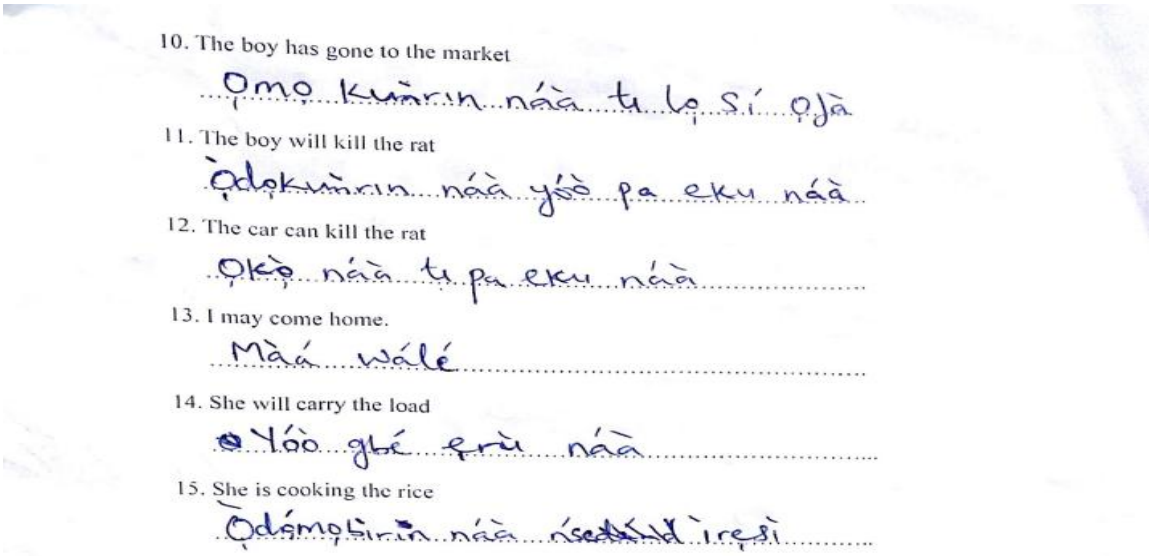
This questionnaire is designed to gather data for the evaluation of a software designed and developed as a result of a BSc. program in computer engineering titled "Development of English to Yorùbá Verbal Group Machine Translation System". Am hereby soliciting your cooperation to honestly answer the question and the data gathered will be used for research purpose alone and treated with uttermost confidentiality.

1. Age 15-20 [ ] 20 – 25 [ ] 25 – 30 [ ] above 30 [ ]
2. Sex Male [ ] Female [ ]
3. Home Town / State.....
4. Educational Leve: SSCE [ ] Undergraduate [ ] Post graduate [ ]
5. Knowledge of Yorùbá orthography (writing) System
   Weak [ ] Average [ ] Excellent [ ]
6. Have you use any Machine Translation System before? Yes [ ] NO [ ]

Please use correct Yorùbá orthography, e.g Ade goes to market - Adé lọ sí ọjà.

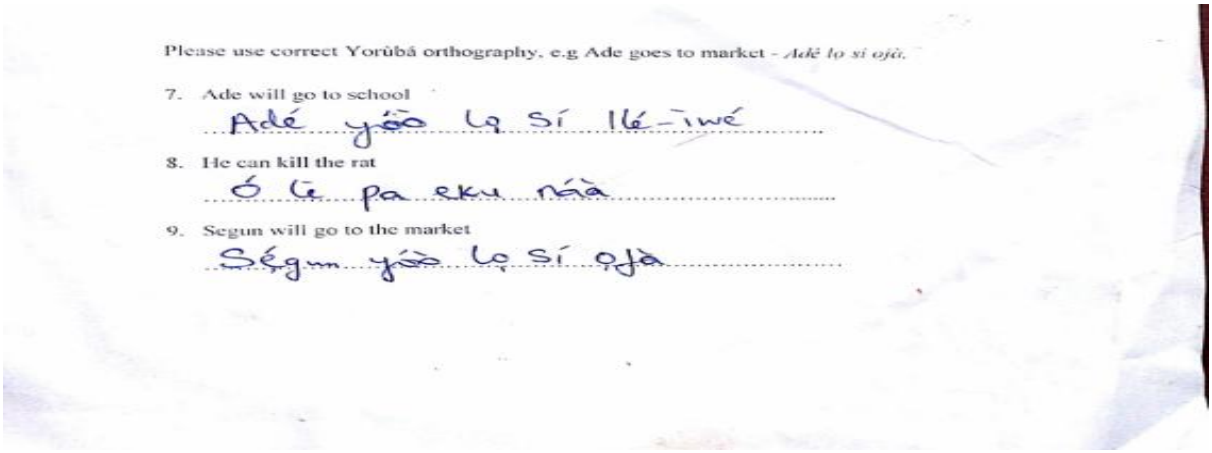
7. Ade will go to school
   .....
8. He can kill the rat
   .....
9. Segun will go to the market
   .....

Sample Questionnaire



10. The boy has gone to the market
Ọmọ kùnrin náà tí lọ sí ọjà
11. The boy will kill the rat
Ọdọkùnrin náà yóò pa eku náà
12. The car can kill the rat
Ọkọ náà tí pa eku náà
13. I may come home.
Màá wálé
14. She will carry the load
Yóò gbé ẹrù náà
15. She is cooking the rice
Ọdámọ̀sirin náà [illegible] ìrẹ̀sì

Figure A: sample of filled questionnaire



Please use correct Yorùbá orthography, e.g Ade goes to market - *Adé lo si ojà.*

7. Ade will go to school
Adé yóò lọ sí ilé-ìwé
8. He can kill the rat
Ó lè pa eku náà
9. Segun will go to the market
Ségun yóò lọ sí ọjà

Figure A1: sample of questionnaire