World Journal of Computer Application and Technology 4(4): 46-57, 2016 http://www.hrpub.org
DOI: 10.13189/wjcat.2016.040402

Development of a Prepositional Phrase Machine Translation System

Safiriyu Eludiora*, Ridwan Atolagbe

Department of Computer Science & Engineering, Obafemi Awolowo University, Nigeria

Abstract The study reported in this paper considered English to Yorùbá machine translation system for prepositional phrase. The prepositional phrase machine translation system is a subset of English to Yorùbá machine translation (EYMT) system. There are issues to address in English to Yorùbá machine translation system. Some of these issues are: serial verb, split verbs, noun phrase, verb phrase, numerals and prepositional phrase to mention few. The prepositional phrase (PP) plays a significant role in EYMT system because it describes the object position in a sentence. The two languages are subject verb object (SVO). In some sentences PP can be in subject and object positions. The object position was considered in the study reported in this paper. The theoretical framework was considered first. Therein the structure of PP was x-rayed and the translation process was modelled and designed. The UML was used to design the system. The flowchart, sequence, use case, and class diagrams were designed using the UML. A bilingual database (lexicon) was built to store words from the source language (English) and its equivalent target language (Yorùbá), and the system translation process model was implemented using the Java programming language. The developed system was tested and the sample outputs were compared with the Yorùbá Google translator outputs. The system performed better than the Yorùbá Google translator in terms of good orthography and syntax.

Keywords Prepositional Phrase, Bilingual Database, Yorùbá Language, Indigenous Languages, Translation Process

1. Introduction

Yorùbá language is spoken by over 40 million people within and outside Nigeria [1]. The Yorùbá language is spoken in Africa, Brazil, Cuba and other parts of the world. However, there is dominance of English language over the language in Nigeria. The other two indigenous languages are: Igbo and Hausa. The standard Yorùbá are taught in the schools (up to University level), used in Media (print and broadcast). The Yorùbá language is endangered and the Yorùbá cultural is gradually going into extinction. There is need for modern day processing tools for the language to catch up with the technological growth the world is experiencing. Again this will increase the audience and peoples' interest in the language.

Yorùbá language is a tonal language like other Nigerian languages. There are several dialects. Most of the African languages are tonal languages. Yorùbá language is a subject verb object (SVO) grammar structure like English language. In some aspects, Yorùbá language exhibits word swapping or re-ordering. Some phrases have these features (word swapping) while some do not. The noun phrase, adjectival phrase and the prepositional phrase have these features. For example, lori tabili naa, on the table. Yorùbá language is head first (tabili comes before the) while English is head last (table is the last word).

There is human language technology development, this area include Automatic Speech Recognition (ASR), Text-To-Speech (TTS) synthesis, Machine Translation (MT) and so on [2]. The machine translation (MT) uses comprehensive bilingual dictionaries to translate the source language (SL) text to its equivalent target language text. The application of linguistics theories, rules, and computer theories enable the development EYMT prepositional phrase.

This is in a bid to mitigate the extinction of Yorùbá language. As the dominance of the English language in Nigeria is quite overwhelming, thereby reducing the development of the major indigenous languages [3]. The aim is to develop a system that can translate English Prepositional phrase (PP) to its Yorùbá equivalent prepositional phrase (text). The prepositional phrase MT system is important because it provides important information about location, descriptions of people and things' positions, relationships, time, and ideas.

Section one introduces the study, related work are discussed in section two. The system theoretical frame work is presented in section three. Sections four and five explain the system design framework, and implementation. The

system results' discussions are presented in section six, and section seven concludes the study.

2. Review Work

"Ref [4]" identifies the difficulties of translating English prepositions (at, in, and on*)* which the Saudi Students were facing when translating them into Arabic. A survey consisting 50 Saudi English Foreign Language (EFL) students (25 males, and 25 females) was conducted. The result revealed that Saudi EFL students face problem-related to the translating of simple prepositions from English into Arabic. Significant differences relating to the performance of both males and females were recorded where females scored higher marks than the males. These findings suggested that acquired skills and abilities involved in translation appeared to be more strongly activated in the English-Arabic tasks in women as compared to the men.

"Ref [5]" proposes a phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bangla. A transliteration module was added to handle Out-Of-Vocabulary (OOV) words. This is especially useful for low-density languages like Bangla for which only a limited amount of training data is available. Furthermore, a special module handling translation of preposition words was implemented to treat systematic grammatical differences between English and Bangla. The improvement of the system was evaluated using the BLEU, NIST, and TER scores with the overall score of the system being 11.7 percent and for short sentences, which was 23.3 percent.

Translation processes for translating English to *Yorùbá* was proposed by [6]. The proposed machine translator can only translate simple sentences. Context-free grammar and phrase structure grammar were used. The rule-based approach was used for the translation processes. Re-write rules were designed for the translation of the source language to the target language [6].

"Ref [7]" experiments the concept of *Yorùbá* verbs' tone changing. For instance, *Adé wọ ilé* Ade entered the house. 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 [8]" did research on split verbs as one of the issues of English to *Yorùbá* machine translation system. The context-free grammars and phrase structure grammar was used for the modelling. Authors used rule-based approach and designed re-write rules for the translation process. The re-write rules are meant for split-verbs' sentences. The machine translator can translate split verbs sentences. For instance, Tolu cheated Taiwo, *Tolú rẹ́ Táíwò jẹ*.

"Ref [9]" proposes 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, the authors observed the problem that does arise when the identity of the doer/speaker cannot be identified in the target language. The Author proposed different representations for he/she/it. *Kùnrin* was proposed for he, *Bìnrin* was proposed for she, and *ǹkan* was proposed for it.

"Ref [10]" proposes a rule-based approach for English to *Yorùbá* Machine Translation System. There are three approaches to machine translation process. The author reviewed these approaches and considered rule-based approaches for the translation process. According to Author, there is limited corpus that is available for *Yorùbá* language this informs the rule-based approach.

"Ref [11]" proposes system that can assist in the teaching and learning of *Hausa*, *Igbo,* and *Yorùba*. The study considered body parts identification, plants, and animals' names. The English to *Yorùbá* machine translation and *Yorùbá* number counting systems were part of the main system. The model was designed to build a system for the learner of the Nigerian three indigenous languages. It is on-going research work.

"Ref [12]" propose web-based English to *Yorùbá* machine translation system. Authors considered a data-driven approach to design the translation process. Context-free grammar was considered for the grammar modelling. The *Yorùbá* language orthography was not properly considered in that study.

"Ref [13]" considers a hybrid approach to English to *Yorùbá* machine translation. The paper only itemised the steps the authors will take in the development of the proposed system. The study is on-going.

"Ref [14]" propose English to *Yorùbá* machine translation system for noun phrase. According to the authors, rule-based approach was used and automata theory was used to analysis the production rules. The system was able to translate some noun phrases. It was evaluated using Nigerian daily news and the system translation accuracy using some phrases was 90 percent.

3. Theoretical Framework

This section introduces the translation abstraction, the phrase grammar and the re-write rules.

Figure 1 shows the PP translation process abstraction, from the source language to intermediate translation to the target language translation. The PP is on the chair, *lórí (ní orí) àga náà*. At the intermediate level the PP was transcribed to word for word and the final translation required word re-ordering. The final translation is *lórí àga náà*.

3.1. Phrase Grammar and Re-write Rules

The English and Yorùbá write rules are illustrated below. The list of acronyms is in table 1. These the acronyms used to replace the English acronyms in the Yorùbá section. Figure is used to explain the context free grammar, meaning

that a sentence or phrase can be translated at the surface level without any hiding meaning. The phrase grammar is used to describe the relationship between the sentence or phrase constituents (words). English and Yorùbá sentence structures are presented in (1) and (2) below. The re-write rules explained the how phrases are derived from noun and verb phrases. The two phrases are realized from the sentence.

The re-write rules in (2) showed Yorùbá that is head-first in the Noun phrase (NP) structure while in English is head-last in a Noun phrase (NP) structure. For example, `the man' is (DetN) *okunrin naa*, that is, (NDet). Also in (5) Yorùbá is head-first in the Adjectival phrase (AdjP) structure while English is head-last in the Adjectival phrase (AdjP) structure For example, the tall boy is (DetAdjP), *omokunrin giga naa*, that is, (NAdjDet). The position of qualifier (tall) does not change in the two languages.

Rule 7 explains the prepositional phrase which is the focus of the study. That is, PP ===> PreNP. This is derived from the whole sentence, the PP re-write rules were designed for the two languages.

Table 1. Lists of Acronyms

English	Yorùbá
NP	Àpólà ọ̀rọ̀ Orúkọ̀ (APỌO)
PP	Àpólà ọ̀rọ̀ Atọ́kùn (APTK)
VP	Àpólà ọ̀rọ̀ ìṣe (APỌI)
ADJP	Àpólà Òrò Àpónlé (APỌA)
PRE	ọ̀rọ̀ Atọ́kùn (ATK)
N	Òrò Orúkọ (ỌO)
PRN	Arópò Òrò Orúkọ (AỌO)
ADJ	Òrò Àpónlé (ỌA)
DET	Asàpéjúwe Ìlò ọ̀rọ̀ orúkọ (AIỌO)

English Sentence Structure (1)

Rule 1 S ===> NPVP
Rule 2 NP ===> DetN
Rule 3 NP ===> DetAdjP
Rule 4 NP ===> PP
Rule 5 AdjP ===> AdjNP
Rule 6 VP ===> VNP
Rule 7 PP ===> PrepNP

Yorùbá Sentence Structure (2)

Rule 1 S ===> NPVP
Rule 2 NP ===> NDet
Rule 3 NP ===> NAdjP
Rule 4 NP ===> PP
Rule 5 AdjP ===> NAdj
Rule 6 VP ===> VNP
Rule 7 PP ===> PrepNP

The PP has six re-write rules each of the two languages as shown in (3) and (4) below. Rule 1 shows that PP is produced from noun phrase and PP can produce prepositional and noun phrase.

English Prepositional phrase structure (3)

Rule 1 NP =➔ PPNP
Rule 2 PP =➔ PRENP
Rule 3 NP =➔ ADJPNP
Rule 4 ADJP =➔ ADJNP
Rule 5 NP =➔ DETNP
Rule 6 NP =➔ N

Yorùbá Prepositional phrase structure (4)

Rule 1 NP =➔ PPNP
Rule 2 PP =➔ PRENP
Rule 3 NP =➔ NPADJP
Rule 4 ADJP =➔ NPADJ
Rule 5 NP =➔ NPDET
Rule 6 NP =➔ N

For example, on the chair, *lórí (ní orí) àga náà* where *ni* becomes *l*. This phrase can be tokenized as follows:

English on the chair

- PP::=Pre NP
- NP::=Det N
- PRE::=on
- DET::=the
- N::= chair

Yorùbá *[lórí àga náà]*

- APTK::=ATK APOO
- APOO::= AIỌO OO
- ATK::= orí
- AIỌO::= náà
- ỌO::= aga

3.1.1. Re-write Rules

The grammar rule is very crucial in the translation process of human languages. The other rules that guided this system's design include:

Rule 1: A prepositional phrase (PP) consists of a preposition and noun phrase (NP). In the case of target language noun (ỌO) comes before determiner (AIỌO). For example,

SL: on<Pre>the<DET>chair**<N>**.
TL: lorí<ATK>àga<ỌO>náà**<**AIỌO**>**

Rule 2: A determiner (if it does exist) must precede an adjective and a noun in SL, but reverse is the case in the TL. For example,

SL: The<DET> tall**<ADJ>** boy**<N>**.
TL: ọmọkùnrin**<**ỌO**>**gíga**<**ỌA**>**náà<AIỌO>.

3.1.2. Parse Tree

Figure 2 and 3 show the English and Yorùbá prepositional phrase structure of the on the chair, *lórí àga náà*. The parse trees show the pictorial view of the prepositional phrase. The essence of re-write rules is to provide structural view of the given PP before the real coding. The rules were tested using JFLAP. Also NLTK could be used as well.

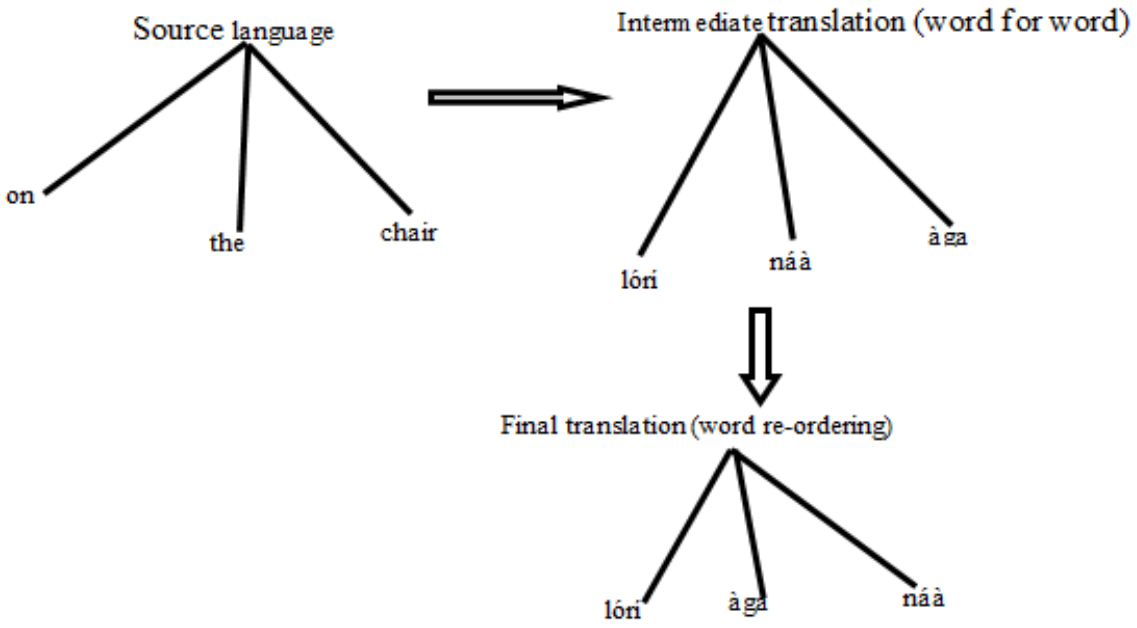

Figure 1. The prepositional phrase translation process abstraction

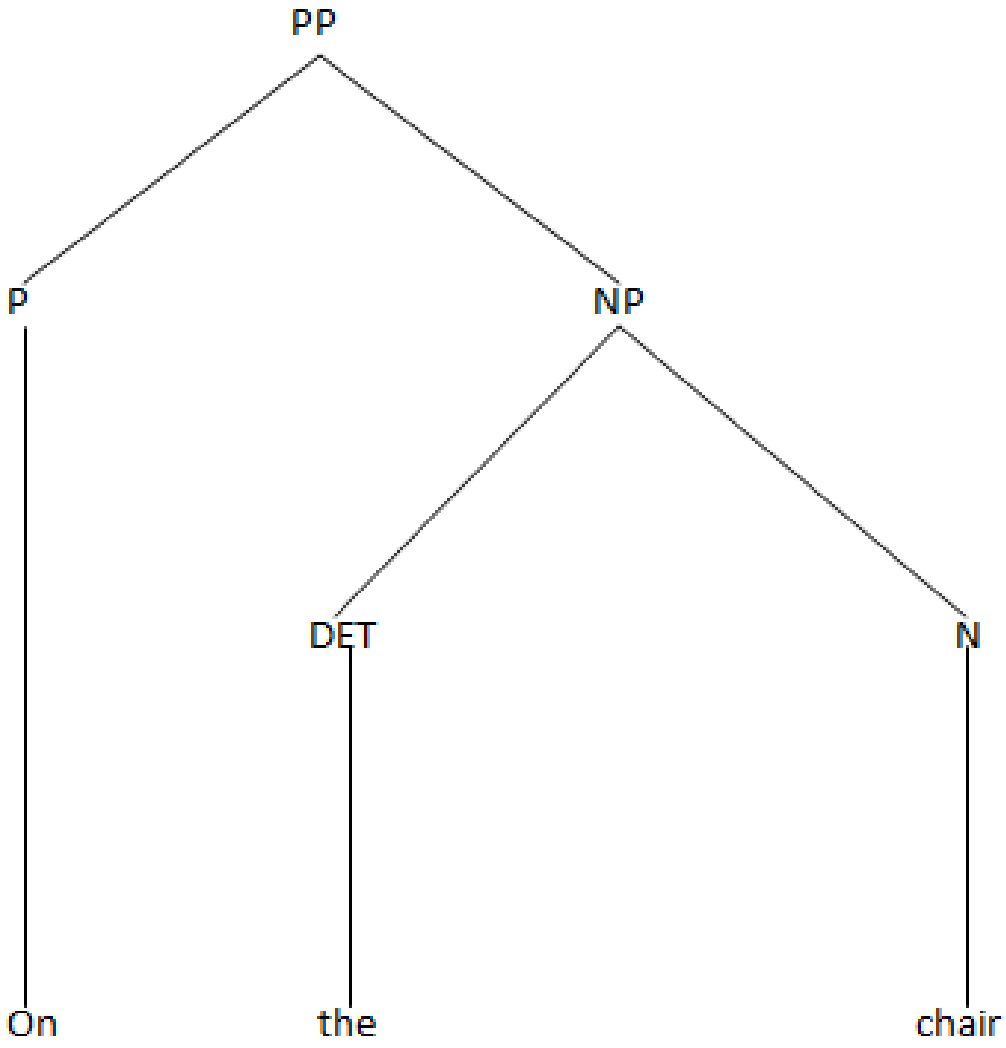

Figure 2. Parse tree for an English prepositional phrase

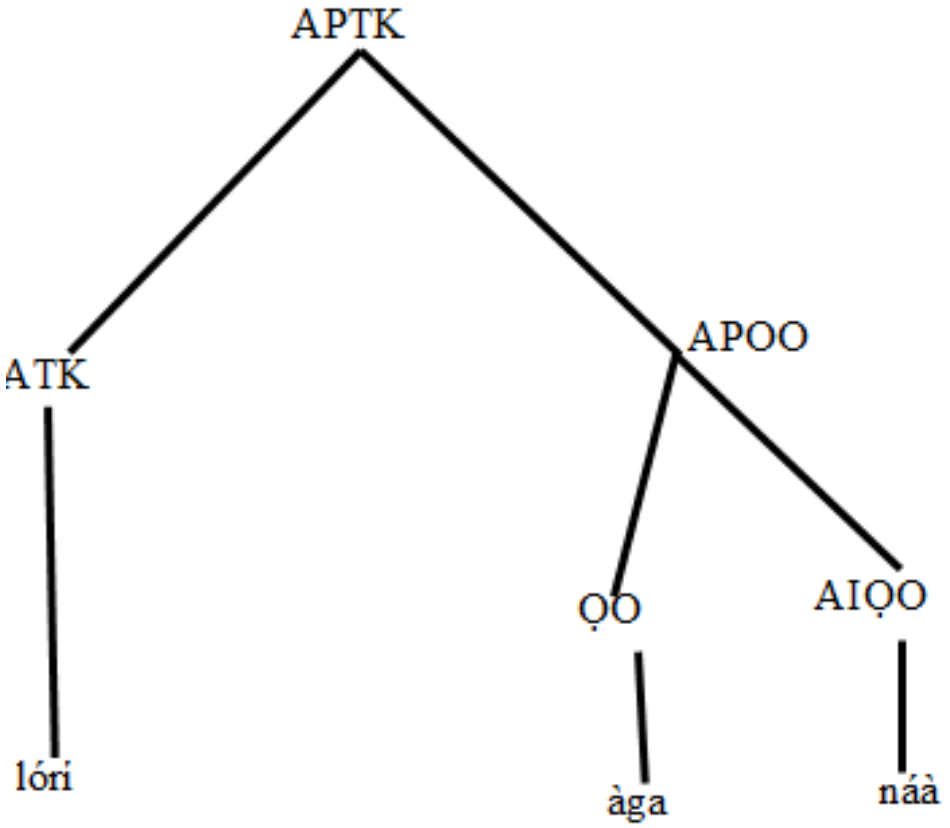

Figure 3. Parse tree for a Yorùbá prepositional phrase

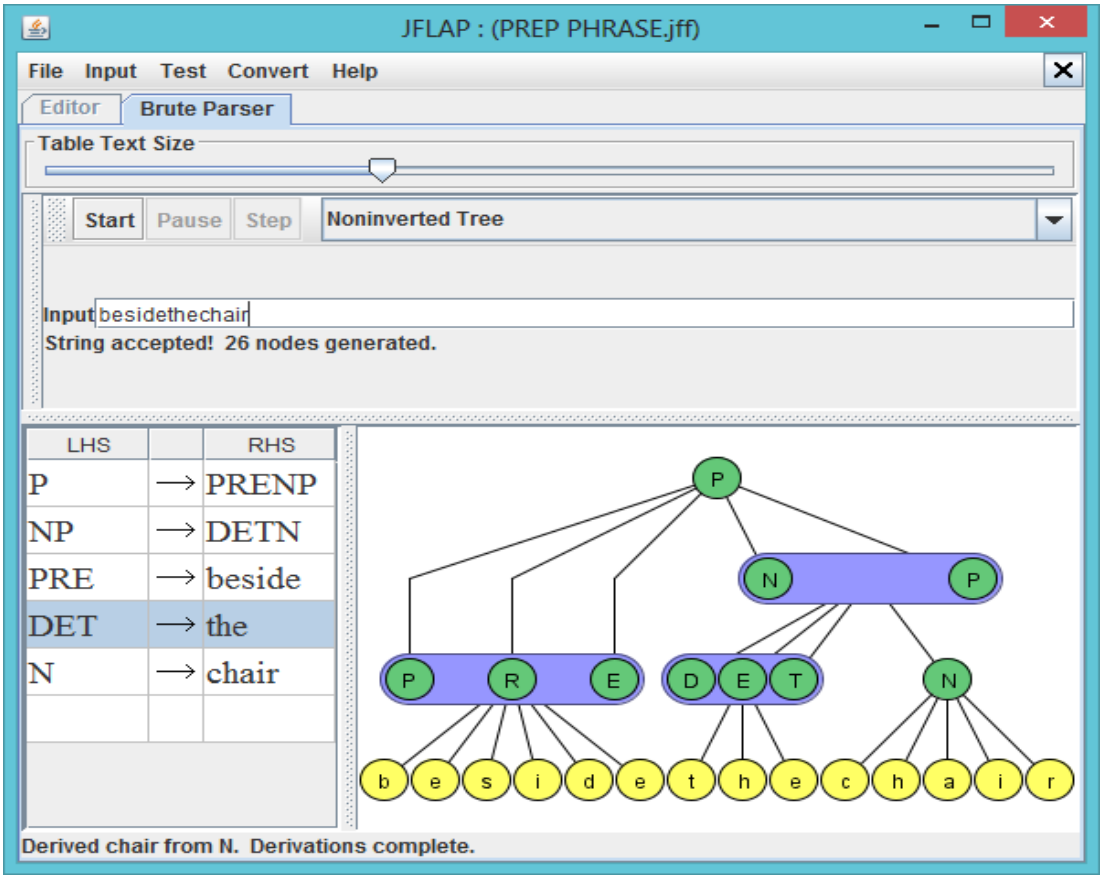

Figure 4. Sample of English prepositional phrase

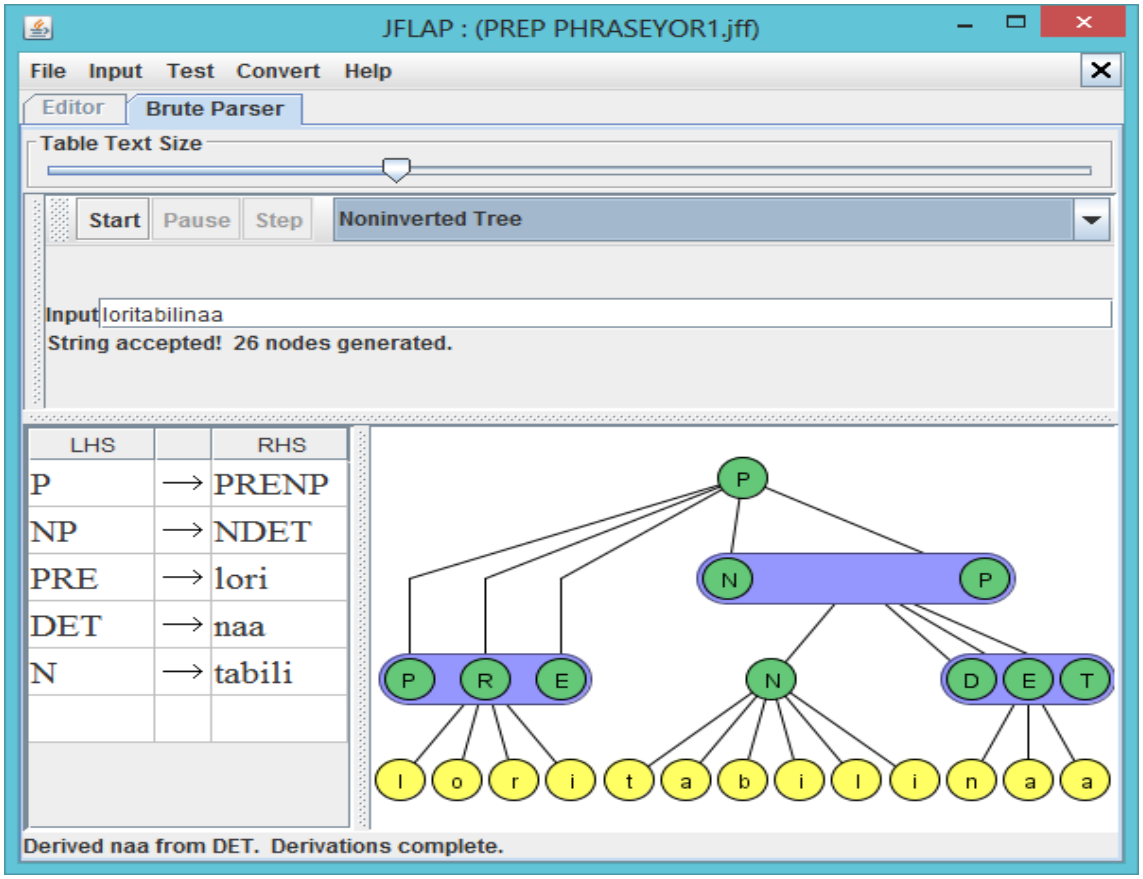

Figure 5. Sample of Yorùbá prepositional phrase

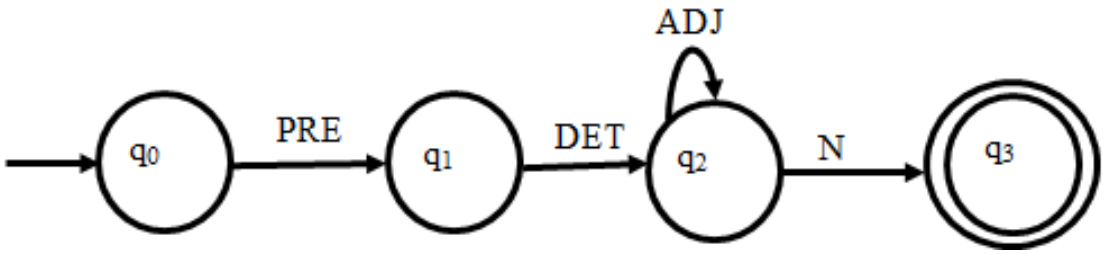

Figure 6. State diagram for the English translation process

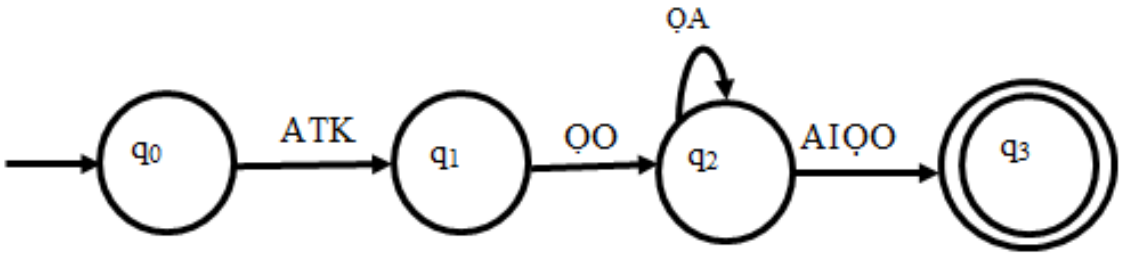

Figure 7. State diagram for the Yorùbá translation process

3.1.3. Re-write Rules Testing

JFLAP was used to determine the correctness of the re-write rules designed in (3) and (4). Figures 4 and figure 5 show the outputs of the SL and TL prepositional phrases. The two language grammars re-write were used for the JFLAP environment. This is to determine whether the PP strings will be accepted or not. The nodes are generated following the re-write rules provided. If the re-write rules follow the sequence of the language grammars it will accept the string being tested, otherwise it will be rejected. Many PP can be tested for the two languages.

3.2. The Prepositional Phrase Translation Process Model

Figure 6 is the state diagram for the English PP translation

process model. Figure 6 describes possible phrases that can be translated from the source language (SL) to the target language (TL). These are possible translation combinations for the English preposition phrase. They are: PREDETN and PREDETADJN. It means that there can be, on the table and on the flat table for example. Figure 7 is the state diagram for the Yorùbá language PP translation process. Figure 7 shows possible combinations of prepositional phrases that can be accepted by the TL. They are: ATKỌOAIỌO and ATKỌOỌAAIỌO. One important thing to note is that, the noun (ỌO) and adjective (ỌA) swapped with the determiner. It shows that Yorùbá language is head first and English language is head last.

4. System Design Framework

This section explains the system design framework. This involves the database design, and system software design.

4.1. System Design

Software system design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements as shown in Figure 8. Figure 8 essentially has six steps. The start (user run the application), English text (input from the user).The lexical analyzer brakes the PP entered by the user into lexemes (lexical items or words), it then compare each lexeme with the content of the database. If the lexemes are in the database, the lexeme will pass to the syntax generator/analyzer. Otherwise, the system will ask the user to add to the database or enter a new PP. The syntax analyzer will use all the re-write rules provided to arrange the PP words and send it to the target language for final translation.

4.2. Database Design

Home domain terminologies were used for the lexicon (database). The lexemes from bi-lingual dictionary in line with are available within home environment. Machine translations systems are usually domain sensitive. The lexemes (data) were divided into two: Data 1 and data 2.

Data 1: Parallel corpus:- Words were collected from both languages.

Data 2: Tagged corpus:- Each word was assigned its appropriate POS tag.

Data 1: Parallel corpus

Figure 9 shows the sample of parallel corpus database of English and Yorùbá languages. In data 1, different lexemes were collected. In data 2, they were broken into different parts of speech (POS) such as noun, verbs and others were tagged. In the database words are separated according to their parts of speech.

Some of lexemes in figure 9 were separated into five parts of speech (e.g. noun, pronoun, verb, adjective, and preposition) and their respective Yorùbá equivalents as shown in tables 2-6 for clarity.

Table 2. List of English pronouns and their Yorùbá equivalents

English	Yorùbá
She/he/it	Ó
they	Àwọn
you	ìwọ/irẹ
we	àwa
them	wọ́n

Table 3. Some English prepositions and Yorùbá equivalents

English	Yorùbá
About	Nipa
Beside	Ní ẹ̀gbẹ́ tàbí lẹ́gbẹ̀
Before	Ìbẹ̀rẹ̀
Behind	Ní ẹ̀hìn tàbí lẹ́hìn
After	Ní ìparí tàbí níparí
Back	ẹ̀hìn
Front	Iwájú
End	òpin

Table 4. English nouns and Yorùbá equivalents (extract).

English Noun	Yorùbá Noun
Boy	Omodé-kùnrin
Girl	Omodé-bìrin
Man	Okùnrin
Woman	Obìnrin
Ade	Adé
Mayowa	Máyòwá

Table 5. Some of English Adjectives and their Yorùbá equivalents

English Adjectives	Yorùbá Adjectives
Tall	Gíga
Small	Kékeré
Old	Àgbàlágbà
Short	Kúkurú
Beautiful/Handsome	Arẹ̀wà
Big	Ńlá

Table 6. List of English determiners and their Yorùbá equivalents

ENGLISH	YORÙBÁ
A	Kan
An	Kan
Some	Díẹ̀
The	Náà

Data 2: Parts of Speech Tagging (Tagged corpus)

Part-of-speech (POS) tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech based on the definition as well as its context. In POS words are manually tagged. Table 7 shows the tagged POS prepositional phrase.

Table 7. POS tag set

English Prepositional Phrase	Yorùbá Prepositional Phrase
on<PRE>the<DET> table<N>	lórí<ATK>àga<QO>náà< AIQO>
at<PRE>the<DET>end< of> the<DET> day<N>	ní<ATK>òpin<ATK> ọjọ́<QO>

4.3. Software Design

Figure 10 shows the sequence diagram, it contains six modules which include Main-GUI, Add-GUI, Translator, Word, DataBaseUtil, and note txt. Figure 10 shows the sequence of PP text among the modules of the system. Figure 11 shows the system class diagram. The class diagram explains the base-line for the system coding. It depicts the interaction between different modules within the system. The five modules are MainGUI, AddGUI, Translator, DataBaseUtil and Word.

5. System Implementation

The system was implemented using Java programming language with the integrated development environment being the NetBeans IDE. Figure 12 shows the user case diagram. The enter text (user is expected to enter the PP), after that user should click translate button for the system to translate. The user gets the outputs from the graphical user interface (GUI).

Figure 13 shows the system GUI. It has three planes. First plane the user enters the English PP and click translate button. The second plane is the plane that transcribes the English PP to Yorùbá PP word for word. The third plane is the final translation that displays Yorùbá PP text. Figure 14 is the sample system outputs.

6. Results Discussion

The system was tested and it results were compared with Yorùbá Google translator's translations. Tables 8-9 show that the system was able to produce translations with correct meanings and appropriate tone marks and under-dots contrary to that of the Google translator system.

Table 8. IFEMTPP outputs

English prepositional phrase	Yorùbá Equivalent (via IFEMT-3)
Behind the door	lẹ́hìn ilẹ̀kùn náà
In the bottle	nínú ìgò náà
For the fight	fún ìjà náà

Table 9. Google translator outputs

English prepositional phrase	Yorùbá Equivalent (via Google translate)
Behind the door	Sile ni enu
In the bottle	Ninu igo
For the fight	Fun awon ija

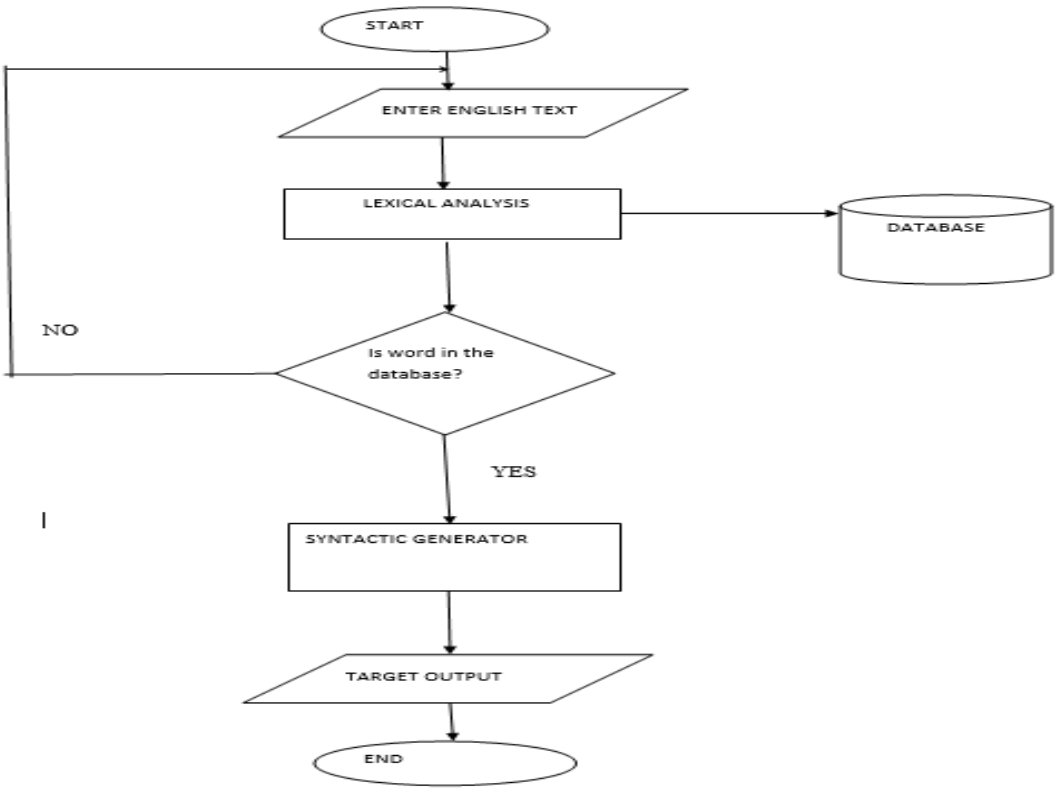

Figure 8. Flow chart for the system

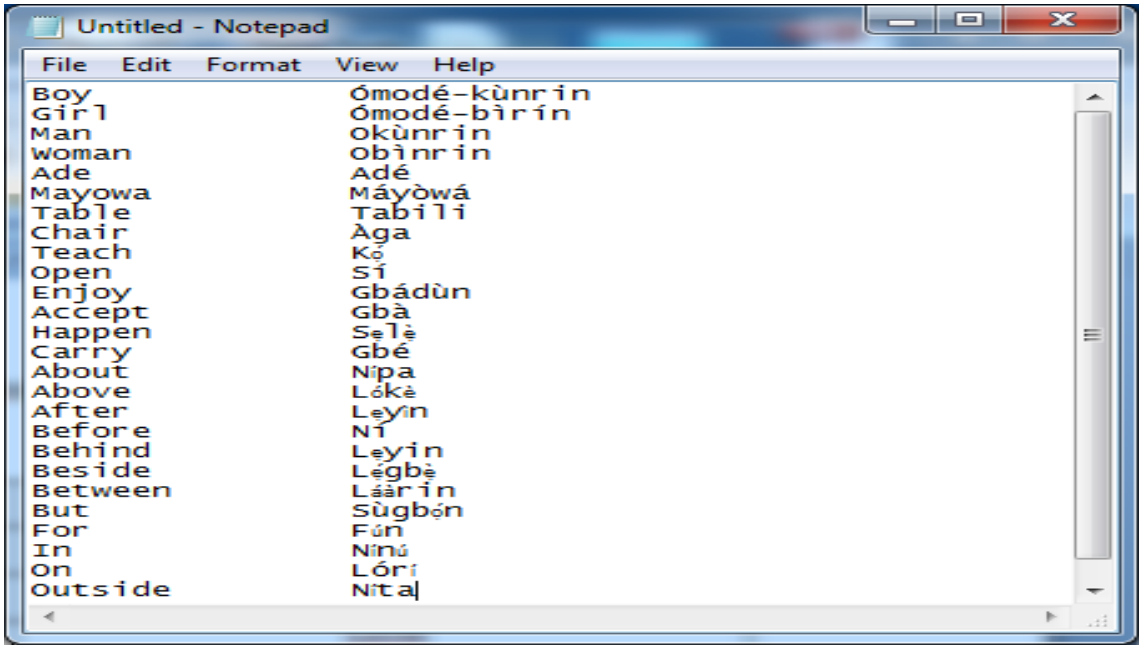

Figure 9. Parallel corpus

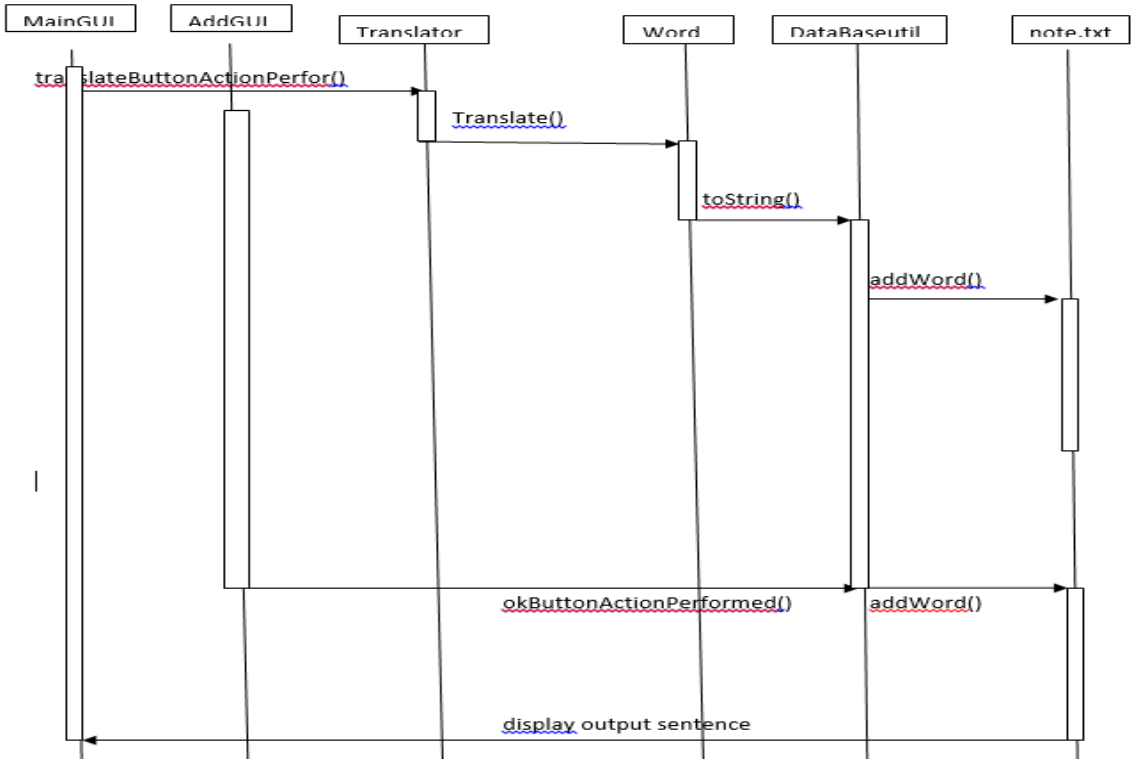

Figure 10. Sequence diagram

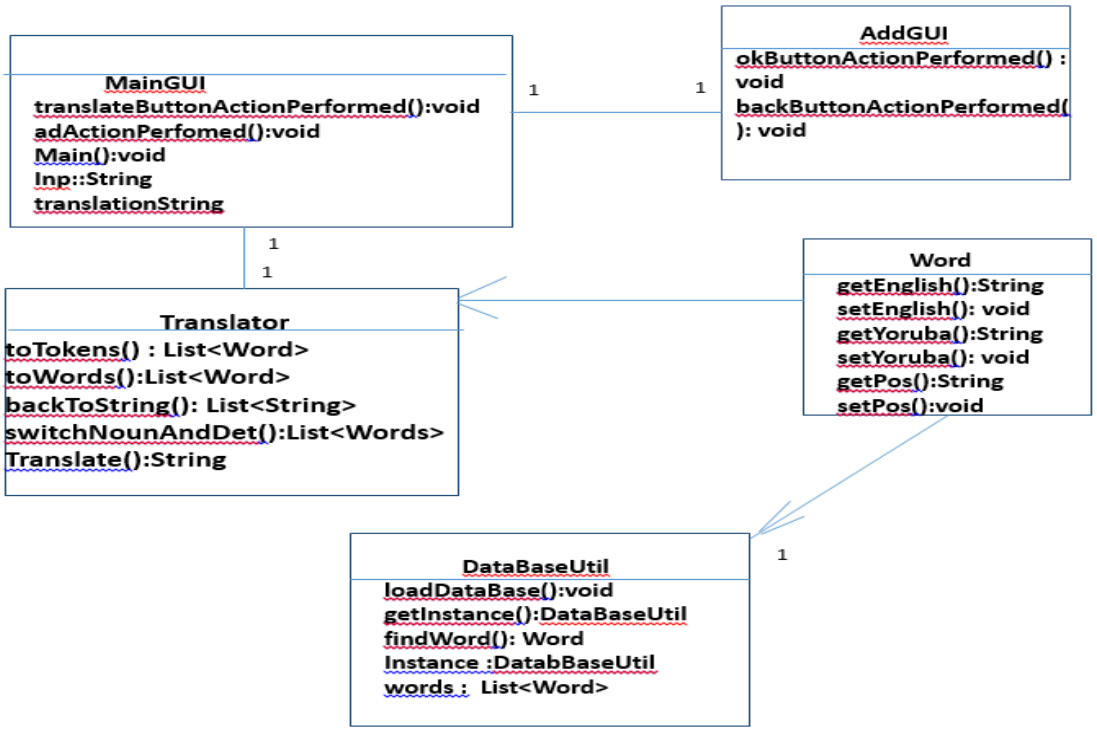

Figure 11. Class diagram

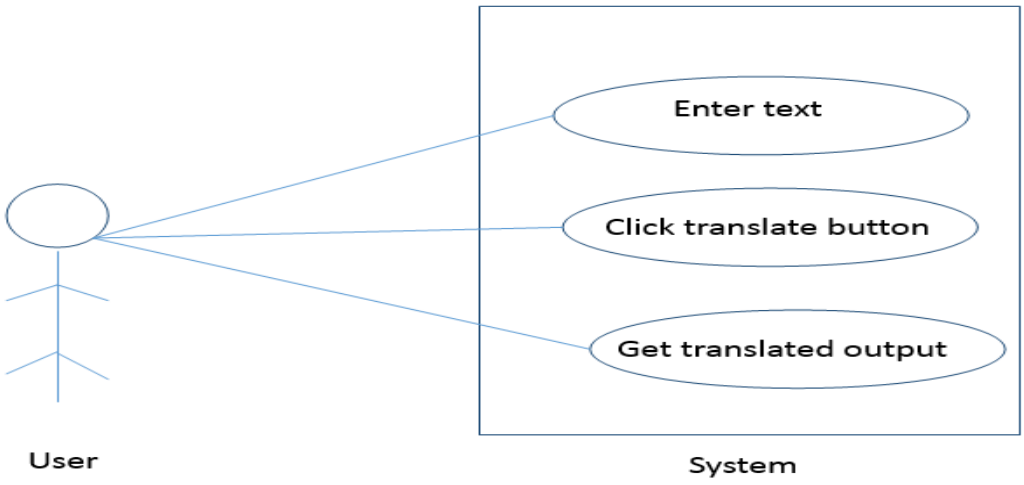

Figure 12. Use case diagram of the system

Figure 13. Sample output 1

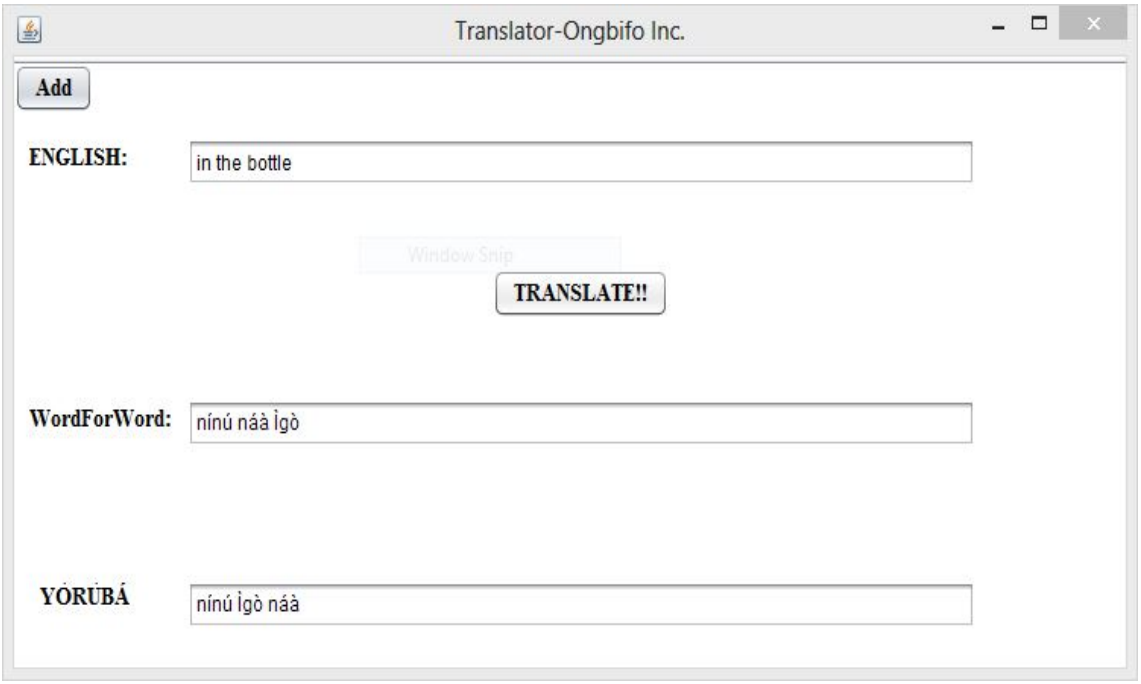

Figure 14. Sample output 2

7. Conclusions

The development of the English to Yorùbá Prepositional Phrase machine translation (EYMT) system was successfully carried out, with system being able to give accurate translations with appropriate tone-marks and under-dots. This prepositional phrase MT was used to experiment some rules and the results gotten will be integrated with the main EYMT system. In future, other phrases and more complex sentences will be studied.

REFERENCES

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

[2] Adegbola T., (2009); Building capacities in human language technology for African languages: Proceedings of the EACL 2009 workshop on language technologies for African languages – AfLaT 2009, pp 53-58.

[3] Abiola O. B., Adetunbi A. O., Oguntimilehin A., (2013); A computational model of English to Yorùbá noun phrases translation system; FUTA journal of research in sciences; vol. 1: 34-43.

[4] Almaflehi N., Saad S.A., (2013) the problem of translating the prepositions at, in and on into Arabic: An applied linguistic approach; Journal for the study of English linguistics; ISSN 2329-7034, vol. 1, No. 2.

[5] Islam Z., Tiedmann J., Eisle A., (2009) English to Bangla phrased-based statistical machine translation; Unpublished M.Sc. thesis; Saarland University; Saarland.

[6] Eludiora S.I., (2014) Development of English to Yorùbá machine translation system; Unpublished Ph.D. thesis; Obáfẹ́mi Awólọ́wọ̀ University, Ile- Ife, Nigeria.

[7] Eludiora, S. I., Agbeyangi, A. O. and Fatunsin, A. (2015a) 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.

[8] Eludiora, S. I., Okunola, M. A and Odejobi, O.A. (2015b): 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.

[9] Eludiora, S. I., Awoniyi, A. and Azeez, I. O. (2015c) 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.

[10] 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.

[11] Odejobi, O. O., Ajayi, A. O., Eludiora, S. I., Akanbi, L. A., Iyanda, I. R. and Akinade, O. A. (2015) A web-based system for supporting teaching and learning of Nigerian indigenous languages, in OAU TekCONF 2015 proceedings, Nigeria, 350-360.

[12] Akinwale, O. I., Adetunmbi, A. O., Obe, O. O., Adesuyi, A. T. (2015) Web-based English to Yorùbá Machine Translation, International Journal of Language and Linguistics, 3(3): 154-159.

[13] Abiola, O. B., Adetunmbi, A.O. and Oguntimilehin, A. (2015) using a hybrid approach for English to Yorùbá text to text Machine Translation System (proposed), International Journal of Computer Science and Mobile Computing (IJCSMC), vol 4, issue 8, 308-313.

[14] Abiola, O. B., Adetunmbi, A.O., Fasiku, A. I. and Olatunji, K.A. (2014) a web-based English Yorùbá Noun phrases machine translation system, an international journal of English and Literature. Vol 5(3) 71-78.