English to Yorùbá Machine Translation System using Rule-Based Approach

Agbeyangi, Abayomi O.
Department of Computer Engng.
Moshood Abiola Polytechnic
Abeokuta, Ogun State-Nigeria.
abayomi_sola@yahoo.co.uk
(correspondent author)

Eludiora Safiriyu I.
Dept. of Computer Sci. & Engng
Obafemi Awolowo University
Ile-Ife, Osun State-Nigeria
eludioraomolola@gmail.com

Adenekan, Olujide A.
Department of Computer Engng.
Moshood Abiola Polytechnic
Abeokuta, Ogun State-Nigeria.
adenekanolujide@yahoo.com

Abstract - Rule-based approach is a good approach for Machine Translation System used for language with lots of grammar which Yorùbá language is one. In this paper we present Rule-based approach to Yorúbà Machine Translation System. The popularity of Yorúbà language among the three main languages in Nigeria calls for the need to computerise the language.

Transfer Rule-Based Machine Translation is use in the development of the System. It was used because it allows us to use manual tagging of the part of speech (POS). Rewrite rules was developed for the two languages (Yorùbá and English). The data was collected from home domain vocabularies. The re-write rule was verified using Natural Language Toolkits (NLTKs) and implement using python programming language.

The system interface gives the user the opportunity to type simple English language sentence and the resulting Yorùbá Translation is displayed. The result shows that the system performance is close to the expert opinion, having considered the scope for which the system is developed. Based on the result gathered there are some issues to address that could be considered in a future work.

Keywords—Machine translation, Yorùbá language, rule-based approach, English language, sentence

I. INTRODUCTION

According to [1], Machine translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. At its basic level, MT systems perform simple substitution of words in one natural language for words in another; but this alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with statistical techniques is a rapidly growing field that has led to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies [2]

According to [3], improved translation quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words falls into a specific part of speech (POS). With the assistance of these techniques, MT has proven useful as a tool to assist human translators and in a very limited number of cases, can even produce output that can be used directly without further processing.

The progress and potential of machine translation has been debated much through history. Since the 1950s, a number of scholars have questioned the possibility of achieving fully automatic machine translation of high quality [4]. Some critics claim that there are in-principle obstacles to automatizing the translation process of what has been achieved is the development of programs which can produce 'raw' translations of texts in relatively well-defined subject domains, which can be revised to give good-quality translated texts at an economically viable rate or which in their unedited state can be read and understood by specialists in the subject for information purposes. In some cases, with appropriate controls on the language of the input texts, translations can be produced automatically that is of higher quality needing little or no revision [5].

The remaining part of the paper is structured as follows: Section 2 explains machine translation approaches; section 3 gives an overview of English and Yorùbá similarities, differences and characteristics. Section 4 gives the system design and implementation, while Section 5 discusses the results. Section 6 concludes the paper.

II. MT SYSTEM APPROACHES

MT systems can be developed using three approaches depending on data (corpora) availability and type of language. The three approaches are: rule-based (RBMT), statistical and Hybrid. There are two types of rule-based machine translation systems: Transfer Rule-Based Machine Translation and Inter-lingual RBMT Systems [6].

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. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation [7]. Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. In less than two decades, SMT has come to dominate academic MT research and has gained a share of the commercial MT market [8].

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies [9]. Several MT companies (Asia Online, LinguaSys, Systran, PangeaMT, and UPV) are claiming to have a hybrid approach using both rules-based and statisticals.

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 [10][11]

Rule based machine translation system, consists of collection of rules called grammar rules, lexicon and software programs to process the rules [12]. In [13], it is defined as systems that use large collections of rules, manually developed over time by human experts, which map structures from the source to the target language. The rules are written with linguistic knowledge gathered from linguists or other means.

When the sentences are inputted (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.

Rules play major role in various stages of the translation: syntactic processing, semantic interpretation, and contextual processing of language. Figure 1 below is the schematic diagram of a rule-based machine translation system.

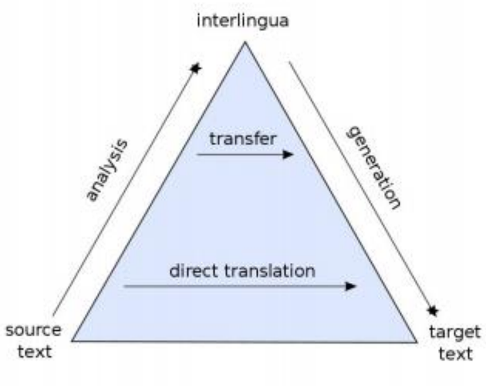


Fig. 1. Rule-based Machine Translation

[a] (source: http://tinyurl.com/pv6sv5j)

Transfer-based machine translation is a type of machine translation based on the idea of inter-lingua and is currently one of the most widely used methods of machine translation [6]. Both transfer-based and interlingua-based machine translation have the same idea; to make a translation, it is necessary to have an intermediate representation that captures the "meaning" of the original sentence in order to generate the correct translation [14]. In interlingua-based MT this intermediate representation must be independent of the languages in question, whereas in transfer-based MT, it has some dependence on the language pair involved. The way in which transfer-based machine translation systems work varies substantially, but in general they follow the same pattern; they apply sets of linguistic rules which are defined as correspondences between the structure of the source language and that of the target language [6].

A. *Stages in Translation*

The methods which are chosen and the emphasis depends largely on the design of the system, however, most translation systems include the following stages:

- Morphological analysis
- Lexical categorization
- Lexical transfer
- Structural transfer
- Morphological generation

Transfer Rule-Based Machine Translation was used in this research because it allows us to use manual tagging of the part of speech (POS). The work is also restricted to simple sentences translation.

III. English and Yorùbá Language

According to [15], English language basically moves from concrete to abstract, while Yorùbá language moves from abstract to concrete. Thus, Yorùbá language can be seen as a complex language to study. It has a lot of cultural entities (Proverb - ewì, oríkì, etc) which cannot we adequately represent in English (e.g işé ni ògùn ìsé).

There are various differences and similarities between English and Yorùbá Language, some of which are discussed here:

- Yorùbá language borrows English language words for most of the words that does not have a Yorùbá equivalent.

  Biro → Bírò, Bread → Búrẹ́dì

- In English Language determinant (e.g the) always come after a noun but in Yorùbá language, determinant always follow noun.

  The<Det> boy<N> → ọmọkùnrin<N> náà<Det>

- Yorùbá language is a tonal language with 3 distinct tones while English is not.

- Most sentences in English language cannot be translated to Yorùbá using word-for-word translation. e.g.

  The boy is coming → Náà ọmọdekùnrin ń bọ̀

  The correct translation must be; ọmọdekùnrin náà ń bọ̀.

The main features of both languages are highlighted in the table below:

TABLE I. Features of English and Yorùbá language

Attributes	English	Yorùbá
Language types	Non-tonal	Tonal eg. ilé, agbà, ọ̀mọ̀
Timing	Stressed e.g. He found it on the street	Syllable e.g. bàbá, dé, bí, bì, igbá
Orthography	Non-Phonetic e.g. enough, fish, pharmacy, farm	Almost Phonetic e.g. ẹgbẹ́, ẹ̀gbẹ, edé, èdè
Language	Large digital	Little digital

resources	resources	resources
Inflectional	Inflectional e.g. waits, likes, goes	Non-inflectional e.g. lọ, ńlọ, tilọ
Grammatical Structure	Subject Verb Object (SVO) e.g. She drives this car	Subject Verb Object (SVO) e.g. ọkọ̀ ayọ́kẹ́lẹ́ yiì l'ówa (translated)

[b.] (source: [20])

IV. METHODOLOGY

The software is designed as a window application. The grammar of the language is written to follow the re-writes rule developed for the Translation process.

Sample of the re-write rules developed for the two languages (English and Yorùbá) are given below:

English language:

S ::= <NP> <VP>

VP ::= <V> <NP> | <NP> <V> | <V> <V>

PP ::= <P> <NP>

V ::= <NP> <V>

NP ::= <DET>< N> | < PRON> <N> | <DET> <ADJ>

Yorùbá language:

S::= <NP> <VP>

VP ::= <NP> <V> | <V> <PP> | <V> | <V> <V>
|<PRON> <V> |<V> <NP> | <VP> <PP>

PP ::= <P> <NP>

NP ::= <N> <PRON> |<N> <DET> |<DET> <PP>
|<N> |<PRON> <N> | <PRON>

S stands for the sentence, NP, PP, VP, N, V, ADJ and DET are the non-terminals. NP is Noun Phrase, PP is prepositional phrase, VP is Verb Phrase, P is preposition, ADJ is Adjective, N is Noun and V is verb. The Left hand side (LHS) is substituted with the Right hand side (RHS) until the terminals are reached.

In figure 2 below the re-write rule for Yorùbá translation is model using Finite State Automaton (FSA). The Automaton is tested with sample input as shown in figure 3.

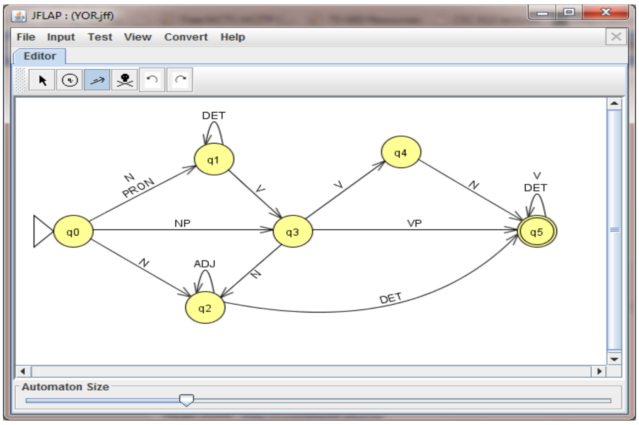
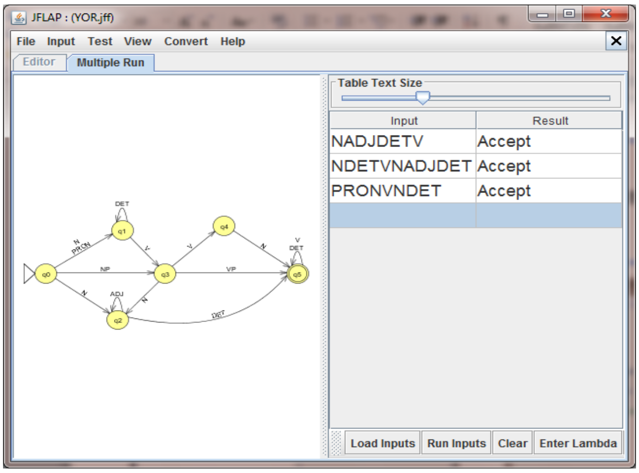


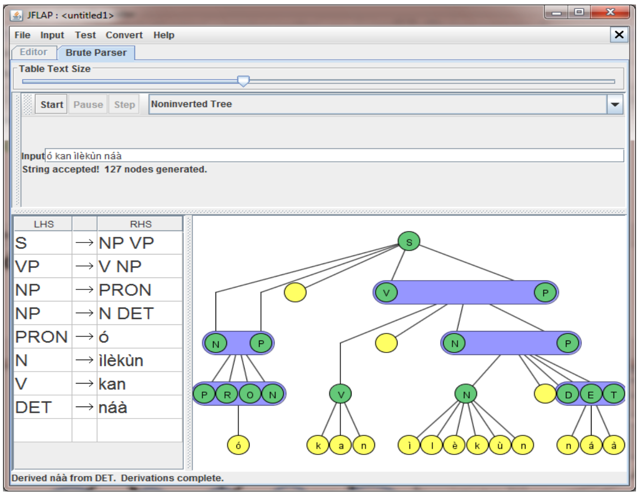
Fig. 2. State Diagram for Yorùbá Re-Write Rule



Input	Result
NADJDETV	Accept
NDETVNADJDET	Accept
PRONVNDET	Accept

Fig. 3. Yorùbá Re-Write Rule with accepted states

Sample sentences are run in figure 4 and 5 using JFAP to show the pattern of the translation parse tree structure.

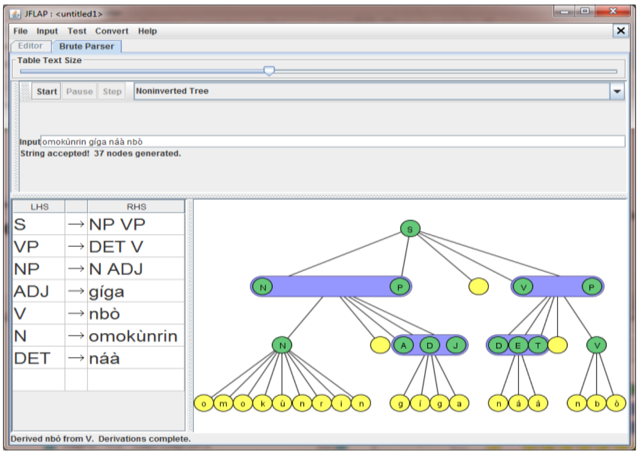


Input ó kan ìlèkùn náà
String accepted! 127 nodes generated.

LHS		RHS
S	→	NP VP
VP	→	V NP
NP	→	PRON
NP	→	N DET
PRON	→	ó
N	→	ìlèkùn
V	→	kan
DET	→	náà

Derived náà from DET. Derivations complete.

Fig. 4. JFLAP sample sentence tree structure 1



Input omokùnrin gíga náà nbò
String accepted! 37 nodes generated.

LHS		RHS
S	→	NP VP
VP	→	DET V
NP	→	N ADJ
ADJ	→	gíga
V	→	nbò
N	→	omokùnrin
DET	→	náà

Derived nbò from V. Derivations complete.

Fig. 5. JFLAP sample sentence tree structure 2

The grammar parse three structure was also model using NLTK parser. NLTK is a very important tool that is used in this research work to easily model the behavior of both language sentence tree structures. Example is show in figure 6 and 7 below.

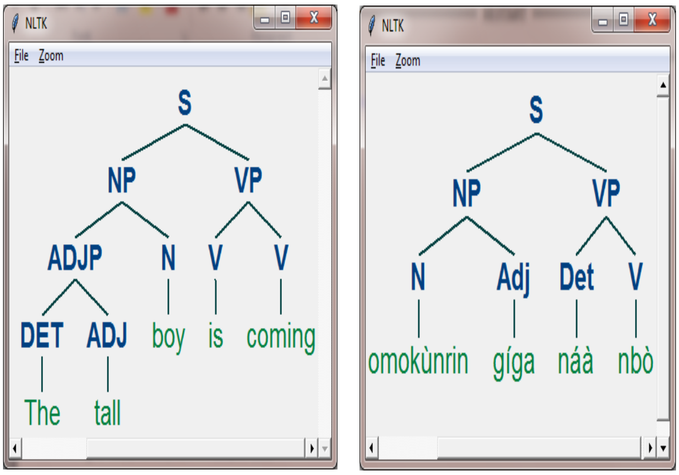


Fig. 6. NLTK sample sentence tree structure 1

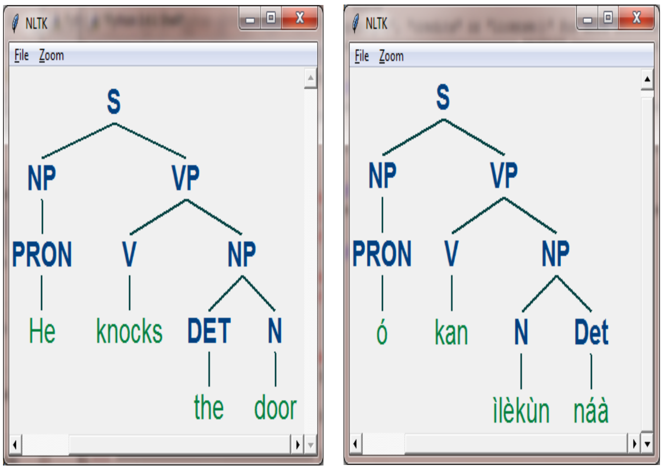
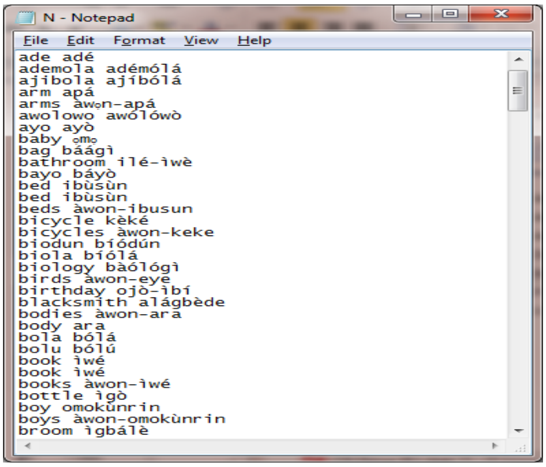


Fig. 7. NLTK sample sentence tree structure 2

A. DATA COLLECTION

The data (corpus) for the research was from simple sentence spoken in the home environment. The sentences are further broken down into their part of speech (POS). The different parts of speech are stored in pairs. Figure 8 shows a sample of the stored part of speech.



```
ade adé
ademola adémólá
ajibola ajíbólá
arm apá
arms àwọn-apá
awolowo awólówò
ayo ayò
baby ọmọ
bag báágì
bathroom ilé-ìwè
bayo báyò
bed ibùsùn
bed ibùsùn
beds àwon-ibusun
bicycle kèké
bicycles àwon-keke
biodun bíódún
biola bíólá
biology bàólógì
birds àwon-eye
birthday ojò-ìbí
blacksmith alágbède
bodies àwon-ara
body ara
bola bólá
bolu bólú
book ìwé
book ìwé
books àwon-ìwé
bottle ìgò
boy omokùnrin
boys àwon-omokùnrin
broom ìgbálè
```

Fig. 8. Nouns in the database

B. DEVELOPMENT TOOLS

The main tools used in this research are:

- Python programming language – this is the core programming environment for the application development.
- NTLK (Natural Language Toolkit) – this is a support kit for python programming language. Its features include: support for parsing, Part of Speech (POS) tagging, corpora design and analyses.
- PyQt – this is also a support kit for the design of the application GUI.
- py2exe – this was used to compile the python codes (.py) to an executable file (.exe).
- NSIS – use basically in building the window installer for the application.

C. REQUIREMENT ANALYSIS

The requirements and specifications of the software are as follow:

- to present a user friendly interface to the user;
- to give the user access to enter simple basic sentences in English language provided the sentence is within the domain covered;
- parse the sentence to understand the structure;
- translate and output the equivalent meaning of the sentences entered in standard Yorùbá language; and
- give the user ability to add to the corpus (database)

The system sequence diagram and the use case diagram are shown in figure 9 and 10 below.

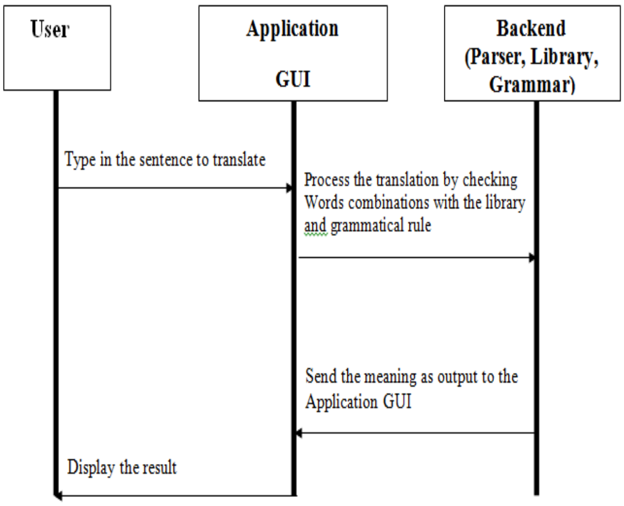

User
Application GUI
Backend (Parser, Library, Grammar)
Type in the sentence to translate
Process the translation by checking Words combinations with the library and grammatical rule
Send the meaning as output to the Application GUI
Display the result

Fig. 9. Sequence Diagram showing the flow of operation

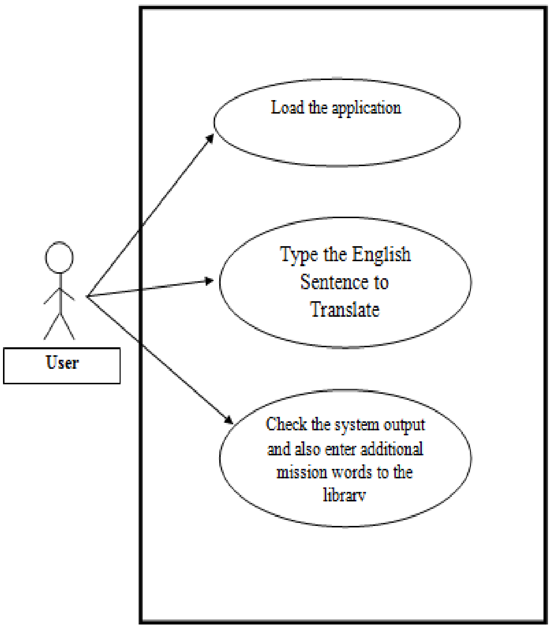
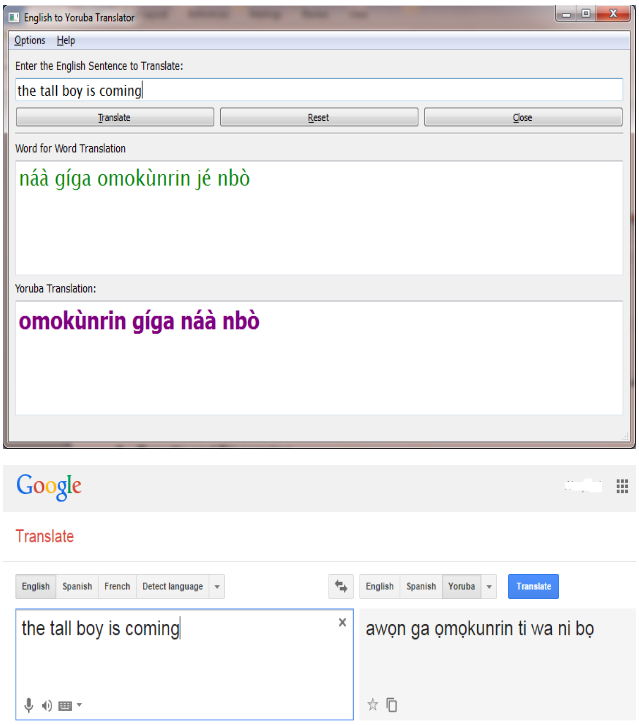

Load the application
Type the English Sentence to Translate
Check the system output and also enter additional mission words to the librarv
User

Fig. 10. Use case diagram of the system

V. Results and Discussion

The system perform is compared with another machine translation system. Google Translator was use since is a standard machine translator that include an English to Yorùbá translation. Figure 11, 12 and 13 below shows the result of the output from our system with output from google tranlator.

The result from the output from Google Translator shows that our system gives a better translation as can be considered as a standard tranlation for Yorùbá languages.



Enter the English Sentence to Translate: the tall boy is coming

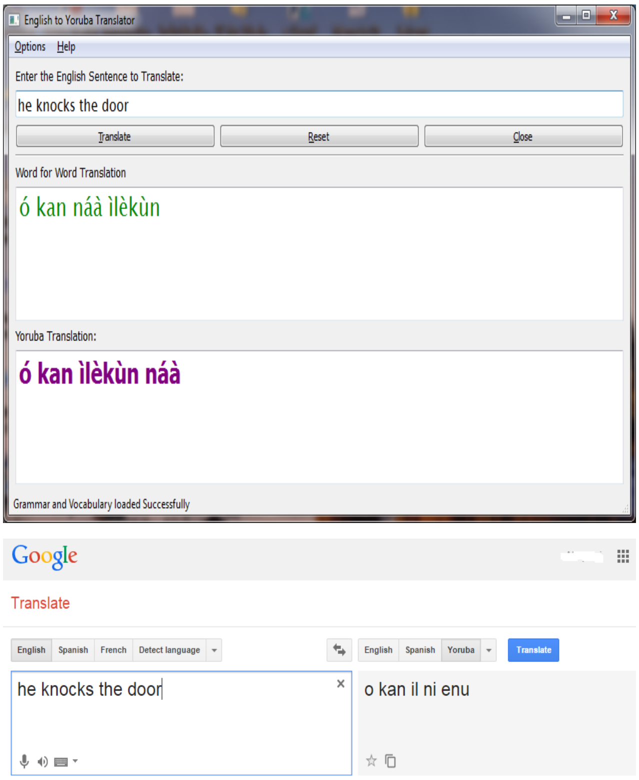
Word for Word Translation: náà gíga omokùnrin jé nbò

Yoruba Translation: omokùnrin gíga náà nbò

the tall boy is coming

awọn ga ọmọkunrin ti wa ni bọ

Fig. 11. Sample output from the system and Google translator



Enter the English Sentence to Translate: he knocks the door

Word for Word Translation: ó kan náà ìlèkùn

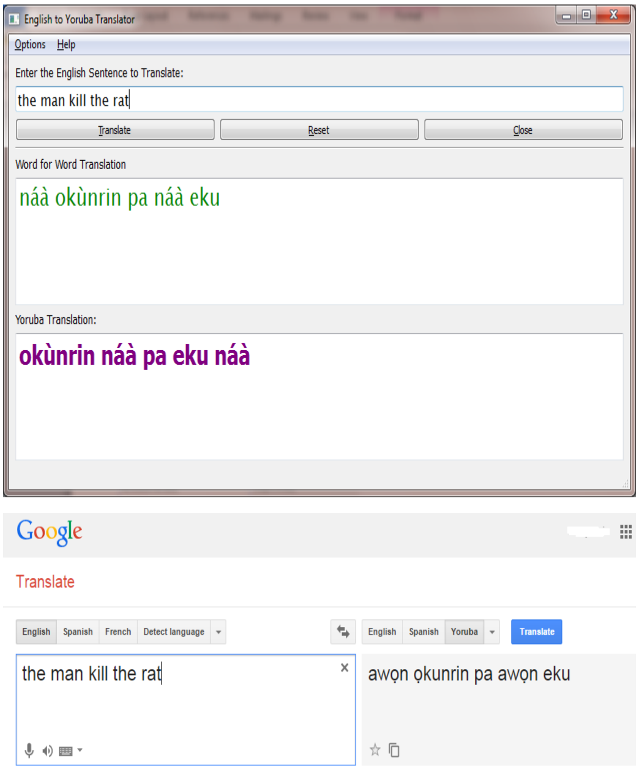
Yoruba Translation: ó kan ìlèkùn náà

Grammar and Vocabulary loaded Successfully

he knocks the door

o kan il ni enu

Fig. 12. Sample output from the system and Google translator



Enter the English Sentence to Translate: the man kill the rat

Word for Word Translation: náà okùnrin pa náà eku

Yoruba Translation: okùnrin náà pa eku náà

the man kill the rat

awọn ọkunrin pa awọn eku

Fig. 13. Sample output from the system and Google translator

VI. CONCLUSION

According to the research conducted in [16], improvement to the translation process can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Thus, to give an automated high-quality translations, there is need to generate more robust re-write rules that can accommodate the translation.

There are lots of work that can be done to improve the quality of the system output inorder to increase its usefullness.

Other areas of note for further research include ability to extent the system capability to translate longer sentences, ability to translate paragraph text and the issue of ambiguities in translation.

REFERENCES

[1] ALPAC (Organization). (1966). Language and Machines: Computers in Translation and Linguistics: a Report. National Academy of Sciences, National Research Council.

[2] Nirenburg, S. (1987). Machine translation: theoretical and methodological issues. Cambridge University Press.

[3] Nagao, M. (1985). Evaluation of the quality of machine-translated sentences and the control of language. Journal of the Information Processing Society of Japan, 26(1), 197-1202.

[4] Bar-Hillel, Y. (1960). The present status of automatic translation of languages.Advances in computers, 1(1), 91-163.

[5] Wilks, Y. (2008). Machine translation: its scope and limits. Springer Science & Business Media.

[6] Hutchins, W. J. (1986). Machine translation: past, present, future (p. 66). Chichester: Ellis Horwood.

[7] Weaver, W. (1955). Translation, in Machine Translation of Languages, MIT Press, Cambridge, MA

[8] Lopez, A. (2007, June). Hierarchical Phrase-Based Translation with Suffix Arrays. In EMNLP-CoNLL (pp. 976-985).

[9] Adam, B. (2009). "AppTek Launches Hybrid Machine Translation Software", http://www.speechtechmag.com/Articles/News/News-Feature/AppTek-Launches-Hybrid-Machine-Translation-Software-52871.aspx . *Accessed 5th August, 2015*

[10] Bond F. (2006). “Introduction to Machine Translation". url: www.cs.mu.oz.au/research/it/nlp06/materials/Bond/mt-intro.pdf. Accessed 14th July, 2015

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

[12] http://language.worldofcomputing.net/ machine-translation/rule-based-machine-translation.html. *Accessed10th August, 2015*

[13] http://www.safaba.com/machine-translation/machine-translation-technologies/rule-based-machine-translation. *Accessed13th August, 2015*

[14] Arnold, A., Sadler, L., and Humphreys, R. (1993) Evaluation: an Assessment. Machine Translation 8 (1/2), 1-24.

[15] Odejobi, T., Owolabi, K., Adegbola, T. (2011). Localising for Yorùbá: Experience, Challenges and Future Direction.

[16] Abu Shquier, M. and T. Sembok, (2008) Word Agreement and ordering in English-Arabic machine translation proceeding of the International Symposium on Information Technology, Aug. 2008, IEEE Xplore Press, USA.

[17] Ahlswede, T., and Lorand, D. (1993) Word Sense Disambiguation by Human Subjects: Computational and Psycholinguistic Applications. In Boguraev, B., & Pustejovsky, J. (eds.), Proceedings of a Workshop Sponsored by the SIGLEX of the ACL. Ohio State University.

[18] Alonso, J. A. (1990) Transfer InterStructure: designing an ‘interlingua’ for transfer-based MT systems. Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (Austin, TX), pp. 198-201.

[19] Boretz S., and Adam, A. (2009) "AppTek Launches Hybrid Machine Translation Software" SpeechTechMag.com.

[20] Poole D., Mackworth, A. Goebel, R. (1998) Computational Intelligence: A Logical Approach. New York: Oxford University Press.

[21] Russell, J. Norvig, P. (2003) Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

[22] Simon, A. (2009) “Yorùbá language, alphabet and pronunciation” http://www.omniglot.com/writing/ Yorùbá language, alphabet and pronunciation.htm.

[23] Agbeyangi A.O., Adenekan O.A., Lawal O.O., & Durosinmi A.E. (2015). Rule-based approach to Yorùbá Machine Translation System: 4th iSTEAMs Research Nexus 2015 International Conference, University of Ilorin, Ilorin, Kwara State, March 11-13, 2015.

[24] Agbeyangi A.O. (2013). Development of a Machine Translation System for simple Yorùbá Sentences. Linguistics and the Glocalisation of African Languages for Sustainable Development: A Festschrift in Honour of Prof. Kola Owolabi.

[25] Eludiora S. I. (2014) Development of a Machine Translation System for English to Yorùbá. Unpublished PhD Thesis, Department of Computer Science and Engineering Obafemi Awolowo University, Ile-Ife, Nigeria.