Web-Based English to Yoruba Machine Translation

Akinwale O. I., Adetunmbi A. O., Obe O. O., Adesuyi A. T.

Computer Science Department, Federal University of Technology, Akure, Nigeria

Email address:
mdtosin@gmail.com (Akinwale O. I.), aoadetunmbi@futa.edu.ng (Adetunmbi A. O.), olubes@gmail.com (Obe O. O.),
atadesuyi@futa.edu.ng (Adesuyi A. T.)

Abstract: The growth of globalization in the world today has increased the rate at which people interact and integrate, thereby increasing the level of international integration from interchange of world views, products, ideas and other aspects of culture. Language differences therefore pose a major barrier to smooth running of these processes. Therefore there is need for existence of system that will help translate between languages. English is a West Germanic language which has become the lingual Franca in Nigeria and 53 other countries. Therefore vital information are written and spoken in English language in Nigeria. Meanwhile,Yoruba language is lagerly spoken in Nigeria with over 40 million speakers in the south-western part of the country and also in parts of Benin republic. This research deals with the translation of English text to Yoruba text using rule based method. Twenty two rules were formulated for the translation which is specified using context free grammar. A bilingual dictionary dataset containing English words and the corresponding translation in Yoruba language was used. The research model was implemented with ASP.net and C# programming languages which has been hosted on http://www.naijatranslate.com. The translator was evaluated to have accuracy of 90.5%.

Keywords: Machine Translation, Rule-based Machine Translation, English Language, Yoruba Language, Computational Rules, Translation System

1. Introduction

English language is the Nigerian lingua franca which is commonly spoken among tribes in the country.This has therefore posed a threat to the survival of indigenous Nigerian languages. Consequently, most children cannot speak their mother tongue. Therefore, experts are agitated that if a child cannot speak his or her mother tongue today, there is probality that in the next 20 to 25 years the sons and daughters of the child may lose the language[1]. This implies that in the next 50 years or more, the fate of Nigerian languages such as Yoruba, would be close to near extinction.The recent policy of Nigeria Federal Ministry of Educationthat made the study of indiginous languages optional in the Senior Secondary Schools do not help matters[1].This research provides a means of preventing extinction of Yoruba language. Also it helps in the flow of globilization by developing a web-based user friendly English to Yoruba Machine Translation System. This system is easily accessible to learning and to teaching the indigenes and anyone interested in Yoruba language. The translator is userfriendly and English words are easily translated to Yoruba words. More so, it assists in understandingthe Yoruba language with English language.

2. Related Works

[2][3] worked on Web based English to Yoruba machine translation. In the research, computational models were formulated using finite state automata, which was used todevelop a web-based translation system for Noun-phrases in English language to Yoruba language. Linguists were consulted and there was a detailed study of the syntactic structures of both languages with emphasis on noun-phrases. Rules were formulated for the generation of Noun-phrases from English to Yoruba which were specified using context-free grammar. Also, [4] worked on Development of an English to Yoruba Machine Translation system. The research work carried out computational analysis of English to Yoruba texts translation process. Rule-based approach was used to carry out the research. The translator was modeled using context-free grammar and re-write rules, Parse Tree and Automata theory-based techniques and design of corresponding software using UML.

Google incorporation language translation service, Google translate, is a system based on statistical machine translation which started in the year 2006 with two languages [5]. It is currently probably the best known online language translation service provider [6]. It performs hundreds of millions of translations every day. Presently, it offers full support for translation between 64 different languages. Google translate is a common existing tool that can translate Yoruba language to other languages and vice versa [7][8]. The efficiency of this research model will be evaluated with Google translate. [9] worked on Using Statistical Machine Translation As A language Translation tool for understanding Yoruba. Translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. Existing software tool kits were used. There was no English -Yoruba parallel corpus, which [5] had to create English - Yoruba parallel corpus.

3. Research Model

Rule based approach of machine translation was used for this system. Specifically, dictionary based type of rule based was used as the main approach which is the most realistic of all the types of machine translation. Figure 1 shows the architecture of the system and the process that every translation will follow.
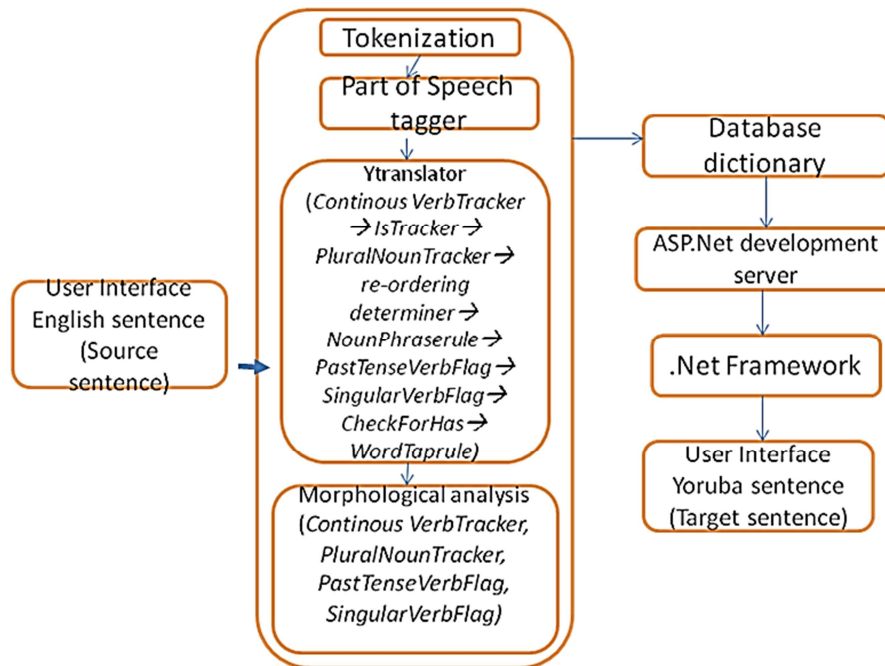

Figure 1. *The System Architecture.*

Table 1. *List of computational rules generated for both English and Yoruba sentences.*

S/N	English rules	Yoruba arrangement of the rule	Translator's rule
R_1	NP = det + N	NP = N + det	Re-ordering determinants
R_2	NP = det + a + N	NP = N + a + det	Noun Phrase
R_3	NP = p + N	NP = N + p	Re-ordering determinants
R_4	NP = p + a + N	NP = N + a + p	Noun Phrase
R_5	V = "is" + a	V = ε + a	"is" = Empty(*IsTracking*)
R_6	V = "is" + lvc	V = "n" + LV	Continuous verb
R_7	V = "is" + det "a"	V = "je" + ε	Det "a" = empty
R_8	V = "is" + det "the" +N	V = "ni" + N + det	(*IsTracking*)
R_9	NP = pn	NP = "awon" + N	Plural Noun
R_{10}	NP = det + pn	NP = "awon" + N + det	Plural Noun
R_{11}	NP = det + a + pn	NP = "awon" + N + a + det	Plural Noun
R_{12}	NP = p + pn	NP = "awon" + N + p	Plural Noun
R_{13}	NP = p + a + pn	NP = "awon" + N + a + p	Plural Noun
R_{14}	V = lvs	V = "maa n" + LV	Singular verbs
R_{15}	V = lvc	V = LV	Continuous verb
R_{16}	V = "has" + det + N	V = "ni" + N + det	*CheckforHas*
R_{17}	V = "has" + det + a + N	V = "n" + N + a + det	
R_{18}	V = "has" + LV	V = "ti" + LV	
R_{19}	V = "has" + "to"	V = "ni" + Lati	
R_{20}	V = "to" + det + N	V = "si" + N + det	
R_{21}	V = V + "d"	V = LV	Past tense verb
R_{22}	V = V + "ed"	V = LV	Past tense verb

From figure 1, tokenization is the first step in the translation process. It is the splitting of the input sentence which is in English language into words which are tokens. Then each token will be tagged with part of speech. 22 computational rules were formulated and they form the basis for the translator, which is called Y-Translator. These computational rules were formed based on some selected English Grammar rules and their arrangement in Yoruba language. The 22 computational rules are represented in table 1.

From the computational rules, production rules based on context free grammar were also formulated for English and Yoruba sentence structure. For English structure, the production rules are as follow:

1. S → NP VP
2. NP → N | p | dN | daN | pN | paN | a | ε
3. VP → V NP
4. V → avLV | av | LV | LVab
5. LV → lvc | lvp | lvs
6. N → sn | pn

The Production rules based on Yoruba sentence structure are as follow:

1. S → NP VP
2. NP → N | p | Np | Nad | Np | Nap | a | ε
3. VP → V NP
4. V → avLV | av | LV | Lvab
5. N → sn | pn

Where 'S' means Sentence, 'NP' means Noun Phrase, 'VP' means verb phrase, 'N' means noun, 'P' means Pronoun, 'd' means determinant, 'a' means adjective, 'V' means verb'av' means auxiliary verb, 'LV' means Lexical verb, 'lvc' means continuous lexical verb, 'lvp' means plural lexical verb, 'lvs' means singular lexical verb, 'sn' means singular noun, 'pn' means plural noun, 'ab' means adverb and 'ε' represents empty.

The computational rules are categorized into nine which comprise of; word replacers, *ContinuousVerbTracker*,

P*luralNounTracker*, S*ingularVerbFlag*, re-ordering determinant, *PastTenseVerbFlag, IsTracker, NounPhraseRule and WordTapRule*. *ContinuousVerbTracker* component recognizes and translates continuous verbs. *PluralNounTracker* recognizes and translate plural noun by removing the suffix's'. *SingularVerbFlag* recognizes and translate singular verbs. Re-ordering determinants recognizes determinants in the English sentence, then re-order the position of its translation in the Yoruba sentence. *PastTenseVerbFlag* recognizes and translate verbs in past tense and translate by retrieving the present form of the verb which exist in the database dictionary in Yoruba language. *NounPhraseRule* identifies, translate and re-arrange noun phrases in a sentence. *WordTapeRule* identifies the appropriate translation in a sentence, for words that have more than one part of speech.

Some of the rules generated for the model will require morphological analysis which are *ContinuousVerbTracker, SingularVerbFlag, PluralNounTracker and PastTenseVerbFlag*. The bound morphemes (Suffix) that the translator recognizes are "–*ing*", "-*ed*", "-*d*" and "-*s*". Meanwhile, the translation of the root word is retrieved from the dictionary.

The dictionary is bilingual that is, it contains words in English with their parts of speech with the equivalent word in Yoruba language. Some data were extracted from the data set to exist independently from the raw data. The data extracted were the "determinants", "auxiliary verbs" and "Pronoun" in English words with their corresponding translation in Yoruba. The raw data was also edited updated with irregular verbs which do not exist in the dictionary ideally and all words are well arranged. The data sets serves as the database structure for the machine translation. Figure 2 shows the structure of the general data set which contains every other word arranged alphabetically, while figure 3 shows the determinant data set.

Figure 2. An Extract of The Database Dictionary (General Data Set).
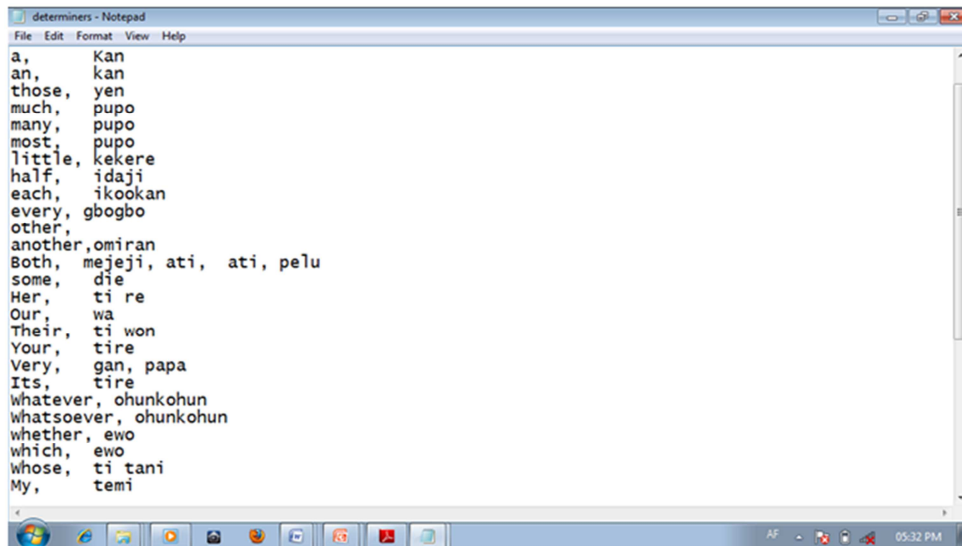
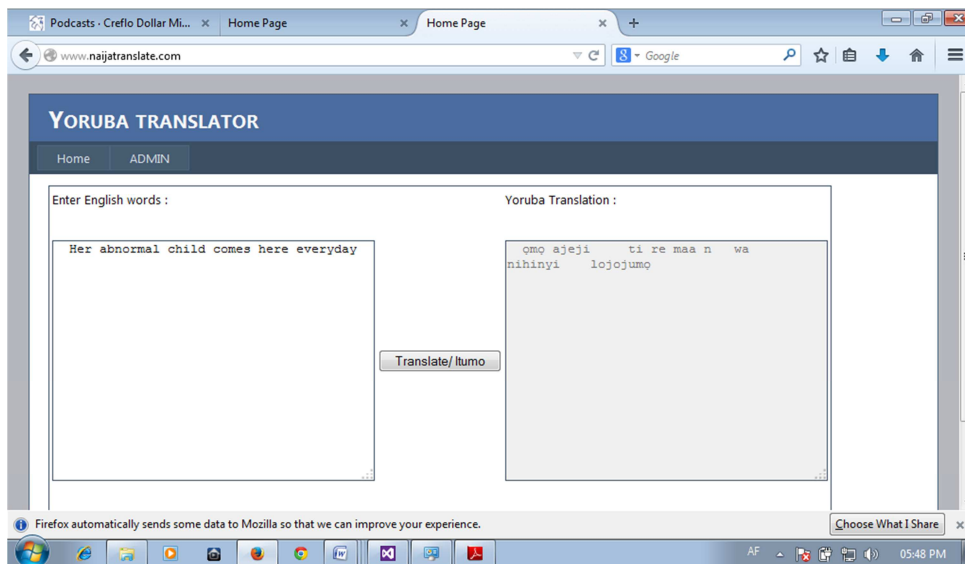**Figure 3.** *An Extract of the Database Dictionary (Determiners).*

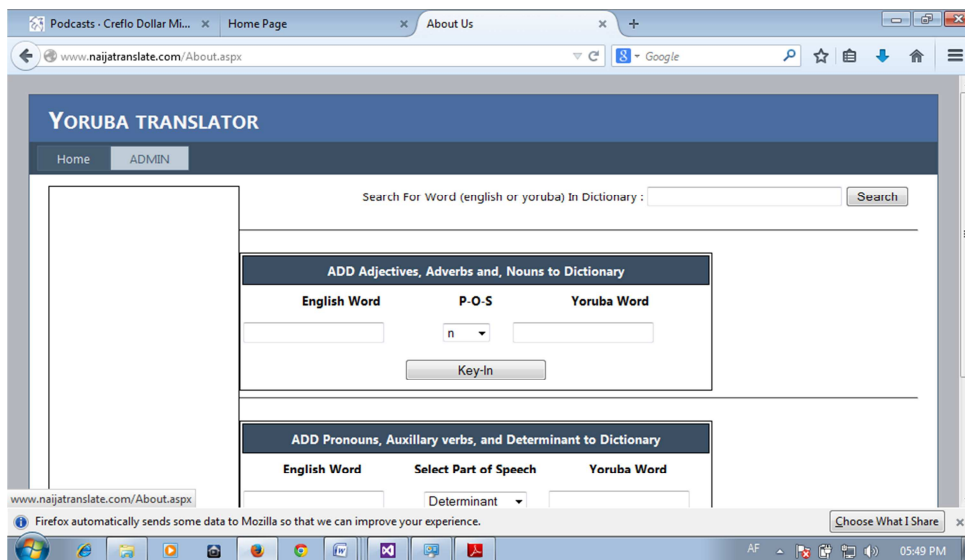**Figure 4.** *The Home Page of The Translator.*

Figure 5. *The Admin Page of The Translator.*

The research model was implemented with VISUAL STUDIO 2012, ASP.net – an interface design tool that makes use of HTML and C# programming language to code the design. Figure 4 shows the snapshot of the home page of the translator while figure 5 shows the snapshot of the ADMIN page which is purposely for updating the data set with new English and Yoruba which does not exist in the data set.

4. System Evaluation

The evaluation was carried out by comparing the efficiency of this translator with Google translate. Table 2 is an extract of the sentences tested on Google translate (http://www.translate.google.com/m) and Y-Translation model (http://www.naijatranslate.com).

Table 2. An Extract of the Sentences translated on The Ytranslator and Google Translate.

Input Sentenes	Y-Translation Model output	Google Translate Output
She wrote easily	Obinrin naa kọẹ lainira	O kọwe awọn isoro
She has written in the book	Obinrin naa kọ ninu iwe naa	O ti kọ ninu awon iwe
The boy hit the dog	ọmọkunrin naa lu aja naa	Awọn ọmọkunrin lu awọn aja
She is a girl	Obinrin naa je ọmọbinrin	O ni kan ọmọbinrin
The lady must be in the car	Iyalode naa gbudọ wa ninu ọkọ naa	Awọn iyaafin gbọdọ wa ni awọn ọkọ
He bought a book for them	Okunrin naa ra iwe kan fun wọn	O si ra iwe kan fun wọn
She bought a weapon for herself	Obinrin naa ra ohun ija kan fun ontikarare	O ra kanija fun ara
Each man must eat their food	Okunrin ikọọkan gbudọ je ounje ti wọn	Kọọkan ọkunrin gbọdọ je wọn ounje
She bought a weapon for him	Obinrin naa ra ohun ija kan fun on	O ra kan ija fun u

Table 3. The Result of the System Based on 200 Sentences.

Translator	Sentences Generated	Correctly Translated	Wrongly Translated	Partially Translated	Partial Translation%	Absolute Accuracy %
Y-Translator	200	181	0	19	100	90.5
Google Translate	200	32	117	51	40.5	16

5. Result of the Evaluation

From the result shown in table 3, the partial translation refers to sentences translated that are grammatically correct but not perfectly translated that is, the translated sentence in Yoruba does not have the same meaning with the English sentence inputed to the system. Partial Accuracy is the percentage of all grammatically correct sentences. The absolute accuraccy is the percentage of senteces translated that are grammaticaly correct with perfect translation of the English sentence inputed. Wrongly translated are sentences that are not perfectly translated and grammatically incorrect in Yoruba language. Figure 6 shows the number of correctly translated, wrongly translated and partially translated for Y-Translator and Google translate. Meanwhile, figure 7 shows the percentage distribution of the accuracy level of both translators.
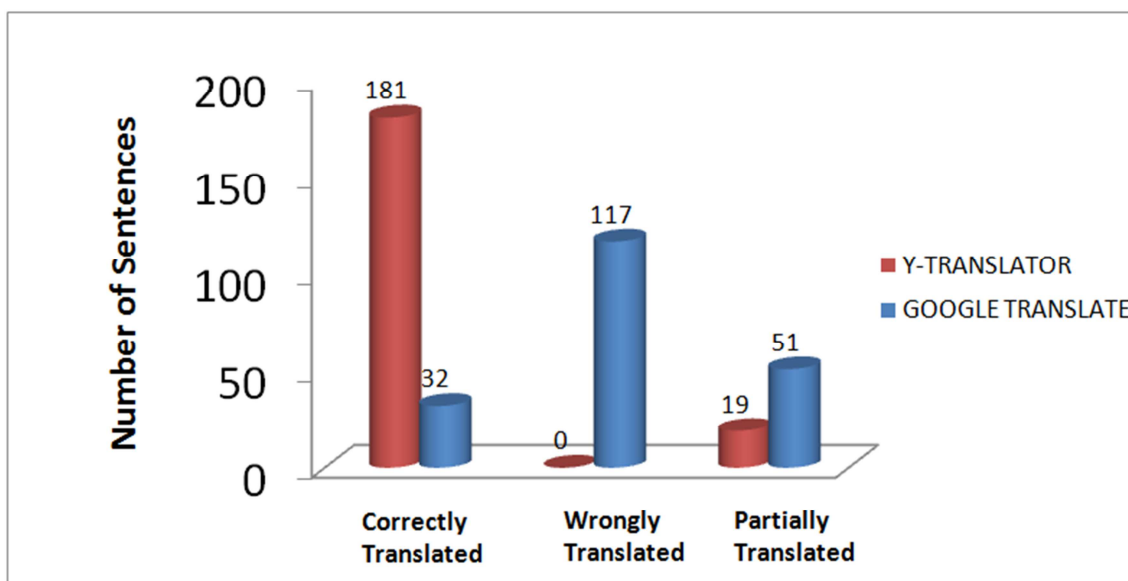

Figure 6. Level of Translation Accuracy Distribution based on 200 Sentences.
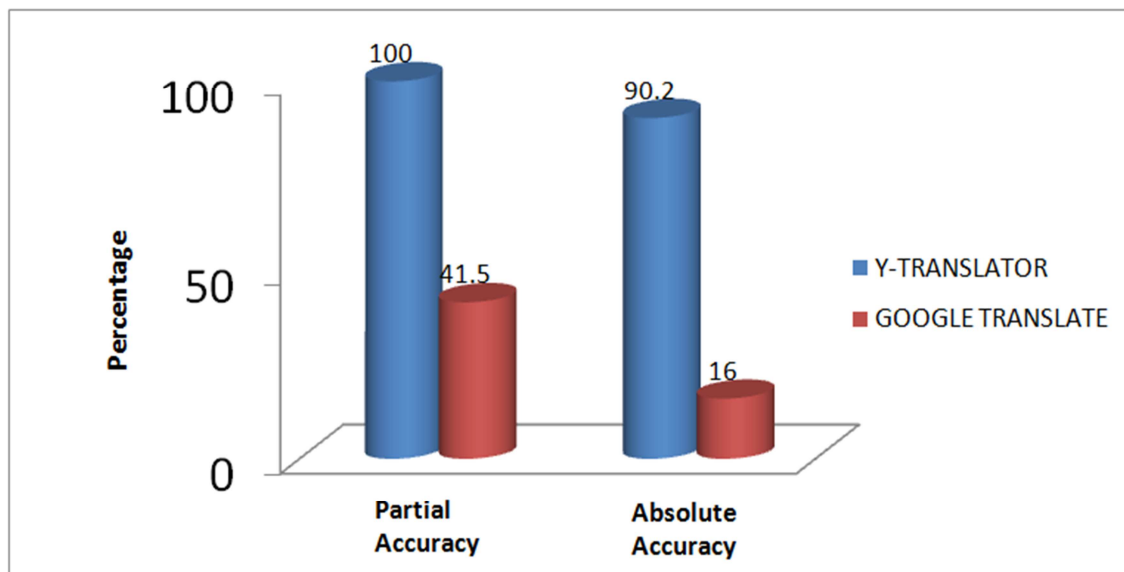
Figure 7. *Percentage Distribution of the Accuracy of the Translators.*

6. Conclusion

The rule based approach to machine translation is still the most realizable method for translationas shown from the result. The translator has been able to translate past tense verbs, continous verb tense and plural noun which will require additional words in Yoruba language. This is an improvement on existing research work on English language to Yoruba language tanslation. The translator has been able to translate a complete sentence in English language to Yoruba langue which is an improvement to the research carried out by Abiola et al. (2014). Also, the translator has a higher translation accuracy than Google translate. There is still room for improvements in this research work. There is need for a rule that can translate English words that have more than one meaning due to their part of speech. That is, a rule that will be able to recognise which translation will be appropriate for such words based on their part of speech in a particular sentence. The translator sometimes have to apply more than on rule to translate a particular sentence. Therefore fewer rules are recommended to avoid confusion for the translator.

References

[1] Chux O. (2013, August 8). *Punch Newspaper*. Retrieved from http://www.punchng.com/feature/gradually-nigerian-languages-are-dying/

[2] Abiola O.B., Adetunmbi A.O., Fasiku A. I. & Olatunji K. A. (2014). A Web-Based English to Yoruba noun-phrases Machine Translation System. *International Journal of English and Literature,*5(3),71-78.http//dx.doi.org/10.5897/IJEL2013.0472.

[3] Adeoye O., (2012). *A Web-Based English to Yoruba Noun-Phrase Machine Translation*. (Masters Thesis). Federal University of Technology, Akure, Nigeria.

[4] Eludiora, S. (2013). *Development of an English to Yoruba machine Translation system*. (Doctoral dissertation) Obafemi Awolowo University, Ile-Ife, Nigeria.

[5] Haiying L., Arthur C. G. & Zhaiqiang C. (2014). *Comparison of Google Translation with Human Translation: Proceedings of the Twenty-seventh International Florida Artificial Intelligence Research Society Conference, University of Memphis, Institute of Intelligent system, Memphis, USA.*

[6] Stephen H. & Carmen P. S. (2010). *Translation and the Internet: Evaluating the Quality of free Online Machine Translators,* Quaderns. Rev. trad. 17, 197 – 209.

[7] Henderson F., (2010). *Giving a Voice to More languages on Google Translate, The official Google Translate blog.* Available: http://googletranslate.blogspot.com/2010/05/giving-voice-to-more-languages-on.html.

[8] Mahsa, M. (2012). *English - Persian Phrase - Based Statistical Machine Translation: Enhanced Models, Search and Training*. (Doctoral dissertation). Massey University, Albany (Auckland), New Zealand.

[9] Yetunde, O. F. and Omonayin, I. (2012). Using Statistical Machine Translation As A language Translation tool for understanding Yoruba. *EIE's 2nd International Conference on.Computing Energy, Networking, Robotics and Telecommunications*. eieCon2012. 86-91. http://dx.doi.org/10.13140/2.1.3522.8454.