Vol. 5(3), pp. 71-78, May, 2014
DOI: 10.5897/IJEL2013.0472
ISSN 2141-2626

http://www.academicjournals.org/IJEL

International Journal of English and Literature

Full Length Research Paper

A web-based English to Yoruba noun-phrases machine translation system

Abiola O.B[1], Adetunmbi A.O[2], Fasiku A.I.[3] and Olatunji K.A[1]

[1]Computer Science Department, Afe Babalola University, Ado – Ekiti, Nigeria.
[2]Computer Science Department, Federal University of Technology, Akure, Nigeria.
[3]Computer Engineering Department, Ekiti State University, Ado – Ekiti, Nigeria.

Received 10 May 2013; Accepted 9 April, 2014

The field of natural language processing enables machines to read and understand the languages human being speaks. There are three major languages in Nigeria: Yorùbá, Igbo and Hausa. Yorùbá, a major Nigeria language spoken by over fifty million people which has the potentials of serving as medium for scientific and technological development deserves more recognition than it is in Nigeria today. Developing a computational model for English language and Yoruba language noun-phrases involve a profound understanding of the syntactic and grammatical features of the two languages as well as their vocabularies since they are not related syntactically and grammatically. Twenty nine rules were formulated for the noun phrase translations which were specified using the context free grammar (CFG). We then modeled and recognized the grammar of the language using the finite state automata (FSA) whose operations was based on the first set techniques. The first sets techniques allow the parser to choose which production rule to apply based on the first input word of an input phrase. We also developed a bilingual lexicon which is made up of words in English language with their corresponding Yoruba counterparts and their equivalent part of speech. The model was implemented using PHP Hypertext Preprocessor (PhP) programming language and my structured query language (SQL) and was tested on four-hundred randomly selected noun-phrases and gives accuracy of 91% which is quite encouraging.

Key words: Natural language processing, English, Yoruba, computational model, noun-phrases, translation system, context-free grammar and finite state automata.

INTRODUCTION

In Nigeria, there are three major Indigenous languages: Yorùbá, Igbo and Hausa. The languages spoken in Nigeria are not evenly distributed, for instance in the South-West part of Nigeria, Yorùbá is largely spoken; Igbo is largely spoken in the South-East part of Nigeria; while in the North-West part of Nigeria, Hausa is largely spoken. (Yusuf et al, 2007). The dominance of the English language is quite overwhelming in Nigeria; this can be seen in practically all domains: government and administration, education, the media, the judiciary, science and technology to mention but few. High government officials avoid using their languages in official contacts even with their own people for fear of being labelled tribal and parochial. In the national and state houses of assembly, English language continues to be the language of debate and record in spite of the fact that

*Corresponding author. E-mail: adeoyetoyin@yahoo.com

for the use of the major indigenous Nigerian languages. (Fabunmi and Akeem, 2005).

The use of computers has so far been greatly restricted only to those people who have some knowledge of the English language. This has resulted in a fast way of killing the major indigenous languages in the country especially the Yoruba language. The Yoruba language is less used among its people because its roles have been taken over by the English language. This prompted the need to develop a machine translation system that will give Yoruba language a public profile in the information technology (IT) world so as to provide a platform for the people to really appreciate the beauty of their indigenous language. (Adeoye, 2012).

Kobomoje (2008), described translation as the transfer of the meaning of a text from one language to another for a new relationship. This implies that translation is not a straight forward case of substituting word(s) in the source language (SL) with the equivalent word(s) in the target language (TL). The translated text must convey the same meaning as the original text meaning. This is to say that, the translator must understand the message that the author of the original text is trying to convey. Translating from one language to another involve a proper understanding of the grammar of the two languages that are involved. Our developed system involves English as the source language and Yorùbá as the target language. These two languages are not closely related in terms of their syntax, grammatical structures and vocabularies. The differences in their syntactical and grammatical structures, made us to first have a profound understanding of the syntactical and the grammatical features of the two languages involved as well as their vocabularies.

The methodology behind the system combines the formulation of some grammatical rules for the generation of noun-phrases in the two languages as well as developing computational models for recognizing the grammar of the language. The grammatical rules that were manually formulated, using some literatures, reflect the most common local syntactic differences between English and Yoruba. These small set of rules turns out to be already sufficient for producing some legible translations of some noun-phrases from some selected documents. The system is first realized, by specifying the syntax of the two languages using CFG. A FSA was then used in modelling and recognizing the grammar of the language.

The operation of the FSA was based on the first sets techniques. The first sets techniques allow the parser to choose which production rule to apply based on the first and the next input word(s) or tokens of any input string. A bilingual lexicon was developed which is made up of words in English language with their corresponding Yoruba counterparts and their equivalent parts of speech.

Yoruba versus English language

The only channel by which human beings abstract reality is language. Yorùbá, (native name èdè Yorùbá, 'the Yorùbá language') is a dialect continuum of West Africa with over 50 million speakers. The Yorùbá language is slipping away from us because of the various trans-national structural revolutions going on in the world today in the name of globalization. Yorùbá language is at the point of death because some of its roles have been taken over by the English language. Although a language only dies when nobody speaks it any more, Yorùbá is yet to die even though people are still speaking it. But the threat of extinction is still solidly there. Yorùbá still exists in Nigeria today because of high-level of illiteracy. If we have a low percentage of literacy, the language will be gone (Fabunmi and Akeem, 2005).

Parents want their children to speak and learn English straight from infancy. The negative effects of the negligence and negligible use of Yorùbá by the élite, has spilling over effects on Yorùbá as a discipline. Many Yorùbá words have virtually disappeared, and taken over by English loan words. Yorùbá unlike English is not a compulsory subject needed to gain an admission into any Nigerian university. This criterion alone always gives English a dominant edge over any of the recognized three indigenous Nigerian languages.

Most of the African languages are tonal languages among which Yorùbá is one. Yorùbá is a tonal language with three level tones: High (Ohùn òkè), Low (Ohùn Ìsàlẹ̀) and Mid (Ohùn àárín), represented with [´], [`] and [¯] respectively. Every syllable must have at least one tone; a syllable containing a long vowel can have two tones. The three level tones determine the meanings that each word has in Yorùbá language. For example, a form that has the same form of vowels and consonants can have different meanings depending on the tones that it has. That is the tonality of a word can totally alter the meaning. The following examples present all the three tonalities in Yoruba language.

(i) Òjó 'personal name' (ii) Òjò 'rain'
(iii) Ojo 'cowardice'
(i) Igba 'two hundred'(ii) Igbá 'calabash' (iii) Ìgbà 'time' (iv) Ìgbá 'garden egg'
(v) Igbà 'climbing rope'

Different analyses of the same input word may result in a different number of outputs and we were able to overcome this by using the "Konyin Nigeria multilingual keyboard" to enter all Yoruba texts to indicate the tone on each syllable of a word demonstrativessubstitute for nouns in some cases and implies a gesture of pointing to something in the situational context. Examples are: 'this, these, that, those'.

The Noun-Phrase

According to (Howard, 1982 and Bamisaye, 2000), the noun-phrase in English is composed potentially of three parts. The head which is the central part and the minimal requirement for the occurrence of a noun-phrase. The other two parts are optionally occurring. The head may be preceded by some pre-modification, and it may be followed by some post-modification.
Noun- phrase can be made up of nouns, noun modifiers, adjectives and the following sub-divisions of the parts of speech: Determiners, numerals and predeterminers.
Determiners: are classes of words that are used with nouns and have the function of defining the reference of the noun in some way. Examples of determiners are:

i. Articles which can either be a definite article or Indefinite article. For definite article we have 'the', Indefinite articles are 'a, an'.
ii. Demonstrative:
iii. Possessive are 'my, your, his, her, its, our their'…
iv. Quantifiers are 'many, few, several'…

b. Numerals are:

i. Cardinal numerals which include 'one, two, three, four, five, six'…
ii. Ordinal numerals which are 'first, second, third, fourth, fifth'…

c. Predeterminers are all, both, half…

Based on the subdivisions described above, the following rules were generated for the translations of English to Yoruba. Noun-phrases which were specified using the CFG. (Table 1 shows the English rules for noun phrases and Yoruba arrangement of the rule (Awobuluyi 1978))

Where R_1 to R_{29} are the rules number: R_1 means rule 1…
N means nouns
Adj means adjectives
Dart means definite article
Inart means indefinite article
Dem means demonstrative
Poss means possessive
Quant means quantifier
PreDet means predeterminer
CardNum means cardinal numerals
OrdNum means ordinal numerals
Nmod means noun modifier this depend on the noun modifying another in phrase

For example, The little child means ọmọ kékeré naa in Yoruba language
NP ⟶ (Dart) (Adj) (N) (Rule 3)
⟶ The (Adj) (N)
⟶ The little (N)
⟶ The little child
The Yorùbá arrangement of the phrase is given as:

NP⟶ (N) (Adj) (Dart)
⟶ Ọmọ (Adj) (Dart)
⟶ Ọmọ kékeré (Dart)
⟶ Ọmọ kékeré naa

We then modeled and recognized the grammar of the language using the FSA whose operations was based on the first set techniques. The first sets techniques allow the parser to choose which production rule to apply based on the first input word of an input phrase.
For example,
If an input phrase is a combination of Dart, Adj and Noun, the system chooses the right production rule in that order for the Yoruba translations.

The proposed model

The major task behind the translation is developing an exhaustive lexicon consisting of the source language words along with its corresponding translated version of the target language. Any word to be translated is checked in the developed bilingual lexicon, if found it is replaced with the translated version stored in the database. The translation system has three main blocks as illustrated in Figure 1 below.

Preprocessing

The process begins with the preprocessing of the SL text. When a user enters an input text (noun-phrase), the input is first stored and then preprocessed to know the number of words present in the text. The system recognized a word whenever a space is encountered, which signifies the end of the word and eliminate the space automatically while carry space signifies the end of the phrase.

Lexical transfer

The lexical transfer performs the transfer of each word in the source text, by assigning to each word of the source text, its corresponding target word counterpart and the equivalent part of speech. This is done with the use of the bilingual lexicon, which is made up of a bilingual exhaustive lexicon of Yorùbá and English words with their

Table 1. English rules for noun phrases and Yoruba arrangement of the rule.

S/No	English rules for noun phrases.	Yoruba arrangement of the rule.
R_1.	NP=Dart+N	NP=N+Dart
R_2.	NP=Inart+N	NP=N+InArt
R_3.	NP=Dart+Adj+N	NP=N+Adj+Dart
R_4.	NP=Inart+Adj+N	NP=N+Adj+Inart
R_5.	NP=Dart+Adj+Adj+N	NP=N+Adj+Adj+Dart
R_6.	NP=Inart+Adj+Adj+N	NP=N+Adj+Adj+Inart
R_7.	NP=Dart+OrdNum+N	NP=N+OrdNum+Dart
R_8.	NP=Dart+CardNum+Adj+N	NP=N+Adj+CardNum+Dart
R_9	NP=Dart+OrdNum+Quant+N	NP=N+Quant+OrdNum+Dart
R_{10}.	NP=Dart+Nmod+N	NP=N+Nmod+Dart
R_{11}.	NP=Inart+OrdNum+N	NP=N+OrdNum+Inart
R_{12}.	NP=Dart+CardNum+N	NP=N+CardNum+Dart
R_{13}.	NP=Dem+N	NP=N+Dem
R_{14}.	NP=Dem+CardNum+N	NP=N+CardNum+Dem
R_{15}.	NP=Dem+CardNum+Adj+N	NP=N+Adj+CardNum+Dem
R_{16}.	NP=Poss+N	NP=N+Poss
R_{17}.	NP=Poss+Adj+N	NP=Adj+Poss+N
R_{18}.	NP=Poss+OrdNum+N	NP=N+Poss+OrdNum
R_{19}.	NP=Poss+CardNum+N	NP=N+CardNum+Poss
R_{20}.	NP=Poss+Adj+Adj+N	NP=N+Poss+Adj+Adj
R_{21}.	NP=Poss+OrdNum+Adj+N	NP=N+Poss+Adj+OrdNum
R_{22}.	NP=Quant+N	NP=N+Quant
R_{23}.	NP=Quant+Adj+N	NP=Quant+N+Adj
R_{24}.	NP=Quant+CardNum+N	NP=Quant+CardNum+N
R_{25}.	NP=CardNum+N	NP=N+CardNum
R_{26}.	NP=OrdNum+N	NP=N+OrdNum
R_{27}.	NP=preDet+N	NP=PreDet+N
R_{28}.	NP=PreDet+Dart+N	NP=PreDet+N+Dart
R_{29}.	NP=Nmod+N	NP=N+Nmod

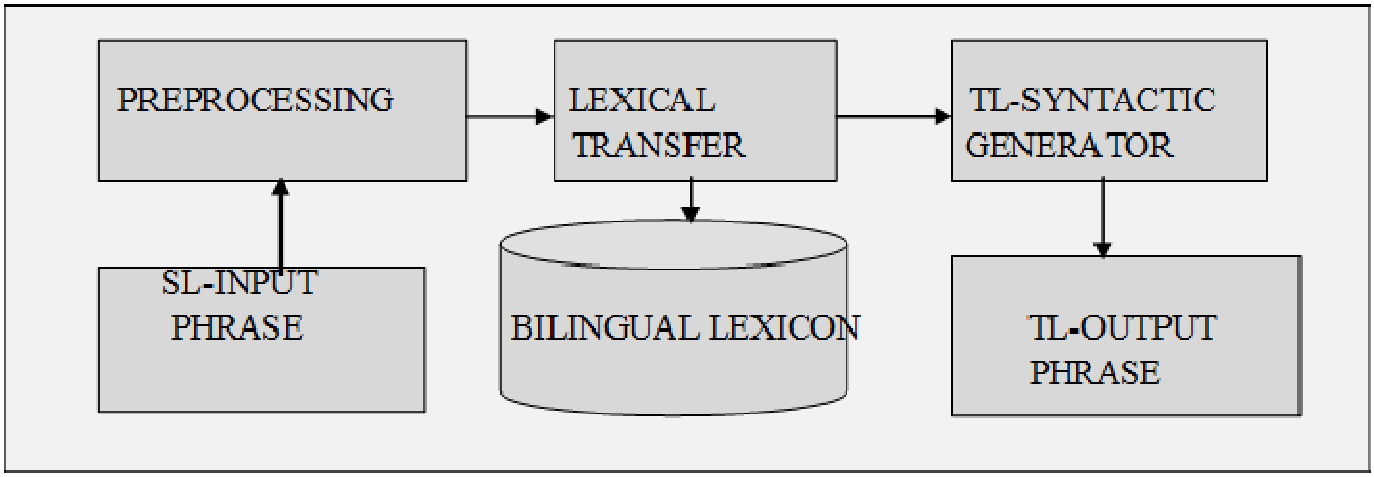

Figure 1. The Translation System Blocks Diagram.

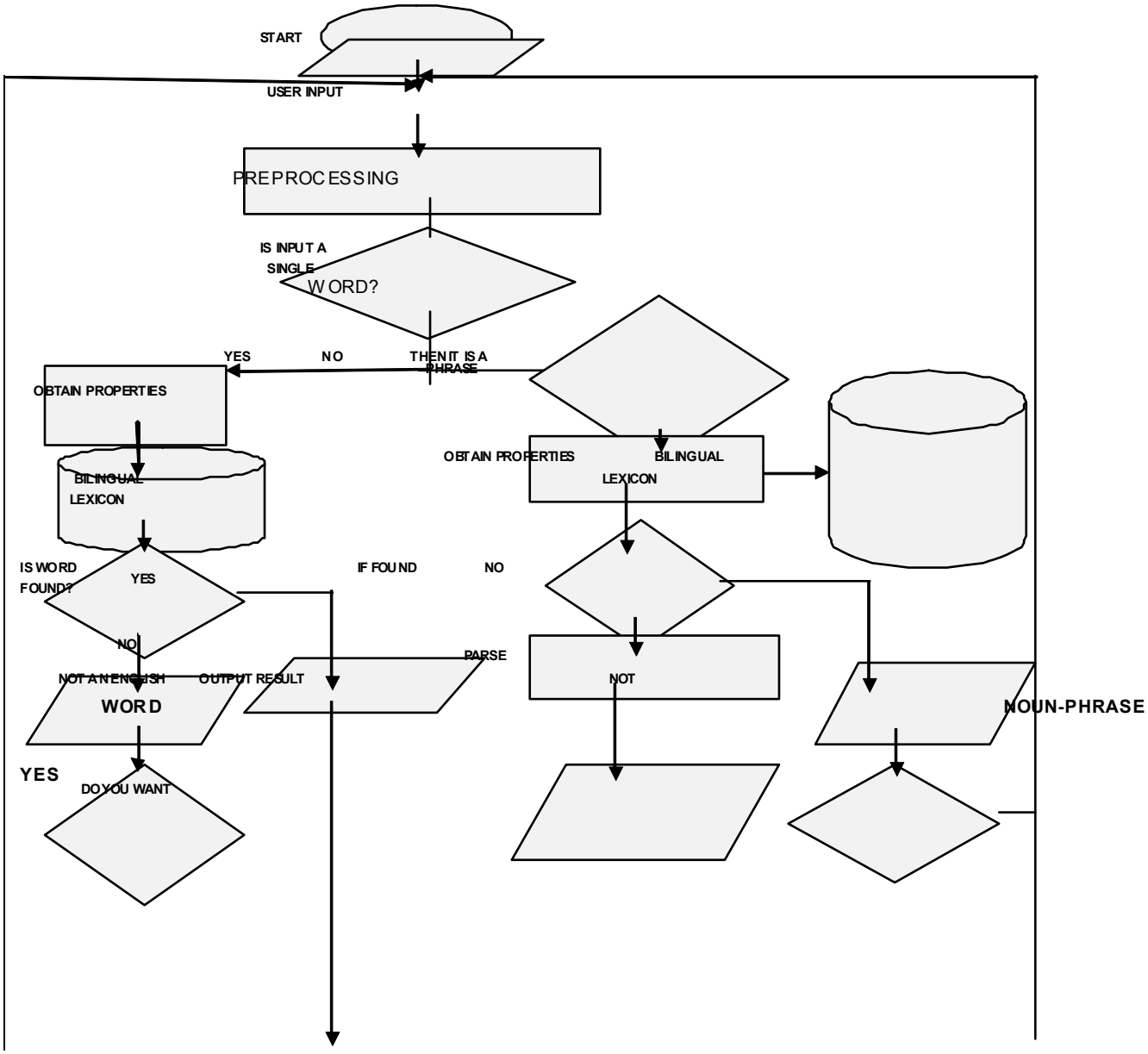

Figure 2. The operational flow of the translation System.

corresponding parts of speech. Undoubtedly, the bilingual lexicon is one of the main bottlenecks of our system and a better dictionary will improve the results significantly.

TL-Syntactic generator

In the target language syntactic generator block, the source language words translated to target language counterparts are processed and the output phrase is produced in the target language.

Figure 2 depicts the operational flow of the system. Once English input is entered, it undergoes pre-processing. If the input is a single word, the properties of the word which include its Yoruba counterpart and the part of speech is obtained from the bilingual lexicon. If the input text is a phrase, the properties for each word in the phrase is obtained from the bilingual lexicon any word that is not found in the bilingual lexicon might not have been stored in the database. The input is then passed on

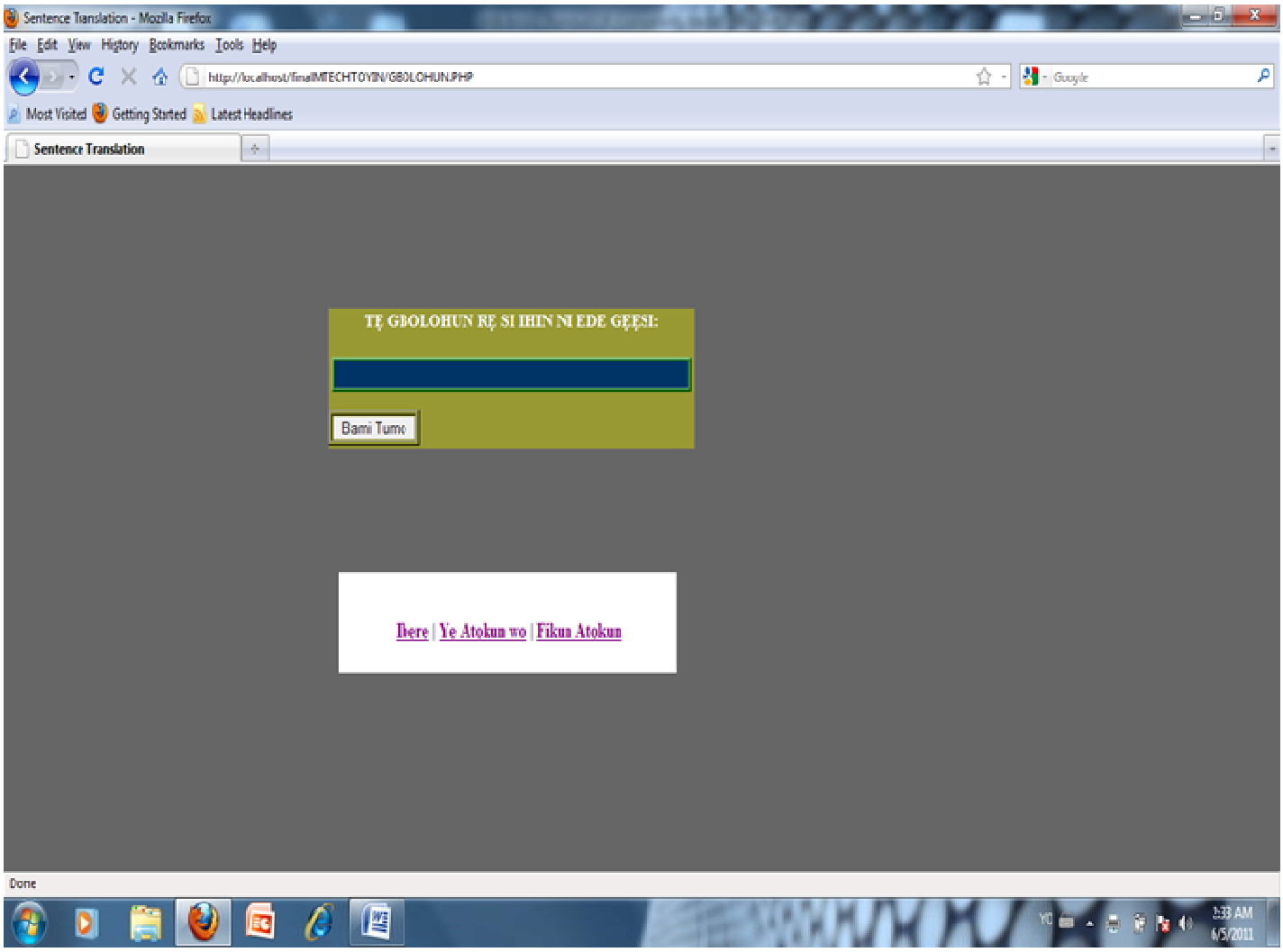

Figure 3. The translation interface.

to the rule engine which applies a collection of lexical and structural transfer rules based on first set techniques in order to parse and transfer as well as to generate the Yoruba translations for all possible words in the phrase. If the input is not parsed, then the input is possibly not a noun-phrase or English words. The system is user friendly and was properly designed. It is aimed at providing technical solutions for the usage of Yoruba language on the web and as well, provides a platform for people to really appreciate the beauty of their indigenous language.

Figure 3 shows the translation interface. The translation interface is named 'ìtúmọ̀ èdè gẹ̀ẹ́sì tí o sòro diẹ̀' when this is clicked, the user is provided with an interface to enter words or noun-phrases in English language so as to obtain the Yorùbá translations. After typing a noun-phrase on the text field, the user clicks the "Bami Tumo"or "translate" button to obtain the meaning of the phrase entered. Any word that is not interpreted might have not been stored. The translation interface also leads to other interfaces on the web. (Figure 3, 4, 5)

System implementation and experimental set-up

We developed a bilingual lexicon which is made up of words in English language with their corresponding Yoruba counterparts and their equivalent part of speech. Noun-phrases were generated randomly from four documents: Daily news papers, the Holy Bible, hymnal and motivational books to calculate the accuracy of the system. A total number of four-hundred noun-phrases

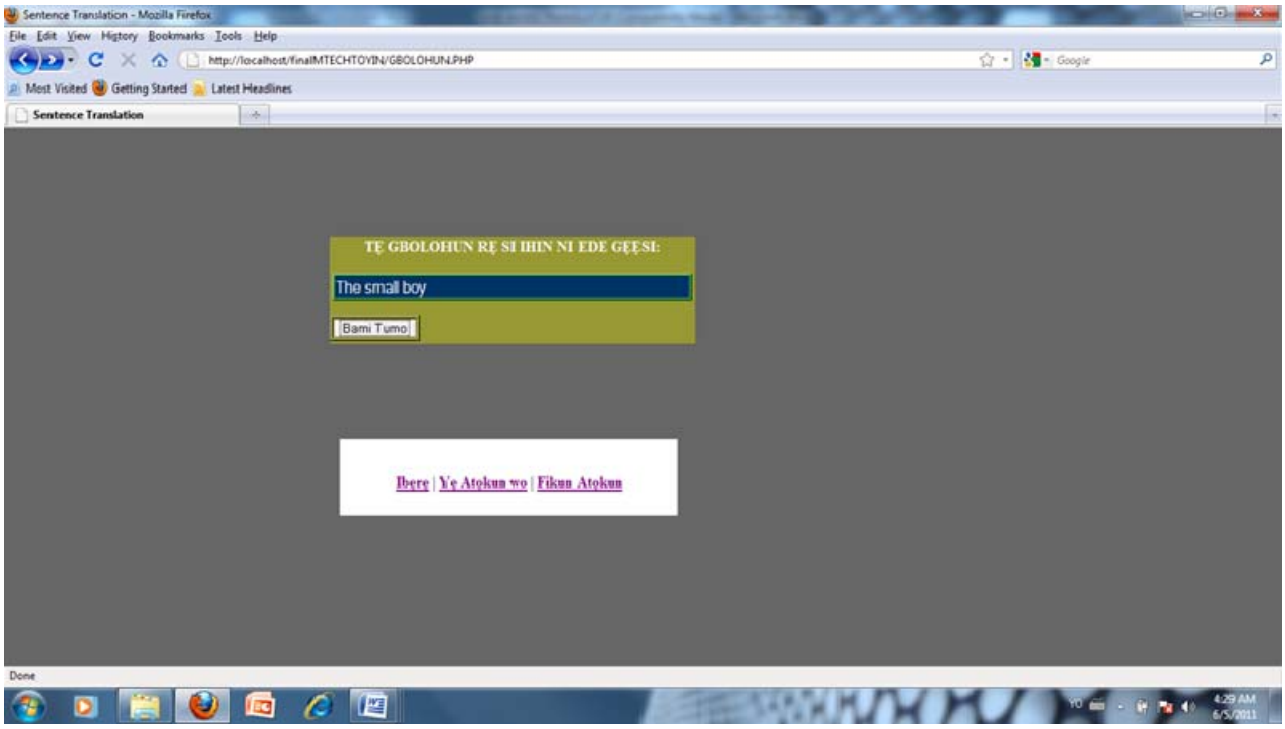

Figure 4. An example of a translation “the small boy”

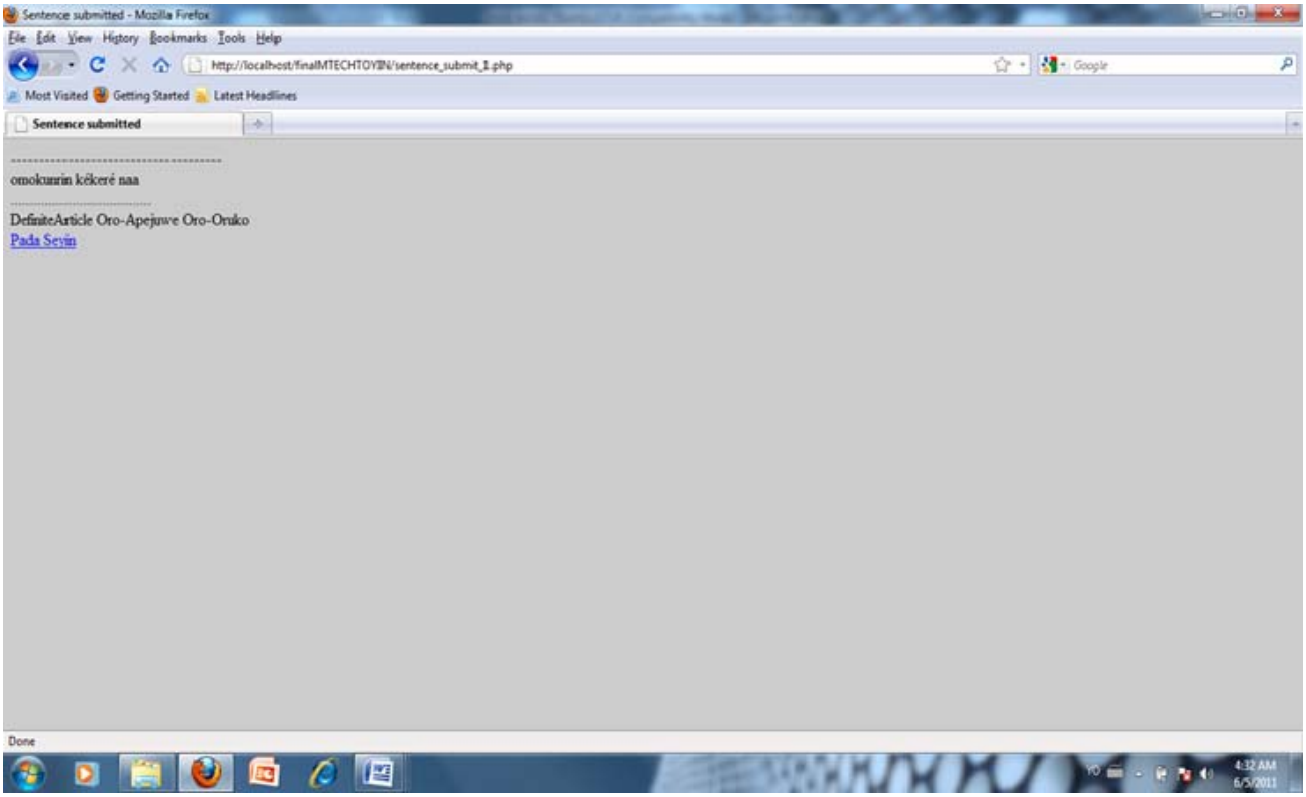

Figure 5. The Yoruba interpretation: “omokunrin kekere naa”

Table 2. Results of the system.

Documents	Phrases generated	Correctly translated phrases	Wrongly translated phrases	Accuracy
DAILY NEWS	160	146	14	91%
HOLY BIBLE	70	65	5	93%
HYMNAL	100	88	12	88%
MOTIVATIONAL BOOKS	70	66	4	94%
TOTAL	400	365	35	91%

were generated as datasets for the system. The following results were obtained and the overall performance accuracy of the system was 91%. This shows that the performance of the translation system is quite good and encouraging. (Table 2)

The model was implemented using PhP programming language and MySQL. The system if fully developed will go a long way in preventing the extinction threat of the Yoruba language.

CONCLUSION AND FUTURE WORK

From the above analysis, it is concluded that the overall accuracy of English to Yoruba noun-phrases machine translation system is 91%. The accuracy can be improved by improving and extending the bilingual lexicon. The current version of our work performs translations of only noun-phrase which is part of a complete sentence and it produces promising and acceptable translations. The system is still under development to achieve higher quality translations; we are hoping to address other phrases that make up a complete sentence and as well use machine learning techniques in our future work. It is hopeful that the model will go a long way at providing a global easy to read guide for all the words and noun-phrases that learners need to communicate with in the language thereby, improving the use of the language among its people. The dying aspects of the language and its culture will equally be preserved by providing technical solutions to its usage. The system will be of immense benefits among the Yoruba people and those that are willing to learn the language.

Conflict of Interests

The author(s) have not declared any conflict of interests

REFERENCES

Adeoye OB (2012). "A Web-Based English to Yoruba Noun-Phrases Machine Translation System", M.Tech Thesis, Federal University of Technology, Akure, Nigeria.

Akshi K (2005). "Design and Development of Translator's Workbench for English to Indian Lang." Translation J. 9(3). Retrieve December, 2010. Source at: http:// accurapid.com/journal/33TWB.htm.

www.nlp.hivefire,com/entity/profile/andrew-mccallum. Retrieved, 2010.

Awobuluyi O (1978). "Essentials of Yoruba Grammar" Published by Oxford University Press Nigeria, Iddo Gate Ibadan. ISBN 0195753003.

Bamisaye OT (2000). "Essentials of English Syntax" Department of English, University of Ado-Ekiti, Nigeria. Published by Balfak Educational Publisher, Ado-Ekiti, Ekiti State ISBN 978-2558-14-04.

Fabunmi AF, Akeem SS (2005) "Is Yoruba an Endagered Language" Nordic Journal of African Studies 14(3): 391-408.

Howard J (1982). "Analyzing English an Introduction to Descriptive Linguistics" City of Birmingham Polytechnic, United Kingdom. ISBN 0-08-028667-4. www.bcu.ac.uk/pme/school of english/staff/howard_jackson. Retrieved, 2011.

Kobomoje A (2008). "Problems of Translations Yoruba and English in Focus"B.Sc Thesis Department of Linguistics and Nigerian Languages, Faculty of Arts, University of Ado-Ekiti, Ekiti State. Nigeria.

Yusuf O (2007). "Basic Linguistics for Nigerian Languages Teachers" Published by Linguistics Association of Nigeria in collaboration with M and J Grand Orbit Communication Limited; and Emhai Press Port-Harcourt. ISBN 978-33527-4-2.