fjrs.futa.edu.ng

FUTA Journal of Research in Sciences

ISSN: 2315-8239

FUTA Journal of Research in Sciences, 2013 (1): 34-43

A COMPUTATIONAL MODEL OF ENGLISH TO YORUBA NOUN-PHRASES TRANSLATION SYSTEM

O.B. Abiola, A.O. Adetunmbi[1] and A. Oguntimilehin

Department of Computer Science, Afe Babalola University, Ado-Ekiti, Ekiti State, Nigeria.

adeoyetoyin@yahoo.com, ebenabiodun2@yahoo.com

[1] Department of Computer Science, Federal University of Technology, Akure, Ondo State.

bayoadetunmbi@yahoo.com

ABSTRACT

The field of natural language processing enables machines to read and understand the languages human being speaks. 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 programming language and MySQL and was tested on one hundred and sixty randomly selected noun-phrases from daily news, and gives accuracy of 91% which is quite encouraging. The system if fully developed will go a long way in preventing the extinction threat of the Yoruba language.

Keywords: Natural Language Processing, English, Yoruba, Noun-Phrases, Translation System, Context-Free Grammar and Finite State Automata.

INTRODUCTION

Natural language processing (NLP) is an area of research and application that gives machines the ability to read and understand the languages human beings speak (Chowdbury, 2005). To translate from one language into another, there is need for a proper understanding of the grammar of the two languages that are involved. In Nigeria, there are three major languages: Yorùbá, Igbo and Hausa. The dominance of the English language in Nigeria is quite overwhelming and the use of the computers has so far been greatly restricted only to those people who have some knowledge of the English language, thereby reducing the development of the major indigenous languages. (Yusuf, 2006). 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 as affirmed by (Bamisaye, 2000).

(a).Determiners: are classes of words that are used with nouns and have the function of defining the reference of the noun in some way. The following are groups of determiners as affirmed by (Bamisaye, 2000):

i. Articles which can either be definite article or indefinite article. For definite article we have 'the', while for indefinite articles we have 'a, an'.

ii. Demonstrative: demonstratives substitute for nouns in some cases and imply a gesture of pointing to something in the situational context. Examples are: 'this, these, that, those'.

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…

The rule of formation of a sentence in English language defers from the rule of formation in Yoruba language. This is because these two languages defers syntactically and grammatically.

In Yorùbá language, the rule of formation of a sentence is given by:

GB → APOR → APOI → APATK

Where GB = Gbólóhùn (Sentence)
APOR = Àpólà ọ̀rọ̀ Orúkọ (Noun-Phrase)
APOI = Àpólà ọ̀rọ̀-Ìse (Verb-Phrase)
APATK = Àpólà atọ́kùn (Prepositional phrase)
(Awobuluyi, 1978)

In English language, the rule of formation of a sentence is given by:

S ⟶ NP ⟶ AUX ⟶ VP

Where S = Sentence
NP = Noun-Phrase
Aux = Auxilliary
VP = Verb Phrase (Bamisaye, 2000).

From the rule of formation of sentences in both languages, as affirmed by (Awobuluyi, 1978) and (Bamisaye, 2000), it is cleared that the noun-phrase known as "Àpólà ọ̀rọ̀ Orúkọ" in Yoruba language, which is the emphasis of this work is just a part of a complete sentence. So we took our time to really study the grammatical structures of noun-phrases from English (the source language) to Yoruba (the target language).

COMPUTATIONAL MODELS

A grammar is a powerful tool for describing and analysing languages. It is a set of rules by which valid sentences in a language are constructed. (Alfred *et al*, 2006). In the design of the translation system depicted in figure 1, the source language text goes through lexical analysis, where each word in the source text is assigned its corresponding target word counterpart and the equivalent part of speech. This is done with the use of the bilingual lexicon, which comprises of words in English language and Yoruba language with their equivalent part of speech (Adeoye, 2012).

The source language words translated to target language counterparts is then processed to produce the output phase in the target language.

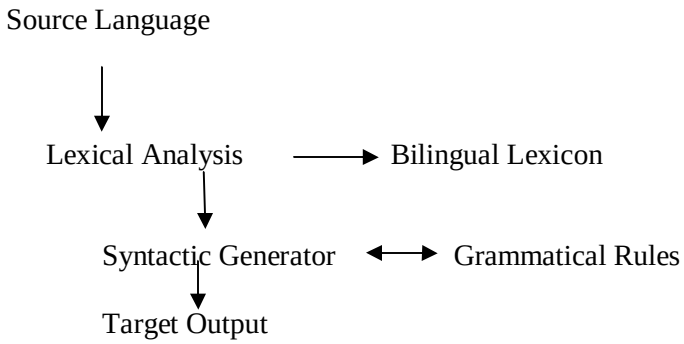


Figure 1: Translation System Phases

All languages are divided into rules. These rules govern the use of words and at a higher semantic level, they also govern their meanings. (Howard, 1982)

In the context of this work, twenty nine rules were formulated for the noun phrase translations which were specified using the Context Free Grammar (CFG). These rules were extracted from literatures. Table 1, summarizes the generated grammatical rules for noun-phrases in English language and the Yoruba arrangement of the rules to form meaningful noun-phrases.

From Table 1 R stand for rules while the index of R indicates the rule number. R_1 means rule1, R_2 means rule 2 to mention but few. Also for the acronyms in the table, N means noun, Adj means adjective, 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 which are the sub-divisions of the part of speech as affirmed by (Bamisaye, 2000).

Table 1: Grammatical Rules Generated for Noun Phrases.

S/No	ENGLISH RULES FOR NOUN PHRASES.	THE YORUBA ARRANGEMENT
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

From the grammatical rules of Table 1, the Yoruba arrangement of each rule shows further that, the English language and the Yoruba language differs syntactically and grammatically.

For example in rule 1, where noun-phrase is a combination of a definite article and a noun, to form a meaningful translation of this phrase in Yoruba language, the noun will have to come before the definite article otherwise, we will have a skewed or meaningless translation.

For example the noun phrase 'The Book' the grammatical structures for this phrase in both languages is are follows:

English Language

NP ⟶ Dart <N>
⟶ The <N>
⟶ The Book

Yoruba Language

NP ⟶ <N> Dart
⟶ Iwe Dart
⟶ Iwe naa

FINITE STATE AUTOMATA SYNTACTIC MODEL

Finite State Automata (FSA) whose operations were based on the first sets techniques was then used in modeling and recognizing the grammar of the language. The first sets techniques allow the parser to choose which production rule to apply based on the first input word of an input phrase. Figure 2 and Figure 3 show the FSAs for English and Yoruba noun-phrases. A finite state automaton is a diagrammatic representation of a regular grammar. An FSA begins from one of the states (called the start state), goes through transitions depending on inputs to different states and end in one certain set of states marking a successful flow of operation (called accept/final states). The start state is usually drawn with an arrow pointing at it from anywhere. Accept or final states are usually represented by a double circle and they are

those states at which the machine reports that the input strings (set of phrases) as processed so far is a member of the language it accepts (Alfred *et al*, 2006).

Grammar G = r_1| r_2| r_3| r_4|...|r_{29}, was considered. Where G is the grammar of the language,
r_1... r_{29} are the rules (sets of productions) involved in the generation of appropriate patterns in the language. The subscripts are the rule numbers for each rule as specified in the grammar of the language. For effective translation, Grammar G, was sub-divided into nine groups based on first sets techniques, this is to allow for easier parsing.
So we have: G = G_1 U G_2 U... U G_9 based on the first non-terminal and terminal token which starts each production rule.
We then have: G_1 = (r_1,r_3, r_5, r_7, r_8, r_9, r_{10}, r_{12}), G_2 = (r_2, r_4, r_6, r_{11}), G_3 = (r_{13}, r_{14}, r_{15}),
G_4 = (r_{16}, r_{17}, r_{18}, r_{19}, r_{20}, r_{21}), G_5 = (r_{22}, r_{23}, r_{24}), G_6 = (r_{27}, r_{28}), G_7 = (r_{25}), G_8 = (r_{26}),
G_9 = (r_{29}).
For the source language English language, FIRST of the subdivisions are stated thus:
FIRST (G_1) = FIRST (r_1, r_3, r_5, r_7, r_8, r_9, r_{10}, r_{12}) = {Dart};
FIRST (G_2) = FIRST(r_2, r_4, r_6, r_{11}) = {Inart};
FIRST (G_3) = FIRST (r_{13}, r_{14}, r_{15}) = {Dem};
FIRST(G_4) = FIRST (r_{16}, r_{17}, r_{18}, r_{19}, r_{20}, r_{21}) = { Poss};
FIRST(G_5) = FIRST (r_{22}, r_{23}, r_{24}) = {Quant};
FIRST (G_6) = FIRST(r_{27}, r_{28}) = {PreDet};
FIRST (G_7) = FIRST(r_{25}) = {CardNum};
FIRST (G_8) = FIRST(r_{26}) = {OrdNum}; and
FIRST (G_9) = FIRST(r_{29}) = {Nmod}
While for the target language, Yoruba language:
FIRST (G_1) = FIRST (r_1, r_3, r_5, r_7, r_8, r_9, r_{10}, r_{12}) = {ọ̀rọ̀-orúkọ};
FIRST (G_2) = FIRST(r_2, r_4, r_6, r_{11}) = {ọ̀rọ̀-orúkọ};
FIRST (G_3) = FIRST (r_{13}, r_{14}, r_{15}) = {ọ̀rọ̀-orúkọ};
FIRST(G_4) = FIRST (r_{16}, r_{17}, r_{18}, r_{19}, r_{20}, r_{21}) = {ọ̀rọ̀-orúkọ};
FIRST(G_5) = FIRST (r_{22}, r_{23}, r_{24}) = {ọ̀rọ̀-orúkọ /ọ̀rọ̀-àpèjúwe };
FIRST (G_6) = FIRST(r_{27}, r_{28}) = {ọ̀rọ̀-àpèjúwe};
FIRST (G_7) = FIRST(r_{25}) = {ọ̀rọ̀-àpèjúwe};
FIRST (G_8) = FIRST(r_{26}) = {ọ̀rọ̀-orúkọ}; and
FIRST (G_9) = FIRST(r_{29}) = {ọ̀rọ̀-orúkọ}
where; Dart means definite article, Inart means indefinite article, Dem means demonstrative, Poss means possessive, Quant means quantifier, PreDet means predeterminer, CardNum means cardinal numeral, OrdNum means ordinal numeral, Nmod means noun modifier, Adj means adjectives (ọ̀rọ̀-àpèjúwe), N means nouns (ọ̀rọ̀-orúkọ).
Figures 2 and 3 represent the finite state automata for English and Yoruba noun-phrases respectively. The state transition is a transition for grammar G with start states: G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8 ,G_9. The automata operate as follows: When the machine receives an input string, it goes through any of the states G_1 to G_9 depending on the first terminal or word of that input string.

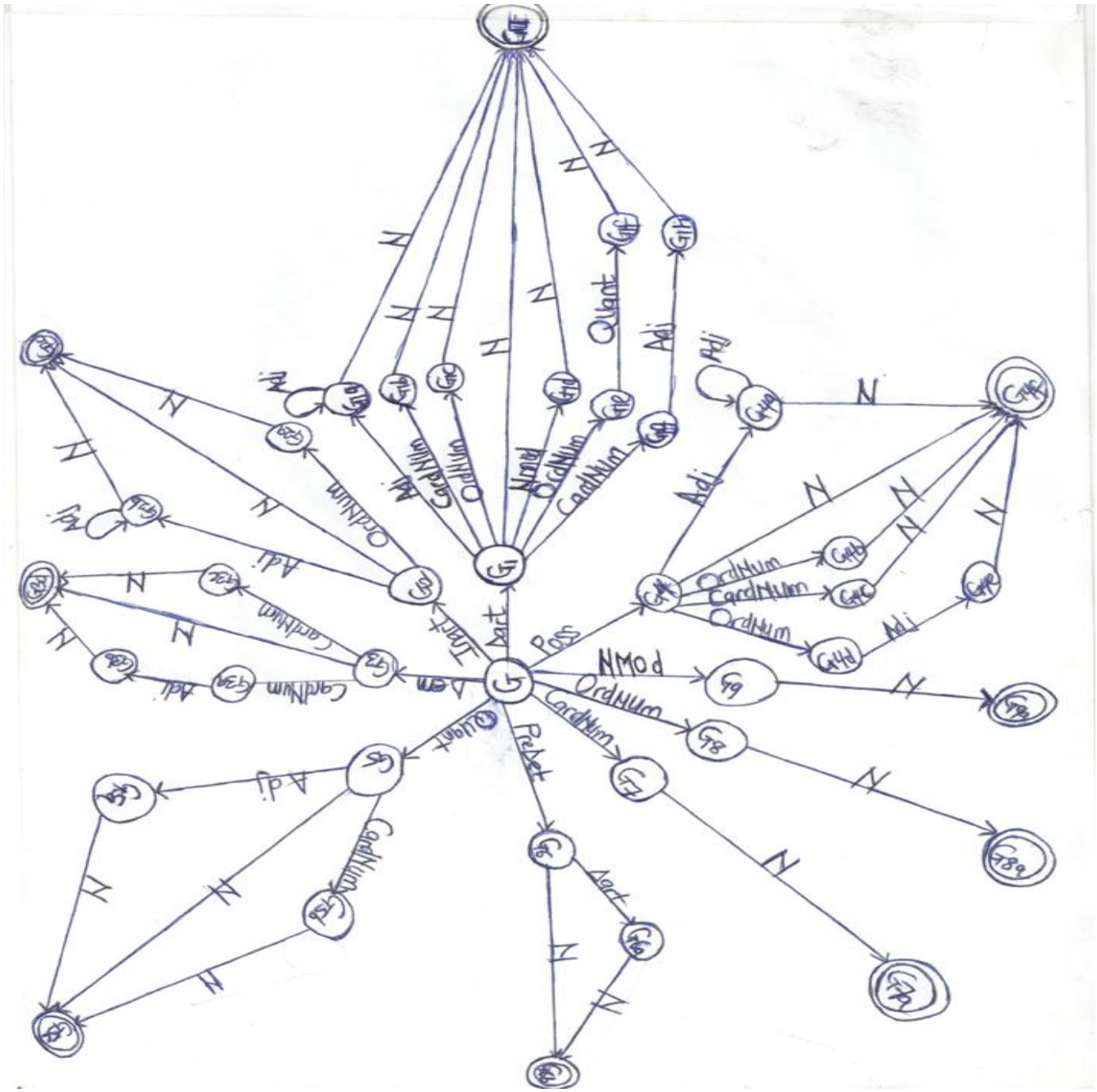


Figure 2: English Noun-Phrase Automaton

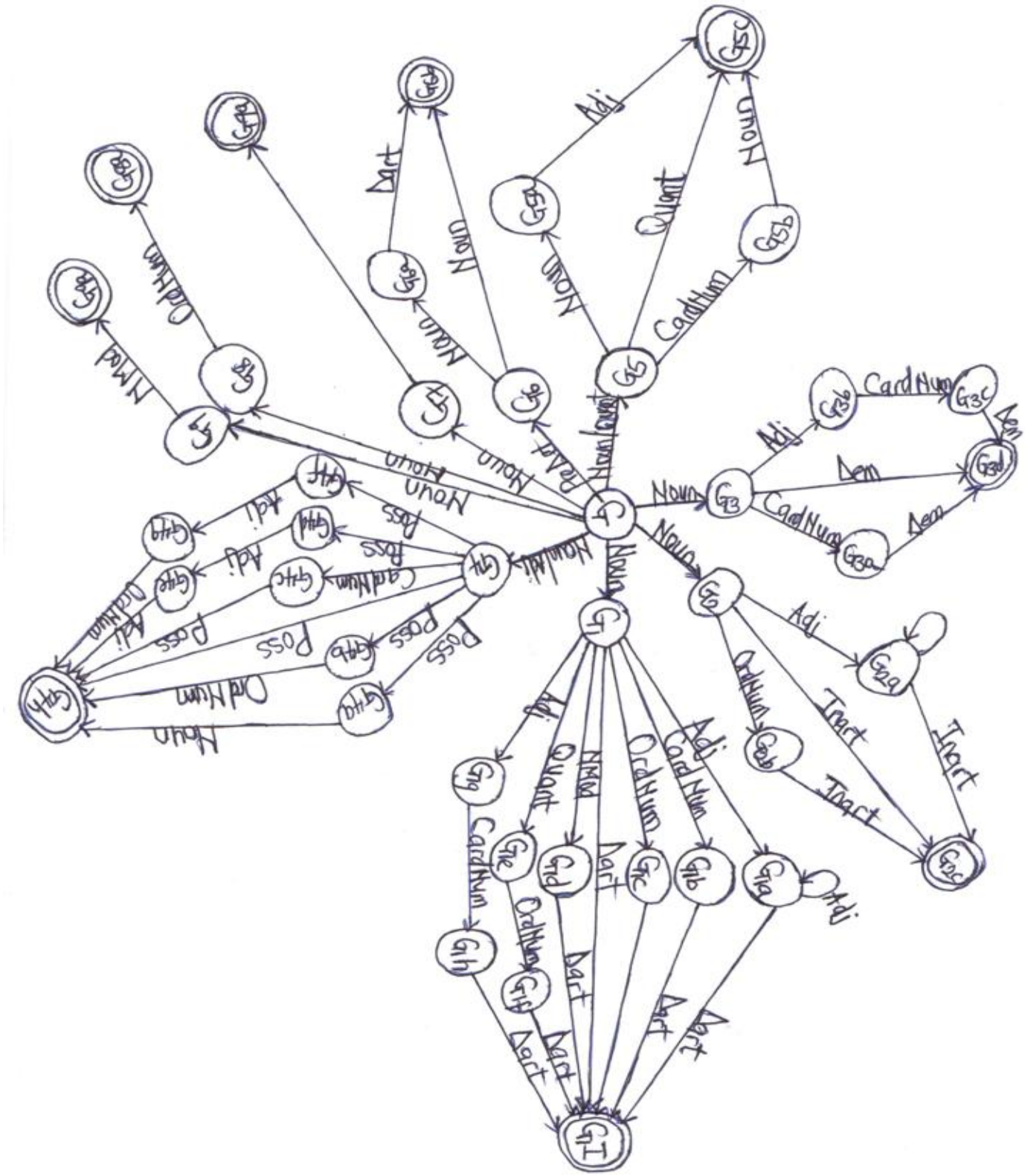


Figure 3: Yoruba Noun-Phrase Automaton

The finite automaton for state (G_1) is illustrated in figure 4 and the operation of the system at **Group One (G_1):** If the first terminal or word of an input string is a definite article, it goes through the state G_1. **G_1** has sets of states {G, G_1 $G_{1a,}$ $G_{1b,}$ $G_{1c,}$ $G_{1d,}$ G_{1e} G_{1f} G_{1g} G_{1h} G_{1i}} and operates on the inputs {Dart, Adj, CardNum, OrdNum, Nmod, Quant, N}. Its starting state is G and the sets of final states are G_{1I}. When the machine receives an input string (NP) with a first word or token 'definite article', it goes through any of the transitions in **G_1,** depending on the follow sets of tokens. For example; when it gets an input such as: {Dart, CardNum, Adj, N}, it goes through the state transition 'G, $G_{1,}$ $G_{1g,}$ $G_{1h,}$ G_{1I}'. If a Dart is found at state $G_{1,}$ the transition move from G_1 to state $G_{1g,}$ and if CardNum is found at that state, it move to state $G_{1h,}$ if at this state, an Adj is found, it move to the final or accept state G_{1I} if a noun is found at the state. Thus the input string will be accepted as a language of the grammar otherwise it will be rejected and lead to skewed this state is explained as follows:

translations if it does not follow any of the transitions. The following are the possible inputs state transitions for group 1 (G_1) in the source language and the corresponding outputs in the target language:

$G, G_1, G_{1a}, G_{1I} \longrightarrow G, G_{1I}, G_{1a}, G_1$
$G, G_1, G_{1a}, G_{1a}, G_{1I} \longrightarrow G, G_{1I}, G_{1a}, G_{1a}, G_1$
$G, G_1, G_{1b}, G_{1I} \longrightarrow G, G_{1I}, G_{1b}, G_1$
$G, G_1, G_{1c}, G_{1I} \longrightarrow G, G_{1I}, G_{1c}, G_1$
$G, G_1, G_{1I} \longrightarrow G, G_{1I}, G_1$
$G, G_1, G_{1d}, G_{1I} \longrightarrow G, G_{1I}, G_{1d}, G_1$
$G, G_1, G_{1e}, G_{1f}, G_{1I} \longrightarrow G, G_{1I}, G_{1f}, G_{1e}, G_1$
$G, G_1, G_{1g}, G_{1h}, G_{1I} \longrightarrow G, G_{1I}, G_{1h}, G_{1g}, G_1$

These possible inputs state transitions in the source language and the corresponding outputs in the target language further reveal the differences in the syntactic and grammatical structures of both languages.

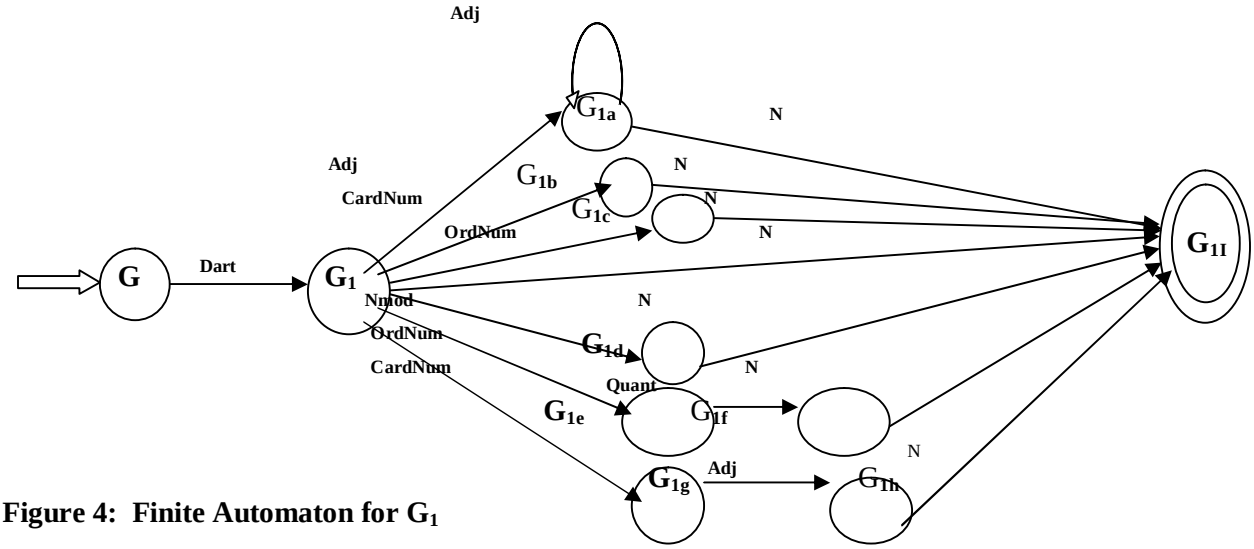
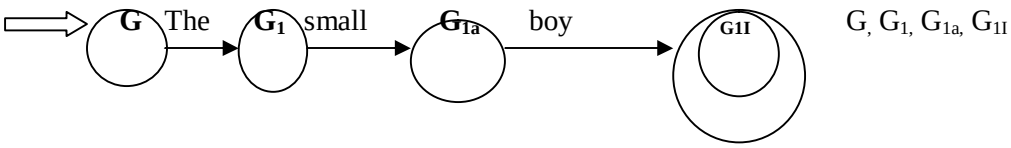
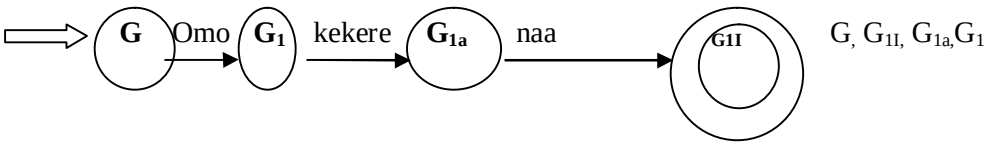


Figure 4: Finite Automaton for G_1

In the first input transition for $G_{1,}$ we have G, G_1, G_{1a}, G_{1I} as an input in the source language that is the English language and $G, G_{1I}, G_{1a,}$ G1 as the output in target language that is the Yoruba language. Below is an example of a noun-phrase illustrating the $G_{1,}$ state.





The decision of which path to choose or which transition to follow is by first sets techniques. If any input string does not follow any of the states transitions $G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8, G_9$ successfully then the string is not a language of the grammar. FSA reads an input string and depending on the input, output true (accept), output false (reject) or gets stuck in an infinite loop and output nothing. Transition function provide more compact Implementation for the work.

THE BILINGUAL LEXICON

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. The lexicon is the store house and it is made up of relations commonly called tables. The table has a name,'Ipele Atọkùn', three columns (fields or attributes) for Yorùbá and English words with their corresponding parts of speech and rows also known as 'records or tuples' that corresponds to individual words and its complete description.

EXPERIMENTAL SET-UP

The model was implemented using PhP programming language and MySQL. It was tested on one hundred and sixty randomly selected noun-phrases from daily news, and gives accuracy of 91% which is quite encouraging. The system if fully developed will go a long way in preventing the extinction threat of the Yoruba language by 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.

Table 2: Results of the system

DAILY NEWS	160
CORRECTLY TRANSLATED PHRASES	146
WRONGLY TRANSLATED PHRASES	14
ACCURACY	91%

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 dying aspects of the Yoruba language and its culture will 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.

REFERENCES

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

Alfred, V.A., Ravi, S., and Jeffrey, D. (2006). "Compiler Principles Techniques and Tools". Published by Addison Wesley Pearson Education Inc.

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

Bamisaye, O.T. (2000). "Essentials of English Syntax" Department of English, University of Ado-Ekiti, Nigeria. Published by Balfak Educational Publisher, Ado-Ekiti, Ekiti State

Chowdbury, G. (2005). "Natural Language Processing". Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK. Source at: www.infotoday.com/books/assist/artist 37.shtml. Retrieved 2010.

Howard, J. (1982). "Analyzing English an Introduction to Descriptive Linguistics" City of Birmingham Polytechnic, United Kingdom.

www.bcu.ac.uk/pme/schoolof english/staff/howard_jackson. Retrieved 2011.

Yusuf, O. (2006) "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.