DEVELOPMENT OF AN ADJECTIVAL PHRASE-BASED SYSTEM FOR ENGLISH TO YORÙBÁ MACHINE TRANSLATION

BY

OLALEYE, TIMILEYIN ENOCH

CPE/12/0891

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING, FACULTY OF ENGINEERING,
FEDERAL UNIVERSITY OYE- EKITI (FUOYE),
EKITI, NIGERIA.

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF ENGINEERING (B.ENG) IN COMPUTER ENGINEERING,

NOVEMBER, 2017

CERTIFICATION

This project with the title

DEVELOPMENT OF AN ADJECTIVAL PHRASE-BASED SYSTEM FOR ENGLISH TO YORÙBÁ MACHINE TRANSLATION

Submitted by

OLALEYE TIMILEYIN ENOCH

Has satisfied the regulations governing the award of degree of

BACHELOR OF ENGINEERING (B.Eng) In COMPUTER ENGINEERING

Federal University, Oye-Ekiti, Ekiti

13/12/2017 ..... 13/12/2017 .....

Engr. Mrs. A.O. Esan **Date**

Supervisor

DEPARTMENT OF COMPUTER ENGINEERING HOD'S SIGN DATE 20/3/18 FEDERAL UNIVERSITY OYE-EKITI

20-03-2018 .....

Dr. I.A. Adeyanju **Date**

Ag. Head of Department

DECLARATION

I Olaleye, Timileyin Enoch hereby declare that this project work carried out is the result of my personal effort under the supervision of Engr. Mrs. Esan of the department of Computer Engineering, Federal University Oye-Ekiti, Ekiti State, as part of the requirement for the award of Bachelor Degree of Computer Engineering, and has not been submitted elsewhere for this purpose. All sources of information are explicitly acknowledged by means of reference.

..... 12/12/2017

OLALEYE, TIMILEYIN E. DATE

DEDICATION

This project work is dedicated to God Almighty the Eternal King of Glory, who has been my reason for living and source of strength and inspiration, I give glory to His Holy name, and my parent who turn my virtual imagination to noble reality.

ACKNOWLEDGEMENTS

I acknowledge the help of God, to the one from whom all mercies flow, the beginning and the end. To the almighty be all glory, honour and adoration forever ever.

I appreciate the effort made by my supervisor; Engr.(Mrs.) A. O. Esan for the motherly love she shows me as well as the Head of Department; Dr. I. A., Adeyanju and all other lecturers of the Department of Computer Engineering; Prof. Omidiora, Dr. J.B. Oladosu to mention a few, for the knowledge they impacted in me. I pray that God will bless you all.

I am indeed, grateful to Dr. Eludiora S. of the department of Computer Science and Engineering OAU, Ife for the help he rendered during my research. God bless you Sir.

I am indeed, indebted to my wonderful parents, Pastor and late Mrs Olaleye for their moral and financial support, advice, words of encouragement, care and prayers. As the lord lives, my daddy you will surely live to eat the fruits of your labour. A special tribute to my priceless Mum, even though you never waited to see me graduate, yet your prayers, words of advice, love and care will forever be remembered. To my adorable siblings: Adesewa, Ibukunoluwa, Fiyinfoluwa and Patience, I say a big thanks to you for the love showered on me. God bless you all.

I also appreciate the love, prayer and support of my lovely uncle; Daddy Gab-Olaleye. I can never forget the family of Pastor and Mrs Olatoye, your love and care is highly commendable. I want to appreciate my friends: Ugwuija Benaiah, Fatokun Oluwatobi, and my room-mate; Jimoh Segun, thanks for your assistance and encouragement. Thanks to all members of CPE Class 17 for being wonderful people.

ABSTRACT

Yorùbá language is gradually going into extinction because most speakers don't know how to write it despite that it is being taught in Primary and Secondary schools in Nigeria. This therefore call for the need of modern day processing tools such as machine translators for the language to catch up with the technological growth the world is experiencing. In the face of rapid globalization, the significance of Machine translation cannot be overemphasized because it can translate the content quickly and provides quality output, thus saving human the stress and time of poring on translating books or looking for human translator. Hence, this research developed an Adjectival phrase-based (ADJP) system for English to Yorùbá Machine Translation.

The data for the developed Adjectival phrase-based (ADJP) system was extracted from locally spoken words and stored in a database. The phrases were broken down into their part of speech (POS) and the database was designed by categorizing all the parts of speech into their different grammatical functions. The corpus was trained to understand the grammatical rules of translation while NLTK parser was used to parse the corpus and test all the rules used as it affects each sentence. Python programming language is the core programming language used in developing the system.

The developed ADJP system was evaluated using human judgement by administering questionnaires to ten respondents. Expert's translated phrases were compared to that gotten from the developed system and the respondents' using the mean opinion score (MOS) technique based on word orthography. Results show that the expert's score was 100 percent while that of the respondents was 76.3 percent and the developed machine translator value was 95.5 percent.

The developed system's correctness is close to that of the Expert, and more accurate than that of the respondents giving accurate translations with appropriate tone-marks and under-dots.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

MT: Machine Translation

SMT: Statistical Machine Translation

RBMT: Rule-based Machine Translation

EBMT: Example-based Machine Translation

SL: Source Language

TL: Target Language

IL: Interlingua

NP: Noun Phrase

AdjP: Adjectival Phrase

Det: Determinant

PP: Prepositional Phrase

PRE: Preposition

APỌO: Àpólàọ̀rọ̀ Orúkọ̀

APTK: Àpólàọ̀rọ̀ Atọ́kùn

APỌI: Àpólàọ̀rọ̀ ìṣe

APỌA: ÀpólàÒròÀpónlé

ATK: ọ̀rọ̀ Atọ́kùn

ỌO: ÒròOrúkọ

AỌO: ArópòÒròOrúkọ

ỌA: ÒròÀpónlé

AlỌO: Asàpéjúweìlòọ̀rọ̀ orúkọ

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND STUDY

Language is the medium of communication. Human language is used in communicating ideas, emotions, feelings, desires, to co-operate among social groups and exhibit habits which can be translated along a variety of channels. Translation is the transfer of the meaning of a text from one language to another. It is a means of sharing information across languages and therefore essential for addressing information inequalities. The work of translation was originally carried out by human translators but its limitations such as high cost, lower speed of translation and insecurity of confidential information led to the development of machine translators (Oladosu, et al., 2016).

Machine Translation (MT) can be defined as a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another (Arnold, Balkan, Meijer, Humphreys, & Sadlery, 1994). The history of machine translation can be traced back to the pioneers and early systems of the 1950s and 1960s, the impact of the ALPAC report in the mid-1960s, the revival in the 1970s, the appearance of commercial and operational systems in the 1980s, research during the 1980s, new developments in research in the 1990s, and the growing use of systems in the past decade (Hutchins, 1994).

Machine Translation (MT) deals mainly with the transformation from one natural language to another language. Natural Language Interface provides the user freedom to interact with the computer in a natural language like English, Yorùbá, Twi, and Hausa or any other language used for day to day communication. (Sangeetha, Jothilakshmi, &

Kumar, 2014). It is an important part of Natural Language Processing in artificial Intelligence which accepts characters of source language and map to the characters of the target language to generate the words with the help of various rules and other learning process techniques (Pankaj & Er.Vinod, 2013). This research designed an adjectival phrase-based system for English to Yorùbá machine translation. The adjectival phrase was chosen because it provides important information about location, description of people and things, positions, relationships, time and ideas.

1.2 PROBLEM STATEMENT

Yorùbá language is endangered because Yorùbá culture is gradually going into extinction. The total dominance of English language over Yorùbá language in almost all human endeavour is also a major challenge. This call for the need of modern day processing tools for the language to catch up with the technological growth the world is experiencing thereby increasing the audience and peoples' interest in the language.

Also, there are limitations associated with human translators which include: high cost, lower speed and insecurity of confidential information (Oladosu, *et al.*, 2016). Hence, a phrase-based English to Yorùbá machine translator would be developed in this research in order to overcome the shortcomings of the human translator thereby increasing peoples' interest in the language.

1.3 AIM AND OBJECTIVES

The aim of this project is to develop an adjectival phrase-based English to Yorùbá machine translation system.

The objectives of the project are;

1. to design an adjectival phrase-based machine translator for English to Yorùbá

2. to implement the system based on the grammar of the two languages using python programming language with PyQt5 (GUI module)

3. to evaluate the developed system using human judgment (mean opinion score).

1.4 SCOPE OF STUDY

This project focused on translation of English adjectival phrases to Yorùbá and the data for the work was obtained from locally spoken words. The performance of the system was evaluated using mean point score based on word orthography.

1.5 RESEARCH METHODOLOGY

The Project methodology include:

- Database creation: the data for this work was extracted from locally spoken words and stored in a database. The phrases were broken down into their part of speech (POS) and the database was designed by categorizing all the parts of speech into their different grammatical functions.
- Training the corpus: the corpus was trained to understand the grammatical rules of translation.
- Designing a parser: NLTK parser was used to parse the corpus and test all the rules used as it affects each sentence.
- Python programming: This is the core programming language used in developing the system.
- Mean point score (human judgement) was used in evaluating the system.

1.6 SIGNIFICANCE OF THE STUDY

In the face of rapid globalization, the significance of Machine translation cannot be overemphasized. This is because machine can translate from English to Yorùbá quickly and provides quality outputs, thus saving human the stress and time of poring on translating books or looking for human translator. Also, the system is comparatively cheap and it guarantees confidentiality of information. The system can be accessed anywhere.

Furthermore, MT finds its application in information retrieval and extraction. Information retrieved can be in form of text, images, spoken documents and broadcast stories while many commercial and government-funded international and national organizations scrutinize foreign-language documents for information relevant to their activities from commercial and economic to surveillance, intelligence and espionage.

CHAPTER TWO

LITERATURE REVIEW

2.1 HISTORY OF MACHINE TRANSLATION

The theory of MT pre-dates computers, with philosophers 'Leibniz and Descartes' ideas of using code to relate words between languages in the seventeenth century (Hutchins, 1993). The early 1930s saw the first patents for 'translating machines'. Georges Artsrouni was issued a patent in France in July 1933. He developed a device, which he called a *'cerveaumécanique'*(mechanical brain) that could translate between languages using four components: memory, a keyboard for input, a search method, and an output mechanism. The search method was basically a dictionary look-up in the memory; therefore, Hutchins is reluctant to call it a translation system. The proposal Russian Petr Petrovich Troyanskii patented in September 1933 bears a resemblance to the Apertium system, using a bilingual dictionary and a three-staged process, *i.e.* first a native speaking human editor of the SL (SL) pre-processed the text, then the machine performed the translation, and finally a native-speaking human editor of the TL post-edited the text ((Hutchins J, 1993); (Hutchins & Lovtskii, 2000)).

After the birth of computers Electrical Numerical Integrator and Calculator (ENIAC) in 1947, research began on using computers as aids for translating natural languages (Hutchins J , 2005). In 1949, Weaver wrote a memorandum, putting forward various proposals (based on the wartime successes in code breaking) on the developments in information theory and speculation about universal principles underlying natural languages. In the decade of optimism, from 1954-1966, researchers encountered many predictions of imminent 'breakthroughs'. In 1966, the Automated Language Processing Advisory Committee (ALPAC) report was submitted, which said that, for 'semantic

barriers', there are no straightforward solutions. The ALPAC report committee could not find any "pressing need for MT" nor "an unfulfilled need for translation (ALPAC, 1966). This report brought MT research to its knees, suspending virtually all research in the United States of America (USA) while some research continued in Canada, France, and Germany (Hutchins J. , 2005). At this time the focuses of MT began to shift somewhat from pure research to practical application using a hybrid approach. Moving towards the change of the millennium, MT became more readily available to individuals via online services and software for their personal computers (Anthony, 2013).

2.2 APPROACHES TO MACHINE TRANSLATION

A machine translation (MT) system first analyses the source language input and creates an internal representation. This representation is manipulated and transferred to a form suitable for the target language. Then at last output is generated in the target language. MT systems can be classified according to their core methodology. Under this classification, two main paradigms can be found: the rule-based approach and the corpus-based approach. In the rule-based approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts is required. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two major classifications of MT systems gave birth to the Hybrid Machine Translation Approach. (Okpor, 2014)

MT is classified into seven broad categories: rule-based, statistical-based, hybrid-based, example-based, knowledge-based, principle-based, and online interactive based methods. The first three MT approaches are the most widely used and earliest methods. At present, most of the MT related research is based on statistical and example-based approaches

2.2.1 Rule-Based Approach

In the field of MT, the rule-based approach is the first strategy that was developed. A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules. Nevertheless, building RBMT systems entails a huge human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers, bilingual dictionaries, source to target transliteration, TL morphological generator, structural transfer, and reordering rules. Nevertheless, a RBMT system always is extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation, and contextual processing of language. Generally, rules are written with linguistic knowledge gathered from linguists. Transfer-based MT, Interlingua MT, and dictionary-based MT are the three different approaches that come under the RBMT category. There are problems associated with the RBMT approach which include: Insufficient amount of really good dictionaries i.e. Building new dictionaries is expensive, some linguistic information still needs to be set manually and it is hard to deal with rule interactions in big systems, ambiguity, and idiomatic expressions (Okpor, 2014).

2.2.1.1 Direct Translation

In the direct translation method, the SL text is analyzed structurally up to the morphological level and is designed for a specific source and target language pair (Noone, 2003) (Dasgupta & Basu, 2008). The performance of a direct MT system depends on the quality and quantity of the source-target language dictionaries, morphological analysis, text processing software, and word-by-word translation with minor grammatical adjustments on word order and morphology. Challenges of a DMT System include are

that it can be characterized as word-for-word translation with some local word-order adjustment. Also, the linguistic and computational naivety of the approach is an issue. From a linguistic point of view, what is missing is any analysis of the internal structure of the source text, particularly the grammatical relationships between the principal parts of the sentences.

2.2.1.2 Interlingua Based Translation

The next stage of progress in the development of MT systems is the Interlingua approach, where translation is performed by first representing the SL text into an intermediary (semantic) form called Interlingua. The advantage of this approach is that Interlingua is a language independent representation from which translations can be generated to different TLs. Thus, the translation consists of two stages, where the SL is first converted in to the Interlingua (IL) form before translation from the IL to the TL. The main advantage of this Interlingua approach is that the analyzer of the parser for the SL is independent of the generator for the TL. There are two main drawbacks in the Interlingua approach. The first disadvantage is difficulty in defining the Interlingua. The second disadvantage is Interlingua does not take the advantage of similarities between languages, such as translation between Dravidian languages. Nevertheless the advantage of Interlingua is it is economical in situations where translation among multiple languages is involved. (Anthony, 2013)

There are the difficulties in defining an Interlingua, even for closely related languages (e.g. the Romance languages: French. Italian, Spanish, Portuguese). A truly universal and language-independent Interlingua has defied the best efforts of linguists over the years. Also, it is difficult to extract meaning from texts in the original languages to create the

intermediate representation and semantic differentiation is target-language specific and making such distinctions is comparable to lexical transfer not all distinctions needed for translation

2.2.1.3 Transfer Based Translation

Because of the disadvantage of the Interlingua approach, a better rule-based translation approach was discovered, called the transfer approach. Recently, many research groups have been using this third approach for their MT system, both abroad and in India. On the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: i) Analysis, ii) Transfer and iii) Generation. In the first stage, the SL parser is used to produce the syntactic representation of a SL sentence. In the next stage, the result of the first stage is converted into equivalent TL-oriented representations. In the final step of this translation approach, a TL morphological analyzer is used to generate the final TL texts (Anthony, 2013).

One of the problems with transfer Based Machine translation approach is that rules must be applied at every step of translation. There are rules for source language analysis (syntactic/semantic), rules for source-to-target transfer and rules for target language generation.

2.2.2 Statistical-Based Approach

The statistical approach comes under Empirical Machine Translation (EMT) systems, which rely on large parallel aligned corpora. Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge and statistical models extracted from bilingual corpora. In statistical-based MT, bilingual or multilingual textual corpora of the source and target

language or languages are required. A supervised or unsupervised statistical machine learning algorithm is used to build statistical tables from the corpora, and this process is called the learning or training (Zhang, 2006). The statistical tables consist of statistical information, such as the characteristics of well-formed sentences, and the correlation between the languages. During translation, the collected statistical information is used to find the best translation for the input sentences, and this translation step is called the decoding process. The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by $p(e\backslash f)$, which is the Probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL. The problem of modeling the probability distribution $p(e\backslash f)$has been approached in a number of ways. One intuitive approach is to apply Bayes theorem. That is, if $p(f\backslash e)$and $p(e)$indicate translation model and language model, respectively, then the probability distribution$p(e\backslash f)\infty$ if $p(f\backslash e)p(e)$. The translation model if $p(f\backslash e)$ is the probability that the source sentence is the translation of the target sentence or the way sentences in E get converted to sentences in F. The language model if $p(e)$ is the probability of seeing that TL string or the kind of sentences that are likely in the language E. This decomposition is attractive as it splits the problem into two sub problems. Finding the best translation is done by picking the one that gives the highest probability, as shown in Equation 2.1.

$$\tilde{e} = \frac{\arg\max p(e\backslash f)}{e \varepsilon e^*} = \frac{\arg\max p(f\backslash e)p(e)}{e \varepsilon e^*} \qquad (2.1)$$

There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model.

2.2.2.1 Word Based Translation

As the name suggests, the words in an input sentence are translated word by word individually, and these words finally are arranged in a specific way to get the target sentence. The alignment between the words in the input and output sentences normally follows certain patterns in word based translation. This approach is the very first attempt in the statistical-based MT system that is comparatively simple and efficient. The main disadvantage of this system is the oversimplified word by word translation of sentences, which may reduce the performance of the translation system (Anthony, 2013).

2.2.2.2 Phrase Based Translation

According to Anthony (2013) a more accurate SMT approach, called phrase-based translation, was introduced, where each source and target sentence is divided into separate phrases instead of words before translation. The alignment between the phrases in the input and output sentences normally follows certain patterns, which is very similar to word based translation. Even though the phrase based models result in better performance than the word based translation, they did not improve the model of sentence order patterns. The alignment model is based on flat reordering patterns, and experiments show that this reordering technique may perform well with local phrase orders but not as well with long sentences and complex orders (Anthony, 2013).

2.2.2.3 Hierarchical Phrase Based Model

By considering the drawback of word-based translation and phrase-based translation, (Chiang, 2005) developed a more sophisticated SMT approach, called the hierarchical phrase based model. The advantage of this approach is that hierarchical phrases have

recursive structures instead of simple phrases. This higher level of abstraction approach further improved the accuracy of the SMT system.

2.2.3 Hybrid-Based Translation

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach, which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use this hybrid-based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better than the previous and has more power, flexibility, and control in translation.

Hybrid approaches integrating more than one MT paradigm are receiving increasing attention. The METIS-II MT system is an example of hybridization around the EBMT framework; it avoids the usual need for parallel corpora by using a bilingual dictionary (similar to that found in most RBMT systems) and a monolingual corpus in the TL (Dirix, Schuurman, & Vandeghinste, 2005). An example of hybridization around the rule-based paradigm is given by Oepen, 2007). It integrates statistical methods within an RBMT system to choose the best translation from a setoff competing hypotheses (translations) generated using rule-based methods (Oepen, et al., 2007)

In SMT, Koehn and Hoang integrate additional annotations at the word-level into the translation models in order to better learn some aspects of the translation that are best explained on a morphological, syntactic, or semantic level (Koehn & Hoang, 2007). Hybridization around the statistical approach to MT is provided by Groves and Way; they combine both corpus-based methods into a single MT system by incorporating phrases from both EBMT and SMT into an SMT system (Groves & A, 2005). A different hybridization happens when an RBMT system and an SMT system are used in a cascade; Simard proposed an approach, analogous to that by Dugast*et al.*,(2007) using an SMT system as an automatic post-editor of the translations produced by an RBMT system (Simard *et al.*, 2007); (Dugast, Senellart, & Koehn, 2007)

2.2.4 Example-Based Translation

The example-based translation approach is based on analogical reasoning between two translation examples, proposed by Makoto Nagao in 1984. At run time, an example-based translation is characterized by its use of a bilingual corpus as its main knowledge base. The example-based approach comes under the EMT system, which relies on large parallel aligned corpora. Example-based translation is essentially translation by analogy. An EBMT system is given a set of sentences in the SL (from which one is translating) and their corresponding translations in the TL, and uses those examples to translate other, similar source-language sentences into the TL. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. EBMT systems are attractive in that they require a minimum of prior knowledge; therefore, they are quickly adaptable to many language pairs. A restricted form of example-based translation is available commercially, known as translation memory. In a translation memory, as the user translates text, the translations are added to a database, and when the

same sentence occurs again, the previous translation is inserted into the translated document. This saves the user the effort of re-translating that sentence, and is particularly effective when translating a new revision of previously-translated document (Anthony, 2013).

2.2.5 Knowledge-Based MT

Knowledge-Based Machine Translation (KBMT) is characterized by a heavy emphasis on functionally complete understanding of the source text prior to the translation into the target text. KBMT does not require total understanding, but assumes that an interpretation engine can achieve successful translation into several languages. KBMT is implemented on the Interlingua architecture; it differs from other Interlingua techniques by the depth with which it analyzes the SL and its reliance on explicit knowledge of the world. KBMT must be supported by world knowledge and by linguistic semantic knowledge about meanings of words and their combinations. Thus, a specific language is needed to represent the meaning of languages. Once the SL is analyzed, it will run through the augmenter. It is the knowledgebase that converts the source representation into an appropriate target representation before synthesizing into the target sentence. KBMT systems provide high quality translations. Nevertheless, they are quite expensive to produce due to the large amount of knowledge needed to accurately represent sentences indifferent languages.

2.2.6 Principle-Based MT

Principle-Based Machine Translation (PBMT) Systems employ parsing methods based on the Principles & Parameters Theory of Chomsky's Generative Grammar. The parser generates a detailed syntactic structure that contains lexical, phrasal, grammatical, and

thematic information. It also focuses on robustness, language-neutral representations, and deep linguistic analyses. In the PBMT, the grammar is thought of as a set of language-independent, interactive well-formed principles and a set of language-dependent parameters. Thus, for a system that uses n languages, one must have n parameter modules and a principles module. Thus, it is well-suited for use with the Interlingua architecture.

PBMT parsing methods differ from the rule-based approaches. Although efficient in many circumstances, they have the drawback of language-dependence and increase exponentially in rules if one is using a multilingual translation system. Another drawback of current PBMT systems is the lack of the most efficient method for applying the different principles. UNITRAN is one of the examples of PBMT.

2.2.7 Syntax Based Model

Syntax is the hierarchical structure of a natural language sentence. Depending on the type of input, syntax-based models can be divided into two broad categories: the string-based systems and tree-based systems.

2.2.7.1 String-Based Systems

String-based systems are MT systems whose input is a string to be simultaneously parsed and translated by a synchronous grammar (Galley et al., 2006). In a synchronous CFG the elementary structures are rewrite rules with aligned pairs of right-hand sides as in equation (2.2)

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle \tag{2.2}$$

Where X is a non-terminal, γ and α are both strings of terminals and non-terminals, and, $\sim$ is a one-to-one correspondence between non-terminal occurrences on the source and target side.

2.2.7.2 Tree-Based Systems

Tree-based systems perform translation in two separate steps: parsing and decoding. A parser first parses the source language input into a 1-best tree T, and the decoder then searches for the best derivation(a sequence of translation steps) d*that converts source tree T into a target-language string among all possible derivations D as shown in equation (2.3):

$$d^* = argmax_{d \in D}\, P(d/T) \tag{2.3}$$

Tree-based systems offer some attractive features. Among these features are: much faster decoding (linear time vs. cubic time), do not require a binary-branching grammar as in string-based models and can have separate grammars for parsing and translation (Huang, Knight, & Joshi, 2006).The tree-based systems can be sub-divided into three, which are: tree-to-string, string-to-tree and tree-to-tree models respectively. Joshua (Li et al., 2009) is an open source toolkit for parsing in syntax-based machine translation.

In string-to-tree model, the input is a string and the output is a parse tree (Galley, et al., 2006) while tree-to-string model parses a tree at the input and outputs a string. The Tree-to-tree model extracts rules using parse trees from *both* side(s) of the bitext. By modeling the syntax of both source and target languages, tree-to-tree model have the potential benefit of providing rules linguistically better motivated.

However, despite the advantages of tree-based systems, they suffer from a major drawback: they only use the 1-best parse tree to direct the translation, which potentially introduces translation mistakes due to parsing errors (Oladosu, et al., 2016).

2.2.8 Forest-Based Translation

Forest-based translation is a compromise between the string-based and tree-based methods because it combines the advantages of both methods. Forest based translation encourages faster decoding and alleviates parse errors. Informally, a packed parse forest, or *forest* in short, is a compact representation of all the derivations (i.e., parse trees) for a given sentence under a context-free grammar (Billot & Lang, 1989). Forest based machine translation mainly extends the tree-string model to forest to string. Forest-to-string translation is an extension of the tree-to-string model because it uses a packed parse forest as the input and outputs a string (Mi, Liang, & Liu, 2008). Forest-to-string models can be described as equation (2.4):

$$e^* = argmax_{d \in D(T), T \in F(f)} P(d/T) \tag{2.4}$$

where f stands for a source string, e stands for a target string, F stands for a forest, D stands for a set of synchronous derivations on a given tree T, and e* stands for the target side yield of a derivation. In order to deal with word order differences in machine translation and to translate differently all the meanings of an ambiguous input in a forest, forest reordering model was proposed by Cmejrek, (2014). He presented a novel extension of a forest-to-string machine translation system with a reordering model as stated in equation (vii):

$$f_{order} = \sum_{0 \le i < j \le n} -\log P_{order}(O_{ij} = O_{ij}^{h} | h) \tag{2.5}$$

Research shows that the method provides improvement from0.6 up to 1.0 point measured by (Ter– Bleu)/2 metric.

2.2.9 Online Interactive Systems

In this interactive translation system, the user is allowed to suggest the correct translation to the translator online. This approach is very useful in a situation where the context of a word is unclear and there exists many possible meanings for a particular word. In such cases, the structural ambiguity can be solved with the interpretation of the user.

2.3 Yorùbá Language and Culture

After a thorough research, it has been discovered that there is insufficient parallel English – Yorùbá corpus, hence English – Yorùbá statistical machine translator is not common (probably Yorùbá Google translator). There are basically three indigenous languages in Nigeria, they are the Hausa language spoken in the northern part of Nigeria, the Igbo is spoken by the Eastern part of the country and the Yorùbá which is spoken in the south-western part of Nigeria (Ninan & Odetunji, 2013). The English language is the official language use in communication in Nigeria and it becomes the language (Eludiora, Agbeyangi, & Ojediran, 2015) of debate and record in spite of the use of major indigenous Nigerian languages. The Yorùbá language (target language) is a tonal language spoken by people of the south- western part of Nigeria, which covers states like Ọyọ́, Ọsun, Ògùn, Òndo, Èkìtì, Lagos, Kogi and Kwara. (Eludiora & Elufidodo, 2016)

2.4 STRUCTURE OF ENGLISH AND YORÙBÁ LANGUAGE

According to (Odejobi, Owolabi, & Adegbola, 2011), 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. ìşé 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.

2.4.1 Phrase Grammar And Re-Write Rules

The English and Yorùbá write rules are illustrated below. The list of acronyms is in table 2.1. These the acronyms used to replace the English acronyms in the Yorùbá section. 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) *okunrinnaa*, 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.

Table 2.1:English Part and relative Yorùbá Translation

ENGLISH	**Yorùbá**
NP	Àpólàọ̀rọ̀ Orúkọ̀ (APỌO)
PP	Àpólàọ̀rọ̀ Atọ́kùn (APTK)
VP	Àpólàọ̀rọ̀ ìṣe (APỌI)
ADJP	ÀpólàÒròÀpónlé (APỌA)
PRE	ọ̀rọ̀ Atọ́kùn (ATK)
N	ÒròOrúkọ (ỌO)
PRN	ArópòÒròOrúkọ (AỌO)
ADJ	ÒròÀpónlé (ỌA)
DET	Asàpéjúwe Ìlòọ̀rọ̀ 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 ADJP 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 Adjectival phrase structure (3)

Rule 1 NP => ADJPNP

Rule 2 ADJP => ADJNP

Rule 3 NP => ADJPNP

Rule 4 ADJP => ADJNP

Rule 5 NP => DETNP

Rule 6 NP => N

Yorùbá Adjectival phrase structure (4)

Rule 1 ADJP =>ADJPN

Rule 2 ADJP=> DETADJ

Rule 3 NP => NPADJP

Rule 4 NP =>PPNP

Rule 5 NP => NPDET

Rule 6 NP => N

For example: the old man. This phrase can be tokenized as follows:

English on the chair

AP: =ADJPN

ADJP: =Det Adj

DET: = an

ADJ: = old

N: = man

Yorùbá/ÀgbàlágbàỌkùnrinKan/

APỌA: =ATK APOO

APOO: = AIỌO ỌA

ỌA: = Àgbàlágbà

AIỌO: = Ọkùnrin

ỌO: = Kan

(Eludiora & Atolagbe, 2016)

2.5 METHODS OF EVALUATING MACHINE TRANSLATION SYSTEM

Traditionally, human judgment is used in evaluating MT systems based on two main criteria: adequacy and fluency. Human judgment of the MT output is expensive and subjective therefore, automatic evaluation measures are a necessity. There are various methods used in evaluating MT systems. Among them are: BLEU (bilingual evaluation

understudy), WER (word error rate), PER (position-independent word error rate) and NIST.

2.5.1 Word Error Rate (WER)

One of the first automatic metrics used to evaluate MT systems was Word Error Rate (WER), which is the standard evaluation metric for Automatic Speech Recognition. WER is computed as the Levenshtein distance (Levenshtein, 1966) between the words of the system output and the words of the reference translation divided by the length of the reference translation. The Levenshtein distance is computed using dynamic programming to find the optimal alignment between the MT output and the reference translation, with each word in the MT output aligning to either 1 or 0 words in the reference translation, and vice versa. Those cases where a reference word is aligned to nothing are labeled as deletions, whereas the alignment of a word from the MT output to nothing is an insertion. If a reference word matches the MT output word it is aligned to, this is marked as a match, and otherwise is a substitution. The WER is then the sums of the number of substitutions (S), insertions (I), and deletions (D) divided by the number of words in the reference translation (N) as shown in Equation (6).

$$WER = \frac{S+I+D}{N} \tag{2.6}$$

2.5.2 MWER (Multi-Reference WER)

The application of WER to more than one reference translation refers to the minimum of the WER scores between the MT output and each reference. In essence, MWER is the WER between the MT output and the closest reference translation. While this allows

WER to be used with multiple references, the references are not combined in any fashion and are not truly exploited by the metric (Nießen, Och, Leusc, & Ney, 2000).

2.5.3 Position-Independent Error Rate (PER)

Position-independent Error Rate or (PER) came into inception to address the word-ordering limitation of WER by treating the reference and hypothesis as bags of words, so that words from the hypothesis can be aligned to words in the reference regardless of position. Because of this the PER OF an MT output is guaranteed to be lower than or equal to the WER of the MT output. This variant has the disadvantage of being unable to distinguish a correct translation from one where the words have been scrambled (Tillmann, Vogel, Ney, Zubiag, & Sawaf, 1997).

2.5.4 BLEU (Bilingual Evaluation Understudy)

BLEU (Bilingual Evaluation Understudy) is the current standard for automatic machine translation evaluation. Like MWER, a key characteristic of BLEU is its direct exploitation of multiple references. The BLEU score of a system output is calculated by counting the number of n-grams, or word sequences, in the system output that occur in the set of reference translations. BLEU is a precision-oriented metric in that it measures how much of the system output is correct, rather than measuring whether the references are fully reproduced in the system output. BLEU could be gamed by producing very short system outputs consisting only of highly confident n-grams, if it were not for the use of a brevity penalty which penalizes the BLEU score if the system output is shorter than the references. (Papineni, Roukos, Ward, & Zhu, 2002)

2.5.5 NIST

NIST is a method for evaluating the quality of text which has been translated using machine translation. It is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST calculates how informative a particular n-gram is (Oladosu, et al., 2016).

2.6 RELATED WORKS

Eludiora, Abayomi, & Fatusin, (2015) Worked on English to Yorùbá Machine Translation System for Yorùbá Verbs' Tone Changing and deduced that in translating English sentences (text) to Yorùbá sentences (text), some Yorùbá verbs change tone from the bilingual dictionary low-tone to mid-tone when they are translated to Yorùbá. They are called tone change verbs. These tone change verbs do pose some challenges in English to Yorùbá machine translation, and Most of the time it changes the meaning of the sentence. These changes usually depend on the positions of the nouns and pronouns in the sentence. The verbs in this category were collected from different Yorùbá sentences that contain tone change verbs. They developed the system using some re-write rules were designed for the two languages. The re-write rules were tested using JFLAP. Apart from re-write rules, there are other grammatical rules considered and the rules affected the Yorùbá translations. The software was designed using unified modelling language (UML). The Rule-based approach was used for the translation. Python programming language was used for the software development. The python has natural language tool kits that are used for the sentence parsing. The system accept English sentence then discover the pattern for the sentence. The system was implemented and tested for twenty tone change verbs within the home domain. The two languages are subject verb object

(SVO) sentence structure with some differences. The results show that the MT system can translate these tone change verbs. The system is efficient in its response time

Agbeyangi, Eludiora, & Adenekan, (2015) developed a system named "English to Yorùbá Machine Translation System using Rule-Based Approach. They deuced that rule-based approach is a good approach for Machine Translation System used for language with lots of grammar which Yorùbá language is one. In their research they laid emphasis on popularity of Yorúbà language among the three main languages in Nigeria calls for the need to computerize the language. They used Transfer Rule-Based Machine Translation 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 (Agbeyangi, Eludiora, & Adenekan, 2015).

Haque, Dandapat, Srivastava, Naskar and Way, (2009) developed English to Hindi Transliteration system based on the phrase-based statistical method (PB-SMT). A PB-SMT model has been used for transliteration by translating characters rather than words as in character-level translation systems. They modeled Translation in PB-SMT as a decision process, in which the translation a source sentence is chosen to maximize. They used source context modeling into the state-of-the-art log-linear PB-SMT for the English—Hindi transliteration task. To improve the system performance, they took

source context into account substantially (Haque, Dandapat, Srivastava, Naskar, & Way, 2009)

Islam, Tiedmann, & Eisle, (2009), Proposed 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 Eludiora, (2014). He proposed a machine translator that 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 (Eludiora S. , 2014).

Eludiora, Agbeyangi, & Fatunsin, (2015) experiment on the concept of Yorùbá verbs' tone changing. For instance, "Adé wo ilé" means "Ade entered the house". In this case, the dictionary meaning of enter in Yorùbá is wole. 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.

Adenekan, Agbeyangi, & Eludiora, (2015) proposed 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 the author, there is limited corpus that is available for Yorùbá language this informs the rule-based approach.

Odejobi, Eludiora, Akanbi, Iyanda, & Akinade, (2015) proposed 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 an on-going research work.

Akinwale, Adetunmbi, Obe, & Adesuyi, (2015) proposed a 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.

(Abiola, Adetunmbi, & Oguntimilehin, 2015) Considered a hybrid approach to English to Yorùbá machine translation. The paper only itemized the steps the authors will take in the development of the proposed system. The study is on-going.

Abiola, Adetunmbi, Fasiku, & Olatunji, (2014) proposed 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.

CHAPTER THREE

METHODOLOGY

3.1 THE APPROACH

In this project, an adjectival phrase-based machine translation system was designed for English to Yorùbá. The phrases were broken down into their part of speech (POS) and the database was designed by categorizing all the parts of speech into their different grammatical functions. The data for the work was extracted from locally spoken words and stored in a database and the corpus was trained to understand the grammatical rules of translation. NLTK parser was used to parse the corpus and test all the rules used as it affects each sentence. Python programming was used in developing the system and Mean point score (human judgement) was used in evaluating the system.

3.2 REQUIREMENT ANALYSIS

The requirements and specifications of the Adjectival Phrase English to Yorùbá Machine Translation system software are as follow:

i. to present a user friendly interface to the user;

ii. to give the user access to enter adjectival phrases in English language provided the phrases is within the domain covered;

iii. formulate a grammar for the English phrase using phrase structure rewrite rule;

iv. translate and output the equivalent meaning of the sentences entered in standard Yorùbá language; and

v. implement a system for the translation based on the grammar of the two languages using python programming language with PyQt5(GUI module)

3.3 DEVELOPMENT TOOLS

The main tools used for the project are:

- JFLAP: this was used to test the re-write rule and grammar using parse trees.
- Python programming language: this is the core programming environment used 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.
- PyQt5: this also supports kit for the design of the application GUI.
- py2exe: this was used to compile the python codes (.py) to an executable file (.exe).

3.4 ARCHITECTURE OF THE DEVELOPED SYSTEM

The Adjectival Phrase English to Yorùbá Machine Translation system is made up of the following;

1. System GUI: this is the user interface that interfaces the User and the translator which the Users can easily type what they intended translating.
2. Translator: this translates the words provided by Users by fetching the corresponding translation provided by the Parser.
3. Parser: Parsing is an important phase which is used to understand the syntax and semantics of any source language sentences confined to the grammar. Parsing is actually the automatic analysis of texts according to any grammar. Parser would parse the words in the database

4. Database: Database is the module used in storing the data used for the translation. The database make up of parallel corpus from Words were collected from both languages. The sentences were broken down into their part of speech (POS). The database was designed by categorizing all the parts of speech into their different grammatical functions.

Figure 3.1 shows the architecture of the system

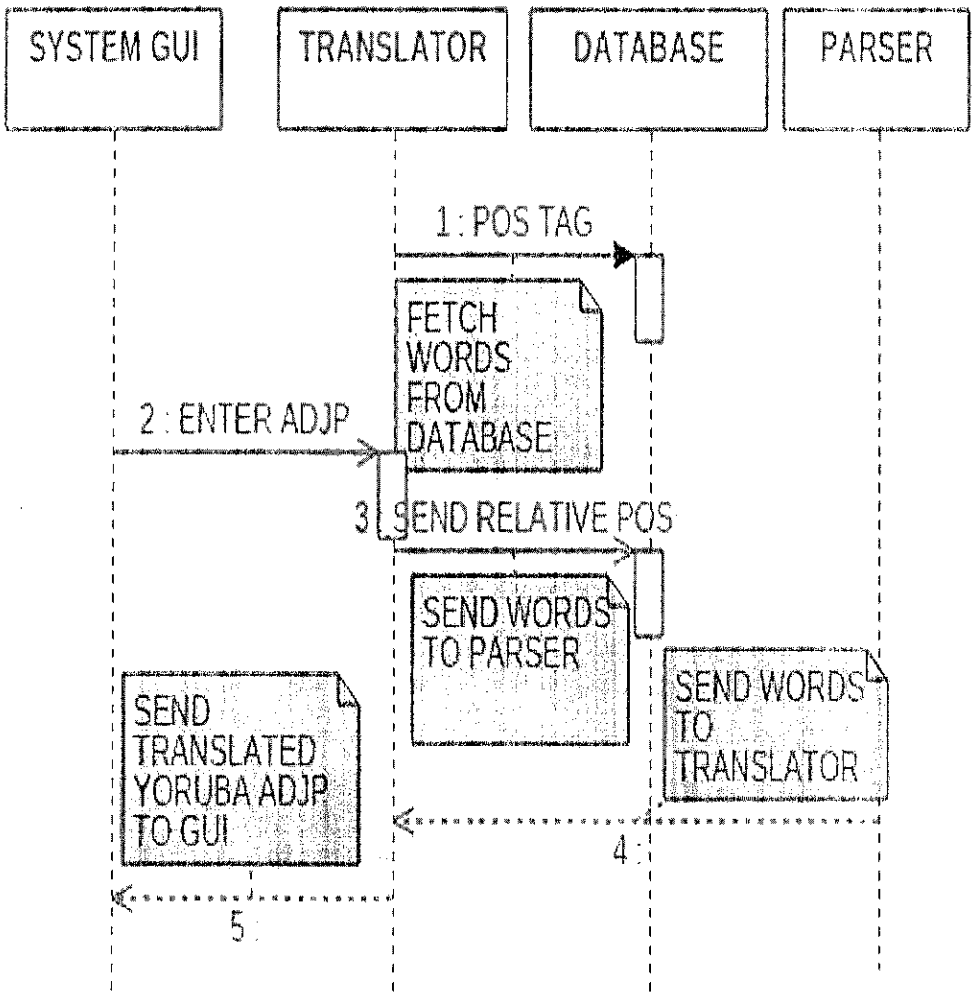

Figure 3.1: Architecture of the System

3.5 THEORETICAL AND SYSTEM FRAMEWORK DESIGN

The theoretical and system framework design involves the system design, database design, and system software design.

3.5.1 SYSTEM DESIGN

The system was design considering all principles and the rules guiding the translation from the source language to the target language. The system design procedure involves that the users are allowed to enter a text in the source language which is the English Language, the texts are broken into token (lexemes). The token is then patterned according to the re-write rules. The re-write rules are designed and developed. The lexemes are fetched from the database. The outputs of the system are then displayed through the Graphical User Interface (GUI)

3.5.1.1 Re-write Testing

The rules that guides the system design are;

Rule 1: An Adjectival phrase (ADJP) consists of adjective and Noun phrase (NP). In the case of target language noun (QO) comes before determiner (AIQO). For example

SL: an<DET>old<ADJ>man<N>.

TL: Àgbàlágbà<QA>Qkùnrin<QO>Kan<AIQO>

Rule 2: A determiner 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<QO>gíga<QA>náà<AIQO>

The JFLAP was used to test the rewrite rules as shown in Figures 3.2 and 3.3. The mode of translation is based on the grammar designed for both English language and Yorùbá language.

3.5.1 Database Design

The data (corpus) for the research was collected from both languages. The sentences were broken down into their part of speech (POS). The different parts of speech are stored in pairs. Table 3.1 and 3.2 shows list of English pronouns and their Yorùbá equivalents, and English Determinants and their Yorùbá equivalents. Figure 3.4, and Figure 3.5, show the Noun and Adjectives respectively.

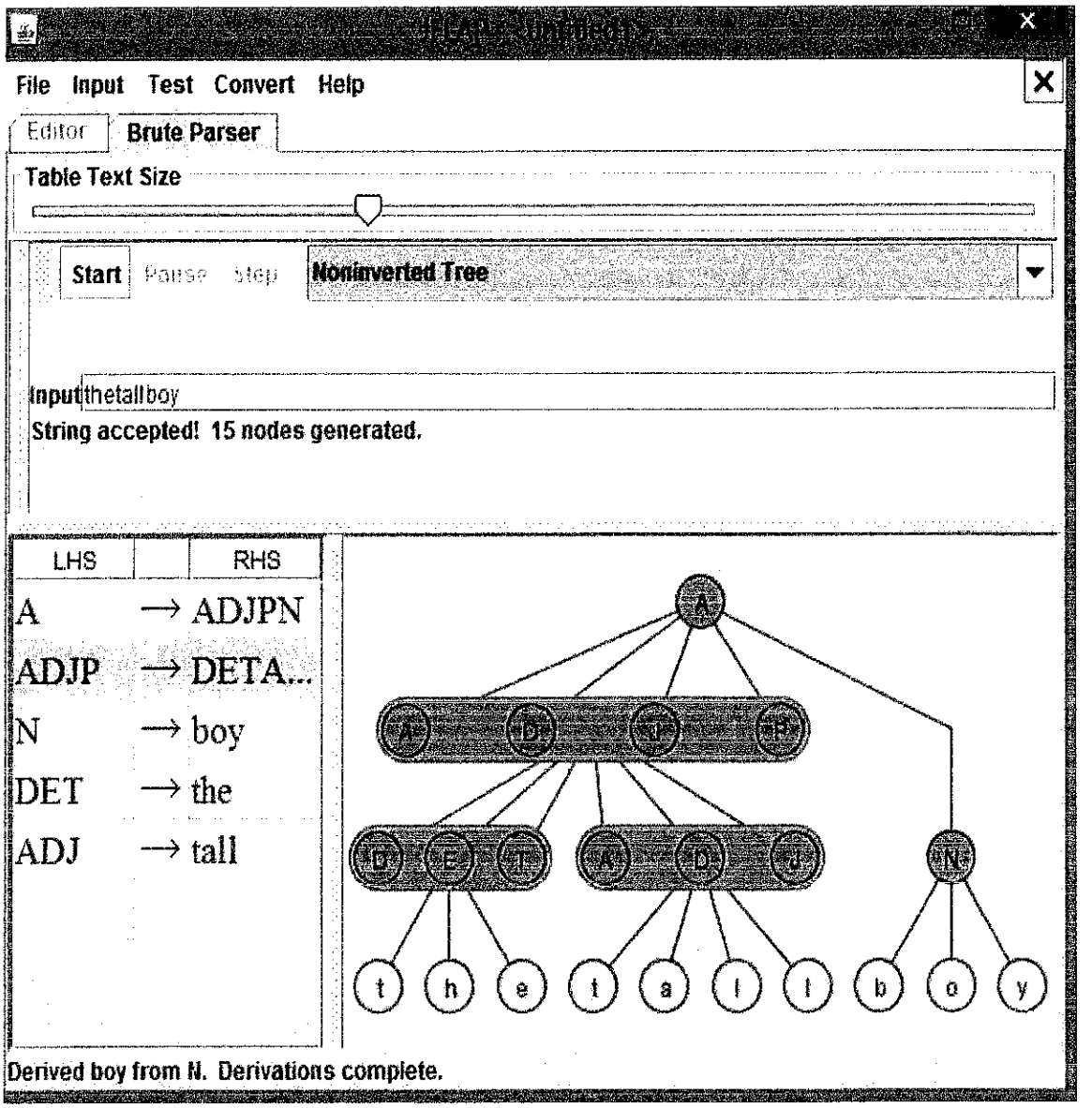

Figure 3.2: English Adjectival Phrase Rewrite Test

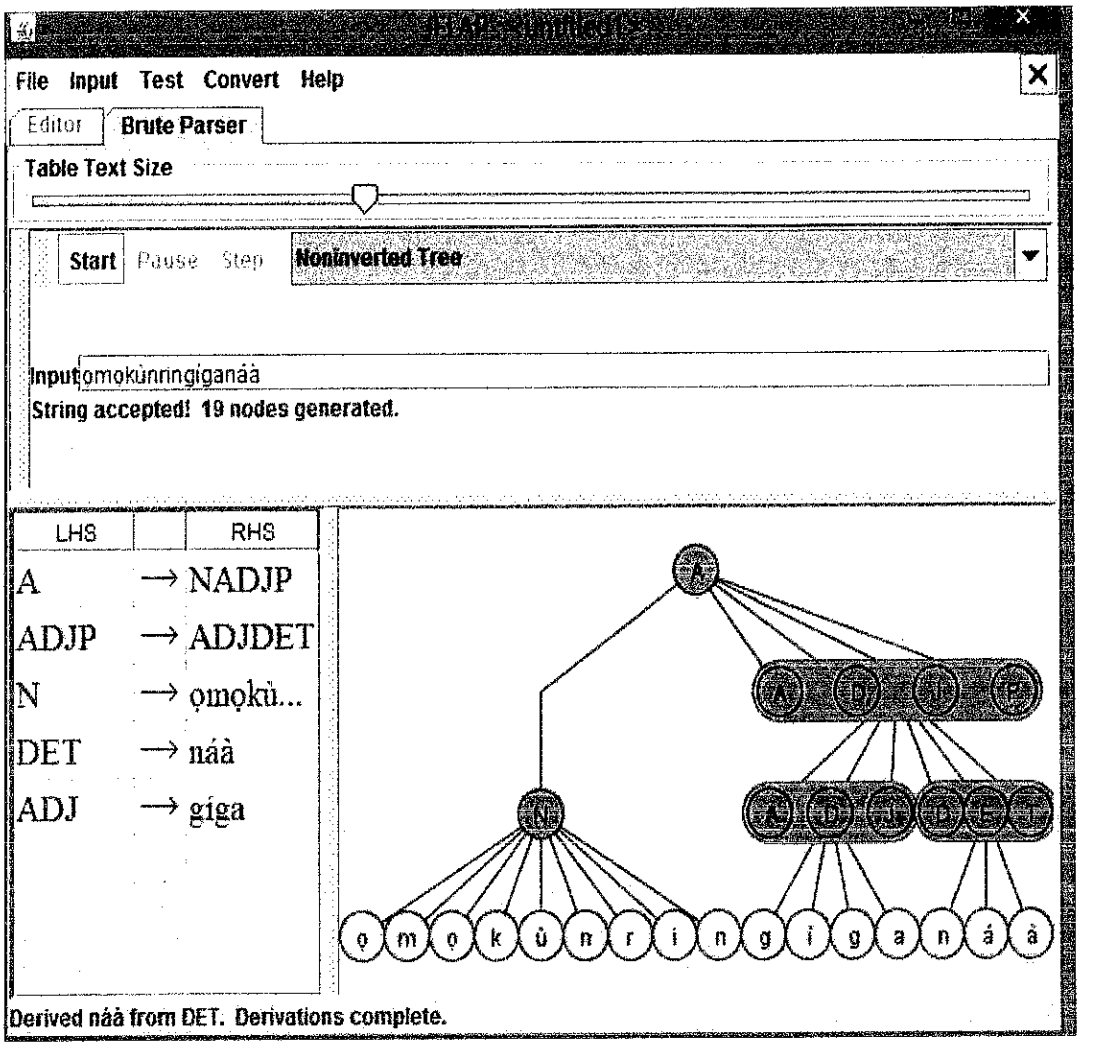

Figure 3.3: Yorùbá Adjectival Phrase Rewrite Test

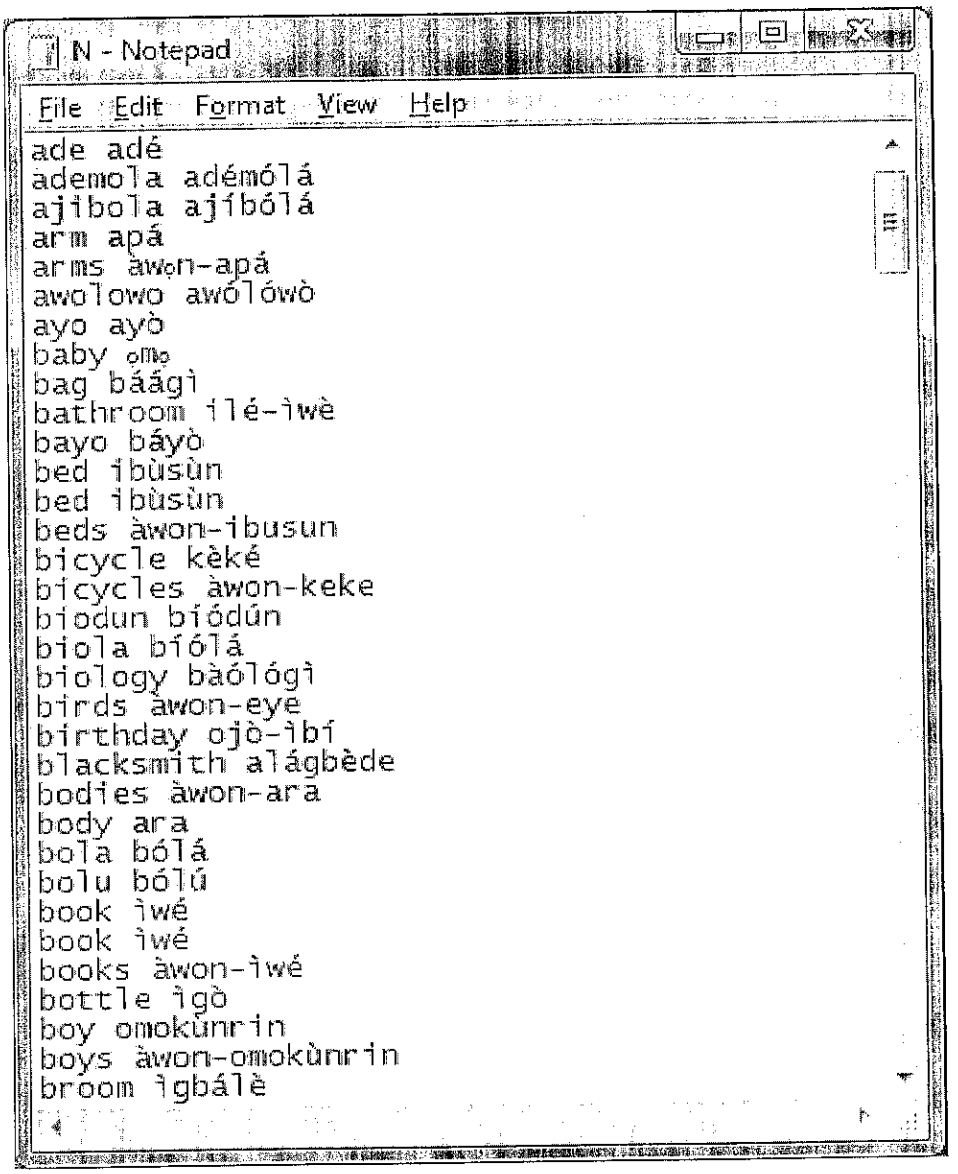

Figure 3.4: Database for noun

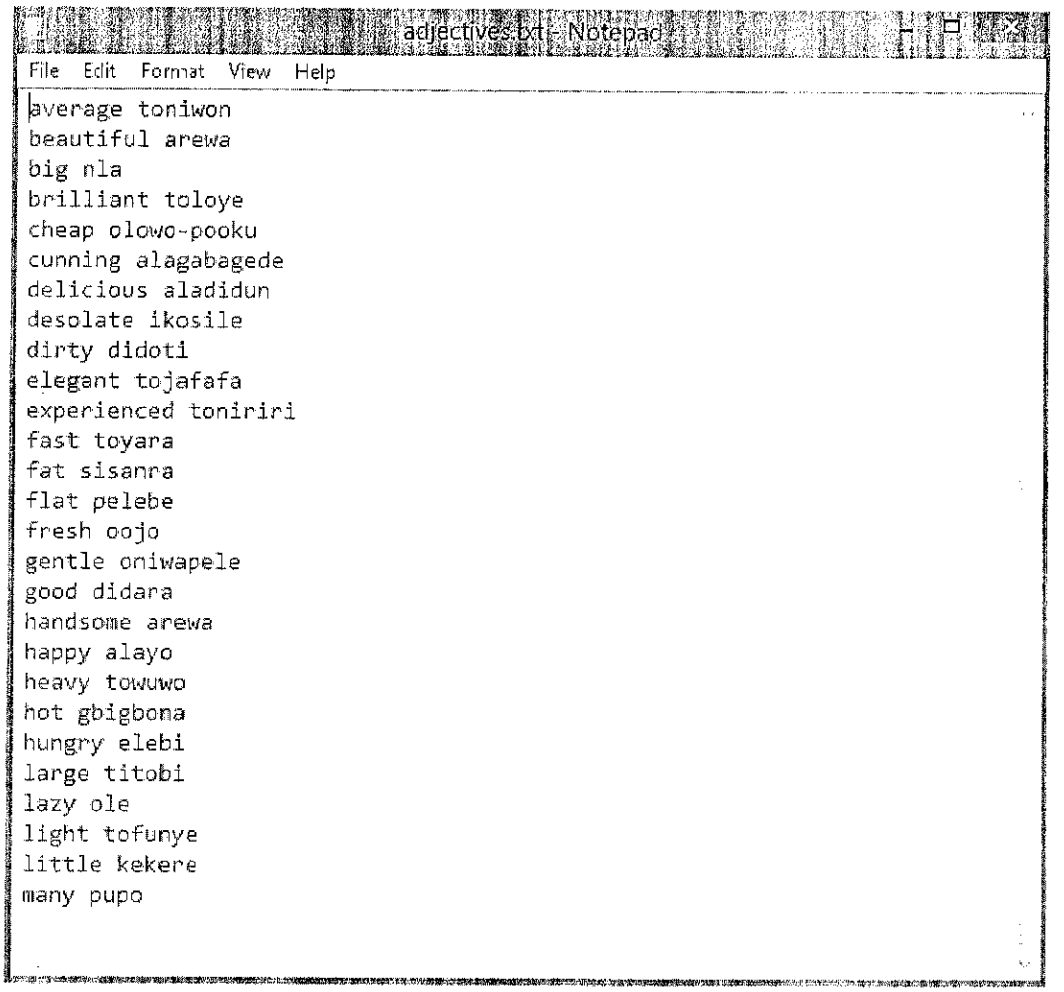

Figure 3.5: Database for adjectives

Table 3.1: List of English pronouns and their Yorùbá equivalents

English	Yorùbá
She/he/it	Ó
They	Àwọn
You	ìwọ/irẹ
We	Àwa
Them	wọ́n

Table 3.2: List of English determinants and their Yorùbá equivalents

English	Yorùbá
A	Kan
AN	Kan
SOME	Díẹ̀
THE	Náà

3.5.2 Adjectival Phrase Translation Process

The English ADJP translation process model is shown in Figure 3.6. Figure 3.6describes possible phrases that can be translated from the source language (SL) to the target language (TL). The ways translation of English adjectival phrase can be combined are: ADJDETN and PREDETADJN. Figure 3.7 is the state diagram for the Yorùbá language ADJP translation process. Figure 3.8 shows possible combinations of adjectival phrases that can be accepted by the TL. They are: ATKQOAIQO and ATKQOQAAIQO. One important thing to note is that, the noun (QO) and adjective (QA) swapped with the determiner. It shows that Yorùbá language is head first and English language is head last.

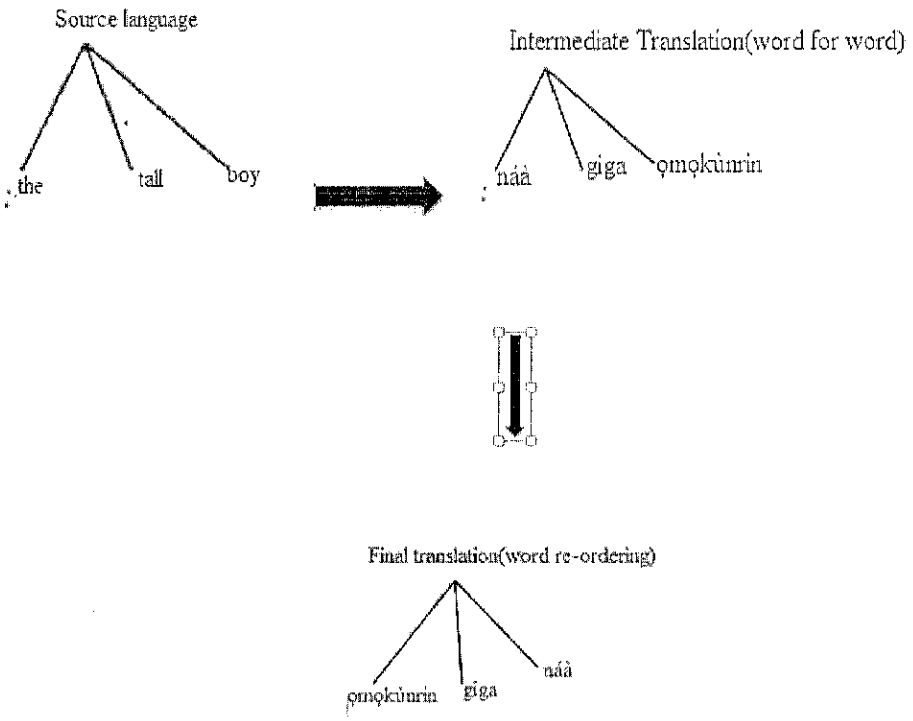

Figure 3.6: Adjectival Phrase Translation Process Abstraction

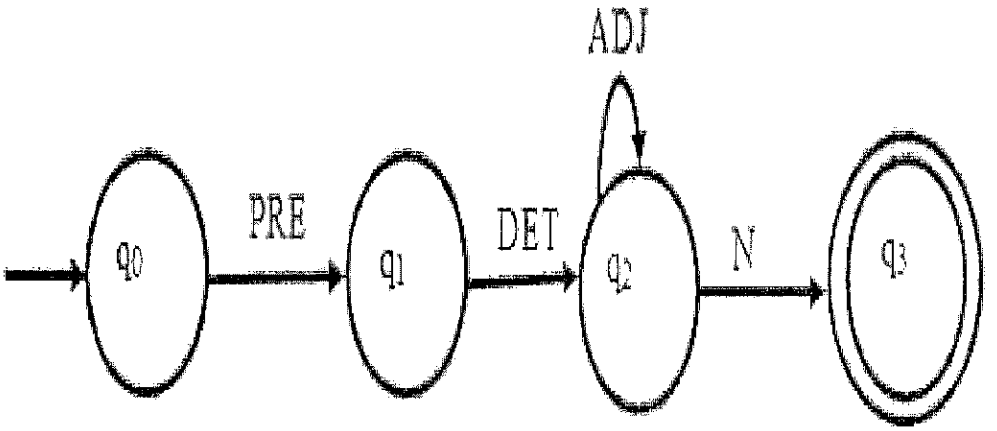

Figure 3.7: State diagram for the English translation process

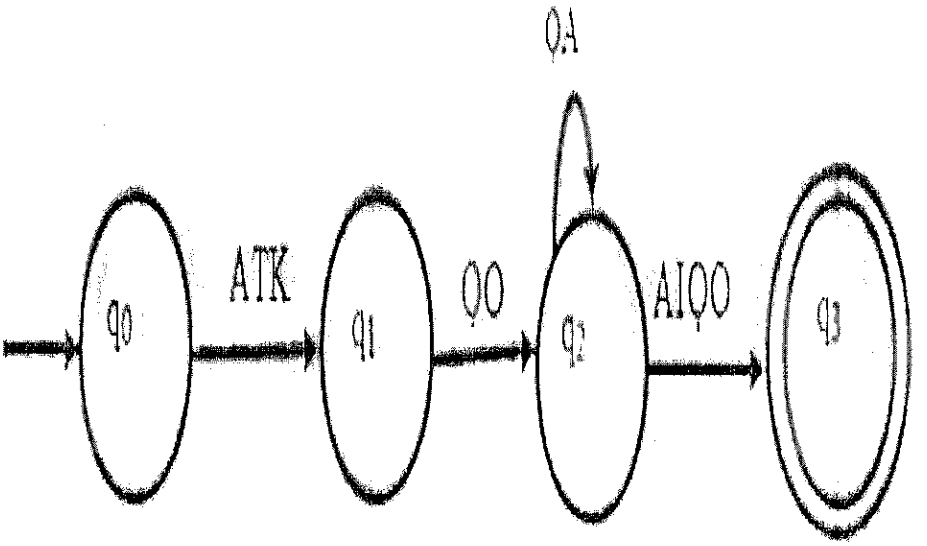

Figure 3.8: State diagram for the Yorùbá translation process

3.6 SYSTEM SOFTWARE DESIGN AND IMPLEMENTATION

The software design can be divided into two different modules; the Graphical User Interface (GUI) which is designed using UML and implemented using python programming language. The GUI has three planes, the first plane is where user enters the adjectival phrase. The second plane display input phrases word for word. The third plane displays the translated Yorùbá adjectival phrase. After the phrases have been typed, the translator module of the code begins to execute. The phrase is broken into lexemes, it then tagged into different parts of speech.

The translator module will accept input sentence from the GUI module then break it down, and send it to the database module to confirm that the lexemes are in the database. However, if the lexemes are not in the database an error message will be generated. The translated sentence to target language is then displayed by the GUI. Python programming language was used in the software coding and the interface of the machine is designed using PyQt5. The lexemes are manually tagged and each word is categorised according to its parts of speech. The Natural Language Tool Kits (NLTKs) was used as the parser module. The translation process is based on the phrase grammar rules built in the source code which implements the re-write rules. The machine translation system has the capability to translate sentences that contains an adjectival phrase from the English Language to Yorùbá language in its textual form.

CHAPTER FOUR

SYSTEM EVALUATION, RESULT AND DISCUSSION

4.1 EVALUATION OF SYSTEM

The developed system was evaluated by administering questionnaires to respondents and the mean opinion score (Human Judgment) approach was used in determining the performance of the system.

4.1.1 The Mean Opinion Score

The Mean opinion score (MOS) is a subjective measurement of people's opinion. The Expert i.e. the professional translator translates the sentences from English language to Yorùbá language. The evaluation was done in other to compare the developed system to experimental subject respondents' and the Expert translations.

4.1.2 Questionnaire Design

The questionnaire designed has simple phrases that consist of adjectival phrases to test the experimental subject respondent on the ability to translate simple sentences. The questionnaire has ten (10) simple adjectival phrases which were used in testing the respondents' translation accuracy based on Yorùbá language orthography and the syntax of the language which is described in term of tone marks and diacritics (dotted vowels and consonant).

4.1.3 Questionnaire Administration

The questionnaires were administered in Ikole-Ekiti, Ekiti state, Nigeria. This area was chosen because there are literate Yorùbá speakers and the questionnaires were distributed among the Yorùbá speakers from the Yorùbá ethnic group.

4.2 SYSTEM OUTPUT RESULT

The sample of the output generated by the system is shown in Figure4.1, 4.2 and 4.3.

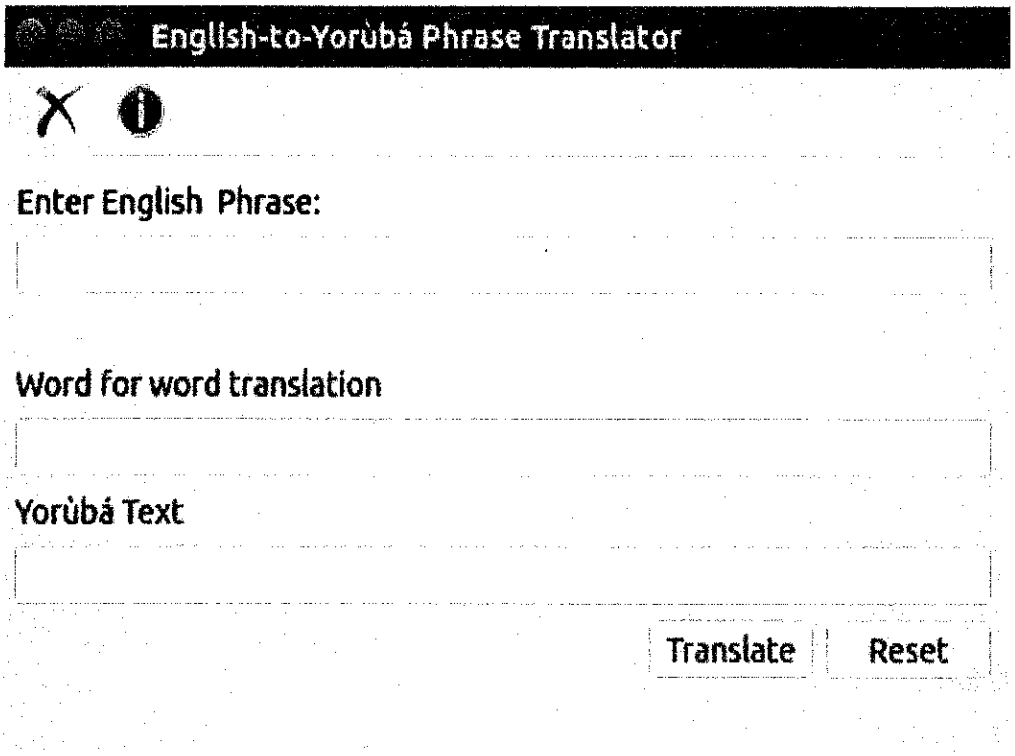

Figure 4.1: System GUI

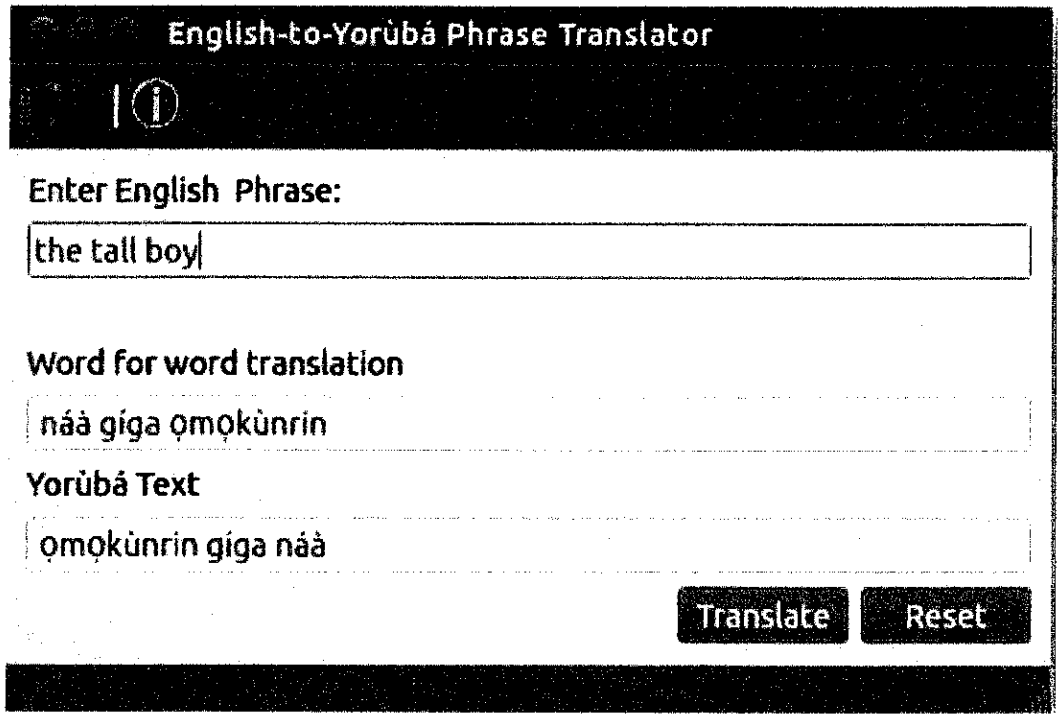

Figure 4.2: System Output Sample 1

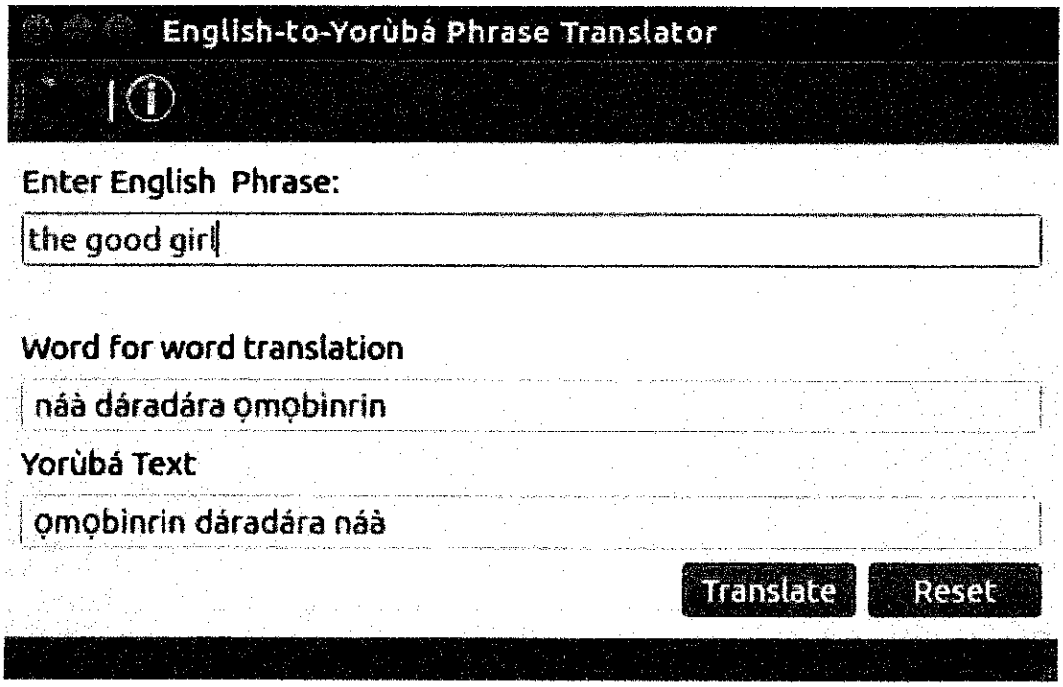

Figure 4.3: System Output Sample 2

4.3 DISCUSSION OF RESULT

The developed system was evaluated to determine its performance and this therefore demonstrates the quality and shortcoming of the developed system based on system accuracy using word orthography (tone marking and under dotting) accuracy. Results show that most of the experimental respondents got the translation correctly while many do not know how to tone mark words when compared to the expert's translated phrases as well as the developed system as shown in table 4.1. Results of evaluation based on word orthography (tone mark and the under dots correctness) using the mean opinion score (MOS) is shown in Figure 4.4. From the graph, it is shown that the expert has the highest score while the developed system has higher accuracy than the experimental subject respondents.

From table 4.1, the expert's percentage accuracy was 100 while the developed system has 95.5 percent accuracy and the result from experimental subject respondents is 76.3 percent. Figure 4.3 depicts that the machine correctness is close to that of the Expert and more accurate than that of the average experimental subject respondents.

Table 4.1: Analysis of Evaluation Results

Phrases	Expert	Respondent Average	Machine
1	100	70	100
2	100	78	100
3	100	80	85
4	100	75	100
5	100	80	90
6	100	70	100
7	100	80	90
8	100	78	90
9	100	77	100
10	100	75	100
Average	100	76.3	95.5

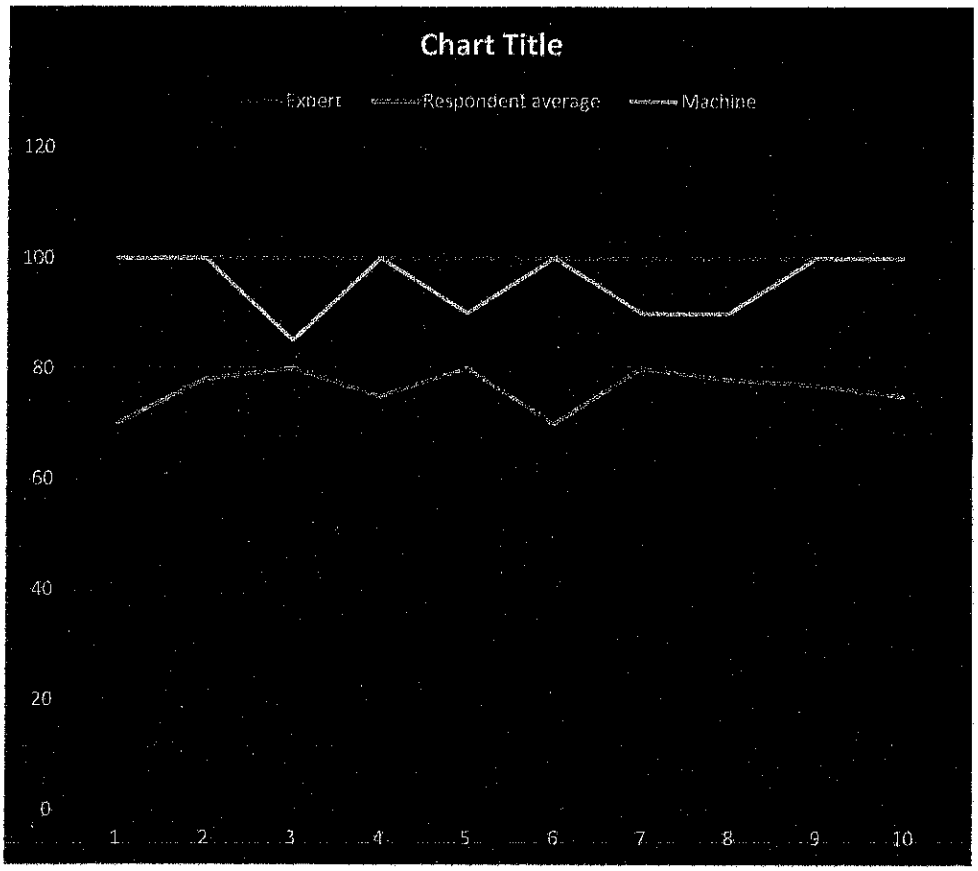

Figure 4.4: Translated phrases orthography accuracy

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

An adjectival phrase-based system was designed in this research to translate English language to Yorùbá. The system was designed to enhance the learning of Yorùbá language with a user-friendly interface.

Results show that the developed system was able to give accurate translations with appropriate tone-marks and under-dots because its accuracy is close to that of the Expert and more accurate than the experimental subject respondents'.

5.2 RECOMMENDATION

The result gotten from this project shows that most people that speak Yorùbá language are not good at writing it. In lieu of this, it is recommended that the school encourage researchers in computational linguistics by funding research in the field and creating community for Machine Translation.

It is also recommended that future researchers work on translation of Adjectival phrases from English to other languages (Hausa and Ibo) in Nigeria.

Finally, with the advent of Neural Network, I will recommend further research in Neural Machine Translation.

REFERENCES

Abiola, O. B., Adetunmbi, A., & 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), 4*(8), 308-313.

Abiola, O. B., Adetunmbi, A., Fasiku, A. I., & Olatunji, K. (2014). a web-based English Yorùbá Noun phrases machine translation system. *International journal of English and Literature, 5*(3), 71-78.

Adenekan, D. I., Agbeyangi, A. O., & Eludiora, S. I. (2015). English to Yorùbá Machine Translation System using rule-based approach. *Journal of Multidisciplinary Engineering Science and Technology, 2*(8), 2275-2280.

Agbeyangi, A., Eludiora, S., & Adenekan, O. A. (2015). English to Yorùbá Machine TranslationSystem using Rule-Based Approach. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, 2275-2280.

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*, 154-159.

Almaflehi, N., & Saad, S. (2013). the problem of translating the prepositions at, in and on into Arabic: An applied linguistic approach. *Journal for the study of English linguistics, 1*(2).

ALPAC. (1966). *Language and Machines: Computers in Translation and Linguistics. A repor.* Washington D.C., 20418 USA: National Academy of Sciences, National Research Council.

Anthony, P. J. (2013, march). Machine Translation Approaches and Survey for Indian Languages. *The Association for Computational Linguistics and Chinese Language Processing, 18*(1), 47-78.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., & Sadlery, L. (1994). *Machine Translation: An Introductory Guide.* London: NCC Blackwell.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, (pp. 65-72). Ann Arbor, Michigan.

Billot, S., & Lang, B. (1989). The structure of shared forests in ambiguous parsing. *In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 143- 151.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, (pp. 263-270). Michigan.

Dasgupta, T., & Basu, A. (2008). An English to Indian Sign Language Machine Translation System. Retrieved from www.cse.iitd.ac.in/embedded/assistech/Proceedings/P17.pdf.

Dirix, P., Schuurman, I., & Vandeghinste, V. (2005). Metis II: Example-based machine translation using monolingual corpora - system description. *2nd Workshop on Example-Based Machine Translation*, (pp. 43-45).

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. *Second Workshop on SMT*, (pp. 220-223).

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

Eludiora, S. I., Agbeyangi, A. O., & Fatunsin, A. (2015). Development of an English to Yorùbá Machine Translation System for Yorùbá Verbs' Tone Changing. *International Journal Computer Application, 129*(10), 12-17.

Eludiora, S., & Atolagbe, R. (2016). Development of a Prepositional Phrase Machine Translation System. *World Journal of Computer Application and Technology, 4*(4), 46-57.

Eludiora, S., Abayomi, A., & O.I, F. (2015). Development of English to Yorùbá Machine Translation System for Yorùbá Verbs' Tone Changing. *International Journal of Computer Applications*, 0975 – 8887.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., WeiWang, & Thayer, I. (2006). Scalable Inference And Training Of Context-Rich Syntactic Translation Models. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,*, 961–996.

Groves, D., & A, W. (2005). Hybrid example-based SMT: the best of both worlds. *ACL Workshop on Building and Using Parallel Texts*, (pp. 183-190).

Haque, Dandapat, Srivastava, Naskar, & Way. (2009, August 7). "English—Hindi Transliteration Using Context-Informed PBSMT:the DCU System for NEWS

2009". *Proceedings of the 2009 Named Entities Workshop ACL-IJCNLP 2009*, pp. 104-107.

Huang, L., Knight, K., & Joshi, A. (2006). Statistical syntax-directed translation with extended domain of locality. *In Proceedings of the 7th Conference of the Association for Machine Translation in the America*, 223-231.

Hutchins, J. (1993).. The first MT patents. *MT News International*, 14-15.

Hutchins, J. (2005). *The history of machine translation in a nutshell.* Retrieved April 14, 2017, from hutchinsweb: http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Hutchins, J. W. (1994). *MACHINE TRANSLATION: A BRIEF HISTORY.* London: Academic Press.

Hutchins, W. J., && Lovtskii, E. (2000). Petr Petrovich Troyanskii (1854-1950): A forgotten pioneer of mechanical translation. In W. J. Hutchins, & E. & Lovtskii, *Machine translation* (pp. 187-221.).

IBM. (1954, january 8). *IBM Archives online: Press release January 8th 1954.* Retrieved April 15, 2016, from IBM Archives online: http://www-03.ibm.com/ibm/history/exhibits/701/701-translator.html.

Islam, Z., Tiedmann, J., & Eisle, A. (2009). *English to Bangla phrased-based statistical machine translation.* Saarland: Saarland University.

Kamal, D., & Goya, I. V. (August 2011). Hybrid Approach for Punjabi to English Transliteration System. *International Journal of Computer Applications* , 0975 – 8887.

Koehn, P., & Hoang, H. (2007). Factored translation models:In Proceedings of the 2007 Joint Conference on Empirical Methods. In NLP and Computational Natural Language. *Joint Conference on Empirical Methods; In NLP and Computational Natural Language*, (pp. 868-876).

Levenshtein, V. I. (1966). *Binary Code Capable of Correcting Deletions,Insertions, and Reversals.* Soviet Physics Doklady.

Mi, H., Liang, H., & Liu, Q. (2008). Forest-based translation. *Proceedings of Association ofComputational Linguistic*, 192-199.

Nießen, S., Och, F., Leusc, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. *2nd International Conference on Language Resources and Evaluation (LREC-2000)*, (pp. 39–45). Athens.

Noone, G. (2003). *Machine Translation : A Transfer Approach.* A Project Report. Retrieved from www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203.pdf.

Odejobi, O. O., Eludiora, S. I., Akanbi, L. A., Iyanda, I. R., & Akinade, O. A. (2015). A web-based system for supporting teaching and learning of Nigerian indigenous languages. *OAU TekCONF 2015 proceedings*, (pp. 350-360). Ife.

Odejobi, T., Owolabi, K., & Adegbola, T. (2011). *Localising for Yorùbá: Experience, Challenges and Future Direction.*

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosen, V., & Flickinger, D. (2007). Towards hybrid quality-oriented machine translation on linguistics and

probabilities in MT. *11th Conference on Theoretical and Methodological Issues in Machine Translation*, (pp. 144-153).

Okpor, M. D. (2014, September). Machine Translation Approaches: Issues and Challenges. *International Journal of Computer Science Issue, 11*(2), 159-165.

Oladosu, J., Esan, A., Adeyanju, I., Omodunbi, B., Olaniyan, O., & Adegoke, B. (2016, September). Approaches to Machine Translation: A Review. *FUOYE Journal of Engineering and Technology, `1*(1), 121-126.

Oladosu, J., Esan, A., Ibrahim, A., Benjamin, A., Olatayo, O., & Bolaji, O. (2016). Approaches to Machine Translation: A Review. *FUOYE Journal of Engineering and Technology*, 120-126.

Pankaj, K., & Er.Vinod, K. (2013). Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 318-321.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a Method forAutomatic Evaluation of Machine Translation. *40th Annual Meeting of the Association for Computational Linguistics*, (pp. 311-318). Philadelphia.

Sangeetha, J., Jothilakshmi, S., & Kumar, R. (2014). An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism. *International Journal of Engineering and Technology (IJET)*.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. *Second Workshop on SMT*, (pp. 203-206).

Tillmann, C., Vogel, S., Ney, H., Zubiag, A., & Sawaf, H. (1997). Accelerated DP Based Search For Statistical Translation. *European Conference on Speech Communication and Technology*, (pp. 2667–2670). Rhodes.

Yulian, H. (2014). Statistical machine translation based on translation rules. *Journal of Chemical and Pharmaceutical Research*, 1628-1635.

Zhang, Y. (2006). *Chinese-English SMT by Parsing.* Retrieved from www.cl.cam.ac.uk/~yz360/mscthesis.pdf.

APPENDIX A

Questionnaire

DEPARTMENT OF COMPUTER ENGINEERING FEDERAL UNIVERSITY OYE-EKITI, IKOLE CAMPUS

Questionnaire on the Design of Adjectival Phrase English to Yoruba Machine Translation System

Dear correspondent,

This questionnaire is designed to collect data for the evaluation of a software designed and developed as a final year project in partial fulfilment of B.Eng. degree in computer engineering titled "Design of Adjectival phrase based English to Yorùbá Machine Translation System". In lieu of this, I hereby solicit for your cooperation to honestly answer the questions and it will be treated as confidential.

Section A

Please tick the appropriate answer.

1. Age 15-20 ( ) 20-25 ( ) 25-30 ( ) above 30 ( )
2. Sex Male ( ) Female ( )
3. State of Origin .....
4. State of Residence .....
5. Educational level: SSCE ( ) Undergraduate ( ) Postgraduate ( )
6. Knowledge of Yorùbá orthography(writing) System
   Weak ( ) Average ( ) Excellent ( )
7. Have you use any Machine Translation System before? Yes ( ) No ( )
8. If above is Yes what is the name of the System .....

Section B

Please use appropriate Yorùbá orthography .e.g.

1. The good girl

   _____

2. A tall boy

   _____

3. Some old book

   _____

4. A lazy boy

   _____

5. The intelligent boy

   _____

6. For young student

   _____

7. The friendly dog

   _____

8. A short skirt

   _____

9. The black boy

   _____

10. The big dress

    _____