Word Sense Disambiguation In English To Yorùbá Machine Translation System

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

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

Ojediran, Damilola T.
Dept. of Comp. Sci. and Engng
Obafemi Awolowo University
Ile-Ife, Osun State-Nigeria

***Abstract*—Ambiguity is a pervasive characteristic of natural language. The specific sense intended is mostly determined by the textual context in which an instance of the ambiguous word appears. In this paper, we examine how words which can be used as both nouns and verbs in a simple subject-verb-object (SVO) English sentence can be disambiguate.**

First, we collate words that are normally used in the home domain according to their different parts of speech. Next, their corresponding Standard Yorùbá translations were obtained. Phrase Structure Grammar was used to determine the structure of the sentence and the position of each parts of speech that make up the sentence. Rewrite rules were also written for the simple English sentences constructed from a combination of these ambiguous words; the words were then stored in a database designed to accept determiners, nouns and verbs. Thereafter, the system was implemented using Python programming language and SQLite3 for the database.

The system was evaluated using mean opinion score approach by ranking the output from the system and comparing the output with the translation given by a group of Yorùbá native speakers. The result showed that the system gave correct translation for each of the ambiguous English sentence inputted and produced a recall of about 90% with respect to the collected corpus. Based on the result gathered, there are some issues to address that could be considered in a future work.

Keywords—Word sense disambiguation, Machine translation, Yorùbá language, sentences, corpus/corpora

I. INTRODUCTION

Word sense disambiguation (WSD) is the process of determining the correct sense of a word in context [1]. In English Language as well as many other Natural Languages, there are certain words which when used in sentences can have different meanings. Mostly common in English Language are words which can be used as both nouns and verbs in a sentence. It has been noted that WSD is a fundamental problem in natural language processing (NLP), and it is important for applications such as machine translation. Ability to identify the sense of a word (i.e. meaning) used in a sentence, when the word has multiple meanings always poses a great challenge. Most state of-the-art WSD systems are supervised classifiers that are trained on manually sense-tagged corpora, which are very time-consuming and expensive to build [2].

One of the successful approaches to WSD is the use of supervised machine learning. However, this approach involves the collection of a large text corpus in which each ambiguous word has been annotated with the correct sense to serve as training data [1]. Due to the rigorous annotation process, only a handful of sense-tagged corpora are publicly available. The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, resolution, coherence, inference etc. Several SENSEVAL conferences have attempted to put Word Sense Disambiguation on an empirically measurable basis in hosting evaluations in which a given corpus of tagged word senses are created using WordNet's senses and participants attempt to recognize those senses after tuning their systems with a corpus of training data (SemEval-2007, Senseval-2).

Studies have been conducted to show the validity of using parallel corpora for word sense disambiguation [3]. The same approach of using parallel corpora was also use in our research.

The scope of this research work is restricted to Home domain (English language spoken in home environment) while the motivation is to develop a system that disambiguates nouns and verbs in English to *Yorùbá* Machine Translation. It will contribute immensely to the advancement of research into building efficient English to *Yorùbá* Machine Translation System.

The remaining part of the paper is structured as follows: Section 2 gives an overview of Yorùbá grammar structure and it's constituent. Section 3 discusses the system design and implementation, while Section 4 discusses the results. Section 5 concludes the paper.

II. YORÙBÁ GRAMMAR

Standard *Yorùbá* (SY) language is the language of trade, education, mass communication and general everyday interaction between *Yorùbá* people. It is a language spoken by over 40 million people, mainly in West Africa. In Nigeria, it is spoken in Lagos, Ọ̀sun, Ògùn, Òndó, Ọ̀yọ́, Èkìtì and Kwara, as well as some part of Kogi State [4].

The *Yorùbá* Alphabet has 25 letters altogether which can be represented in both upper and lower case. It also has eighteen consonants and seven oral vowels. It also has five nasal vowels. Furthermore, *Yorùbá* has three level tones: high, mid and low represented with [´], [¯] and [`] respectively. Tones usually occur on vowels. The three level tones determine the meanings that each word has in *Yorùbá language*. For example, a word that has the same form (i.e. vowels and consonants) can have different meanings depending on the tones. Table 1 below shows examples of these words.

TABLE I. EXAMPLES OF WORDS WITH DIFFERENT TONES

Yorùbá Word	English Meaning
Igba Igbá Ìgbà Igbà	‘two hundred’ ‘calabash’ ‘time’ ‘climbing rope’
ọkọ ọkọ́ ọ̀kọ̀ ọkọ̀	‘husband’ ‘hoe’ ‘spear’ ‘vehicle’

A. *Yorùbá Morphology*

Yorùbá has some productive methods of word derivation. The main morphological processes in the language include: affixation, compounding and reduplication.

1) *Affixation*

Yoruba uses prefixation and infixation to derive new words. Each of the *Yorùbá* oral vowels (except /u/ in the standard dialect) can be used as a prefix to derive a new word. Each of the usable six oral vowels – a, e, ẹ, i, o, ọ - has two forms as a prefix: mid toned and low toned. They are attached to verbs to derive nouns.

a) *Low toned prefixes*

ọ` + dẹ` ‘to be soft’ = ọ̀dẹ` ‘idiot’

ì + ṣẹ́ ‘to break’ = ìṣẹ́ ‘poverty’

è + rò ‘to think’ = èrò ‘thought’

à + rè ‘to go’ = àrè ‘wonderer’

b) *Mid toned prefixes*

ẹ + rù - ‘to carry’ = ẹrù ‘load’

ọ + dẹ - ‘to hunt’ = ọdẹ ‘hunter’

e + wé - ‘to wrap’ = ewé ‘leaves’

o + dì - ‘to fold’ = odì ‘malice’

2) *Compounding*

Yoruba also derive new words by combining two independent words:

ẹran ‘meat’ + oko ‘farm’ = ẹranko ‘animal’

ìyá ‘mother’ + ọkọ ‘husband’ = iyakọ ‘mother-in-law’

3) *Reduplication*

Yorùbá derives nominal items/adjectives from verbs through a partial reduplication of verbs. New nouns can also be derived by a total reduplication of an existing noun. E.g.

jẹ ‘to eat’ = jíjẹ ‘edible’

se` ‘to cook’ = síse` ‘cooked’

ọmọ ‘child’ = ọmọọmọ ‘grand-children’

Yorùbá language is strictly SVO, and the 3s object simply copies the vowel of the preceding verb, an iconic representation of the extension or completion of the verbal activity, as in the following:

- ó fà á - ‘He pulled it.’
- ó sí í - ‘He opened it.’[4]

III. METHODOLOGY

The sentences were broken down into constituents which are unit words of the sentence. The rewrite rules produced using the Phrase Structure Grammar for the sentence is as follows:

S ::= <NP> <VP>

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

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

where S is the sentence. NP, VP, N, V and DET are the non-terminals. NP is Noun Phrase, VP is Verb Phrase, N is Noun and V is verb. The operation is such that the Left hand side (LHS) is substituted with the Right hand side (RHS) till the terminals are reached. Below in figure 1 is the Finite State Automaton for the grammar.

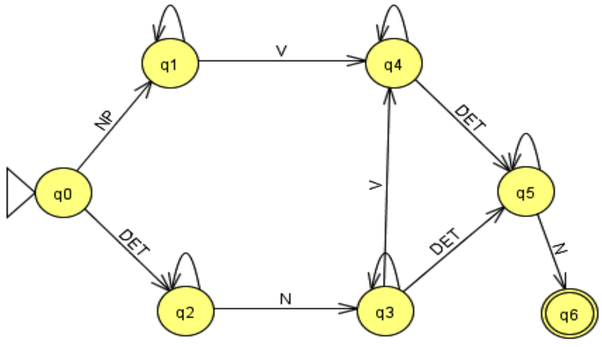


Fig. 1. Finite State Automaton Model for the grammar

The grammar rules play a major role in the translation process of the system. The rules used were gotten manually from simple English sentences formed which had combination of words that could be used as nouns and verbs. These rules were verified and parsed using NLTK (Natural Language Took Kit) before applied in the software development process. The outputs from the parser are shown in figure 2 to 4.

Rule 1: Any word preceded by a noun or pronoun is a verb

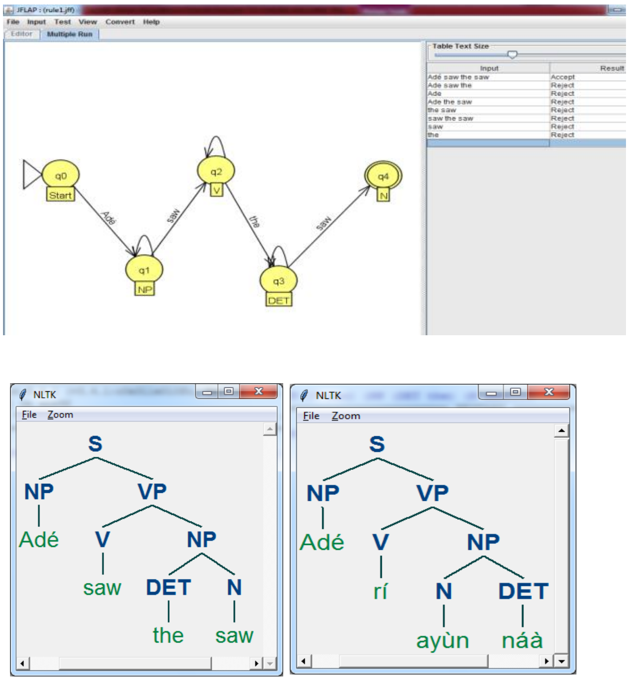
Adé saw the saw

| | | |

N V DET N

Standard *Yorùbá* Equivalent: *Adé rí ayùn náà*

Here verb 'V' being preceded by noun 'N' which validates the rule.



Input	Result
Adé saw the saw	Accept
Ade saw the	Reject
Ade	Reject
Ade the saw	Reject
the saw	Reject
saw the saw	Reject
saw	Reject
the	Reject

Fig. 2. FSA and NLTK output for rule 1

Rule 2: Any word preceded by a determiner is a noun

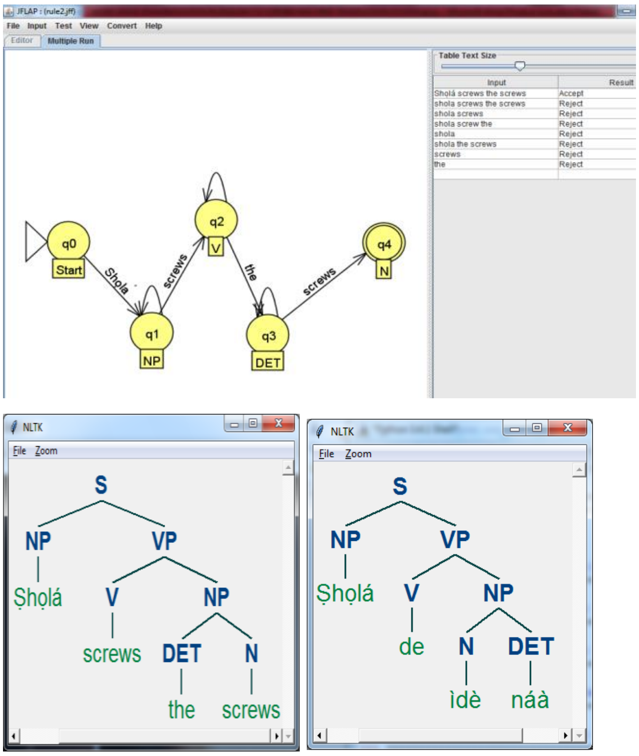
Ṣholá screws the screws

| | | |

N V DET N

Standard Yorùbá Equivalent: Ṣholá de ìdè náà

Here, determinant 'DET' precedes noun 'N' which validates the rule.



Input	Result
Sholá screws the screws	Accept
shola screws the screws	Reject
shola screws	Reject
shola screw the	Reject
shola	Reject
shola the screws	Reject
screws	Reject
the	Reject

Fig. 3. FSA and NLTK output for rule 2

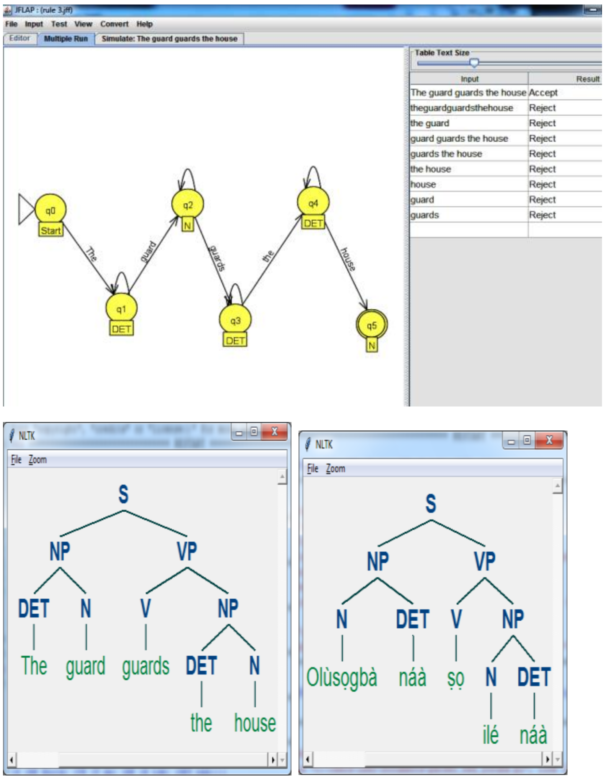
Rule 3: If a word precedes a determiner, it is a verb

The guard guards the house

| | | | |

DET N V DET N

Standard Yorùbá Equivalent: Olùsọ́gbà náà ṣọ́ ilé náà Here, 'V' precedes 'DET'.



Input	Result
The guard guards the house	Accept
theguardguardsthehouse	Reject
the guard	Reject
guard guards the house	Reject
guards the house	Reject
the house	Reject
house	Reject
guard	Reject
guards	Reject

Fig. 4. FSA and NLTK output for rule 3

These rules were used for the implementation and it only accommodates SVO sentences. Table 2 shows the tagging pattern used for the system. In the table, words were grouped according to their parts of speech along with a unique POS tag to identify them if they are used in a sentence.

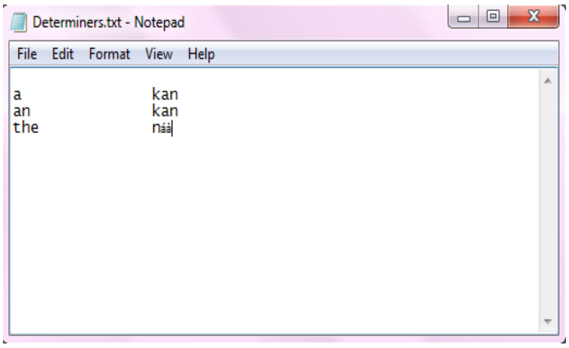
The database for this project is responsible for storing English words and their equivalent Standard Yorùbá translation. These words are then retrieved for use by the software when user seeks to translate an inputted English sentence.

TABLE II. PART OF SPEECH TAG SET

Tag	Meaning	Examples
N	Noun	saw, guard, nails, presents, balance, post, screws, records, suit, tie, walk, boy, girl, man, woman, Jide, Shola, Ayo
PRN	Pronoun	he, she, it, they, we
V	Verb	saw, guard, nail, balance, post, record, suit, tie, walk
DET	Determiner	the, a, an
ADJ	Adjective	big, large, small, beautiful, handsome

IV. DATA COLLECTION

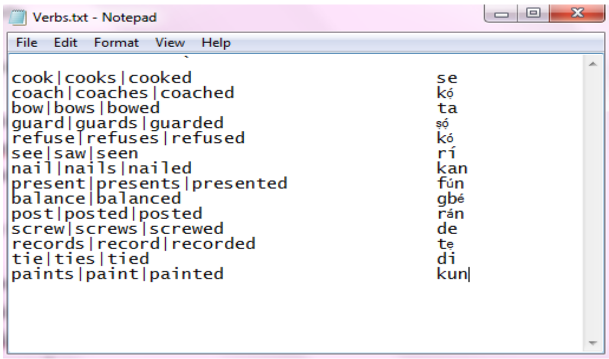
Words used in this research are limited to commonly used words in homes and offices. The Yorùbá translations were gotten from English/Yorùbá parallel dictionaries. Figure 5 shows the list of the determiners and their equivalent Yorùbá translations which were collected and stored in the database.



```
a                 kan
an                kan
the               náà
```

Fig. 5. List of Determinants in database

Figure 5 shows the list of verbs in their inflectional usage and their equivalent Yorùbá translations.



```
cook|cooks|cooked                          se
coach|coaches|coached                      kọ́
bow|bows|bowed                             ta
guard|guards|guarded                       ṣọ́
refuse|refuses|refused                     kọ́
see|saw|seen                               rí
nail|nails|nailed                          kan
present|presents|presented                 fún
balance|balanced                           gbé
post|posted|posted                         rán
screw|screws|screwed                       de
records|record|recorded                    tẹ
tie|ties|tied                              di
paints|paint|painted                       kun
```

Fig. 6. List of Verbs in database

It can be seen from figures 6 that words spelt using the same alphabet in English Language (source language) have different meanings in *Yorùbá* Language (target Language) as explained previously.

V. DEVELOPMENT TOOLS

The system design follows the architecture of a window application where the applications serve as a link between the user and the corpus (database). The main tools used are:

- Python programming language – this is the core programming environment for the application design.
- NTLK (Natural Language Toolkit) – this is a support kit for python programming design. Its features include: support for parsing, Part of Speech (POS) tagging, corpora design and analyses.
- SQLite – this is used in building the database.

A. *Requirement Analysis*

The requirements and specifications of the system are as follow:

- to present a user friendly interface to the user;
- to give the user access to input simple SVO sentences in English language provided the sentence is within the domain covered;
- Check for ambiguity in the sentence entered, translate and output the equivalent meaning of the sentences entered in standard Yorùbá language;
- give the user the ability to add to the corpus (database)

Unified Modelling Language (UML) was used for modeling the software both structurally and behaviorally. The UML sequence diagram for the system shows the GUI Module which sends the sentence to the Tagger which tags each word according to their part of speech. The Tagger then sends the tagged words to the Disambiguator where the nouns and verbs in each sentence are disambiguated using the rewrite rules. The Disambiguator then sends the words to the Translator that converts each word to their equivalent *Yorùbá* word, reorders them and sends them back to the GUI for output to the user. The sequence diagram for the system is depicted in figure 7.

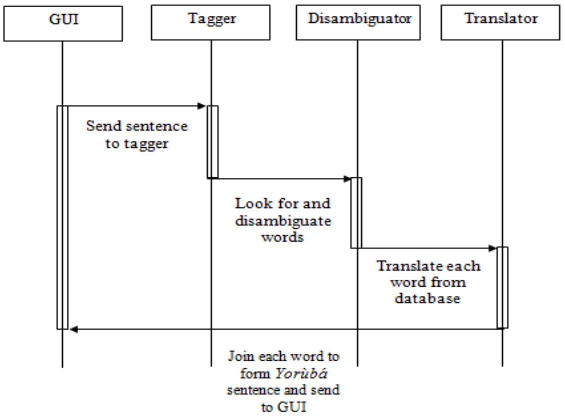

GUI
Tagger
Disambiguator
Translator
Send sentence to tagger
Look for and disambiguate words
Translate each word from database
Join each word to form *Yorùbá* sentence and send to GUI

Fig. 7. Sequence Diagram of the System

The UML use case diagram for the system is also shown in figure 8 below

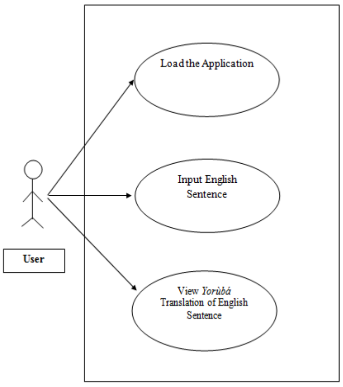


Fig. 8. Use Case Diagram of the System

B. *System Modules*

The modules can also be called classes or functions which perform various operations leading to the output of the translation. They include;

- Part of Speech (POS) Tagger: It tags each word in the word-array to its corresponding part of speech, and then returns the pattern to the parser to rearrange.

- Parser: It contains different rules of parsing listed in a select case statement. The cases that were selected for parsing depend on the pattern of the part of speech (POS) arrangement. Once a case is selected, the necessary reordering to the target language will be done. To complete the process, the pattern will sent to a function that will map the POS to their equivalent *Yorùbá* meaning.

- Lexicon: The lexicon contains a list of words which can be used to form the basic sentences to be used by the system. It is a collection of both English and Standard *Yorùbá* words.

VI. RESULTS AND DISCUSSION

From our discussion in the methodology, the aim of the research work is to develop a system and not to investigate or compare a system. Therefore, mean opinion score is use to test the outcome of the system. The program testing follows three distinct stages:

a) ***Unit Testing:*** *This involved the testing of the system at each stage of development to ensure it performs expected. This is basically done by the developer and several others who assisted the developer in testing the system.*

b) ***Integration Testing:*** *This involved the testing of each module of the system to ensure proper interaction among them. This is done by the system developer.*

c) ***System Testing:*** *This involves testing the entire system on different PC's in other to meet the requirement of user and to make sure that the application is working correctly as expected.*

The mean opinion score (MOS) was recorded after several attempt by users, other researchers in the field. The result shows that the system scored 4 (Very Good) (i.e rating from 1-bad to 5-excellent). Figure 9 shows the GUI of the system while figure 10 and 11 give examples of translation from the system and from Google translator.

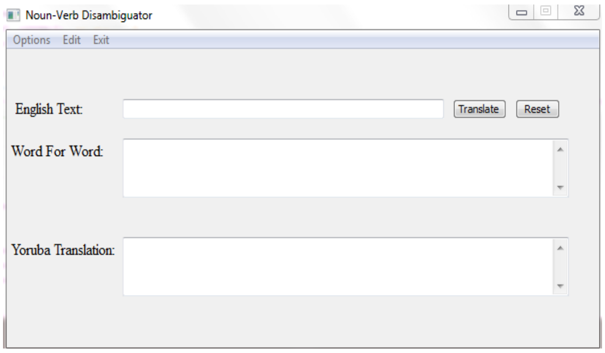
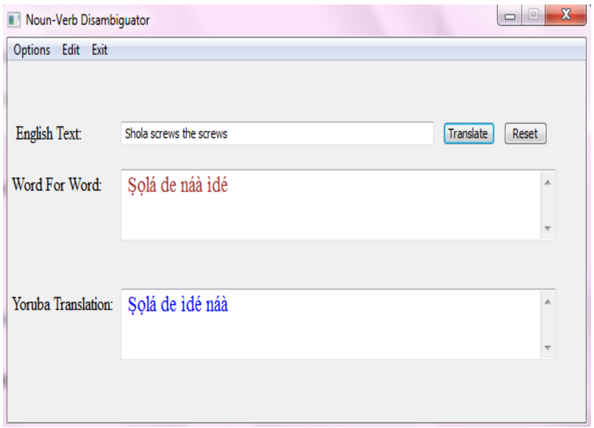


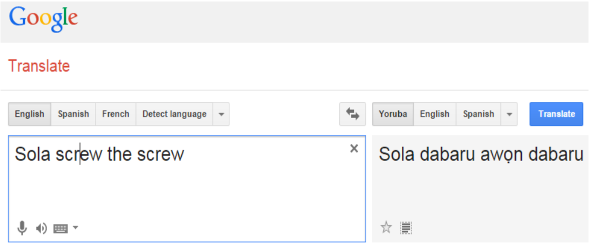
Fig. 9. Graphical User Interface (GUI) of the system



English Text: Shola screws the screws

Word For Word: Ṣọlá de náà ìdé

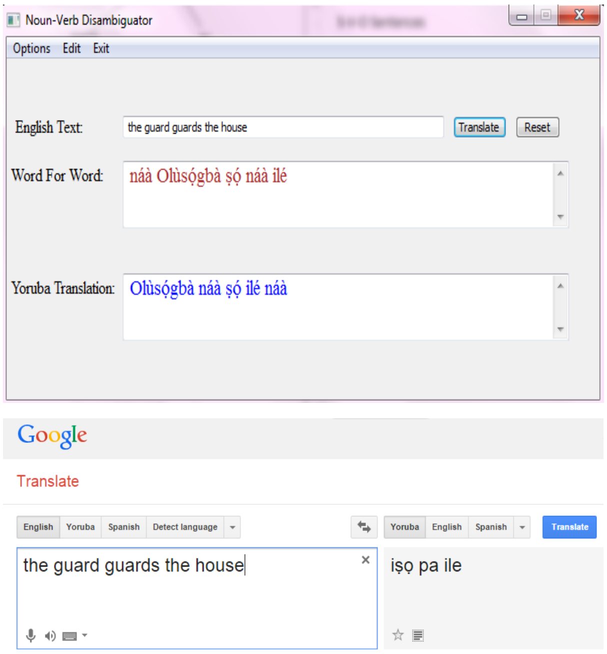
Yoruba Translation: Ṣọlá de ìdé náà



Sola screw the screw

Sola dabaru awọn dabaru

Fig. 10. Sample Output of the System and Google Translator



English Text: the guard guards the house

Word For Word: náà Olùsọ́gbà ṣọ́ náà ilé

Yoruba Translation: Olùsọ́gbà náà ṣọ́ ilé náà

the guard guards the house

iṣọ pa ile

Fig. 11. Sample Output of the System and Google Translator

VII. CONCLUSION

The need to achieve computational ability of indigenous languages is a progressive one and it is expedient. Therefore, a lot of research is ongoing towards the development a fully functional English to *Yorùbá* Machine Translation System. According to [5], improvement to the translation process can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Thus, The Word Disambiguation System developed will greatly help in contributing towards the ongoing research conducted on developing efficient English to *Yorùbá* Machine Translation by reducing the level of ambiguity during translation of SVO sentences.

Other area of note for further research is an improvement on the system to extend beyond basic SVO sentences.

REFERENCES

[1] Chan Y.S. and Ng H.T. (2005). Scaling Up Word Sense Disambiguation via Parallel Texts. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005), pages 1037–1042, Pittsburgh, Pennsylvania, USA.

[2] Agirre E. and Edmonds P. (2006). Word Sense Disambiguation. Algorithms and Applications. Text, Speech and Language Technology. Springer, Dordrecht.

[3] Ide N., Erjavec T. and Tufis D. (2002). Sense discriminationwith parallel corpora. . In ACL-2002 Workhop on Word Sense Disambiguation: Recent Successes and Future Directions, pages 54–60, Philadelphia.

[4] Adegbola, T.; Owolabi, K. and Odejobi O.A. (2011), "Localising for Yoruba: Experience, Challenges and Future Direction. Proceedings of Conference on Human Language Technology for Development, Alexandria, Egypt. Bibliotheca Alexandrina, pp. 7-10.

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

[6] Chéragui, M.A. (2012). "Theoretical Overview of Machine Translation". Proceedings ICWIT

[7] Diab M. and Resnik P. (2002). An Unsupervised Method for Word Sense Tagging Using Parallel Corpora. In Proceedings of ACL, pages 255–262.

[8] Hutchins W.J. and Somers H.L. (1992). "An Introduction to Machine Translation" Academic Press, London.

[9] Hutchins W.J. (1995). "Machine Translation: A Brief History." In *Concise history of the language sciences: from the Sumerians to the cognitivists*. Edited by E.F.K. Koerner and R.E. Asher. Oxford: Pergamon Press, 1995. Pages 431-445.

[10] Hoste V., Hendrickx I., Daelemans W., and van den Bosch A. (2002). Parameter Optimization for Machine-Learning of Word Sense Disambiguation. Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems, 8:311–325.

[11] Tufis D., Ion R., and Ide N. (2004). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pages 1312–1318, Geneva, Switzerland, August. Association for Computational Linguistics.

[12] van Gompel M. (2010). UvT-WSD1: A Cross-Lingual Word Sense Disambiguation System. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 238–241, Uppsala, Sweden. Association for Computational Linguistics.