DOI: 10.36108/ujees/3202.50.0210

UNIOSUN Journal of Engineering and Environmental Sciences. Vol. 5 No. 1. March. 2023

DEVELOPMENT OF A XML-ENCODED MACHINE-READABLE DICTIONARY FOR YORUBA WORD SENSE DISAMBIGUATION

Adegoke-Elijah, A., Jimoh, K. and Alabi, A

Abstract: The development of the disambiguation component of a Yorùbá to English machine translation system is hindered by several factors. One of these is the lack of machine readable sense inventory for ambiguous Yorùbá words. This study addressed the problem by developing a machine readable dictionary for ambiguous Yoruba verbs. To achieve this, ambiguous Yorba verbs and their translations were collected from existing bilingual dictionaries. The collected lexicons were transformed into machine readable format using the Extensible Markup Language (XML) Format. The accuracy of translation of the machine readable dictionary was evaluated using mean opinion score, with a score of 4.37 over the scale of 5. This study covered the total number of ninety-three (93) monosyllabic verbs with two hundred and forty-one (241) senses, which gives a coverage of 69.5% of the ambiguous monosyllabic verbs in Yoruba Language. The sense inventory was also used as a component of a Yoruba Word Sense Disambiguation system, and an accuracy of 94.6% was achieved. This study concludes that the digitized data can increase the accuracy of Word Sense Disambiguation component of a Yorùbá to English machine translation system.

Keywords: Dictionary, Disambiguation, Machine-readable, Sense Inventory, Yoruba Language, Extensible-Mark-up Language, XML

I. Introduction

Human languages are inherently ambiguous such that a given word may have more than one meaning [1], this is referred to as lexical ambiguity. Although human beings possess the innate ability to automatically determine the intended meaning of an ambiguous word within a given context, this task remains arduous for a machine. Word Sense Disambiguation (WSD) is an important task in human language technology [2].

A WSD system determines the contextual meaning of an ambiguous word using computational approach [3]. Such systems are therefore expected to possess essential components required to identify the intended sense of an ambiguous word. These components are knowledge source, a classification algorithm and a sense inventory.

A knowledge source can be described as the component that provides the needed knowledge or rules needed by the machine for the disambiguation task; examples of this are corpora, dictionaries, ontologies etc. A classification algorithm can be described as a model which makes use of the knowledge gained from a given source and some feature(s) derived within the context of an ambiguous word to determine its meaning, examples of this are Naïve Bayes algorithm, Lesk algorithm etc.

Adegoke-Elijah, A
(Redeemer's University, Ede, Nigeria)
Jimoh, K
(Osun State University, Oshogbo, Nigeria)
Alabi, A
(Oduduwa University, Ipetumodu, Nigeria)
Corresponding Author:
Email: adegoke-elijaha@run.edu.ng

A sense inventory is the component which clearly defines all the possible meanings of an ambiguous word, examples of lexical resources that do this are dictionaries, thesaurus and other lexical databases.

A typical WSD system is made of three components as shown in Figure 1

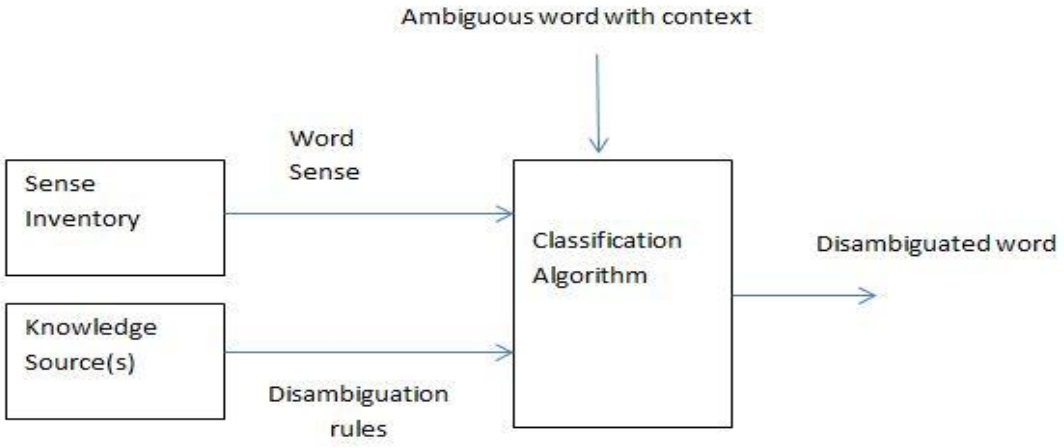

Figure 1: A Typical Sense Disambiguation System.

Several studies have been made in the past to develop sense inventory for different languages, especially English. These are mainly in form of dictionaries that enumerate the possible meanings that a word can possess. These dictionaries are usually written by Lexicographers whose main aim of writing the dictionaries is to examine all the senses of a word from the psycholinguistic view, and not primarily for natural language processing. These dictionaries could be in two forms: paper-based dictionaries and machine readable dictionaries. The paper –based form uses the primitive approach of listing word entries and their senses in a textual format on the pages a book. While this approach is available for many languages in the world, paper –based dictionaries are not directly useful for language engineering because computational machines cannot read directly from them.

Machine readable dictionaries are presented in electronic format, and can be read directly by a machine. Longman Dictionary of Contemporary English (LDOCE) [4] is a middle sized and machine- readable dictionary comprising of 45,000 word entries, with 65,000 meanings. This is made up of 23,800 noun entries with 37,500 senses, 7,921 verbs with 15831 senses and 6,922 adjectives with 11,371 senses. Each of the entries contains information such as definition, usage examples, grammatical information, usage labels in the forms of codes and comments, subject field code indicating the field of interest in which a meaning is related and semantic codes which either classify nominal meanings or express selectional restrictions for the complementation of verbal and adjectival meanings. The unique code system is formed using usage codes, subject field codes and semantic codes. Some of the early researches that made use of LDOCE for word sense disambiguation are [5] and [6]. Wordnet [7] is a hand coded online lexical database developed at Princeton University. In

Wordnet, words and collocations are grouped in Synset. Synset are group of words considered to be synonymous. Wordnet contains 155,287 words arranged into 117,659 synsets. The synsets are made up of 82,115 nouns, 13,767 verbs, 18,156 adjectives and 3,621 adverbs. According to [7], two expressions are synonymous in a given linguistic context C, if the truth value of the context is not altered when one is substituted for the other. This definition therefore necessitates that only words from the same lexical category can form synsets. Each of the wordnet synsets is related to other synsets by means of semantic relations. The semantic relations are antonymy, hyperonymy, hyponymy, meronym, tryponymy and entailment. At present, Wordnet is the the most utilized and best known resource for disambiguating English Language word senses [3] [8]. It can therefore be considered as the de-facto standard for English WSD. One example of such, is a noun disambiguation system reported in [9]. Following the success of application of Wordnet in English, new versions of the resource have been created for other languages such as Dutch, Italian, Spanish, French and German. All these exist within the EuroWordNet framework. Despite the wide acceptability of WordNet, the most cited limitation is its fine-grainedness of its sense distinction, which has limited its performance for disambiguation of word senses. The sense distinction, are often well beyond what could be needed in many language processing applications [10]. To address this most cited problem in [7], [11] did a study that reduced the granularity by developing a coarse gain sense inventory of WORDNET, and thus improve the accuracy of existing WSD algorithms above 80%.

Apart from the use of sense inventory to address linguistic problems, [12] developed a specialized inventory for the ambiguity resolution of Acronyms in Spanish Clinical documents. The study applied Schwartz and Heartz Algorithm to Spanish Subset of MEDLINE to create sense inventory that is made of 51 Clinical Specialties, with a total of 3603 acronyms.

Yoruba is the main language being spoken in the south-western parts of Nigeria [13], and the uniqueness of the Yoruba people has received a lot of attention from different research in the past decades [14]. Despite the usefulness of a sense inventory for the computational disambiguation of ambiguous Yorùbá words, this tool is presently unavailable. The lack of standard sense inventory for the Yorùbá language constitutes a major challenge to *Yorùbá* language engineering, especially the development of a Yoruba Sense Disambiguator and in extension, machine translator, because according to [15], WSD is essential in machine translations.

The aim of this study is do develop a machine readable dictionary containing Yoruba monosyllabic verbs and their possible translations in English Language. The developed tool will serve as a sense inventory for a Yorùbá word sense disambiguation system.

II. Materials and Methods

This section discusses the methods and the tools used in this study. These are divided into three major components data collection, data preparation and its digitalization. Data collection process involves the collation of words used in this stud. The data preparation section involves the transformation of the collected data into format suitable for the subsequent processing, while data digitization

process involves converting the text into machine readable. The workflow for the proposed system is shown in Figure 2.

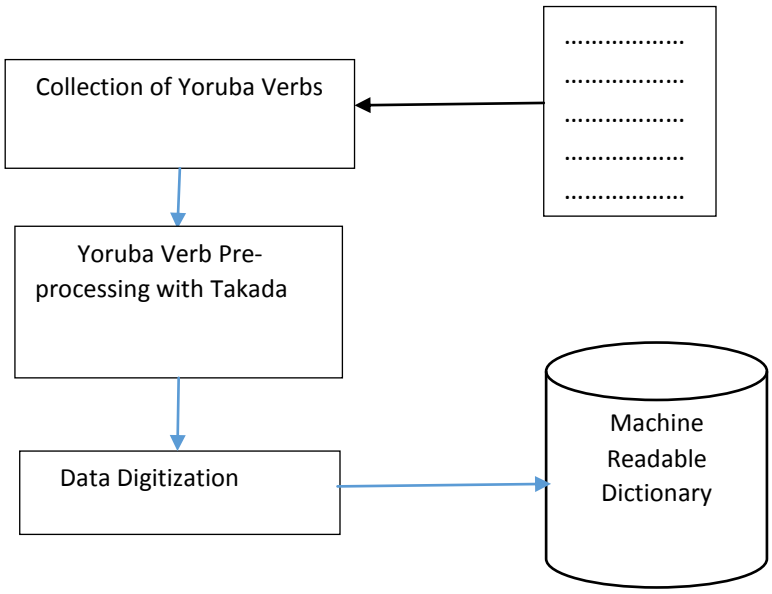

Figure 2: Process Flow for the Proposed System

Each of these processes is discussed below.

i. Data collection

Ambiguous monosyllabic Yorùbá words were collected from [16] and [17]. [16] is a pocket size non-readable bilingual dictionary. The dictionary translates words from English to Yorùbá language and vice-versa. The Yorùbá to English section, which is of interest to this study, is made of about 11,500 words which belong to different grammatical class. The grammatical classes listed are noun, adjective, adverb, pronoun, conjunction, transitive and intransitive verbs and preposition. The dictionary contains English language definitions of the listed Yorùbá words. While there are few words that are monosemous, most of the entries are polysemous. Apart from the definitions for each word and senses, the entry also contains usage examples for each of the definitions of the words. [17] is also a non-machine readable bilingual dictionary for Yorùbá words. The work addressed only monosyllabic Yorùbá words. Just like [16], the dictionary translates words from Yorùbá to English language and vice-versa. However, his study did not tag any of the words with the respective parts of speech. Another obvious difference between [16] and [17] is the word sense distinctions. While [17] used fine-grained sense distinction, [16] only addressed the coarse grain sense of the entries. For example, while [17] enumerated fourteen possible senses for *bà*, [16] classified the senses into six, with four senses representing the verbal meaning and two senses denoting the prepositional sense of the word.

A total number of ninety-three (93) ambiguous verbs were collected. Each of the verbs is defined using three fields. These are: the possible number of senses which gives the possible number of English translations for each Yoruba word, the translation field which

enumerates the possible translations, and the case field which gives example of how each of the senses of the word can be used in meaningful sentences. For example, ro is a Yorùbá word that has three possible number of sense. The possible translations are *weed*, *drip* and *pain*. Each of the senses is mapped to example sentences *Ade ro oko*, *Omi ro si ile* and *Apa n ro mi* respectively.

ii. Data preparation

Apart from orthography, every syllable in a Yorùbá work can be distinguished by means of its tone, and the combination of one or more syllable produces words. Diacritical marks are used for two main purposes in Yorùbá language. Firstly, the use of under-dot diacritic comes with some Yorùbá alphabets such as *s*, *e*, and *o*. Secondly, they are used to specify the tone used in the pronunciation of a given syllable. Any syllable marked with an acute sign is pronounced with a high tone, the syllables marked with grave sign are pronounced with low tone. Lack of any of signs on a syllable shows the syllable is pronounced with mid tone. Despite the importance of diacritics in the language, the standard keyboards contain only alphabets without the needed diacritics. The need to pre-process the data by adding necessary marks therefore arises. In this study, *takada* software was used. The software has a graphical user interface which provides a text area for entering text, and appending the texts with appropriate diacritical marks through the click of the appropriate buttons on the toolbar. With Takada, the collected texts were therefore pre-processed with the appropriate diacritics. For example, the usage sentences described in section (ii) were pre-processed into *Adé ro oko, Omi ro sí ilẹ and Apá ń ro mí.*

iii. Data Digitization

Data digitalization is the process of translating data from paper-based model, which is not in machine readable format, to a machine readable format. One of the ways of achieving this is through the use of Xtensible Markup Language (XML). A document type definition (DTD) describes the acceptable building blocks of an XML documents. It defines the document structure with a list of legal elements and attributes. XML elements are the building blocks of an XML document; they behave as containers which hold texts, elements, attributes, media objects or all of these. An XML document contains one or more elements, with the scope defined by start and end tags. An attribute defines a property of the element. It associates a name with a value, which is a string of character. The DTD for the lexical database under consideration is shown in Figure 3 and this is followed by the description of the elements in the XML document.

- **Lexical database tag**

The tag represents the root element of the lexical database. It has the attribute *category*, which indicates whether the resource is monolingual, bilingual or multilingual. In this case, the value of the category attribute is bilingual. The *creation-date* attribute indicates the date of creation of the database. The *encoding* attribute indicates the method of encoding the languages under consideration. In the case of this database, the value of *unicode* is chosen due to the diacritics present in the source language.

```
<! DOCTYPE Dictionary
[
<!ELEMENT dICTIONARY(Lenguage, Entry)>
<!ELEMENT Language>(sourcelanguage, targetlanguage)>
<!ELEMENT ENTRY(WORD)>
<!ELEMENT WORD (Sense)>
<!ELEMENT Sense(Case,translation)>
<!ELEMENT sourcelanguage (#PCDATA)>
<!ELEMENT targetlanguage(#PCDATA)>
<!ELEMENT case(PCDATA)>
<!ELEMENT translation(#PCDATA)>
<ATTLIST ENTRY entryid CDATA #REQUIRED>
<ATTLIST WORD senses CDATA #REQUIRED>
<ATTLIST SENSE senseid CDATA #REQUIRED>

]>
```

Figure 3. Document Type Definition for the Sense Inventory.

The attribute indicates the title of the lexical resources which has the value Yorùbá -to-English-verb-translation database.

- **Language Tag**

This is the element that specifies the language under consideration. It contains the element source-language and target-language. The source language is Yorùbá while the target language is English.

- **Entry Tag**

This is the element that defines each of the lexical items covered in the database. It has the attribute *entry-id* which assigns numeric identification number to each of the words to be described.

- **Word Tag**

This is the element within the entry tag that specifies the word to be described. It has the attribute *sense* which specifies the number of possible translation of the word.

- **Sense Tag**

The sense tag defines each of the possible translations of the Yorùbá word using the translation and case tags. The sense tag has *sense-id* and *POS* attributes. The *sense-id* associates a unique number to each of the senses of the word under consideration. The *POS* attribute specifies the part of the speech of that sense of the word. The *translation* tag specifies the English translation of the sense, while the *case* tag gives usage example of the sense of the

```
<entry entryid="001"> <word senses="4">bá</word>
<sense senseid="1" POS="V"> <translation>Help</translation>
< case > Ó bá mi sé </case> </sense >
< sense senseid="2" POS="V" >   < translation > Befall < /translation >
< case > Ibi bá wọn < /case > < /sense >
< sense senseid="3" POS="V" >  < translation > Catch up with</translation>
< case > Bọ́lá bá Tọ́pẹ́ < /case > </sense >
< sense senseid="4" POS="V" >
< translation > Hit < /translation >
< case > òkúta náà bá mi< /case > < /sense >
< entry >

<entry entryid="002" >    <word senses="3" > bà < /word >
< sense senseid="1" POS="V" > < translation > Ferment < /translation >
<case> Kòkó náà ti bà < /case >
< /sense >  < sense senseid="2" POS="V" > <translation > Perched < /translation >
< case > ẹiyẹ náà bà sí ori Igi < /case > < /sense     </entry>

<entry entryid="003" >< word senses="2" > bẹ́< /word >
<sense senseid="1" POS="V" ><translation > Jump < /translation >
<case>ọmọ náà bẹ́ < /case >  </sense>
<sense senseid="2" POS="V" > <translation >Burst </translation>
<case> ọrá náà ti bẹ́ </case> </sense>     </entry >

<entry entryid="004" > < word senses="2"> bẹ </word >
< sense senseid="1" POS="V" >   <translation> Peel </translation >
<case> Bọ́lá  bẹ isu  </case >
</sense >  <sense senseid="2" POS="V" >  <translation> Too forward < /translation >
<case> ọmọ náà bẹ < /case > </sense ></entry >

<entry entryid="005" >  <word senses="2" > bì </word >
<sense senseid="1" POS="V" >  <translation > Vomit < /translation >
<case> Bọ́lá bì  </case >  </sense >
<sense senseid="2" POS="V" >  <translation > Change colour </translation >
<case > Aṣọ náà ń bì</case > </sense ></entry >
```

Figure 4. Sample Entries from the Sense Inventory

III. Results and Discussion

i. Evaluation and Result

To evaluate the accuracy of the dictionary, a graphical user interface was developed using Tkinter library of the Python Programming Language. The translation accuracy, which refers to the ability of the system to correctly translate the Yoruba words to English language, was tested using Mean Opinion Score (MOS) on the scale of 1-5 (1-Poor,2-Fair, 3-Average, 4-Good, 5- Excellent). Questionnaires were given to ten (10) native Yoruba language speakers, who have also acquired English Language as their second language. The study considered subjecting the respondent to manually evaluate all the 93 verbs covered in this study to be daunting, and therefore opted to take few entries which could be estimated to be a good representation of all the entries. To select the entries to be evaluated, the first entry was selected, the next three entries skipped; then the fifth entry is selected while entries 6, 7 and 8 are skipped. This process is continued until the last entry of the dictionary was reached. Using this method, a total of twenty-four (24) verbs were selected for the evaluation. The selected verbs and their number of possible of senses are shown in the Table 1 below:

Table 2 shows the grading of the translation accuracy of the developed machine readable dictionary by the ten (10) respondents, referred to as User1-User10. All the users graded the translation accuracy of the 23 selected verbs, and the average of all the grades assigned to each of the selected verbs by each user was computed to calculate each user's overall evaluation of the system. For example, User 1, 2, 3 and 4 graded the system 4.29,4.29,4.38,4.46 respectively. The average of the grades of all the user gives us the Mean Opinion Score of the System, which is 4.37 on a scale of 1-5.

Entry No	Word	No of Senses
1	*Bá*	3
5	*Bì*	2
9	*Dá*	3
13	*dùn*	2
17	*Fún*	3
21	*Gbé*	2
25	*Há*	3
29	*Jí*	2
33	*kán*	2
37	*Kú*	2
41	*Lé*	3
45	*Ná*	2
49	*pẹ́*	2
53	*Rá*	2
57	*rẹ*	2
61	*rọ*	4
65	*Sá*	5
69	*Sí*	3
73	*Sú*	2
77	*tẹ́*	3
81	*Tu*	3
85	*Yá*	3
89	*Yé*	2
93	*yún*	2

The developed machine readable dictionary provides a rich resource for Yoruba Language Engineering. The tool was used in [18], for the resolution of translation ambiguity in a Yoruba to English machine translation system. It was used in conjunction with an ontology, and the accuracy of 94.6% was achieved. The flow chart showing how the machine readable dictionary was used for the disambiguation process is shown in Figure 5.

Table 1: System Evaluation

Entry No	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10
1	5	4	5	5	4	5	4	5	4	4
5	5	4	5	5	4	4	5	5	4	5
9	4	5	4	5	5	4	4	5	5	4
13	5	5	4	5	4	4	5	5	4	4
17	3	4	4	3	4	4	4	3	4	4
21	4	5	4	4	5	5	4	5	4	5
25	5	4	5	5	5	4	5	4	3	4
29	4	3	4	5	4	5	5	4	4	5
33	5	4	3	5	4	5	4	5	4	5
37	4	5	4	3	4	5	4	5	4	3
41	4	3	5	4	5	5	5	3	5	4
45	4	5	4	4	5	4	3	4	3	4
49	4	4	5	5	4	3	5	5	4	4
53	5	4	5	5	4	5	4	5	4	5
57	4	5	4	5	4	5	5	4	5	4
61	5	4	5	4	4	5	5	4	5	4
65	4	5	4	5	4	3	5	4	5	4
69	4	5	3	4	4	5	5	4	4	4
73	3	4	4	4	5	4	5	5	5	4
77	4	5	5	4	4	5	5	4	5	5
81	5	4	5	4	5	4	5	5	4	4
85	3	4	5	5	4	4	5	5	5	4
89	5	4	4	5	4	5	4	5	5	4
93	5	4	5	4	4	5	4	5	4	3
Average Score	4.29	4.29	4.38	4.46	4.29	4.46	4.54	4.5	4.29	4.17
Rounded Mean Opinion Score	4.29	4.29	4.38	4.46	4.29	4.46	4.54	4.5	4.29	4.17
	4.37									

Using the method described above, not less than ninety-three (93) monosyllabic verbs were covered in this study, with each of the verbs having between two to five senses. This gives a total of three hundred and forth-one senses. According to [12], the total number of ambiguous verbs in Yoruba Language is one hundred and forty-one. This study therefore covers 69.5% of the verbs.

ii. Discussion of Results

The coverage of this study is low when compared with other studies that made use of corpora. For example, [19] developed an

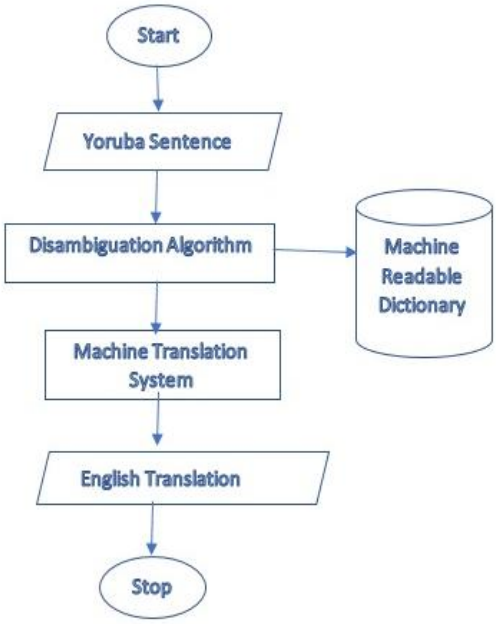

Figure 5: The use of the Machine Readable Dictionary for WSD

English-Macedonian machine readable dictionary consisting of 23,296 translation pairs (17,708 English and 18, 343 Macedonian terms) using parallel corpora and open source statistical machine translation tools. A subset of the produced dictionary was manually evaluated and showed accuracy of 79.8%. [20] developed GLAWI which is a large XML-encoded machine readable dictionary automatically extracted from Wiktionnaire. Wiktionnaire is a French-based edition of Wiktionary. GLAWI contains 1, 341, 410 articles. Yoruba is a resource-scarce language that lacks basic text processing resources, and training corpora which are essential for the language processing. This includes parallel corpora which could aid the translation of word pairs between languages, and facilitate the development of a bilingual dictionary. This lack of relevant resources contributed to the limitation of this study. Further studies aimed at developing a large-scale Yoruba to English machine readable dictionary containing thousands of word pairs, using corpora and other possible tools, will therefore be considered in future studies.

IV. Conclusion

Dictionaries are one of the most useful lexical resources. This is because they provide robust lexical and semantic information to assist natural language processing application. More specifically, word sense disambiguation systems rely on the features of the words found in the context of an ambiguous word. Yoruba Verb sense disambiguation depends on the semantic information of the arguments found in the context of the ambiguous verb for its ambiguity resolution. This study has developed a tool capable of aiding the resolution of lexical ambiguity of Yoruba verbs in the context of a machine translation system.

References

1. Navigli, R. "Natural Language Understanding: Instructions for (present and future use" *In proceeding of International Joint Conference on Artificial Intelligence,* Stockholm, Sweden, July 13-19, 2018 pp. 5697-5702.
2. Piantadoshi, S, Tily H. and Gibson E "The Communicative Function of Ambiguity in Language". *Cognition*, Vol 122, Number 3, 2012 pp. 280-291.
3. Navigli, R. (2009) "Word Sense Disambiguation: A survey" *ACM Computing Surveys*, Vol. 41, Number 2, 2009. pp. 10-64.
4. Proctor, P. *Longman Dictionary of Contemporary English.* Longman Group Ltd, Harlow, Essex, England, 1978.
5. Cowie, J, Guthrie, J, & Guthrie, L. "Lexical Disambiguation using Simulated Annealing" *In Proceedings of the 14th conference on*

Computational Linguistics- Nantes, France, August 23-28, 1992, pp 359- 365.

6. Lesk., M. "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a pine cone from an ice-cream cone" *In Proceedings of the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada, June 1, 1986. pp. 24-26.

7. Miller G., Beckwith, R, Fellbaum, C, Gross, D & Miller, J. "Introduction to Wordnet: An online Lexical Database.*" International Journal of Lexicography*, Vol. 3, Number 4, 1990. pp. 235-244.

8. Abd-Rashid, A, Abdul-Rahman, S, Yusof, N. N & Mohamed, A "Word Sense Disambiguation using Fuzzy Semantic Bases String Similarity Model", *Malaysian Journal of Computing*-Vol. 3, Number 2, 2018, pp. 154-161.

9. Buscadi, D, Rosso P. & Masulli F. "The upv-unige-CIAOSENSO WSD System". *In Proceedings of Senseval-3 , The Third International Workshop On The Evaluation Of Systems For The Semantic Analysis*, July, 2004, pp. 77-82.

10. Ide, N and Veronis, J "Introduction to the Special Issue on Word Sense Disambiguation: the State of Art." *Computational Linguistics*, Vol. 24, 1998, pp 2-40.

11. Lacerra, C., Bevilacqua, M., Pasini, T., & Navigli, R. "CSI: A coarse sense inventory for 85% word sense disambiguation" In *Proceedings of the AAAI conference on artificial intelligence* Volume 34, Number 5, April 3, 2020, pp. 8123-8130.

12. Pomares-Quimbaya, A., López-Úbeda, P., Oleynik, M., & Schulz, S. " Leveraging PubMed to create a Specialty-based Sense Inventory for Spanish Acronym Resolution". In *Digital Personalized Health and Medicine*, 2020, pp. 292-296.

13. Oladele M., Adepoju T. , Olatoke O. & Ojo, O. (2020) "Offline Yorùbá Handwritten Word Recognition Using Geometric Feature Extraction And Support Vector Machine Classifier" *Malaysian Journal of Computing*, Voume 5, Number 2, pp. 504-514.

14. Aina, S, Sholesi K, Lawal , A., Okegbile, S & Oluwaranti, A "Gesture Recognition System For Nigerian Tribal Greeting Postures Using Support Vector Machine" *Malaysian Journal of Computing,* Volume 5, Number 2, pp. 609-618.

15. Pu, X., Pappas, N., Henderson, J., & Popescu-Belis, A. "Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation." *Transactions of the Association for Computational Linguistics*, Volume 6, 2018, pp. 635-649.

16. University Press Plc, *A Dictionary of the Yorùbá language*. University Press, Ìbàdàn, 2011.

17. Adéwọlé. L. *A Bilingualized Dictionary Of Yorùbá Monosyllabic Words*. Montem Paperbacks, Akure. 2005

18. Adégòkè-Elijah, A. "Design and implementation of lexical ambiguity resolution for a standard Yorùbá-to-English machine translation system" *Unpublished PhD. thesis*, Obafemi Awolowo University, Ile-Ife, Nigeria, 2018.

19. Saveski, M & Trajkovski., I. (2010) "Development of an English-Macedonian Machine Readable Dictionary by Using Parallel Corpora." *In International Conference on ICT Innovations*, Springer, Berlin, Heidelberg, September 12-15, 2010 , pp. 195-204.

20. Sajous, F and Nabil, H.. "GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary". *eLex,* Herstmonceux, United Kingdom. halshs-01191012, 2015, pp 1-23