Building Ontology for Yorùbá Language

Okediya Theresa
Department of Computer and Information Science
Covenant University
Ota, Nigeria
theresa.okediya@stu.cu.edu.ng

Afolabi Ibukun
Department of Computer and Information Science
Covenant University
Ota, Nigeria
ibukun.fatudimu@covenantuniversity.edu.ng

Iheanetu Olamma
Department of Computer and Information Science
Covenant University
Ota, Nigeria
olamma.iheanetu@covenantuniversity.edu.ng

Ojo Sunday O.
Inclusive African Indigenous Language Technology Institute
Pretoria, South Africa
prof.Sunday.ojo@afrilt.institute

Abstract

Natural Language Ontology (NLO) provides a formal specification of linguistic semantics knowledge implicit in a natural language. Such a NLO could facilitate a shared understanding of the linguistic semantics system of the language that enhances accuracy of language semantics modelling in Natural Language Processing (NLP). This paper presents the construction of a general purpose ontology for Yorùbá language, one of the under-resourced African languages.Taking as input popular Yorùbá terms obtained from online books, blogs, social websites, and Yorùbá dictionary, the Ontology was constructed, and a prototype implementation made, using the Protégé ontology development tool. Ontology validation and evaluation were done using an automated reasoner.It is emnivisaged that such Yorùbá language ontology will contribute to the development of digital resources for the language, towards its long-term preservation.

1 Introduction

Natural Language Ontology (NLO) provides a formal specification of the most basic categories and relations used in describing a natural language, with the aim of uncovering the ontological categories, notions, and semantic structures that are implicit in the use of the language. It facilitates a shared understanding of the linguistics semantics system of a natural language, and can serve as an input into language modelling to minimize reality-model semantic gap, in Natural Language Processing (NLP). It can also facilitate both the knowledge sharing of annotated linguistic data and the searching of disparate language corpora (Benaissa, Bouchiha, Zouaoui, & Doumi, 2015). Also, in specific terms, an African language, such as Yorùbá, is not only a mirror into the mind of the people group, but also a mirror into their culture and history. Just as they carry their history in their genes, so do they carry same in their language. Hence, the need for a Yorùbá NLO, such as proposed in this paper, is aimed at leveraging the digital development of the under-resourced language. This is towards rendering the language, not only a wider visibility, for its upliftment to academic and scientific status through sound linguistic research.

Due to the increase in the digital textual document, more works have been done and still ongoing to capture the large volume of information that comes from a variety of languages in which only a handful possess the Natural Language Processing (NLP) resources required for developing modern language technologies, researchers have in time made effort to represent different languages such as English, Arabic, French among others (Benaissa et al., 2015; Onyenwe, Hepple, Chinedu, & Ezeani,

2018) but most African languages are still very much under-resourced, one out of the numerous under-resourced languages is Yorùbá which is a language spoken by about thirty-three million people of the South-west, Nigeria(Olúmúyìwá & Aládésanmí, 2017). Yorùbá is believed to have originated from the Igala people about 2000 years ago(Afolabi, Daramola, & Adio, 2014). Out of the 36 states in Nigeria, nine are occupied by Yorùbás which are: Lagos, Ògùn, Ọ̀yọ́, Ọ̀ṣun, Òǹdó, Èkìtì, Kwara, Kogi and Edo States. Across these states, there are different dialects of the language. The dialects are subsumed into five major dialect areas namely: North-West Yorùbá(NWY), North-East Yorùbá(NEY), Central Yorùbá(CY), South-East Yorùbá(SEY) and South-West Yorùbá(SWY). Noteworthy is the fact that this language is spoken worldwide in other nations like Benin, Togo, Ghana, Cote d'Ivoire, Sudan, and Sierra-Leone. Speakers of this language are also in Brazil, Cuba, Haiti, the Caribbean Islands, Trinidad and Tobago, UK and America as well(Ayeomoni, 2012; Olúmúyìwá & Aládésanmí, 2017).

However, like any other African cultural heritage, the Yorùbá language is endangered in the face of inter-ethnic interaction, westernization, and globalization(Hassan, Odéjóbí, & Ògúnfolákàn, 2013). It is therefore of importance to have such a popular language well represented online. Ontology is used to handle information at a semantic level and also play a major role in the semantic web, with this technology, programs, and software agents have access to use the content resources available on the World Wide Web(Lakel & Bendella, 2015), thereby enhancing users' access to information. In view of this study having a well-defined ontology will improve natural language understanding, natural language processing and natural language generation of Yorùbá language.

The aim of this paper is to build a well-defined, lexical ontology for Yorùbá language to be used in NLP system. To achieve this there is need to acquire the knowledge necessary to create the ontology, to identify the concepts to represent, to represent these concepts as classes, to define the conceptual relations and to implement the ontology itself(Bautista-Zambrana, 2015).

The remaining part of this paper is arranged as follows: section 2 gives an overview of the language and related works. Section 3 describes the methodology. Section 4 presents the actual implementation of the work and Section 5 gives the conclusion and recommendation.

2 Related works

Ontologies are used to represent knowledge, an ontology can be used in different fields of knowledge. It can be domain bound which implies the ontology represents knowledge elicited from a specific domain. Different researchers have worked to develop ontologies for different purposes. A domain ontology was developed in (Afolabi et al., 2014) for Nigeria's history, a semi-automated approach was used, the ontology itself was implemented using Protégé software. Similarly,Dramé et al.(2014) proposed a method to construct a bilingual domain ontology, the method uses two approaches: learning ontology from text and reusing existing terminological resources. Rani, Dhar, & Vyas(2017) likewise proposed a modelby exploring two topic modelling algorithms for the purpose of determining the statistical relationship between document and terms and build a topic ontology and ontology graph with little human intervention. Even better, Kethavarapu & Saraswathi(2016) generated data from webpages to build a dynamic ontology using a similarity measure and ontology creation module to generate the Web Ontology Language(owl) file. Also, Alruqimi & Aknin(2019) presented an algorithm for deriving a domain-specific ontology from folksonomy tags, the algorithm takes a domain name as input and produces the corresponding domain ontology as output.

Ontology needs to be evaluated after been created, different evaluation methods have been used in literature, Raad & Cruz(2018) highlighted some evaluation methods which include Gold Standard-based, Corpus-based, Task-based, Criteria based, Structure-based and Complex and Expert based. Lakel & Bendella (2015) proposed a combined approach to improve the process of automatic co-construction of ontologies from a corpus. Expert approach was used to evaluate the ontology in (Dramé et al., 2014; Hassan et al., 2013),Alruqimi & Aknin (2019) used a corpus-based approach, Afolabi et al. (2014) combined the gold standard-based and task-based approach to evaluate the ontology created.

As opposed to conventional ontology, Lexical ontologies are "not based on a specific domain, but they are intended to provide structured

knowledge about lexical issues (words) of a language by linking them to their meanings" (Benaissa et al., 2015).Benaissa et al. (2015) modelled a lexical ontology after the WordNet ontology, the Arabic verb was used as input and Markov Clustering algorithm was used to identify similar verbs. Also,Ishkewy, Harb, & Farahat (2014) developed a software module called Azhary, which is a lexical ontology for Arabic language, the ontology was evaluated using the gold standard-based approach and Arabic WordNet was used as the gold standard. Ontologies have also been constructed for the different domain in Yorùbá, Hassan et al. (2013) described an engineering process of building an ontology for Yorùbá cultural heritage using Formal Concept Analysis for the design, the ontology was implemented with Protégé software and validated using domain experts and ontology experts approach. However there is no lexical ontology for this language yet, hence the reason for this work.

3 Research Framework

3.1 Requirements for the NLO

The major purpose of the ontology is to define the semantic relationship between words in Yorùbá language, which will make information retrieval, automatic text analysis easier and make Yorùbá language available and accessible for digital processing. The architecture of the system is shown in Figure 1. The major use cases of the ontology include:

Knowledge Driven Application: Software that requires knowledge represented in the ontology.

Users: a user interacts with the ontology through a Graphical user interface(GUI) by generating queries. Or a programmer that uses the ontology to create an application using any programming language of choice.

Domain Expert: the domain expert supply the relevant knowledge needed to construct the ontology through their documented materials.

3.2 Data Source

In the cause of this research so far, there was no standardized corpus found for Yorùbá language hence the use of different data sources. Some of the terms were gotten from the Yorùbá dictionary, however only few were selected for the reason that many are no more in use for everyday language. In addition words were retrieved from the internet. Yorùbá words site and some other blogs. The terms were downloaded, saved in Excel spreadsheet and input into the protégé software. The words in the corpus are Yoruba language words. In the language there are seven vowels: [a], [e], [ẹ], [i], [o] [ọ], [u] and four to five nasal vowels: an, ẹn, in, ọn and un. The language has 18 consonants: [b], [d], [f], [g], [gb], [h], [j], [k], [l], [m], [n], [p], [r], [s], [ṣ], [t], [w], and [y](Awoyale, 2008).

Mostly verbs (Ọrọ̀-ìṣe) in Yorùbá language are monosyllabic and monomorphemic examples are wa, lo, je, mu, mu gba and so on while nouns (Ọrọ̀-orúkọ) are polysyllabic and polymorphebic which most times use combinations of the monosyllabic/monomorphemic verbs as stems. Other part of speech represented in Yorùbá language they are: Àpèjúwe (Adjective), Àpólà Àpèjúwe (Adjectival Phrase), Àpọ́nlé (Adverb), Àpólà Àpọ́nlé (Adverbial Phrase), Atọ́kùn (Preposition), Àpólà Atọ́kùn (Prepositional Phrase), Àpólà Ọrọ̀-orúkọ (Noun Phrase), and Àpólà Ọrọ̀-ìṣe (Verb Phrase)

The language is essentially tone-driven which help to deal with Homographs. Take the word "igba" which can mean (plate, two hundred, time, garden egg) it also interesting that unlike some languages, the context of use may not necessarily be used to detect the meaning of a word, take the sentence:

Mu igba wa:
"Bring the plate"
"Bring two hundred"
"Bring the garden egg"

There are three distinct tones used in the language: low, mid and high. Only low (marked with a grave sign) and high (marked with an acute sign) tones are marked on top of the vowel, while the mid tone is left unmarked. "igba" in the sentence above when toned low has only one meaning:

Mu igba wa: "Bring two hundred"

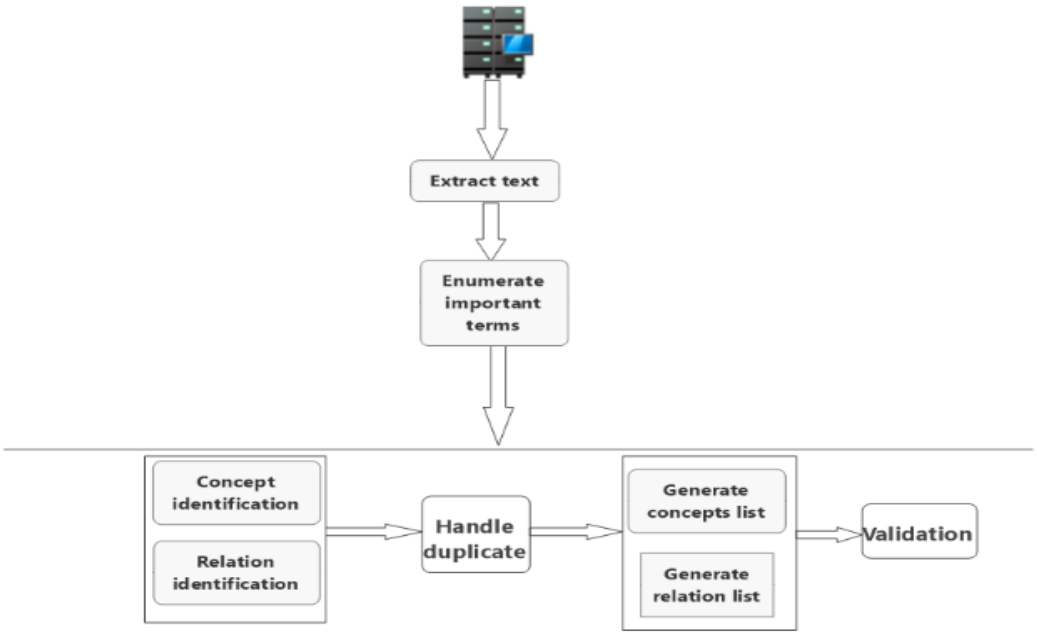

Figure 1: Architecture of the Ontology

3.3 Word and Relation Extraction

The lexical entry to this ontology is the Yorùbá language part of speech, some words can have more than one entries if they have morphological variants such as plural of nouns and inflected form of verb(Staab & Studer, 2009). For example:The verb "wa" which is "come" can have other entries which will point to it as root word, those words won't exist as complete or separate individual, the words include: owa(there is), owa(he came), ewa(telling an elderly person to come, wonwa(they came).

The lexical entries can relate to one another through the following ways(Ishkewy et al., 2014):

- Synonym: B is a synonym of A, if A and B has the same meaning.
- Hypernym: B is a hypernym of A, if A is a (kind of) B.
- Hyponym: B is a hyponym of A, if B is a (kind of) A
- Meronym: B is a meronym of A, if B is a (part of) A
- Holonym: B is a holonym of A, if A is a (part of) B
- Antonym: B is an antonym of A, if A is an (inverse) of B.

3.5 Ontology Building

An ontology development usually encompasses several tasks and **Erreur ! Source du renvoi introuvable.** shows the task in order. Four stages are relevant to the construction of the Yorùbá language ontology, the first is to extract text from different sources as earlier stated in section 1.0, second stage is to identify the concepts and their relations, third phase is to handle duplicates, the exact duplicates(Hassan et al., 2013) are automatically blocked by Protégé while the quasi-exact duplicates and implicit duplicates were manually handled. Finally, validation is done to check for the consistency of the ontology after duplicates have been removed that is to check whether or not all of the statements and definitions in the ontology are mutually consistent. This is achieved using the HermiT reasoner tool in Protégé

4 Implementation and Results

4.1 Protégé OWL Implementation

The ontology implementation was done using Protégé 4.2. There are different ontology languages with different facilities, XML, RDF, RDF(S), OWL and more. However, OWL offers better advantages over others, aside from being the most recent development in standard ontology

languages it also has a richer set of operators - e.g. intersection, union, and negation. It is based on a different logical model which makes it possible for concepts to be defined as well as described.

4.2 Class and Relations description

The concepts were identified from the sources earlier stated, relations were defined across these concepts, and the concepts were arranged hierarchically in a top-down manner as shownin **Erreur ! Source du renvoi introuvable.**that is a more general concept first followed by subclasses. Polysemy deals with relatedness in meaning while Homonymy deals with unrelatedness in meaning. The example below shows homonymous relationship (Babarinde, 2018):

(a) Adé **pa** okùn – 'Ade **sets** rope trap'
(b) Adé **pa** àlọ́ - 'Ade **gives** riddles'
(c) Adé **pa** ìtàn - 'Ade **narrates** a story'
(d) Adé **pa** irọ́ - 'Ade **tells** lies'

Individuals in classes can be related to each other as shown in Figure 2.

4.3 Yorùbá Ontology Validation and Evaluation

According to (Raad & Cruz, 2018), Ontology evaluation is a problem of assessing a given ontology from different perspectives such as accuracy, completeness, conciseness, adaptability, clarity, computational efficiency and consistency. Any evaluation method uses any combination of the criteria earlier listed.

The ontology is compared with Azary, an Arabic
lexical ontology in **Erreur ! Source du renvoi introuvable.**. The ontology constructed was validated using an automated reasoner called HerMiT in Protégé. A reasoner considers the following criteria to assess the performance of an ontology; consistency, satisfiability, and subsumption.

Table 1: NLO and Azhary lexicon

Lexicon	Azhary	YLO
Synonyms relation	Yes	Yes
Hyponym relation	Yes	Yes
Hypernym relation	Yes	Yes
Meronymreation	Yes	Yes
Antonym	Yes	Yes
Happens-before relation	No	Yes
Polysemy	No	Yes
Homonymy	No	Yes

There are different reasoners used to check for the consistency of an ontology but HerMiT does not just determine the consistency of an ontology but can also identify hierarchical relationships between the classes, and much more. The methodology it uses is the hypertableau calculus and it provides the faster way of ontology classification.(Abburu, 2012).
Below is the overall working of the reasoner:

Input : Yorùbá Language Ontology(YLO)

Step1: IF ∃ Model_of_YLO THEN goto step2
ELSE
State = inconsistent

Step2: FOR EACH A in YLO DO
IF ∃ Model_of_YLO SUCH THAT x belongs to A
State = satisfiable
Step3: ∀ class A and B in YLO
Check: IF A IsIn B THEN
State = subsumption

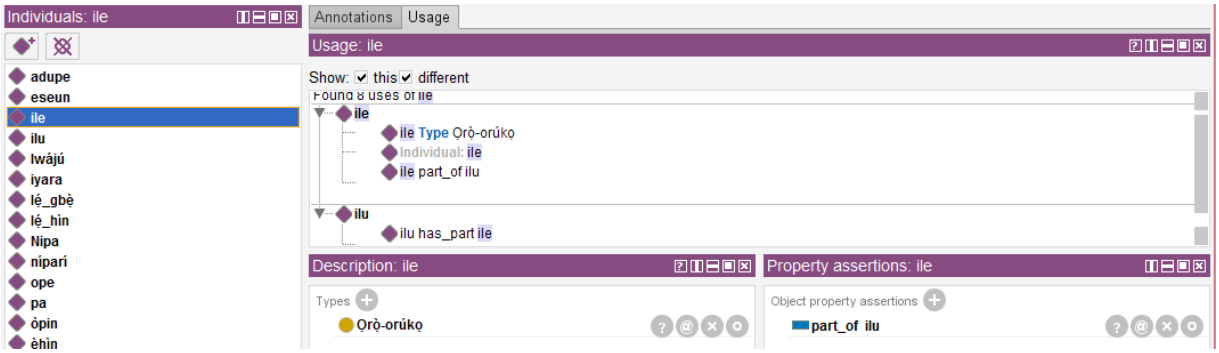

Figure 2: An excerpt of the ontology relative to Homonymy

5 Conclusion and Limitations

The chances for Yorùbá semantic analysis is little since there is no Yorùbá lexical ontology for linguist researchers to depend on, therefore this paper presented the construction of a lexical ontology for the Yorùbá language, using a description logic reasoner the validity of the ontology was tested. The primary use is in automatic text analysis and artificial intelligence applications, it will also support advancement of Natural Language Understanding, Processing and Generation. Moreover, it will make the Yorùbá Language available and accessible for digital processing and sustain the Yorùbá culture in the face of technological advancement. There was no available and well defined corpus for Yorùbá language found so far in the cause of this research which limited the accuracy and consistency of the terms used, also some lexical entries were seen as duplicates because they have the same form as existing ones, this reduced entries.

References

Abburu, S. (2012). A Survey on Ontology Reasoners and Comparison. *International Journal of Computer Applications*, *57*(17), 33–39. https://doi.org/10.5120/9208-3748

Afolabi, I., Daramola, O., & Adio, T. (2014). *Developing Domain Ontology for Nigerian History*. *8*(April), 30–39.

Alruqimi, M., & Aknin, N. (2019). Bridging the Gap between the Social and Semantic Web: Extracting domain-specific ontology from folksonomy. *Journal of King Saud University - Computer and Information Sciences*, *31*(1), 15–21. https://doi.org/https://doi.org/10.1016/j.jksuci.2017.10.005

Amar, F. B. Ben, Gargouri, B., & Hamadou, A. Ben. (2016). Generating core domain ontologies from normalized dictionaries. *Engineering Applications of Artificial Intelligence*, *51*, 230–241. https://doi.org/https://doi.org/10.1016/j.engappai.2016.01.014

Awoyale, Y. (2008). *Global Yoruba Lexical Database v. 1.0*. 1–49.

Ayeomoni, O. M. (2012). A Lexico-syntactic Comparative Analysis of Ondo and Ikale Dialects of the Yoruba Language. *Theory and Practice in Language Studies*, *2*(9), 1802–1810. https://doi.org/10.4304/tpls.2.9.1802-1810

Babarinde, O. (2018). Lexical Ambiguity in Yoruba: its Implications for Second Language Learners. *Journal of Languages, Linguistics and Literary Studies (JOLLS)*, *5*(June), 264–272. Retrieved from http://www.jolls.com.ng

Bautista-Zambrana, M. R. (2015). Creating Corpus-based Ontologies: A Proposal for Preparatory Work. *Procedia - Social and Behavioral Sciences*, *212*, 159–165. https://doi.org/https://doi.org/10.1016/j.sbspro.2015.11.314

Benaissa, B., Bouchiha, D., Zouaoui, A., & Doumi, N. (2015). Building Ontology from Texts. *Procedia Computer Science*, *73*(2015), 7–15. https://doi.org/https://doi.org/10.1016/j.procs.2015.12.042

Bermejo, J. (2007). A simplified guide to create an ontology. *Madrid University*, *DRAFT*, 1–12. Retrieved from http://tierra.aslab.upm.es/documents/controlled/ASLAB-R-2007-004.pdf

Dramé, K., Diallo, G., Delva, F., Dartigues, J. F., Mouillet, E., Salamon, R., & Mougin, F. (2014). Reuse of termino-ontological resources and text corpora for building a multilingual domain

ontology: An application to Alzheimer's disease. *Journal of Biomedical Informatics*, *48*, 171–182. https://doi.org/https://doi.org/10.1016/j.jbi.2013.12.013

Hassan, J. A., Odéjóbí, O. A., & Òǵúnfolákàn, B. A. (2013). *Ontology Engineering in Yorùbá Cultural Heritage Domain*. *6*(5).

Ishkewy, H., Harb, H., & Farahat, H. (2014). Azhary: An Arabic Lexical Ontology. In *International journal of Web & Semantic Technology* (Vol. 5). https://doi.org/10.5121/ijwest.2014.5405

Kethavarapu, U. P. K., & Saraswathi, S. (2016). Concept Based Dynamic Ontology Creation for Job Recommendation System. *Procedia Computer Science*, *85*, 915–921. https://doi.org/https://doi.org/10.1016/j.procs.2016.05.282

Lakel, K., & Bendella, F. (2015). Dynamic Evaluation of Ontologies. *Procedia Computer Science*, *73*, 16–23. https://doi.org/https://doi.org/10.1016/j.procs.2015.12.043

Olúmúyìwá, T., & Aládésanmí, Ọmọ́bọ́lá Agnes. (2017). Written Literature in an African Language: An Examination of Interrogative Sentences in Fágúnwà's Novels. *Journal of Siberian Federal University. Humanities & Social Sciences*, *9*(12), 3025–3036. https://doi.org/10.17516/1997-1370-2016-9-12-3025-3036

Onyenwe, I. E., Hepple, M., Chinedu, U., & Ezeani, I. (2018). A Basic Language Resource Kit Implementation for the Igbo NLP Project . *ACM Transactions on Asian and Low-Resource Language Information Processing*, *17*(2), 1–23. https://doi.org/10.1145/3146387

Raad, J., & Cruz, C. (2018). *A Survey on Ontology Evaluation Methods*.

Rani, M., Dhar, A. K., & Vyas, O. P. (2017). Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, *63*, 108–125. https://doi.org/https://doi.org/10.1016/j.engappai.2017.05.006

Staab, S., & Studer, R. (2009). Handbook on Ontologies, Second edition (International Handbooks on Information Systems) 2009. In *International Handbooks on Information Systems*.