A COMPUTERIZED IDENTIFICATION SYSTEM FOR VERB SORTING AND ARRANGEMENT IN A NATURAL LANGUAGE: CASE STUDY OF THE NIGERIAN YORUBA LANGUAGE

Enikuomehin A. Oluwatoyin
Dept. of Computer Science
Lagos State University, Lagos Nigeria

ABSTRACT: *The context of Understanding has continued to be a major attraction to researchers in Natural Language Processing. This is built on the theory that language can be used effectively if it is understood and can be analyzed and as such, most Natural Language Processing research tend towards the belief that the human brain has a section dedicated for language analysis and understanding therefore, human ambiguity which, remains the major difference between natural and computer languages, can be modeled using appropriate man machine modeling tools since programming languages are designed to be unambiguous, that is, they can be defined by a grammar that produces a unique parse for each sentence in the language. The paper evaluates the classification process for a Natural language 'the Yoruba language' and presents a new method by which the language can be transformed into a computer understandable language using its morphological identification framework. Result shows that the approach is admissibly in line with known benchmarks. The paper recommends that non tonal language can also be experimented using the defined approach.*

Keyword: Morphology, Natural language, Yoruba, Model, Tonal Language

INTRODUCTION

Natural Language understanding continues to be the major concern in research on natural language. This is compounded with efforts in automating the generation of such language. Natural Language generation and natural language understanding are key areas in the domain of natural language processing (NLP) and recent research has included areas like computational linguistics, bilingual transformation amongst others. These are subsets of the larger research area coined Artificial Intelligence (AI). They are aimed at making the computer performs many tasks as maybe done by human. NLP crosses across Computer sciences, Information science, Economic Intelligence, Artificial intelligence and Linguistic with themajor concern being developing systems that can generate interaction between computer and human. Natural language processing techniques have been useful in generating automated tools for language understanding and processing in ways in which the human elements in human communication can be understood.

Obviously, for an appropriate automated system to be built, the developer must understand how human uses knowledge in acquiring, storing and manipulating languages. These include the understanding of the language itself and the understanding of the language constituents. This is necessary because Natural language processing deals broadly with the understanding, analysis, manipulation and/or generation of Natural language in way usable for the computer. NLP research tends towards the belief that the human brain has a section dedicated for language analysis and understanding. Subsequently, this has been given as reasons why we understand that a child cannot learn a language over a little or limited number of time. Learning is a computer based process and this shows that such process is similar to the complexity of the human brain. Furthermore, research on brain injuries has shown that patients with brain problems and injury have challenges with speech and language understanding. This is notably common with those with problem on the left hemisphere. Recall that the brain is divided into two groups, the left and right hemisphere, the left hemisphere, generally called the logical brain is involved in language analysis and understanding. In 92% of the people, language is represented in the left hemisphere [1] [2]. Natural Language processing tools allows the computer to communicate with human using everyday language. There is no difference between NLP and computational linguistics other than the fact that the latter is concerned with how computational methods can be used to aid the understanding of human languages. This paper is majorly concerned with the processes in computational linguistic since language rely on people's ability to use their knowledge and inference ability to properly resolve ambiguities. Ambiguity remains the major difference between natural and computer languages, because programming languages are designed to be unambiguous, that is, they can be defined by a grammar that produces a unique parse for each sentence in the language. In this paper, we evaluate the classification process for a Natural language 'the Yoruba language' and present a method by which the language can be transformed into a computer understandable language using its morphological identification framework.

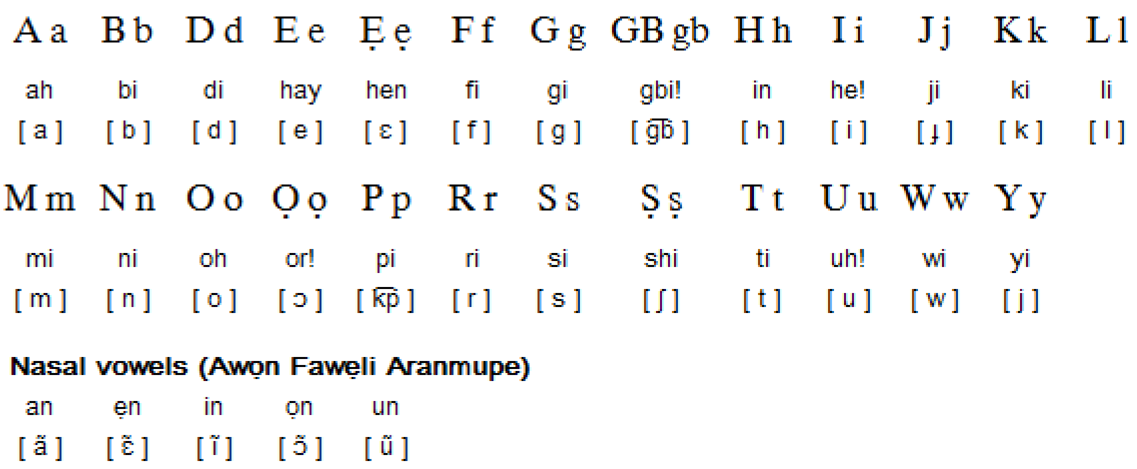
THE YORUBA LANGUAGE

Yorùbá is a tonal language spoken natively by about thirty million people in Nigeria and in the neighboring countries of the Republic of Benin and Togo. In Nigeria, Yorùbá speakers reside in the Southwest region in states such as Oyo, Ogun, Osun, Ondo, Ekiti, Lagos, Kogi and Kwara states. Yorùbá is a Kwa language, which belongs to the Yoruboid group under the Niger-Congo phylum. It has three basic but significant tones [3]. One of the effects of the large number of Yorùbá speakers and their geographic spread is the emergence of geography-bound linguistic variations. Yorùbá is a dialect continuum including several distinct dialects [4] Estimates of the total number of Yorùbá dialects vary from twelve to twenty-six [5]. The differences inherent in these dialects are marked in the areas of pronunciation, grammatical structure and vocabulary. There are other dialects found all over West Africa. In the Republic of Benin, Yorùbá dialects include Ketu, Nago, Ije, Ajase, Idaitsa, Tsabe; while Ana and Itsa are two of the dialects found in Togo. Some Yorùbá dialects are also found in the African Diaspora, especially the Caribbean. The dialect of Yorùbá used in Brazil is called Nago, while the one used in Cuba is referred to as Lucumi. It is however possible to classify Yorùbá dialectal forms, found in Nigeria, into five regional groupings: North-West Yorùbá (NWY); North Eastern Yorùbá (NEY); Central Yorùbá [2]; South-West Yorùbá (SWY); South-East Yorùbá (SEY). Phonological, lexical and grammatical variations are the hallmarks of these groupings since there are variant degrees of mutual intelligibility among the 'geographic' dialects found in each group. A consensus standard form has however evolved and is recognized as the form for writing and teaching the language. This form, relatively close to the SWY, is understood by speakers of all the different dialects and it continues to serve the communicative purpose of all speakers.

In the 1960s through the 1970s, various orthography committees were set up, by both government and academic groups, to consider and subsequently review the standard orthography for the language. Significant reviews were done based on the report of the orthography committee in 1966. It is primarily the basis for the creation and introduction into schools of the standard Yorùbá orthography and hence for the standard Yorùbá language. The standard form of Yorùbá is the type of Yorùbá learned at school, and spoken (or written) mostly by educated native speakers to addressees who speak different dialects [6]. The language is one of the three national languages in Nigeria. It is an honor due, in part, to the socio-political importance and sheer number of its speakers.

This has translated into socio-linguistic developments for the language in terms of its study and usage in the country. As with many other African languages, the earliest study of the Yorùbá language was done by missionaries interested in translating the scriptures for evangelical purposes. The peculiar outcome of these efforts (coupled with the abolition of slavery) was the emergence of the writing and studying of the Yorùbá language and culture among settled free slaves in Sierra Leone. Known as Aku, these Yorùbá people did pioneering work on the writing and studying of the language, such that Yorùbá became one of the first West African languages to have a written grammar and dictionary in 1849 [7]. Since then, work has continued in and on the language until today. There are many grammars, dictionaries and literary texts in the language today. Since words are formed by combination of alphabets, we present the Yoruba alphabet in the following section.

The Yoruba alphabet



A a	B b	D d	E e	Ẹ ẹ	F f	G g	GB gb	H h	I i	J j	K k	L l
ah	bi	di	hay	hen	fi	gi	gbi!	in	he!	ji	ki	li
[a]	[b]	[d]	[e]	[ɛ]	[f]	[g]	[g͡b]	[h]	[i]	[ɟ]	[k]	[l]

M m	N n	O o	Ọ ọ	P p	R r	S s	Ṣ ṣ	T t	U u	W w	Y y
mi	ni	oh	or!	pi	ri	si	shi	ti	uh!	wi	yi
[m]	[n]	[o]	[ɔ]	[k͡p]	[r]	[s]	[ʃ]	[t]	[u]	[w]	[j]

Nasal vowels (Awọn Fawẹli Aranmupe)

an	ẹn	in	ọn	un
[ã]	[ɛ̃]	[ĩ]	[ɔ̃]	[ũ]

THE MORPHOLOGICAL ANALYSIS OF THE YORUBA LANGUAGE

The Yoruba language is a semi–agglutinative language. That is, a language in which words are made up of linearly sequential morphemes with component representing a meaningful morpheme. Agglutination is a process in linguistic morphology derivation in which complex words are formed from strings of morphemes with each having a single grammatical or semantic meaning. Mostly, the word order is subject-object-verb, though because the language is verse, other word order applies.

The order of the language can be represented in the following tree diagram:

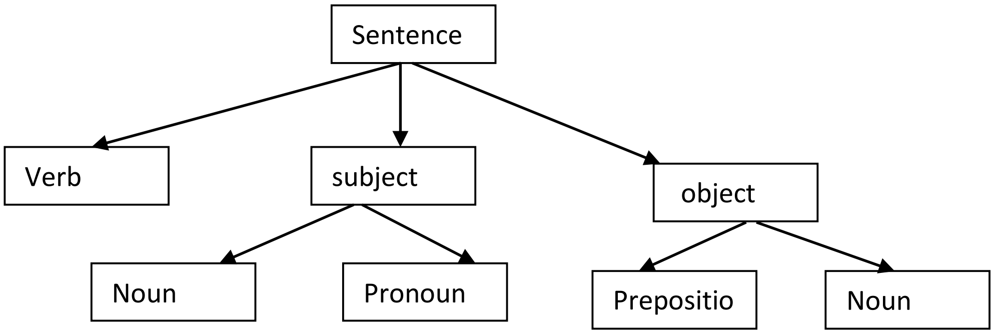


Figure 1: Word order VSO

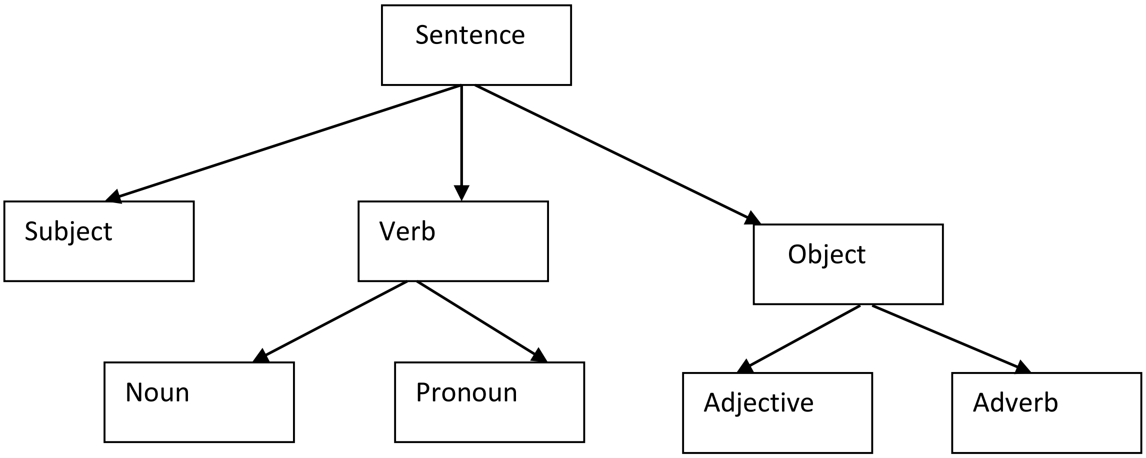


Figure 2: Word and SVO (As in o je ewa . the bought beans) [8]

Morphological analysis is an essential part of Natural language processing which is very crucial to language understanding and computerization. It primarily concerns with the process of segmenting words into their component morphemes and the assignment of grammatical information to grammatical categories. The Yoruba language is rich in the context of building morphologies. Suffix and prefix additions are acceptable in the language. These tools are used in the process of developing of a morphological analyzer. Methods such as root driven, the suffix stripping are also used frequently in the buildup. In NLP, there are two basic morphological tools: the morphological analyzer and the morphological generator. The analyzer will return its root/stem word along with its grammatical information depending upon its word category. Because Yoruba is a tonal language, it has a local morpheme per word ratio which makes it an isolating language, in which case, a single or compound word gives a complete sentence. Consider the Yoruba words. Ra: (buy), ra: (rotten), ra: (disappear), ra(roast/grill) etc, thus low morphemes or two or three letters or few letters can make a complete sentence with a meaning. Such subject verb combinations also exist aside the earlier discussed SVO, example O ra baba, (he gives praises/ o ra baba (He bought capper) example cited from [9]. The development of a morpheme analyzer has evolved through different approaches[10],[11]. Thus, Yoruba, being a language that can permit several ranges of morphological changes, partially due to the wealth of morphemes permissible in the language, will not permit the use of Brute force method for appropriate morpheme generation. A better approach for the generation of morpheme is the use of suffix stripping algorithm. Suffix stripping is a major algorithm used in stemming, particularly necessary in the area of information retrieval and linguistic morphology. Suffix stripping is much simpler to maintain than the Brute force algorithm. Morphology is not equally prominent in all spoken languages, what one language expressed morphologically may be expressed by a separate word or left implicit in another language. For example, English expresses plural nouns by means of morphology but Yoruba uses separate word expressing the same meaning[12] in English boys, in Yoruba awon okunrin. However, languages such as English, Vietramese, Yoruba has morphology playing a major role and are thus called analytic. Yoruba language permits both suffix and prefix stripping. This can be justified by the Bible below: (See Bible attached)

Application:

Morphological Analyzer is a software component that is capable of detecting morphemes in a piece of text. In English, text area usually processsed before it is indexed and one of the common pre-possessing step in stemming. Morphological tools will only be appropriate only when the necessary stemming has taken place. [13] shows that, there are not majorly known algorithms available for stemming products however many stemming algorithms that exist are language based. The approach in this paper is to use an algorithm that implements statistical natural language processing to learn about the syntactic structure of the language in terms of morphemes (morphosyntactic structure) and use the derived knowledge to detect morphemes. This, unlike many other work in which an algorithm is recommended for a language, this paper considers that the best algorithm can only be efficient when the system has the capability to learn the morpheme structure of the natural language using statistical means.

Morphemes and Root words

In any language, morphology provides the first step for identifying any word as applicable in the dictionary, this is called the base, root and stem word. The word appears without infections on it. Then the infections can now be used to identify the actual meaning in relation to the context in which the word in used. In the Yoruba language, several words have more than one meaning in speech making. This is an elements of tonal languages, such that the tone determines what the speaker implies; that is ra (buy) and ra (disappear). We implement the approach by developing a tokenization algorithm for this purpose. The choice of tokenization strategy is attributed to the fact that tokenization is the first step for language processing task ranging from part of speech tagging, machine translation, to information retrieval and extraction. It is simpler for inflectional languages like English and Yoruba where two word with space can make the same meaning unlike languages like thai, Lao where space is usually used after character. Various tokenization exist for various language .

The algorithm presented for this work follows from the expectation functions presented as follows:

$$F(M1, M2\ldots..Mn) = \prod_{i-i}^{n} p(mi)$$

Mi = i-th morpheme in a morpheme string,

P(Mi) = the frequency of Mi in the corpus

The algorithm for the tokenization usable for morpheme identification in Yoruba is given as;

1. Find all morphemes that matches the beginning of the string from a list of all free morphemes
2. On the complexion of step(1), the isolation data are used to store all possible combinations of input string, number of tokens and the identification in each combination.
3. If no match is found, strip a single character from the input string and store as array
4. Choose one with least number of token
5. Then concatenate each single character with previous token
6. Check if the meaning has been affected
7. If affected, reverse to 3 then stop else complete run.

The above algorithm is a functional algorithm that tokenizes and supports the processes of prefixation and suffixation. Suffixation is commonly used in infectious language. For the purpose of suffixation, the algorithm will be:

1. Reverse token list order when prefixes are fixed
2. On suffix list, generate another list containing identified token
3. Then search for match and generate a concatenated form
4. Rearrange the token list order

Similarly, a statistical pattern for conditional random field can be used as an undirected graphical models trained to maximize the probability first introduced by Lattery [14] such as natural language processing with observations W = $w_1, w_2, \ldots\ldots w_n$ and Y = $y_1, y_2 \ldots\ldots\ldots y_n$, we generate an Error function using the form:

$$WER = \frac{(\text{Subs} + \text{Dels} + \text{Ins}) * 100\,\%}{No.Of\ \ words\ incorrect\ \ sentence}$$

where Subs – No.of wrong words substituted in the word corpus.
Dels – No. of correct words deleted in the word corpus
Ins – No. of new words inserted in the word corpus

The perplexity values evaluated for the test data were high for the politics corpus than the news corpus. This was due to the small vocabulary size and high OOV words in the politics corpus, when compared to the news corpus. The number of unigrams, bigrams and trigrams were high for both the corpora. The performance of the speech recognition system was further enhanced by the use of distance based language model, dependency- based language model, class-based language model and enhanced morpheme-based language model as discussed in the next section.

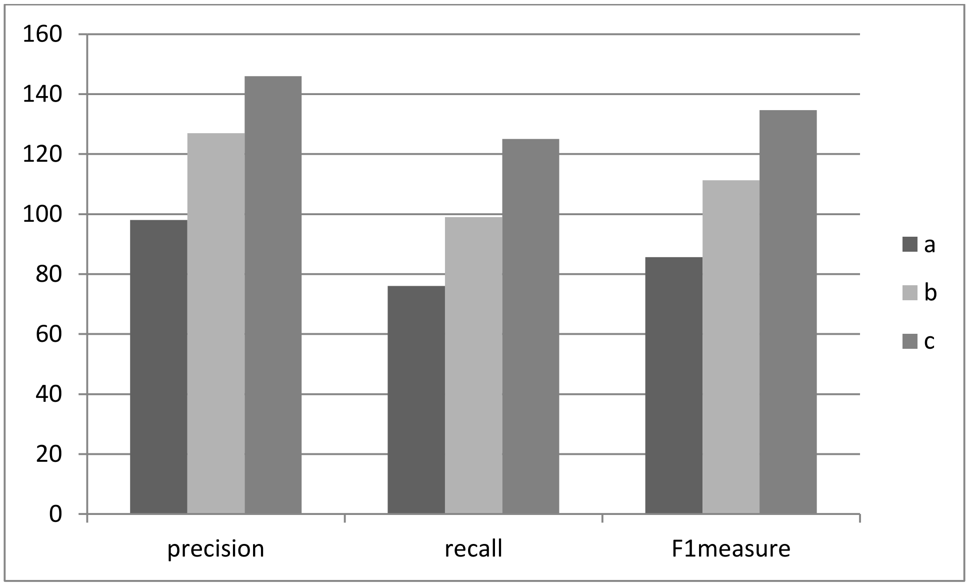
RESULTS

On applying the proposed algorithms stated above, the result obtained over 2,400 Yoruba words is shown below:

Table I. Scoring Table

	Model		
		WER(%)	**Morph Score(2.4k)**
1	Word	25.11	2400
2	MP	24.33	1299
3	Morph	24.26	1291
4	MP no AMB	20.3	2186
5	Lexical word	19.2	2182

The precision, recall and F1 score for suffix stripping can be given in the following diagram:



These results is considered with benchmark for other languages such and LAO and CHAO. Further research in this domain may require the consideration of space reduction models for an enhanced Information Retrieval System, as example, awon okunrin, meaning boys, a retrieval system will distinguish “awon” as separate from ‘okunrin’ whereas in English such words are just pluralized as boys or in some cases men from man. Precision is taken in this study as the amount of tokenized item that are correct and returned by the total amount of tokenized object returned. Recall is taken as number of corrects items returned by the total number of items in the corpus.
F1 – Score = 2* precision * Recalls/(Precision + Recall).

Morphological analyzers can be integrated into the system of language understanding with several application in the Natural Language Processing subsector. Yoruba is a SVD or VSO word order language and computationally, each root word can be generated from the inflectional verb. Yoruba is not too rich inflectionally however several inflections will at most case, change the meaning of the word.

CONCLUSION

Morphological synthesizer of the Yoruba language can be implemented by algorithm as shown above. The paper discusses the complexity of tonal languages in term of computation and meaning generation as a feature common with semi – agglutinative language. In this paper, we implemented an algorithm for developing a morphological synthesizer and the result obtained shows that the method adopted is satisfactory and the research is a further contribution in the development of automated language learning via a system for automatic language understanding. Using a hybrid method to suffix stripping processes in the development of appropriate synthesizer proved to be an efficient method for identifying the morphological categories of a given Yoruba word. A sample of the implemented in JAVA program is also demonstrated.

REFERENCE

1. Dennis, M. and H.A. Whitaker, *Language acquisition following hemidecortication: Linguistic superiority of the left over the right hemisphere.* Brain and language, 1976. **3**(3): p. 404-433.
2. Cohen, L., et al., *Visual word recognition in the left and right hemispheres: anatomical and functional correlates of peripheral alexias.* Cerebral cortex, 2003. **13**(12): p. 1313-1333.
3. Ojo, A. *A global evaluation of the teaching and learning of Yorùbá language as a second or foreign language*. in *Selected Proceedings of the 36th Annual Conference on African Linguistics*. 2006.
4. Bamgboṣe, A., *A short Yoruba grammar*. 1968: Heinemann.
5. Olúránkinse, O., *Euphemism as a Yor [ugrave] bà folkway.* African Languages and Cultures, 1992. **5**(2): p. 189-202.
6. Oyetade, S.O., *A sociolinguistic analysis of address forms in Yoruba.* Language in society, 1995. **24**(04): p. 515-535.
7. Okolo, B.A., *The History of Nigerian Linguistics A Preliminary Survey.* 1981.
8. Adedjouma Sèmiyou, A., J.O. Aoga, and M.A. Igue, *Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family.* Res. Journal of Computer & IT Sciences, 2013. **1**(1): p. 2-5.
9. Allen, A., *FOOD AND CULTURE: continuity and change in the Yoruba of West Africa and their diasporas.*
10. Pesetsky, D., *Complementizer-trace phenomena and the nominative island condition.* The Linguistic Review, 1982. **1**(3): p. 297-343.
11. Posetsky, D., *Russian morphology and lexical theory.* 1979.
12. DELLA TOSCANA, I.C., *II-DAI CASTRA TARDOANTICHI AI CASTELLI DEL SECOLO X.*

13. Finkel, R. and G. Stump, *Principal parts and morphological typology*. Morphology, 2007. **17**(1): p. 39-75.

14. Lafferty, J., A. McCallum, and F.C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. 2001.