NEW FINDINGS IN WEST AFRICAN LANGUAGES AND LITERATURE

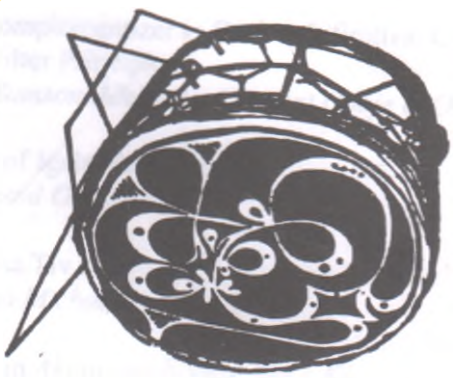
In commemoration of the ***50th Anniversary of the West African Linguistic Society (WALS)***

Oye Taiwo and *Lendzemo Constantine Yuka*

NEW FINDINGS IN WEST AFRICAN LANGUAGES AND LITERATURE

In commemoration of the
50th Anniversary of
the West African Linguistic Society (WALS/SLAO)

Oye Taiwo
&
Lendzemo Constantine Yuka



West African Linguistic Society (WALS)
La Société de Linguistique de L'afique de L'ouest (SLAO)

Published by ***the West African Linguistic Society (WALS)/***
La Société de Linguistique de L'afique de L'ouest (SLAO)

ISSN: 0022-5401

First Published 2016

Cover: Map of African language families and some major African languages
.www.nationaonline.org

Table of Contents

32

Rule-Based Machine Translation: An Interface between Formal and Natural Language Syntax

Clement Odoje

Abstract
The principles which govern ways words can be combined together to form phrases and sentences in natural language is known as syntax while formal syntax is not a matter of experience (unlike natural language), but stipulations in order to provide a specified set of strings in a computer programming language. The focus of this paper therefore, is to explore linguistics as the dual planes of theory and practice, by interrogating how PROLOG was used to capture English/Yorùbá natural language syntax in a rule-based machine translation. The study reveals that the machine was able to generate sentences, break sentences into phrases and words in a bid to translate them in both languages.
Keywords: *PROLOG, Natural Language, Formal Language, Syntax, Machine Translation*

Introduction

The study of linguistics as the scientific study of language has gone beyond its core or general study. Owolabi (2006:6) explains that as a result of the overlapping interests between linguistics and other disciplines, new (sub) branches of linguistics have emerged such as computational linguistics, neurolinguistics, clinical linguistics to mention but a few. This study is subsumed under computational linguistics.

Computational Linguistics is a discipline that spans Linguistics and Computer Science. It is concerned with the computational models of human cognition (Uszkoreit, 2000:1). Meaning that, a computer will be able to capture native speakers' intuitive/ tacit knowledge of language, so that a person can dialogue with a computer system. McGuigan (2003:1) explains that the main concern of computational linguistics is to become better at the automation aspects of translation, generation, speech and comprehension. Thus, McGuigan's views have attracted research interest such as:

- Speech recognition
- Speech synthesis
- Automated grammar correction system
- Machine translation
- Speech interpretation and so on

The activities of computational linguistics began about five decades ago in US and Europe but very new and recent in Nigeria. Students of Linguistics and Computer Science are currently exploring the discipline.

Machine Translation

Machine translation (MT) is a complete mechanization process, that is, a mechanical system without the intervention of either a pre- or post-editor. The outputs of MT must be satisfactory with regard to both semantic accuracy and intelligibility (Reifler 1954:1). Hutchins (2001:5), quoting Holmström (1951), explains further the meaning and the limitation of the machine thus:

> The application of natural language processing[61] to the field of translation is known as machine translation or MT. Although the importance of developing accurate and efficient MT software is most obvious in more linguistically diverse areas of the world than the United States, MT development is a well-researched technology both here in the U.S. and

[61] Natural Language Processing, Computational Linguistics and Human Language Technology have the same focus though called differently from different disciplines.

> abroad. For the layperson, it may seem that computers would be well-suited to the problems involved in translating a text from one language to another. Unfortunately, this is not the case. Because human languages are largely dependent on human experience and intuition, it has proven nearly impossible to emulate human language capabilities in a machine.

While Awobuluyi (2010) sees MT in line with Holmström, he is also of the view that MT is one of the significant contributions of linguistics to technology and ICT. He said:

> ...another operation which researchers would like computer to be able to perform. That operation is known as machine translation, and as its name implies, it involves getting computers to translate well-formed and fully idiomatic written expressions in one language into well-formed and equally idiomatic corresponding expression in another language... computer's judgment on human language may be patently incorrect which shows how difficult it is for now to get machine to accurately replicate all mental operation that human beings perform apparently without much effort. (Awobuluyi 2010: 34-35)

Regardless of the high demand for translation and its needs, the field of MT has barely changed since its emergence in the fifties (Hutchins 2001:1). This reinforces the fact that MT is difficult and challenging; however, it is also a rewarding exercise which is advancing slowly but surely.

As mentioned earlier, the activities of developing MT in Africa and Nigeria in particular is of recent. Only very few Yorùbá-English Rule-Based MT is available and no Statistical Machine Translation (SMT) is available in the language pair yet because there are no enough corpus for its development. SMT relies on large volume of parallel translated materials of the language pairs before close to acceptable translation could be achieved. There are very few translated materials for African languages except Arabic language whose usage goes beyond Africa. This may be because African languages are said to be resource-scarce languages from technological point of view (see Pauw, Wagacha and Schryver, 2011), meaning that African languages are languages that have small or economically disadvantaged user base which are typically ignored by the commercial world (Chan and Rosenfeld 2012). African Languages resource scarcity may inform why there are more Rule-Based MT than other forms of MT in Africa.

MT Approach Adopted

There are three main approaches to MT: dictionary based MT, rule based MT and corpus based MT. Tripathi and Sarlchel (2010) explain that dictionary based approach uses entries of language dictionaries whereby dictionary equivalent words are used for translated verse. They report that the first generation of MT in the 1940s to mid 1960s was entirely based on machine readable or electronic dictionaries. This approach may translate word for word and some phrases but will not translate sentences. There is more to sentence translation than word substitution. Most of the other translation approaches utilize bilingual dictionaries with grammar rules.

Rule-Based approach incorporates linguistics rules to translation algorithms. Aside from millions of dictionaries for the language pair; morphological, syntactic and semantic information about the source and the target language is logically adapted to computer algorithm for translation. While Corpus-Based approach is referred to as empirical approach to MT by Khalilov (2009). Since 1989, corpus approach for machine translation has emerged one of the widely explored areas in machine translation. Because of high level of accuracy achieved during the translation this method has dominated over other approaches. Some of the methods of this approach are: statistical machine translation, example based machine translation, and context based machine translation (Tripathi and Sarlchel 2010).

This paper adopts rule-based approach to MT because the researcher is interested in both formal and natural language syntax. Although earlier researchers have based their adoption of the

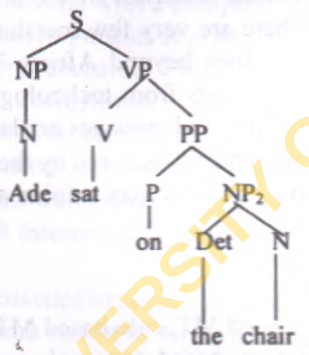
approach on the scarcity of electronic equivalent translated resources which could not be disputed but much more than that, how can we capture differences in the structures of these languages in a formal language so as to achieve translation?

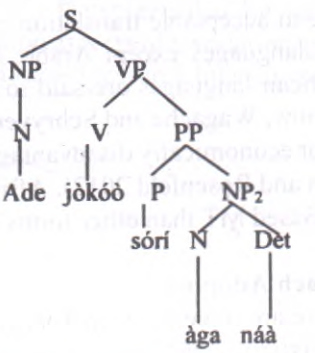
Yorùbá Rule-Based Machine Translation

As mentioned earlier, Computational Linguistics is new in Africa and especially Nigeria. Awofolu (2002)[62] is the first researcher of Yoruba-English MT. The system adopted Rule Based approach to MT because Awofolu (2002) claims that Yoruba electronic resources were scarce. JAVA was used to code the structure of both Yorùbá and English sentences. Rowland (1969) *Teach Yourself Yoruba* was used as a base for grammatical information with spellchecker and statistical analyzer based on neighboring word lexical categories. The work was a pacesetter using syntactic and semantic analyzing algorithms. But the concern with the work is the structure of the sentences and linguistic feature of the lexical entries, which themselves are connected to the deficiency of *Teach Yourself Yoruba* in the light of current syntactic discussion like Awobuluyi (1978) Bamgbose (1990) among others. Awofolu (2002) also did not report how the system was able to handle structural and word ambiguity which is one of the major challenges for Rule-Based MT.

Another research that is worthy of note is Eludiora, Salawu, Odejobi and Agbeyangi (2011). They adopted python for the rule-base MT. POS (Part of Speech) tags were used to identify the class of each lexical entry and parser parsed each sentence to its phrasal and word level categories. The problem with this approach is that the parsing could be highly confusing looking at its linguistic representation. For example consider example (1) an extract from Eludiora, Salawu, Odejobi and Agbeyangi's (2011) presentation.

1





Parsing sentences like the above may not reflect the true linguistic structures of the language. From the structure, one cannot actually say what the head of the NP_2 (Object) projection is because the projections are at the same level which is against merge approach to syntactic analysis. More so, the structures state that P and PP project to NP respectively. This is untrue and is not a correct representation of the languages structural architecture. This is mis-informing which needs an urgent review. Though, difference in the structural headship in the Noun Phrase was achieved in that *my father* is not translated as *mi bàbá* but *bàbá mi*.

The linguistic structural difference between English and Yorùbá languages

Yorùbá is a tonal language like many African languages. It has three distinctive tones; high mid and low. These tones have their lexical, syntactical and semantic functions (see Ladefoged 1974, 1982, Abercrombie 1967, Clement and Goldsmith 1984, Owolabi 1989 and Ufomata 2004).

[62] It should be note that we did not have access to complete report of Awofolu, our assessment is based on three pages report found online.

For example, Owolabi (2013) state categorically the structures where there are tonal manipulations in Yorùbá thereby reinforcing the essences of tone in the language. The tonal manipulations are to be considered in the translation process. Intonation as seen in English may not be of interest since the focus of this paper is not on speech syntheses. Although, emphasis is placed on tone which are marked on Yoruba orthography and if not, they may impede meaning or total loss of sense in such construction. Consider example 1 below

1. Mo fọ abọ
   1st sg V plate/dish

The meaning of the sentence (1) could not be ascertained not until the tone on the verb (fọ) is determined. If the tone is low tone (fọ̀) then it means to wash (i.e to wash a dish) but if it is high tone (fọ́) then it means to break. When this verb has direct DP object the inherent low tone has to change to mid tone because of the direct DP object as observed by Owolabi (2013:) , then we have sentences like:

2a. Mo fọ abọ́
1st sg wash plate I washed plate

But consider a sentence like (2b) below where the DP object is as a result of the lengthening of the verb, the verb retains its inherent low tone while the lengthened vowel has high tone.

2b Mo fọ̀ ọ́
1st sg wash 3rd sg I washed it

If the verb has high tone, the meaning or interpretation will be different from (2a) above; it could be interpreted as *to break* as seen in (3a) below:

3a Mo fọ́ abọ́
1st sg break plate I broke plate

You will now observe that if we lengthened the last syllable of the verb to derive the third person pronoun object like (2b) above, there will be a change on the tone of the pronoun.

3b. Mo fọ́ ọ
1st sg break 3rd sg I broke it

(3b) unlike (2b) has mid tone which is informed by the tone on the verb. Now that the tone on the verb is high, the tone on the lengthened vowel is mid tone. It should also be noted that there is non-existence of any word in Yorùbá as *abọ* as seen in (1) above. Although there are words like *abọ́* (plate/dish) as used in the examples above and abọ̀ (object) which is a technical term in Yorùbá syntax. You will observe that tone distinguishes the meaning of both *abọ́* and *abọ̀*.

Another structural difference in the languages is the representation of tense and aspect. English has three tenses: present, past and future which is marked by modal auxiliary and affixing them to verbal elements through affix hopping rule. For example consider example (2 and 3) below:

2 Bola is tall 3 Bola prest+be tall

(2) is a normal expression in its present tense for while (3) shows how the present tense is attached to the next verbal element "be" which resulted to (2).

Yoruba on the other hand has two tenses: future and non-future. Future tense in Yorùbá is marked by future tense marker like: yóò, ó, máa and ní (the negative form) and non-future tense inferred from the meaning of the verb. For example Omamor (1984) stipulates that active verbs connote past tense in their interpretation while stative verbs do not specify time. This brings about the distinction between example (4 and 5).

4 Adé wà ní ìpàdé
Adé is prep meeting 'Adé is in a meeting'.

5 Adé jẹ iṣu
Adé eat yam 'Adé ate yam'

(4) above does not specify time whether the meeting is concluded or on-going unlike (5) whose interpretation connote that the action has been concluded before it is reported.
Another structural difference in the languages is the arrangement of DP. Though, both languages are head first languages but a difference is observed in the DP. For example consider (6&7) below

6. Ilé pupa náà 7. The red house.

(7) is the translation of (6). They are both DPs. While NP in the DP of (7) is in suit but that of Yoruba is as a result of NP raising to functional position following DP hypothesis. Although, this hypothesis is widely accepted by western scholars, some African scholars are rejecting its positions for the analysis of African languages; example is Ilori (2010). The concern of this paper is not the scholastic arguments but how these differences in the structures of English and Yoruba are capture for translation using PROLOG.

What we did was to declare each sentence in line with its structure and rules using context free grammar[63]. This we did for each language in its language file and then merge each sentence with its translated equivalent sentence in the target language for translation.

Formal Language Syntax

Syntax is usually associated with the rules (or grammar) governing the composition of texts in a formal language that constitute the well-formed formulas of a formal system. In other words Syntax defines what sequences of symbols are valid; syntactic validity is independent of any notion of what the symbols mean. For example, a context-free syntax might say that A: = B + C is syntactically valid, while A: = B +; is not.

Interaction of Formal and Natural Language Syntax in PROLOG for Translation

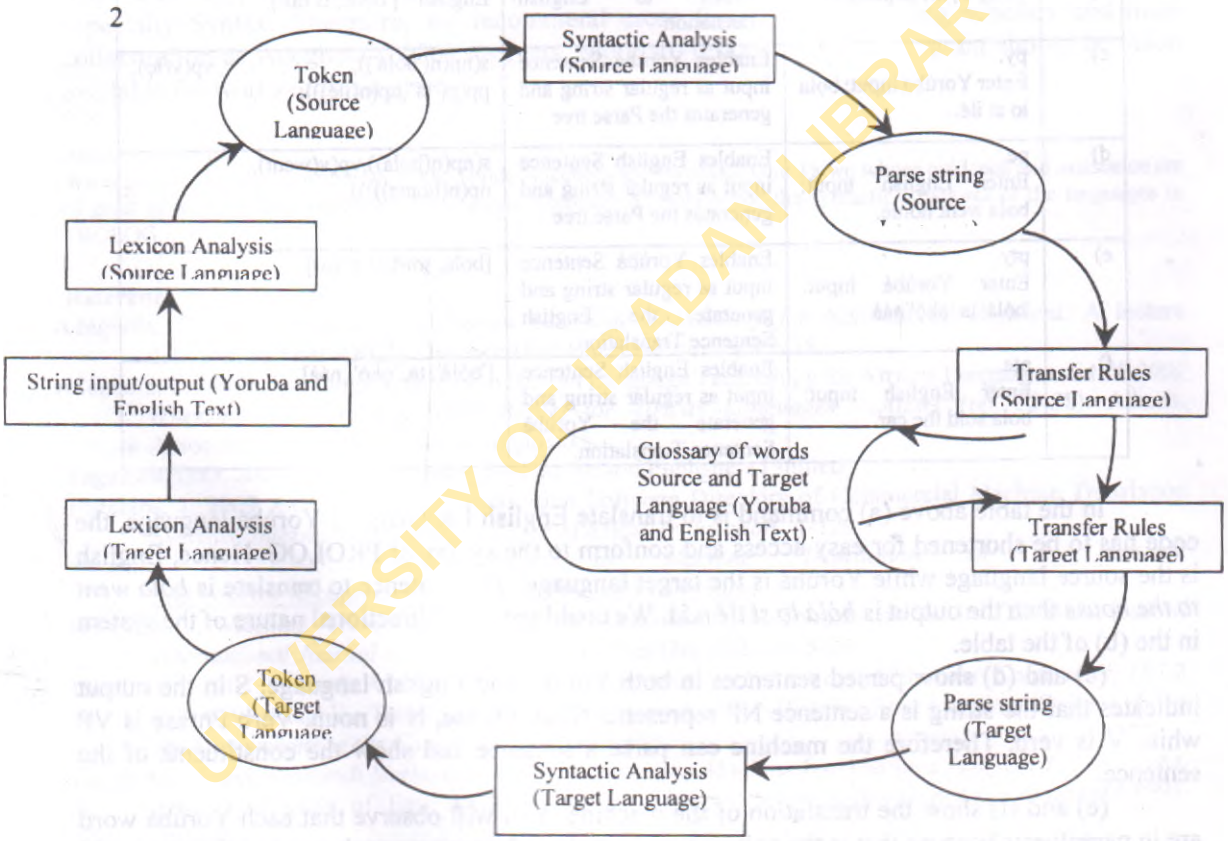
PROLOG is one of the computer logic programming languages. It is run by making a statement in the logical language, which the PROLOG interpreter tries to prove, given the premises in the program. One of the main purposes in developing PROLOG was to create a language in which phrase structure and semantic-interpretation rules for a natural-language question-answering system could be easily expressed (Pereira and Shieber 2002:2). For the purpose of this research Context Free Grammar was used (A context-free grammar is a set of recursive rewriting rules (or *productions*) used to generate patterns of strings) because it look more like Transformational Generative Grammar. Five files were created where formal and natural language syntax interact; whereby syntactic rules or natural language is coded in syntactic rules of formal language in the file created. They are: dictionary file (d.pl), Yorùbá language file (y.pl), English language file (e.pl), Transfer file (t.pl) and Read_line file. Each of these files is interconnected with transfer file.

[63] A context Free Grammar is a set of recursive rewriting rules (or production) used to generate pattern of string.

The dictionary file consists of each lexical item and its equivalents in both languages. Each lexical entry is tagged with its word class for necessary easy call e.g. N for Noun, V for Verb, P for Preposition, Det for Determiner etc. Yorùbá file contains each lexical entry of Yorùbá as well as various rules of forming phrases and sentences in the language (see Appendix 1 for details). The same is to the English file (see Appendix 2 for details). Transfer file connects all the files and that is where the translation actually takes place. It takes a sentence in source language and merges it with its equivalent in the target language. Because the rule is not about mere word substitution sentence like "Bọ́lá ga" is not translated as "Bọ́lá tall" rather "Bola is tall". Read-line assists with translating a string of sentence, say a paragraph or a passage from a file to another file. The sentences that the system cannot translate are posted to another file for proper analysis.

It is worthy of note that this system can only translate basic sentences in the language pairs and any lexical item that is not in the dictionary of the system will not be translated. This makes the exercise rule cumbersome. The diagram (2) below shows the flow of information in the system

2



As observed in the diagram (2) above, all processes start and end at the string input and output for both languages. Depending on the language used as source language, the words in the sentence will have to go through lexicon analysis irrespective of the category of the word (d.pl i.e the dictionary file). If one of the words in the sentence string is not available in the lexicon, the process will be aborted. But if all the words in the sentence string are in the lexicon, then the process proceeds to syntactic analysis (y.pl i.e. Yorùbá language file and e.pl i.e. English language file). At this point, the machine checks if the sentence string conforms to the syntactic rules of the source language. If the sentence conforms to the syntactic rules of the source language, then the process moves to the next stage otherwise error is written on the output string. The next stage is parsing string where the source sentence is parsed into phrase and word categories. Now that the sentence is parsed into

word categories, the parsed structure goes to the transfer rules (t.pl) where the language differences are handled, and then the glossary dictionary is consulted for word substitution. The substituted words return to target language transfer rules and from there, it goes to target language parser. The syntactic structure of the target language has much to do with the parser and the lexicon before the translated output comes out. This is because the syntactic structure of the source language has to merge with that of the target language and differences in structures has been checked. Therefore the translation process is cyclic. To demonstrate how this system works, consider the table below:

Table

S/N	Input Command	Description	Output Sentence
a)	e2y([bola, went, to, the, house],Yorùbá).	English to Yorùbá Translation	Yorùbá= [bólá, lọ, sí, ilé, náà]
b)	y2e(['bọ́lá', ga],English).	Yorùbá to English Translation	English= ['bọ́lá', is tall].
c)	py. Enter Yorùbá input: bọla lọ sí ilé.	Enables Yorùbá Sentence input as regular string and generates the Parse tree	s(np(n('bọ́lá')), vp(v(lọ), pp(p('sí',np(n(ilé))).
d)	pe. Enter English input: bola went home.	Enables English Sentence input as regular string and generates the Parse tree	s(np(n(bola)),vp(v(went), np(n(home))))).
e)	pty. Enter Yorùbá input: 'bọ́lá' ta 'ọkọ̀' náà.	Enables Yorùbá Sentence input as regular string and generate the English Sentence Translation.	[bola, sold, the, car],
f)	pte. Enter English input: bola sold the car.	Enables English Sentence input as regular string and generate the Yorùbá Sentence Translation.	['bọ́lá', ta, 'ọkọ̀', náà]

In the table above (a) command is to translate English Language to Yorùbá language, the code has to be shortened for easy access and conform to the syntax of PROLOG. Hence, English is the source language while Yorùbá is the target language. The sentence to translate is *bola went to the house* then the output is *bólá lọ sí ilé náà.* We could see the bidirectional nature of the system in the (b) of the table.

(c) and (d) show parsed sentences in both Yorùbá and English language. S in the output indicates that the string is a sentence NP represents Noun Phrase, N is noun, Verb Phrase is VP while V is verb. Therefore the machine can parse a sentence and show the constituents of the sentence.

(e) and (f) show the translation of the machine. You will observe that each Yorùbá word are in parenthesis because that is the only means to capture the tone on each word so that we could capture the structure of the language.

Deficiency of the Work

As much as structural ambiguity which is well known to Rule based MT was captured, this exercise has not been able to capture tonal manipulation in Yoruba as Owolabi (2006:134) observed hence sentence like (8) below are realized in the machine has (9).

8. Bọ́lá ta iṣu
Bọ́lá sold yam

9. Bọ́lá tà iṣu
Bọ́lá sold yam