

Womb Grammars: A constraint solving model for learning the grammar of Yorùbá

Ife Adebbara

University of British Columbia

Cognitive Systems

ife.adebbara@mail.ubc.ca

Abstract

We address the problem of inducing the grammar of Yoruba from that of English. We adopt an efficient and linguistically-savvy constraint solving model for an under-resourced language, Yorùbá, from the grammar of English. Our model - Womb Grammars (WG) parses a subset of noun phrases of the target language Yorùbá, from the grammar of the source language English. Our solution is straight forward and only requires a correct property grammar of the source language, a lexicon of the target language and a set of representative input phrases in the target language. This is extensible to and useful for other low resource languages where availability of large corpora is a challenge. Our proposed methodology adapts Womb Grammars in parsing phrases of the target from the grammar of English which is described as properties between constituents. Our model is implemented in CHR_G (Constraint Handling Rule Grammars).

Keywords: property grammars, constraint handling rules grammar, constraint handling rules, failure-driven parsing, womb grammars, Yorùbá

Yorùbá

A fi èro Womb Grammars hàn, èyí tí ó n fa gírámà ède Yorùbá láti gírámà ède gèèsí. Èro yí yìò lo gírámà gbólòhùn óro orúko ède gèèsí láti fi pín èya gbólòhùn óro orúko ède Yorùbá fún íwónba ápeere tí a fi se àlàyé isé yí. Isé yí ko nira fun èro Womb Grammars rárá. A kàn nílò gírámà ède gèèsí èyí tí a pè ní "property grammar", iwé itúmò ède Yorùbá àti ápeere gbólòhùn óro orúko ède Yorùbá. Yorùbá nikan kó ni isé yí wúlò fún a tún le l fún àwo ède mírà tí kò ní àkosílè ède púpò bíi ède Yorùbá.

1. Introduction

Womb Grammars (WGs) a novel grammar induction technique, developed by Dahl and Miralles (Dahl and Miralles, 2012; Becerra et al., 2013; Dahl et al., 2012). induces the grammar of a *target* language from the grammar of a *source* language. The WGs paradigm describes a language's phrases in terms of constraints or properties between pairs of direct daughters of a phrasal category called properties. WGs extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing.

In this paper, WGs are used to induce a subset of noun phrases in Yorùbá from that of English (Adebbara and Dahl, 2016). We assume that the grammar of the source language is correct and that the lexicon and input phrases of the target language are correct and representative of noun phrases in the target language. We adopt a constraint satisfaction approach whereby every phrase of the target language is tested for satisfaction and unsatisfied constraints provide a lead for the reconstruction of the target grammar using the source grammar. We use a context free grammar of the target language to evaluate the correctness of our parser.

Our results in applying and adapting the WG model for inducing Yorùbá noun phrases show that this model compares favourably with others in solving the grammar induction problem: it combines linguistic formality with efficient problem solving, and can transfer into other languages, including languages in which tones have a grammatical and / or semantic function.

The rest of this paper is divided as follows: Section 2 describes the motivation for this work, Section 3 provides a

description of the Yorùbá language. In Section 4, we describe the Linguistic Background of this work. We introduce the concept of property grammars and Womb Grammars in Section 5 and 6 respectively. In Section 7 we explain our results and conclude in Section 8.

2. Motivation

Language endangerment and death has been of serious concern in linguistics and language policy making. Close to seven thousand languages are currently spoken in the world, the majority of which are understudied and endangered. It has been said that an alarming 50 to 90 percent of languages will be extinct by the end of the century (Romaine, 2017).

For various reasons, some speakers of many minor, less studied languages may also learn to use a different language from their mother tongue and may even stop using their native languages. Parents may begin to use only that second language with their children and gradually the transmission of the native language to the next generation is reduced and may even cease. As a result, only the elderly in such communities may use the native language, after a while, there may be no speakers who use the language as their first or primary language and eventually the language may no longer be used at all. Thus, a language may become extinct, existing perhaps only in recordings, written records and transcription and languages which have not been adequately documented completely disappear.

Linguists cannot keep up with the study of these languages even for educational purposes, and there is a growing need for their automatic processing as well, since the amount of

text sources grows much faster than humans can process them. To make matters worse, most linguistic resources are poured into English and a handful of other first world languages, leaving the vast majority of languages and dialects under-explored. Clearly, automating the discovery of an arbitrary language's grammar model would render phenomenal service to the study and preservation of linguistic diversity.

Scientifically, we wanted to explore to what extent the parsing-as-constraint-solving paradigm of NLP problem solving could buy us a great degree of linguistic descriptive formality without sacrificing efficiency, in the realm of grammar induction and in particular for inducing Yorùbá, which is severely under-resourced and endangered.

3. The Yorùbá Language

Yorùbá belongs to the Yoruboid group of the Kwa branch of the Niger-Congo language family, which cuts across most of sub-Saharan Africa. It is a tonal dialect-continuum comprising about 20 distinctive dialects and spoken by over 30 million people in the western part of Nigeria (Gbenga, 1994). Niger-Congo is the largest of the five main language families of Africa. The others being Nilo-Saharan, Afro-Asiatic, Khoisan and Austronesian (mainly found in the nation of Madagascar).

Yorùbá is one of the three regional (national language contained in the constitution) languages in Nigeria and is said to be the most studied African language. Yorùbá is spoken by more than 20 percent of the population of Nigeria. The two other national languages are Hausa and Igbo, both of which are also regional languages in the north and south-eastern parts of the country respectively.

3.1. The Sociolinguistic Situation of Yorùbá in Nigeria

Despite the seemingly large population of Yorùbá speakers, according to (Ayo, 1993) Yorùbá has been classified as a "deprived" language. This is influenced by the language policy of Nigeria which favours the use of English above all indigenous languages. It also pays lip service to for instance the national language policy of education, which states that the mother tongue or the language of the immediate community must be adopted as the language of education in primary schools, and English should only be introduced at a later stage. Other language policies in other domains that encourage the use of mother tongues are also not adopted. So that the regional or national status of Yorùbá and other regional languages is theoretically but not fully implemented in practice (Abidemi and Segun, 2005).

English is the language of the elite and fluency in English is synonymous with a good education. As a result, many parents, even those who are barely educated or not educated at all, ensure that their children are taught in English right from the elementary classes. In most schools, indigenous languages are referred to as vernacular and are prohibited. Violation usually attract fines and many times corporal punishment. Bilingualism is also believed to affect children's ability to attain competence in English and thus parents avoid speaking mother tongues at home for fear of raising

children with poor communication in English. Many children therefore many can neither speak, read nor write in Yorùbá and many do not even understand the language at all.

All the afore mentioned language situation have influenced the lack of adequate language development and consequently resulted in, resource scarcity of Yorùbá.

4. Linguistic Background

4.1. Data Collection

Data collection for this research has been a combination of introspective and empirical collection methods. This approach of data collection was employed in order to ensure that our model is realistic, correct and robust. Introspection has proven to be the most reliable process of data collection and also very useful for building models which require high level linguistic competence such as this WG model (Chomsky, 1957). Introspection has been that of the author who has formal training in linguistics and is also a native speaker of Yorùbá. We have double-checked our introspective conclusions by consulting as well seven other native speakers of Yorùbá, four of which also have formal graduate level training in linguistics.

Data collected have also been compared with two existing grammars of Yorùbá. The first by Ayò Bámgbósé (Ayo, 1966) and the other by Awóbùlúyí (Oladele, 1978). It was important to observe these existing grammatical descriptions of Yorùbá, considering that they are one of the earliest contributions of native speakers who have formal linguistic training to the description of the Yorùbá grammar.

4.2. Strategies for Part of Speech Parsing

Assigning correct part-of-speech tags to each input word explicitly indicates some inherent grammatical structure of any language and a wrong part-of-speech tag will distort the grammatical structure of a language. We adopt a rule-based (derived from linguistic rules) approach. Rule-based approach though rigorous and requiring a great amount of high level linguistic skills, yield good results for any language, including those like Yorùbá which have been identified as resource scarce as well as having a less fixed word order structure.

The tagsets were developed using the Penn Tree Bank of Yorùbá (Yiwola, 2008) as well as judgments of native speakers of Yorùbá who have formal linguistic training. The tagsets were also compared to the grammars of Ayò Bámgbósé (Ayo, 1966) and Awóbùlúyí (Oladele, 1978).

We use the following tagsets: *noun* e.g *ajá* (dog), *pronoun* e.g *àwon* (they), *proper-noun* e.g *Ayò*, *determiner* e.g *kan* (a), *quantifier* e.g *gbogbo*(every) and *adjective* e.g *dúdú*(black). We further define features for each word in order to provide a fine-grained definition of each word tag. We *Number, Gender, Tone, Person, Definitiveness, and Case*. These features have been carefully chosen to ensure that our model accounts for the unique traits of Yorùbá.

5. Property Grammars

The idea of constraint is present in modern linguistic theories such as Lexical Functional grammars (LFG) and Head-

driven Phrase Structure grammars (HPSG). However, constraint satisfaction, a way of implementing constraints, is not really incorporated in the implementation of these theories. Thus, we use a formalism called Property Grammar (PG)(Blache and Rauzy, 2012) which is based completely on constraints: all linguistic information is represented as properties between pairs of constituents, which allow parsing to be implemented as a constraint satisfaction problem. For example in the PG framework, English noun phrases can be described through a few constraints such as precedence (a determiner must precede a noun, an adjective must precede a noun), uniqueness (there must be at most one determiner), exclusion (an adjective phrase must not coexist with a superlative), obligation (a noun phrase must contain the head noun), and so on. Instead of resulting in either a parse tree or in failure as traditional parsing schemes do, such frameworks characterize a sentence through the list of the constraints a phrase satisfies and the list of constraints it violates, so that even incorrect or incomplete phrases will be parsed. Moreover, it is possible to relax some of the constraints by declaring relaxation conditions in modular fashion.

6. Womb Grammars

Womb Grammar (Adebara et al., 2015; Ife and Veronica, 2015; Philippe, 2005) was presented in two versions: *Hybrid Womb Grammars*, in which the source language is an existing language for which the syntax is known, and *Universal Womb Grammars*, in which the source syntax is a hypothetical universal grammar of the authorsâ own devise, which contains all possible properties between pairs of constituents. We adopted the Hybrid Model.

The general WG model can be described as follows: Let L^S be the source language. Its syntactic component will be noted L_{syntax}^S . Likewise, we call the target language L^T and its lexicon (L_{lex}^T). If we can get hold of a sufficiently representative set of phrases in L^T that are known to be correct (a set where our desired subset of language is represented), we can feed these to a hybrid parser consisting of L_{syntax}^S and L_{lex}^T . This will result in some of the sentences being marked as incorrect by the parser. An analysis of the constraints these “incorrect” sentences violate can subsequently reveal how to transform L_{syntax}^S so it accepts as correct the sentences in the corpus of L^T —i.e., how to transform it into L_{syntax}^T by modifying the constraints that were violated into constraints that accept the input.

For instance, let $L^S = English$ and $L^T = Yorùbá$, and let us assume that English adjectives always precede the noun they modify, while in Yorùbá they always post-cede it (an oversimplification, just for illustration purposes). Thus “a red book” is correct English, whereas in Yorùbá we would more readily say “iwe pupa kan” (book, red, a).

If we plug the Yorùbá lexicon and the English syntax constraints into our WG parser, and run a representative corpus of (correct) Yorùbá noun phrases by the resulting hybrid parser, the said precedence property will be declared unsatisfied when hitting phrases such as “iwé pupa kan”. The model transformation module can then look at the entire list of unsatisfied constraints, and produce the missing syntactic component of L^T ’s parser by modifying the con-

straints in L_{syntax}^S so that none are violated by the corpus sentences.

6.1. Modified parsing that calculates both failure and success explicitly

Some of the necessary modifications are easy to identify and to perform, e.g. for accepting “iwé pupa kan” we only need to delete the (English) precedence requirement of adjective before noun (noted $adj < n$). However, subtler modifications may be in order, after some statistical analysis in a second round of parsing: if in our L^T corpus, which we have assumed representative, *all* adjectives appear after the noun they modify, Yorùbá is sure to include the reverse precedence property as in English: $n < adj$. So in this case, not only do we need to delete $adj < n$, but we also need to add $n < adj$.

Previous models of WGs (Adebara et al., 2015; Ife and Veronica, 2015; Ife and Dahl, 2015) focused on failure driven parsing, under the assumption that failed properties are usually the complement of those satisfied, so they can be derived from the failed ones if needed and in general, a grammar is induced by repairing failures. However our more in depth analysis in the context of Yorùbá has uncovered the need for more detail than simply failing or succeeding, as in the case of conditional properties. We therefore now use a success conscious *and* failure conscious approach for inducing the grammar of our target language, Yorùbá. Each input phrase of the target language is tested with all relevant constraints for both failure and success. This makes the model slightly less efficient than if we only were to calculate failed properties, but of course the gain is in accuracy. Efficiency is still guaranteed by the normal Constraint Handling Rules Grammar (CHRG) way of operating: rules will only trigger when relevant, e.g. if a phrase is comprised of only a noun and an adjective, it will not be tested with for instance precedence(pronoun, determiner) or any other constraint whose categories are different from those of the input phrase. We keep a list of all properties that fail and another for those that succeed together with the features of the categories of each input phrase and their counts. It is important to state that constituency constraints are tested only for success. This is because we are interested in checking that our target grammar shares similar constituents with our source language and testing for failure will be irrelevant for these constraints. We also are able to induce constituents present in the target grammar that are not in the source grammar.

7. Results and supporting evidence of correctness

Our results so far have been consistent with linguistic research of Yorùbá grammar. We also use phrases generated by a Context Free Grammar (CFG) which we developed as the input phrases in our WG model and our induced grammar have shown similarities with the CFG. It is important to state that despite equivalences that our induced grammar share with the CFG subset, our induced grammar explicitly encodes more information than the CFG. This is because phrase structure representations such as context free grammars use a unique explicit relation hierarchy, that

encode constituency information so that other information such as linearity, obligatoriness, dependency, etc. are implicit. On the opposite, constraint-based representations, such as our property grammar encode explicitly all these relations. (Blache and Rauzy, 2012). We summarize our results into constituency, precedence, requirement, dependency and obligatoriness.

- 1 **Constituency:** Our constituency results show that in Yorùbá, nouns, pronouns, proper-nouns, adjectives, quantifiers and determiners are allowable categories in noun phrases as described in literature (Peter, 1970; Oladele, 1978; Ayo, 1966) as well as in our CFG.
- 2 **Precedence:** We induce two conditional precedence properties (although conditional precedence(adjective, noun) has two different conditions.), and nine precedence properties. Our conditional precedence imply that there are two orderings, in pronouns and nouns and adjectives and nouns. This we found evidence for in literature(Peter, 1970). However, we do not induce a property for quantifier and noun. This is because, there exist no known pattern responsible for the difference in order, which is consistent with research claims (Oladele, 1978).
- 3 **Requirement:** We do not induce any requirement between nouns and determiner. This is because of a lack of pattern in features where the requirement property succeeds and where it fails. This is consistent with our CFG which has rules where nouns occur without determiners as well rules where nouns occur without determiners. This conclusion is also presented in research (Ayo, 1966; Oladele, 1978; Oladiipo, 2009).
- 4 **Dependency:** Our model also does not induce dependency rules. This is because there was no unique feature present with instances where dependency failed and when it succeeded. This again is supported by (Ayo, 1966) but not explicit in our CFG.
- 5 **Obligatoriness:** Obligatoriness succeeded in all input phrases, showing that at least one of noun, pronoun and proper-nouns are compulsory constituents of Yorùbá. Our CFG also shows these three constituents occur at least once in all rules.

8. Conclusion

We have shown the simplicity with which Womb Grammars automatically induces the grammar of Yorùbá from that of English despite the peculiarities in the grammar of Yorùbá that can make this very difficult which makes our model very useful in language development and language documentation. Our system automatically transforms a user's syntactic description of a source language into that of a target language, of which only the lexicon and a set of representative sample phrases are known. While demonstrated specifically for English as source language and Yorùbá as target language, our implementation can accept any other pair of languages for inducing the syntactic constraints of one from that of the other, as long as their description can be done in terms of the supported constraints.

9. Bibliographical References

- Abidemi, F. F. and Segun, S. A. (2005). Is Yorùbá an endangered language? *Nordic Journal of African Studies*, 14(3):18–18.
- Adebara, I. and Dahl, V. (2016). Grammar induction as automated transformation between constraint solving models of language. In *Proceedings of the Workshop on Knowledge-based Techniques for Problem Solving and Reasoning co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, New York City, USA, July 10, 2016.
- Adebara, I., Dahl, V., and Tessaris, S. (2015). Parsing with partially known grammar. In *Agents and Artificial Intelligence - 7th International Conference, ICAART 2015, Lisbon, Portugal, January 10-12, 2015, Revised Selected Papers*, pages 334–346.
- Ayo, B. (1966). A grammar of Yoruba. vol. 5.
- Ayo, B. (1993). Deprived, endangered, and dying languages. *Diogenes*, 41(161):19–25.
- Becerra, L., Dahl, V., and Miralles, E. (2013). On second language tutoring through womb grammars. Accepted for publication at IWANN 2013, June 12-14, Tenerife, Spain.
- Blache, P. and Rauzy, S. (2012). Hybridization and tree-bank enrichment with constraint-based representations. In *Proceedings of LREC*.
- Chomsky, N. (1957). Syntactic structures [text]. In *Walter de Gruyter*, page 117.
- Dahl, V. and Miralles, J. E. (2012). Womb grammars: Constraint solving for grammar induction. In J. Sneyers et al., editors, *Proceedings of the 9th Workshop on Constraint Handling Rules*, volume Technical Report CW 624, pages 32–40, Department of Computer Science, K.U. Leuven.
- Dahl, V., Miralles, E., and Becerra, L. (2012). On language acquisition through womb grammars. In *7th International Workshop on Constraint Solving and Language Processing*, pages 99–105.
- Gbenga, F. J. (1994). *The Yoruba Koiné: Its History and Linguistic Innovations*, volume 6. Lincom Europa.
- Ife, A. and Dahl, V. (2015). Domes as a prodigal shape in synthesis-enhanced parsers.
- Ife, A. and Veronica, D. (2015). Shape analysis as an aid for grammar induction.
- Oladele, A. (1978). *Essentials of Yoruba grammar*. University Press Plc Nigeria.
- Oladiipo, A. (2009). Analyzing Yoruba bare nouns as dp. *Lagos Notes and Records*, 15(1):30–55.
- Peter, O. (1970). The essentials of the Yoruba language.
- Philippe, B. (2005). Property grammars: A fully constraint-based theory. In *Proceedings of the First International Conference on Constraint Solving and Language Processing, CSLP'04*, pages 1–16, Berlin, Heidelberg. Springer-Verlag.
- Romaine, S. (2017). The impact of language policy on endangered languages. In *Democracy and human rights in multicultural societies*, pages 217–236. Routledge.
- Yiwola, A. (2008). Global Yoruba lexical database.