TRIANGLE. *Language, Literature and Computation*
Volume 14, 2016, pp. 1–36
DOI: 10.17345/triangle14.1-36
ISSN: 2013-939X
http://revistes.urv.cat/index.php/triangle

Using Womb Grammars for Inducing the Grammar of a Subset of Yorùbá Noun Phrases

IFE ADEBARA*

Simon Fraser University, Burnaby, Canada

Abstract. We address the problem of inducing the grammar of an under-resourced language, Yorùbá, from the grammar of English using an efficient and, linguistically savvy, constraint solving model of grammar induction –Womb Grammars (WG). Our proposed methodology adapts WG for parsing a subset of noun phrases of the target language Yorùbá, from the grammar of the source language English, which is described as properties between pairs of constituents. Our model is implemented in CHRG (Constraint Handling Rule Grammar) and, it has been used for inducing the grammar of a useful subset of Yorùbá Noun Phrases. Interesting extensions to the original Womb Grammar model are presented, motivated by the specific needs of Yorùbá and, similar tone languages.

Keywords: property grammars, constraint handling rules grammar, constraint handling rules, womb grammar, computational grammar, syntax, grammar induction, Yorùbá.

1 Introduction

Grammar induction is a data driven approach to learning grammars, and, analyzing languages.

* **Author's address**:
Department of Computing Science
Simon Fraser University
8888 University Drive, Burnaby, Canada
E-mail iadebara@sfu.ca

We adopt a novel grammar induction technique, called, Womb Grammars (WG), developed by Dahl and Miralles (2012a). WG employ a property based methodology, which, relies on the grammar of a source language to induce the grammar of a target language. WG 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 a fragment of noun phrases in the target language. It adopts a constraint-satisfaction approach whereby input phrases of the target language are 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.

In this paper, WG are used to induce a/the grammar of a subset of noun phrases in Yorùbá, from that of an English grammar with same coverage (Dahl & Miralles 2012b). We use English as our source language and, Yorùbá as the target language. English was chosen as the source language for various reasons, especially because, it is a language the author can speak and, write fluently. English also has a wide audience, and, we believe that it will be easier for a wider audience to understand the WG model using a language: English, whose structure is well known. Using a well known language eliminates the need for the reader to in addition to understanding the WG model, learn and, understand the structure of two less known languages (Source and target languages). Also, because English is a language which has been well studied by linguists, many linguistic resources are available and, easily accessible, and, this made creating a correct property grammar (details in section 4.1) easier than using any language which has less resources easily accessible. The author's formal training in linguistics and, native speaker competence in Yorùbá also made Yorùbá a desirable language as the target. Since we need a correct grammar of both the source and, target language, English and, Yorùbá languages were thus, the most accessible languages for this task. In fact, the success achieved from using English to induce the grammar if Yorùbá, reveals that, it is possible to use WG for any pair of languages, including unrelated languages, although with some modifications.

The novel approach of WG, needs neither a pre-specified model family, nor, parallel corpora, nor any of the typical models of machine learning, and, works for more than just specific tasks such as disambiguation (Dahl & Miralles 2012a). It automatically transforms a given (task-independent) grammar description of a source language, from just the lexicon of the target language, and, a representative input set of correct phrases in the target language. The syntactic descriptions of the source and, the target language subset addressed are stated in terms of linguistic constraints (also called "properties" in linguistic literature) between pairs of constituents, although for the

target language constraints we depart from the classic formalism (Blache 2004) in view of Yorùbá motivated extensions. This allows us a great degree of modularity, in which constraints can be checked efficiently through constraint solving.

Using linguistic information from one language to describe another language has yielded good results, however, it was used for tasks like disambiguating the other language (Burkett & Klein 2008), fixing morphological or syntactic differences by modifying tree-based rules (Nicolas *et al.* 2009), and, not for syntax induction which is our focus. The approach adopted usually requires parallel corpora, although, there exists an exception (Cohen & Smith 2010) where a bilingual dataset is used to train parsers for two languages simultaneously. This is accomplished by tying grammar weights in the two hidden grammars, and, is useful for learning dependency structure in an unsupervised empirical Bayesian framework.

Our results in applying and, adapting the WG framework 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.

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 2002).

For various reasons, some speakers of many minor, less studied languages may 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 voice recordings, written records and, transcription and, languages which have not been adequately documented completely disappear.

Linguists cannot keep up with the study of the endangered 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 support 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 explore to what extent the parsing-as-constraint-solving paradigm of natural language processing problem solving could achieve 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-continnum comprising about 20 distinctive dialects and, spoken by over 30 million people in the western part of Nigeria (Fagborun 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 170 million people who make up the population of Nigeria (the largest single black nation on earth). The two other national languages are Hausa and, Igbo, both of which are also regional languages in the north and, southeastern parts of the country respectively. The Yorùbá language is a koine (Fagborun 1994), a process involving dialect mixing, levelling, and, simplification (Trudgill 1986). Standard Yorùbá which can also be referred to as the standard koine is a compendium of several dialects, mainly dialects of Ọ̀yọ́, Ìbàdàn, Abẹ́òkúta, and, Lagos, which were major trade centers of early Church Mission Society missionary activities.

Apart from the standardized koine, there are twenty other dialects of Yorùbá: Ọ̀yọ́, Ìjẹ̀ṣà, Ìlá, Ìjẹ̀bú, Oǹdó, Wo, Owe, Jumu, Iworro, Igbonna, Yagba, Gbedde, Ẹ̀gbá, Akono, Aworo, Bunu (Bini), Èkìtì, Ìlàjẹ, Ìkálẹ̀, Awori. These dialects are largely

mutually intelligible, albeit with some variations in vocabulary and, phonology and, were largely spoken by different groups of people who, though tracing their descent to their common progenitor (Oduduwa), do not consider themselves as one people. Although there are several dialects of Yorùbá, it is important to mention that the present research is based on the 'standard' Yorùbá. This is because, 'standard' Yorùbá is the only variety referred to as Yorùbá and, the only variety taught in schools in Yorùbá land and, diaspora. It is the language of the media and, of official government business. It is described as a resource scarce language (Adeyanju *et al.* 2015) due to the limited number of texts available in Yorùbá.

3.1 The Sociolinguistic Situation of Yorùbá in Nigeria

Despite the seemingly large population of Yorùbá speakers, according to Bámgbósé (1993) it is a deprived language. Yorùbá has also been classified as a seriously endangered language (Fabunmi & Salawu 2005). This is according to Wurm (1998) who categorises language status into:

(i) potentially endangered - there is a decline in the use of the language by children
(ii) endangered - youths have experienced a decline in the use of the language and children experience language loss
(iii) seriously endangered - competent speakers are about 50 years old and, above
(iv) moribund - competent speakers are very old, and, many are dead
(v) extinct - when the last speaker dies;

The status of Yorùbá is influenced by the language policy of Nigeria, which favours the use of English above all indigenous languages. It also pays lip service, for instance, to 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 (Fabunmi & Salawu 2005).

Colonialism imported the English language to Nigeria and, English has since been adopted as the official language in government, education, and, all official businesses. 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 can neither speak, read nor write in Yorùbá, and, many do not even understand the language at all.

In addition, the great linguistic diversity in Nigeria discourages people from speaking Yorùbá as often as they may have. Well over 450 languages are said to be spoken in Nigeria, therefore necessitating the use of English as the lingua franca (Less educated people use the Nigerian Pidgin (an English based creole of some sort) as a medium of communication, and, some educated people use the Nigerian pidgin in less formal environments) (Adegbija 2004).

English is also the language spoken in offices, corporate organizations, and, the use of Yorùbá has greatly declined, resulting in lower competency among many users, code-mixing, code switching, and, several other language loss trends. In addition, although Yorùbá is still widely used in media and, there exist radio and, television programs in Yorùbá, as well as newspapers, they receive their patronage from the older generation since there always exist the English versions, which receive a greater patronage.

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

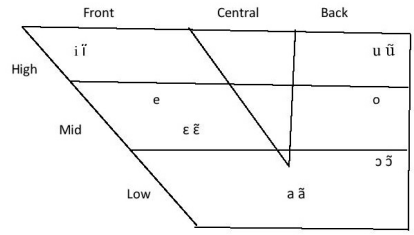
3.2 Yorùbá Phonology

The phonology of Yorùbá, which is how phonemes (Unit of distinctive sound that account for meaning distinction) are strung together to form words, comprises of seven oral vowels, five nasal vowels, three syllabic nasals and, seventeen consonants. Oral vowels are produced from the oral orifice, while nasal vowels are produced through the nasal orifice. Nasalized vowels, those vowels produced with the nasal and, oral orificies, are known to occur through an assimilation process when vowels preceed nasal consonants. All the seven oral vowels can be nasalized.

Pitch, the perceptual correlate of vocal cords' frequency, is distinctive in Yorùbá. This distinctive/phonemic pitch is called tone. For example, the word *rá*(H) (disappear), *rà*(L) (buy) and, *ra*(M) (rub) are distinguished by the three tones in Yoruba:

High-tone, Mid-tone and Low-tone. High-tone is orthographically represented with acute accent, the Low-tone with grave accent and, the mid is unmarked. These tones can interact to form rising and, falling tones for instance *yìí* (this) and, *nàà* (the). Vowels and, syllabic nasals are the tone bearing units in Yorùbá.

Consonant clusters are not permitted in Yorùbá but long vowels are possible.



		Labial	Alveolar	Palatal	Velar	Labiovelar	Glottal
stops	Voiceless		t		k	k͡p	
	Voiced	b	d		g	g͡b	
Affricates	Voiced			d͡ʒ			
Fricatives	Voiceless	f	s	ʃ			
Nasals		m	n	ŋ			
Approximants	lateral		l				
	Central		r	j		w	h

Fig. 1. Yorùbá Vowel and Consonant Chart.

3.3 Yorùbá Morphology

Typologically, Yorùbá is an analytic language, that is, grammatical relations are expressed with little or, no use of inflection. Tense is not marked morphologically, and, agreement relation is limited. It is a combination of isolating, that is, it has a low morpheme per word ratio, and, agglutinating through the use of prefixes, interfixes, reduplication (full and, partial reduplications) and, compounding. Prefixation and, reduplication are the most common word formation processes in Yorùbá. Word formation processes in Yorùbá are also derivational and, words changed through inflections are usually through tonal variations (Awoyale 1988). Due to the syllable structure of Yorùbá, compounding usually involves vowel ellision, coalescense, epenthesis, and, compensatory lengthening to resolve vowel hiatus and, consonant clusters.

3.4 Yorùbá Orthography

The Yorùbá language has been written since the early 19th century. However, there have been several changes in the orthography since then (Fagborun 1989; Olumuyiwa 2013).

Orthographically, the Yorùbá alphabet consists of 25 letters and, uses the familiar Latin characters. Nasalized vowels are represented with a 'n' after the vowel. Syllabic nasals are written as 'n', except when they occur before 'b', where they are represented as 'm'.

Although tones occur on syllables, they are orthographically marked on vowels and, syllabic nasals. Yorùbá syllables are open; this means they end in vowels. The syllables are formed by a single vowel, consonant plus vowel, or, a syllabic nasal.

The present standards were established in 1974, however, there remains a great deal of contention over writing conventions, spelling, grammar, the use of tone marks and, some other linguistic formalities. For the purpose of this paper, we have adopted the generally accepted formalities of the standard language according to grammars by Bámgbósé (1966) and Awobuluyi (1978). These include full tone marking including tone marks on syllabic nasals.

Aa	**Bb**	**Dd**	**Ee**	**Ẹẹ**	**Ff**	**Gg**	**Gbgb**
[a]	[b]	[d]	[e]	[ɛ]	[f]	[g]	[gb]
Hh	**Ii**	**Jj**	**Kk**	**Ll**	**Mm**	**Nn**	**Oo**
[h]	[i]	[ʤ]	[k]	[l]	[m]	[n] [ŋ]	[o]
Ọọ	**Pp**	**Rr**	**Ss**	**Ṣṣ**	**Tt**	**Uu**	**Ww**
[ɔ]	[kp]	[r]	[s]	[ʃ]	[t]	[u]	[w]
Yy							
[j]							

Fig. 2. The Yorùbá Alphabet and the Sounds they Represent.

3.5 Constituent Structure of Yorùbá Noun Phrases

As in English, the canonical word order is SVO, e.g, as in ọmọ ra oúnjẹ (child buy food). However, there are several interesting differences between the constituent structure of Yorùbá and, English. Five of these are within the scope of noun phrases

and, have been handled in this work. Details of how we address these differences are in section 7.

(1) In the arrangement of constituents, Yorùbá has a determiner final position in structure and, determiners are not obligatory in noun phrases (Ajiboye 2006). So that *ajá kékeré* (dog small) and, *ajá kékeré kan* (dog small a) are felicitous in Yorùbá.

(2) Yorùbá has a superstructure, from which 'adjectives' are derived (Awobuluyi 1978; Bámgbósé 1966; Bowen 1858; Awoyale 2008; Ogunbowale 1970). For instance the word *pupa* (red) has two entries in the dictionary (Awoyale 2008). The first entry is as a noun in *Pupa dára ju dúdú lọ* (red good more black than) which means red is better than black, and, as a verb in *ọmọ́ pupa* (child red) which means, child is red. It is important to understand that in the example given for the first entry, *pupa* and, *dúdú* are nouns because both words can be replaced with the pronoun *'ó* 'he/she/it', so that we can say *ó dára ju ú lọ*. The *ó* is derived as *ú* after *ju* through an assimilation process because Yorùbá abhors vowel hiatus. In cases of plural and, honorific nouns (just like *tu* and, *vous* in French), we will have *Pupa dára ju àwọn dúdú lọ* (red, good more plural, black than), *Pupa dára ju wọ́n lọ* (red good more them (or singular honorific) than)'. In the second example however, *ọmọ́ pupa* has also gone through a vowel deletion and, elongation through an assimilation process, which turned *ọmọ ó pupa* to *ọmọ́ pupa*. The mid tone *o* was deleted and, the high tone *o* was elongated. Although the *ó* (which is a pronoun but functions as a determiner) in that sentense is implied but will not be articulated by native speakers of Yorùbá. *ọmọ pupa* 'red child' however is an example where *pupa* performs a modifier or, adjective function to the noun *ọmọ*. Several other words like "kékeré" 'small', *nlá* 'big', *púpọ̀* 'plenty' are examples of some nouns or, verbs that can function as adjectives. For the purpose of this paper, nouns or, verbs that modify other nouns are referred to as adjectives.

(3) In Yorùbá, agreement is morphologically absent and, as a result, plurality is shown syntactically through the occurrence of a pronoun *àwọn* 'they' before a noun as in *àwọn ọmọ* (they child) which means 'children'; addition of cardinals or, adjectives after a noun as in *ọmọ méjì* (child two) implies 'two children', addition of an elided form of the pronoun *àwọn*, before a demonstrative as in *ọmọ wọ̀n yẹn* (child they), that means 'those children' or, reduplication of certain 'adjectives' after the noun as in *ajá dúdú dúdú* (dog black black) (Ajiboye 2006). It is important to note that the words used for plurality in Yorùbá have their own distinct meanings and, are used in certain contexts to depict plurality.

For instance *Ilé ńlá ńlá* (house big big) will denote several houses but *esè kòǹgbà kòǹgbà* (leg big big) will not be a plural expression despite both phrases having identical syntactic categories (Ogunbowale 1970).

(4) Yorùbá doesn't have a formal feature for indicating gender and, when gender distinction is mentioned in Yorùbá grammar, it is a translation equivalence, or, it is with reference to the structure of English and, not Yorùbá (Bámgbósé 1966). The use of words like *akọ* (male), *abo* (female), *obìnrin* (female human being), *ọkùnrin* (male human being) can be used to express gender in words like: *Abo kìnìhún* (lioness), *ọmọ obìnri* (child female) or, 'daughter', *akọ ẹlẹ́dẹ̀* (pig) while the pronoun "o" is used to refer to 'he/she/it' (Bámgbósé 1966). We explain how we use gender features in section 8.7.

(5) Another interesting distinction between Yorùbá and, English structure is the position of adjectives in phrases. Many adjectives occur after nouns, but it is also common to have adjectives before the noun. For instance, *òkú* (dead) always occurs before a noun as in *òkú ẹran* (dead meat). Adjectives that describe a part of something always occur before a noun as in *ìdajì oúnjẹ* 'half food'. Adjectives that describe the attribute of a person or, thing usually occurs before a noun as in *kékeré ẹkùn* (small leopard), but there are some exceptions where having the adjective after the noun are accepted as correct (Ogunbowale 1970). We describe our results, which are consistent with this linguistic claim in section 9.

Although, we cannot discount language universalities, these exceptions which can cause difficulty in simple comparisons have been managed in very interesting ways.

4 Linguistic Background

4.1 Data Collection

Data collection for this research has been a combination of introspective and, empirical collection methods. The collection process commenced with the development of a context free grammar of noun phrases in Yorùbá. This grammar was developed by linguists who also have native competence in Yorùbá. The first context free grammar was developed by the author who is also a linguist and, the grammar was extended by the other linguists Dr. Tunde Adegbola (African Languages Technology Initiative), Dr. Demola Lewis (University of Ibadan), Samuel Akinbo (University of British Columbia). The grammar was then used to generate phrases which were used as input phrases in our Womb Grammars model. In all, fourty five phrases were generated. Native speakers of Yorùbá who do not have formal linguistic training were

invited to check the phrases generated by the context free grammar in order to ascertain that those phrases are intuitively correct to the regular user of the language.

We developed a context free grammar because there is no known treebank of Yorùbá and, Penn Treebank only has a dictionary of words.

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 as previously mentioned, has been contributed by the author who is a native speaker of Yorùbá and, also has formal training in linguistics. We have double-checked our introspective conclusions by consulting seven other native speakers of Yorùbá, three 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 (1966) and, the other one by Awóbùlúyí (1978). It was important to observe these existing grammatical descriptions of Yorùbá, considering that they are among the earliest contributions of native speakers who have formal linguistic training to the description of the Yorùbá grammar (Chelliah & de Reuse 2010). Ajiboye's (2006) description of Yorùbá noun phrases has also been very useful for verifying the grammar induced by our model, especially in relation to plural marking.

4.2 Strategies for Part of Speech Parsing

Accuracy of part-of-speech tagging is a critical and, fundamental building block for many computational linguistic tasks including grammar induction. 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. Rule-based, data driven and, hybrid methods for part-of-speech tagging have been extensively described in literature (Wang & Li 2011). Rule-based methods are those derived from linguistic rules (Brill *et al.* 1990), data-driven methods are derived from statistical analysis of language data (Merialdo 1994), and, hybrid methods are a combination of rule driven and, data driven approaches (Sun & Bellegarda 2011).

Data-driven approaches have yielded very good results, albeit for Indo-European languages and, other data rich languages like English, Dutch, German, etc. However, data driven approaches have error rates which are usually reducible by only a few

percentage, and, also work poorly with languages having a less fixed word order (Voutilainen 2012).

Rule-based part-of-speech tagging on the other hand is the oldest approach that uses hand-written rules for tagging. Rule based taggers depend on dictionaries or, lexicons, that contain word forms together with their associated part-of-speech labels, as well as context-sensitive rules to choose the appropriate tags during application. The dictionary or, lexicon is consulted to get possible tags for each word. Hand-written context-sensitive rules are used to identify the correct tag if a word has more than one possible tag. The linguistic features of the word and, other words surrounding it are analyzed for disambiguation purposes. For example, a tagger for English will have a rule such as: if the preceding word in a phrase is a determiner then the last word in the phrase must be a noun.

We adopt a rule-based 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 we have identified as resource scarce in section 3 and, 3.1, as well as having a less fixed word order structure in section 3.4.

The tagsets were developed, with the use of Awoyale (2008), of Yorùbá as well as native speakers of Yorùbá who have formal linguistic training. The tagsets were compared to the grammars of Ayò Bámgbósé (1966) and, Awóbùlúyí (1978), one of the earliest grammar descriptions of Yorùbá by native speakers who have formal linguistic training.

We use the following tagsets:

i noun - ajá (dog), ẹran (goat), etc
ii pronoun - àwọn (they), ìwọ (you), etc
iii proper-noun - Ayò, Bámgbósé, etc
iv determiner - kan (a), náà (the), etc
v quantifier - gbogbo (every), ìdajì (half), etc
vi adjective - dúdú (black), pupa (red), etc

We further define features for each word in order to provide a fine-grained definition to each word tag. We use the following features:

i Number
ii Gender
iii Tone

iv Person
v Definitiveness
vi 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) (Pollard & Sag 1994). However, constraint-satisfaction, a way of implementing constraints, is not really incorporated in the implementation of these theories. We use a formalism called Property Grammar (PG) (Blache 2004), which is based completely on constraints: all linguistic information is represented as properties of pairs of constituents, which allows parsing to be implemented as a constraint-satisfaction problem. The set of properties forms a system of constraints expressed over categories. They represent different kinds of information (e.g. agreement, morphology or, semantics) that can be used typically for contextual restrictions. PG differs from HPSG and, LFG in that it does not belong to the generative syntax family, but to the model-theoretic syntax one (Blache 2000).

Constraints represent information and, the first benefit of representing a problem using constraints is that of partially solving it, so this allows for partial solutions to be found even when a full solution is not available. These constraint-based or, property-based theories, such as Property Grammars (PG) (Blache 2004) evolved from Immediate Dominance/Linear Precedence (IDLP), which unfolds a rewrite rule into the two constraints: (1) of immediate dominance (expressing which categories are allowable daughters of a phrasal category), and, (2) linear precedence (expressing which of the daughters must precede which others).

For example in the PG framework, English noun phrases can be described using constraints such as: precedence (a determiner must precede a noun, an adjective must precede a noun), requirement (a singular non-generic noun must be used with a determiner), obligation (a noun phrase must contain the head noun), and, so on. The linguistic structure (e.g phrase, sentense, clause etc) is characterized by a finite list of the constraints it satisfies and, a list of constraints it violates, so that even incorrect or, incomplete phrases will be parsed. For instance, the phrase the professor

emeritus, which be characterized as satifying obligation, requirement, determiner precedence but will not satify the adjective precedence. This differs from traditional parsers which characterize a linguistic structure with either a parse tree or, a failure. Using PGs also enables us relax certain constraints by declaring conditions under which those constraints should be relaxed. For instance, the obligation rule of English can be relaxed to allow pronouns and, proper nouns instead of only a noun because pronouns and, proper nouns can also function as the head of a phrase. Dahl and Blache (2004) encode the input PG into a set of CHRG rules that directly interpret the grammar in terms of satisfied or, relaxed constraints and, a syntactic tree is the implicit result. Womb Grammars which is an adaptation of this framework into grammar transformation (Dahl & Miralles 2012b; Christiansen 2005) induces a language's syntactic structure from that of another.

Presently, the PG formalism comprises the following seven categories (we adopt the handy notation of Duchier *et al.* (2013), and, the same example here):

Constituency $A : S$, children must have their categories in the set S
Obligation $A : \triangle B$, at least one B child
Uniqueness $A : B!$, at most one B child
Precedence $A : B \prec C$, B children precede C children
Requirement $A : B \Rightarrow C$, if B is a child, then also C is a child
Exclusion $A : B \not\Rightarrow C$, B and, C children are mutually exclusive
Dependency $A : B \sim C$, the features of B and, C are the same

Example 1. For example, if we denote determiners by D, nouns by N, proper nouns by PN, verbs by V, noun phrases by NP, and, verb phrases by VP, the context free rules $NP \rightarrow D\ N$ and, $NP \rightarrow N$, which determine what a noun phrase is, can be translated into the following equivalent constraints: $NP : \{D, N\}$, $NP : D!$, $NP : \triangle N$, $NP : N!$, $NP : D \prec N$, $D : \{\}$, $N : \{\}$.

In the PG formalism, A refers to the phrase or, sentence being tested. In our case, A refers to Noun phrases, B and, C are categories such as nouns, adjectives, etc. These properties contain the basic syntactic information. Other properties can be added if necessary, in particular for integrating knowledge coming from other linguistic domains or, for particular phenomena, like long distance dependencies.

The use of a dependency property makes it possible to express a dependency grammar using the formalism of property grammars. This is possible because, just like dependency grammars, dependency property is concerned with syntactico-semantic

relations. For instance np(dependence(adjective,noun)) is concerned with using features (described in section 4.2) to determine the gender and, number relationship between adjectives and, nouns, and, other dependency information can be modelled in similar fashion. The features are introduced for each word in the lexicon using the CHRG symbol word/3 (details in section 8.2). However, all the subsets of categories involved in the description of a given input can be characterized by the constraint system (i.e. the grammar). The constituency property contains dependency information that tell us each category can be associated to another.

Thus, there is a general notion of characterization formed by the set of properties that succeed and, those that fail which together characterize an input. The characteristics of an input corresponds to the result derived from the constraint system at the end of the parse. The result of a parsed property is either satisfied or, unsatisfied. The interest of such a conception is that every property in the grammar states some constraint on well formedness. It can be the case that all constraints are satisfied, but this is not an imperative condition. All kinds of input can consequently receive a characterization.

6 Background on Womb Grammars

6.1 Womb Grammars

Womb Grammar Parsing was motivated by the need to aid the world's linguist in describing the syntax of the many languages that are not being studied for lack of adequate resources. It induces the grammar of a target language from the grammar of a source language. The WG paradigm describes a language's phrases in terms of constraints over properties of pairs of direct daughters called properties. WG extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing.

WG was presented in two versions: (1) *Hybrid Womb Grammars*, in which the source language is an existing language for which the syntax is known, and, (2) *Universal Womb Grammars*, in which the source syntax is a hypothetical universal grammar of the authors' own devise, which contains all possible properties of pairs of constituents. Womb Parsing has the novel approach of addressing a problem which more usually is solved as a machine learning problem through constraint solving.

The grammar of our source language, English is described as properties of constituents as described by Blache (2004). We use the hybrid version of Womb Grammars, to which we feed the following English grammar:

Constituency *NP* : *determiner, noun, adjective, pronoun, proper-noun, quantifier*;
Obligation *NP* : $\triangle$*noun*; $pronoun$; $proper - nouns$
Precedence *NP* : *determiner* $\prec$ *noun*; *NP* : *determiner* $\prec$ *adjective*; *NP* : *pronoun* $\prec$ *noun*; *NP* : *pronoun* $\prec$ *adjective*; *NP* : *adjective* $\prec$ *noun*; *NP* : *quantifier* $\prec$ *determiner*; *NP* : *quantifier* $\prec$ *pronoun*; *NP* : *quantifier* $\prec$ *adjective*; *NP* : *quantifier* $\prec$ *noun*
Requirement *NP* : *noun* $\Rightarrow$ *determiner*
Dependency *NP* : *quantifier* $\sim$ *noun*; *NP* : *adjective* $\sim$ *noun*; *NP* : *determiner* $\sim$ *noun*

The nouns used in the noun phrases are proper nouns, common nouns, abstract nouns, pronouns etc and the features such as number, gender, etc are encoded in the features of the word/3 (more details in section 4.2 and 8.2). WG extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing. The first implementation of WG (Dahl & Miralles 2012b) allowed greater efficiency by adopting a failure-driven approach where only failed constraints where analyzed, rather than explicitly calculating all successful and, unsuccessful constraints. However in adapting the model into Yorùbá, we have found it more efficient to explicitly calculate satisfied constraints as well. This is partly because in order to arrive at as nuanced a description as needed for Yorùbá, we had to extend the Property Grammar model to accommodate what we call **conditional constraints**: those constraints which have failed in some phrases and, succeeded in others, but also have a unique pattern responsible for this behaviour (more on this later).

The general WG model can be described as follows:

Let L^T be the target language and, its lexicon (L^T_{lex}). Let E^T be a given expression (a meaningful and, grammatical expression) of a target language L^T. Likewise, let L^S be the source language. Its syntactic component will be denoted L^S_{syntax}. Let Seq^S be a sequence of words of L^S, which is a glossing correspondence i.e a literal transliteration (does not need to be meaningful and, or, grammatical) of E^T and, E^S be an expression of the language L^S which is a meaningful translation of E^T. The translation expression E^S in L^S of E^T in L^T should be meaningful and grammatical. If we can get hold of a sufficiently representative set of phrases, E^T, in L^T that are known to be correct (a set in which our desired subset of the target language will be represented), we can feed these to a hybrid parser consisting of L^S_{syntax} E^T, and, L^T_{lex}. 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^S_{syntax} so that the parser accepts as correct the

sentences in the corpus of L^T—i.e., how to transform L^S_{syntax} into L^T_{syntax} by modifying the constraints that were violated into constraints that accept the input. Seq^S and E^S are thus the gloss and translation expression of E^T, incoporated to enable intelligibility for non-speakers of the target language. Figures 3 and 4 respectively show the problem and, our solution in schematic form.

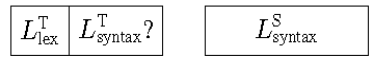


Fig. 3. The Problem.

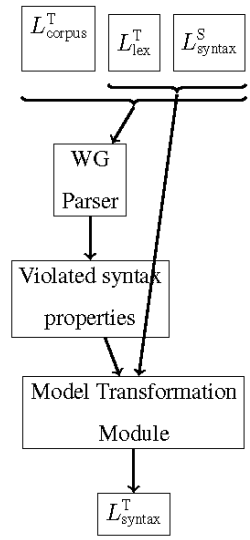


Fig. 4. The Solution.

An Example

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 are always postpositioned

(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 precedence property of L^S_{syntax} will be declared unsatisfied when hitting phrases such as “ìwé pupa kan”. The 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 constraints in L^S_{syntax} so that none are violated by the corpus sentences.

Some of the necessary modifications are easy to identify and, to perform, e.g. for accepting “ìwé pupa kan” we only need to delete the (English) precedence requirement of adjective before noun (denoted by $adj < n$). 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 of English: $n < adj$. So in this case, not only do we need to delete $adj < n$, but we also need to add $n < adj$.

6.2 Constraint Handling Rule Grammar (CHRG)

Constraint Handling Rule Grammars (CHRG) are a constraint-based grammar formalism added on top of Constraint Handling Rules (CHR) (Frühwirth 1998) similar to the way Definite Clause Grammars are implemented over Prolog. A CHRG works as a bottom-up parser, and, the formalism provides several advantages for natural language analysis. A notable advantage is that the notation supports context-sensitive rules that may consider arbitrary symbols to the left and, right of a sequence (Christiansen 2005).

CHRG (CHR Grammars) strives to incorporate as much as possible of CHR’s expressibility. This includes **propagation**, **simplification**, and, **simpagation** rules and, the capacity for a grammar rule to make considerations for arbitrary grammar symbols that may occur to the left and, right of a sequence that is supposed to match a given nonterminal. Ambiguity is not a problem for CHRG, all different parses are evaluated in parallel in the same constraint store –without backtracking. This results in tagger-like grammar rules that can be used for disambiguating simple and, ambiguous context-free grammar rules, and, provides also a way to handle coordination in natural language. In case of errors, the parser is robust enough to deliver

as its result those subphrases that have been recognized. Since the advent of CHR (Frühwirth 1998) and, of its grammatical counterpart CHRG (Christiansen 2005), constraint-based linguistic formalisms can materialize through fairly direct methodologies.

In a nutshell, CHRG rules rewrite constraints into other constraints, that are subject to possible checks described in a rule's guard and, stated as Prolog calls.

Their general format is:

$$\alpha \ -\backslash \ \beta \ /- \ \gamma \ \texttt{::>} \ G \ \texttt{|} \ \delta \texttt{.}$$

This rule is called a *propagation* rule and, it means, if α occurs before β, and, β before γ, and, G condition is fulfilled, then store δ in the constraint store. The part of the rule preceding the arrow `::>` is called the head, G the guard, and, δ the body; $\alpha, \beta, \gamma, \delta$ are sequences of grammar symbols and, constraints so that β contains at least one grammar symbol, and, δ contains exactly one grammar symbol which is a nonterminal (and, perhaps constraints); α (γ) is called *left (right) context* and, β the *core* of the head; G is a conjunction of built-in constraints as in CHR. If left or, right context is empty, the corresponding marker is left out and, if G is empty (interpreted as **`true`**), the vertical bar is left out. The convention from DCG is adopted that constraints (i.e., extra-grammatical stuff) in head and, body of a rule are enclosed by curly brackets. Special grammatical operators are provided for gaps and, parallel matching (parallel matching is used when it is required that more than one pattern matches the string to be recognized). Gaps and, parallel match are not allowed in rule bodies. A gap in the rule heads is denoted "..." while parallel match is denoted by "$$". Gaps are used to establish references between two long distant elements while a parallel match is useful when more than one pattern is needed to match the string to be recognized (Dahl & Miralles 2012b).

A *simplification (grammar) rule* is similar to a propagation rule except that the arrow is replaced by `<:>` and, the grammar symbols in the head are deleted.

$$\alpha \ -\backslash \ \beta \ /- \ \gamma \ \texttt{<:>} \ G \ \texttt{|} \ \delta \texttt{.}$$

A **simpagation (grammar) rule** is similar to a simplification rule, except that one or, more grammar symbols in its head are prefixed by a ! symbol, which means that such a grammar symbol is not removed. It can also be fully integrated with Prolog and, a single file can conveniently incoporate prolog, CHR and, CHRG codes.

$$!\alpha \ -\backslash \ !\beta \ /- \ !\gamma \ \texttt{<:>} \ G \ \texttt{|} \ \delta \texttt{.}$$

Constraint solving as an implementation paradigm for WG allows us to convey linguistic constraints and, also test them in a modular fashion. The results are also concise.

7 Some Implementation Details

Our CHRG implementation adapts the original implementation (Christiansen 2005) where the appropriate WG constraints are entered in terms of a constraint g/1, whose argument stores each possible grammar property. For instance, our English grammar hybrid parser for noun phrases includes the constraints below and, some others described in section 6.1:

g(np(obligatority([noun, proper-noun, pronoun]))) - at least one noun or, pronoun or proper-noun child
g(np(constituency(determiner))) - children must have their categories in the set
g(np(precedence(determiner, adjective))) - $determiner$ children precede $adjective$ children
g(np(requirement(noun, determiner))) - if $noun$ is a child, then also $determiner$ is a child
g(np(dependence(determiner, noun))) - the features of $determiner$ and, $noun$ are the same

and, so on.

These properties are weeded out upon detection of a violation by CHRG rules that look for them, e.g. an input noun phrase where an adjective precedes a noun may provoke deletion of the constraint g(precedence(noun, adjective)) when the following CHRG rule applies:

```
!word(C2,F1,_):(N1,_N2),...,
!word(C1,F2,_):(_, N4),
{g(np(precedence(C1,C2)))}, {all(A, B)}<:>
{fail(np(precedence(C1,C2)), F1, F2, N1, N4, A, B)}.
```

Fig. 5. Simpagation Rule.

The parser works by storing each Yorùbá word in the lexicon in a CHRG symbol word/3 along with its category and, features (i.e. word(pronoun,[[plural, n/a], personal, both, third-person,[low,mid]], won)). Since the CHRG parse predicate stores

and, abstracts the position of each word in the sentence, this simpagation rule in Figure 5 is triggered when a word of category C2 comes before category C1, given the existence of the grammar constraint that C1 must precede C2 and, the phrase boundaries in which the word is found[2]. Each of the properties dealt with has similar rules associated with it.

When the parse predicate is invoked, the constraints of the source grammar are tested with the input phrases of the target language. Constraints that are violated at least once are output as unsatisfied with the number of times they are unsatisfied and, the features involved. Constraints that are satisfied on the other hand are output as satisfied with the number of times they are satisfied, the features of the words involved and, the phrase boundaries of those two words.

8 Methodology

8.1 Modified Parsing that Calculates both Failure and Success Explicitly

Previous models of WG focused on failure driven parsing, under the assumption that failed properties are usually the complement of those satisfied, so satisfied properties 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 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 the words in each input phrase and, the counts of the

[2] Recall that in CHRG syntax the symbols prefixed with exclamation marks are kept, while the ones without are replaced by the body of the rule. The :(A,B) are used to retrive the word boundaries, ... signifies a gap greater or equal to zero, and, all(A,B) retrieves the phrase boundaries.

property (The count of a property is the number of times it occurs, in either the failed list, or, the succeeded list, and, each time a property is added to any of these lists, the count is updated for that property). It is important to state that while other constraints are tested for success and, failure, 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. More of this in section 8.5.

8.2 How we Handle Tones

Yorùbá is a tone language, therefore it was imperative to ensure that tones are properly represented. We adopt a full tone marking approach (as described in section 3.2) so that each word is marked with its tones in order to avoid ambiguities. The tones are also stored as features, so that for instance the word tútù (cold) that has two syllables, one with a high and, the other with a low tone. Each syllable respectively is stored as "High-low" in the features of the word. For instance:

[tútù]::> $word(adjective, [[neutral, n/a], descriptive, both, [High-low]]$, tútù).

In the example, the category of the word is adjective, the features of the word are enclosed in square brackets and, the word is tútù. These tones are used to infer conditional properties in the second phase of parsing. So that if, for instance, a property is found to succeed only if it is around words with certain tones, such property will be said to be conditional. The conditions of such a property will be the tones which form a unique pattern in the success of the property. It is important to note that the tones are not used in isolation of other features and, a property will be said to be conditioned on tones alone, if, tones are the only patterns responsible for succeess or, failure. The addition of tones makes our approach extensible to many other languages including tone languages. In the case of non tone languages, the tones will not be necessary and, can be marked with n/a or may be replaced with any other traits useful for the language.

8.3 Glossing of Phrases

The system provides a record of the English gloss of every word in the target phrases and, this record is consulted during parsing to produce the English equivalent for each word in the input Yorùbá phrase. The gloss is rendered below every Yorùbá phrase, and, the free translation in English comes immediately after. As in:

$< 0 >$ àwọn $< 1 >$ ọmọ $< 2 >$ púpọ̀ $< 3 >$

they	child	plenty
plenty	children	

Fig. 6. Phrase Transliterations.

8.4 How Conditional Properties are Induced

The original Property-based model, as we have seen, succeeded in detecting and, signalling mistakes in the sentence, without blocking the analysis. In this first model, which is parsing-oriented, incorrect sentences could be "accepted" through declaring some constraints as relaxable. For instance, while from the context-free grammar rules shown in section 5, we wouldn't be able to parse "the the book" (a common mistake from cutting and, pasting in word processors), but this is possible, if we relax the uniqueness of determiner constraint in the constraint-based formulation.

Relaxation can be made conditional (e.g. a head noun's requirement for a determiner can be made relaxable in case the head noun is generic and, in plural form, as in "Lions sleep at night"). The failure of relaxable constraints is signalled in the output, but does not block the entire sentence's analysis. Implementations not including constraint relaxation capabilities implicitly consider all properties as relaxable. And in fact, when exploiting constraint-based parsing for grammar transformation rather than for parsing, this is exactly what we need, since in an unknown language any constraint may need to be "relaxed" and, even if needed, corrected.

For that reason, we have considered it more appropriate, in the context of grammar induction, to state "exceptions" as part of a property itself, rather than separately in a relaxation definition. Thus we have created conditional properties, e.g. "conditional precedence(pronoun, noun)" which expresses that a pronoun preceeding a noun is conditional. We give the condition which says: "if the pronoun is plural and it is a personal pronoun and, can be used in place of both animate and, inanimate nouns". This means that a pronoun preceeds a noun only if the condition holds and, the opposite ordering, if otherwise. We define a property as conditional if it happens that there occurs a pattern for which the property fails for some phrases and, succeeds for others. All constraints that fail and, succeed are tested in this phase. We find these properties by searching for features which are unique to the failure and, or, success of these constraints.The features which are responsible for this difference in behaviour form the condition under which such property succeeds or, fails.

Let us note that requirement(noun, determiner) is in fact a conditional property of English and, the condition under which this requirement property will not hold is if the noun is generic and, plural. So that while count and, or, non generic form of a noun require a determiner (e.g “The boy” and, “The boys”), a plural generic noun does not require a determiner (e.g Chickens lay eggs).

8.5 Failed and Satisfied Properties

For each phrase parsed, the failed and, or, satisfied properties are explicitly output. This enables us to identify which specific properties are satisfied for each phrase and, those which are unsatisfied. We also explicitly retrieve the word boundaries of every word analysed during parsing as well as the phrase boundaries where the words are found. This is basically to eliminate ambiguity that can occur in a situation where a single property fails or, succeeds more than once in a phrase, or, if the model is extended to include other phrases, for example a verb phrase, because the model for now is limited to noun phrases. The boundaries before the semicolon represent the pair of word boundaries while the second boundary after the semicolon represents the phrasal bondaries.

```
< 0 > àwọn < 1 > ọmọ < 2 > púpọ̀ < 3 >

    they      child    plenty
    plenty  children

    Succeeded Property: np(obligatority([noun,proper-noun,pronoun]))1-2; 0-3

    Succeeded Property: np(obligatority([noun,proper-noun,pronoun]))0-1; 0-3

    Succeeded Property: np(constituency(pronoun))0-1; 0-3

    Succeeded Property: np(constituency(adjective))2-3; 0-3

    Succeeded Property: np(constituency(noun))1-2; 0-3

    Succeeded Property: np(precedence(pronoun,noun))0-2; 0-3

    Succeeded Property: np(precedence(pronoun,adjective))0-3; 0-3

    Failed Property: np(precedence(adjective,noun))1-3; 0-3

    Failed Property: np(requirement(noun,determiner))1-2; 0-3
```

Fig. 7. Parse Results of a Sample Phrase.

In Figure 7, the parse results show that obligatority succeeds twice because there's a noun and, a pronoun in the phrase, three constituency rules succeed because there are three different categories and, two precedence rules succeed. The failed requirement and, precedence properties are also printed.

8.6 Inducing Properties not Present in the Source Language

We also have an additional functionality that finds all precedence and, constituency constraints that are absent in our source but present in our target language.

```
prec(precedence(X,Y)):-
english_properties(List),
not(member(g(np(precedence(X,Y))), List)),
not(member(g(np(precedence(Y,X))), List)),
X \= Y.

word(C1,F1,_):(N1,_),...,word(C2,F2,_):(_,N2),
{others}, {all(A, B)}::> prec(precedence(C1,C2))
|{succeed(np(precedence(C1,C2)),F1, F2, N1, N2, A, B)}.
```

Fig. 8. Inducing Precedence Properties Present in Target Phrase but Absent in Source Grammar.

Figure 8 shows that during parsing, every pair of words with categories `C1` and, `C2`, features `F1` and, `F2` and, word boundaries `N1` and, `N2` are tested for precedence if and only if, the precedence rule of those words' categories doesn't already exist in any precedence rule of the property grammar of English. We define a predicate `prec` which checks that the proposed precedence rule and, its converse is not a member of the property grammar of English while we use `others`, a CHRG constraint to trigger this precedence rule. We do this in order to ensure that we do not create redundancies in a bid to infer precedence rules that occur only in the target language.

The model has the capacity to test the precedence constraints by checking if such a constraint succeeds for all relevant phrases and, if the converse succeeds for any input phrase. If we find that the converse does not succeed in any of the input phrases, this constraint is induced as a property of the target language, else we search for a unique feature to ascertain if it should be induced as a conditional feature. In the case of constituency, like we explained in section 8.1, we are only interested in finding the constituent and, not the number of times it occurs or, otherwise. If a constituent not present in the source grammar is found at least once, in the input phrase of the

target language, then the constituent is induced as part of the grammar of the target language.

8.7 Parsing each Property in our English Subset

Parsing with Dependence Constraints

Due to the structure of Yorùbá grammar with regards to gender and, number, detailed in section 3.5, we initialize all words with neuter gender and, dual number, save a few words which have their number inherent in their meaning and, function (words such as ọ̀pọ̀lọpọ̀ (numerous), àwọn (third person plural). We also implement a predicate that creates a variant for each number or, gender tag based on the environment in which the word is found. This successfully handles the syntactic process of defining number and, gender in Yorùbá. The dependency rule of English is then tested for satisfaction and, otherwise.

This parsing strategy clearly accomodates the existence of plurality in Yorùbá without a previous knowledge of the peculiarities of Yorùbá's syntactic strategies for number, as well as gender which is in reference to English.

However, though plurality and, gender are present in Yorùbá grammar and, are seen to succeed in parse results, the presence of failed dependence relations is further analyzed to determine if the dependency constraints should be a conditional property in Yorùbá. Results show that there is no pattern under which failure or, success occurs. This lays credence to research findings by linguists who have studied the Yorùbá grammar (Awóbùlúyí 1978; Bámgbósé 1966; Ajiboye 2006).

Parsing with the Obligatority Constraint

Previous models of Womb Grammars failed if a noun was absent; the obligatority constraint in our version of WG only fails if a phrase doesn't have a noun or, pronoun or, proper noun constituent as described in section 8.4. Obligatority succeeds if the converse occurs. This idea eliminates the possibility of obligatority failing because it contains a personal noun which is a valid obligatory category.

```
word(C1,F1,_):(N1,N2),{g(np(obligatority(C)))},
{all(A,B)}::> member(C1,C)|
{succeed(np(obligatority(C)),C1,F1,N1,N2,A,B)}.
```

Fig. 9. Rule for Succeeding Obligatority.

In Figure 9, we declare a list of categories `C` which contains a noun, proper-noun and, pronoun. The obligatority constraint is satisfied if `C1` is a member of `C`.

This approach enables us to check if pronouns and, proper nouns in Yorùbá perform the same head function as they do in English. Our results indicate that proper nouns and, pronouns can indeed function as the head of a phrase as in English.

Parsing with Precedence Constraints

In parsing a precedence constraint, we use the precedence rules of English. A precedence rule is satisfied if the target phrase has exactly the same ordering as English. For instance in Yorùbá, every occurrence of `precedence(det,n)`, `precedence(det,adj)` fails while `precedence(adj,n)` fails in certain input phrases and, succeeds in other input phrases (details of this are given in Appendix 1). This gives us adequate information to make a valid conclusion that in Yorùbá determiners always come after nouns and, after adjectives. However, with the presence of both failed and, successful instances of adjectives and, noun orderings, and, a pattern of features responsible, we can only make conclusions that Yorùbá nouns and, adjectives have different orderings. The system materializes this in our output as:

```
-conditional precedence(adjective,noun):-
adjective preceeds noun if adjective is plural,
quantity-uncountable, can be used for both animate and inanimate
nouns
```

Fig. 10. A Conditional Precedence Property.

Our results in Figure 10 reiterate linguistic findings by Yorùbá grammarians (Awóbùlúyí 1978; Bámgbósé 1966).

```
word(C1,F1,_):(N1, _),...,word(C2,F2,_):(_, N4),
{g(np(precedence(C1,C2)))}, {all(A, B)}::>
{succeed(np(precedence(C1,C2)), F1, F2, N1, N4, A, B)}.
```

Fig. 11. Precedence Rule for Checking Success.

In Figure 11 we show from our code that precedence constraints are satisified if and, only if a there exists `precedence(C1,C2)` rule in English, and, an input phrase where `C1` preceeds `C2`. The categories `C1`, `C2` and, the features `F1`, `F2` of the words are stored in a list if the precedence rule is satisfied. We also retrieve the word boundaries and, phrase boundaries as described in section 8.5.

Parsing with Constituency Constraints

A success driven approach is adapted for parsing constituency. This is because we are only interested in the presence of the constituent and, not the number of times it occurs in parsed phrases. Previous Womb Grammar formalisms didn't test for constituents and, this resulted in adding constituents that are absent in the target grammar to the target grammar simply because they exist in the source grammar. We also have a function that can induce constituents that are in the target language but absent in the source, more of this in section 8.6.

```
word(C1,F1,_):(N1, N2),{g(np(constituency(C1)))},
{all(A, B)}::>
{succeed(np(constituency(C1)), F1, N1, N2, A, B)}.
```

Fig. 12. The Constituency Rule.

In Figure 12, every time a word is found, the category of the word is tested with the constituency rule of English. If the category of the word is found in the English properties, then it is satisfied and, added to a succeeded properties list and, subsequently induced. We also store the word boundaries and, phrase boundaries as in section 8.5

Parsing with Requirement Constraints

We test requirement by checking the constituent that co-occur with another constituent in a phrase.

```
check_waits \ wait(Prop, F1, F2, N1, N2, A, B), g(Prop) <=>
fail(Prop, F1, F2, N1, N2, A, B).
check_waits <=> true.

word(Required,F1,_):(N1,N2),
{g(np(requirement(Required,Requiring)))},{all(A, B)} ::>
{wait(np(requirement(Required,Requiring)), F1, [], N1, N2, A, B)}.
{wait(np(requirement(_,Requiring)),_,_,_,_,_,_)},
!word(Requiring,_,_)<:>true.
```

Fig. 13. Rules to Test Requirement Failure.

In Figure 13, we test whether requirement fails by checking that as daughter of the same mother a `Requiring` category doesn't occur with a `Required` category. We achieve this by saving the constraint `wait/7` everytime we find a word with a `Requiring` category and, there exists a property that says `Requiring` category requires a `Required` category. We continue to `wait` till the entire phrase is parsed, if peradventure `Required` will be a category in the phrase which should not necessitate a failure. If no `Required` is found within that phrase, the category `Requiring` and, its features `F2`, its word boundaries `N1` and, `N2` and, the phrase boundaries `A`, `B` where `Requiring` was found are added to the failed properties list; if `Required` is found, we delete `wait/7`.

```
assign(A,B,C,D,A,B,C,D).

!word(X,FX,_):(N1,_N2),...,!word(Y,FY,_):(_,N4),
{g(np(requirement(C1,C2)))},{all(A,B)} <:>
X = C1, Y = C2 -> assign(X, FX, Y, FY, Requiring, FRg, Required, FRd);
X = C2, Y = C1 -> assign(Y,FY,X,FX,Requiring,FRg,Required,FRd) |
{succeed(np(requirement(Requiring, Required)), FRg, FRd, N1, N4,A,B)}.
```

Fig. 14. The Rule to Test Requirement Success.

In testing for success as in Figure 14, if a daughter `Requiring` of a phrase requires a sibling `Required`, we must check that in that phrase (e.g. np) when `Requiring` appears as a daughter, `Required` is also a daughter of that phrase. In the implementation, we check if a `Requiring` and, `Required` occur in the same phrase. The order of the categories is not important here, as we are simply checking if it occurs anywhere within a phrase. We store the features `F1`, `F2` of the word to induce conditional properties in the second phase of parsing. We also store the word and, phrase boundaries which were retrieved as in section 8.5.

9 Results and Supporting Evidence of Correctness

Our results so far have been consistent with linguistic research of Yorùbá grammar which we prove for a fragment of NPs. We also use phrases generated by our Context Free Grammar (CFG), which we wrote for a fragment of NPs considered as the evaluation data in our WG model and, the induced grammar have shown similarities with the CFG. Details of the induced grammar are given in Appendix 1.

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, context free formalisms are unable to directly accommodate extra information such as features, and, their related linguistic constraints, as their grammar symbols can only be simple identifiers, with no possibility of carrying extra arguments, as logic grammar symbols can, to transmit and, consult such extra information.

We summarize our results into constituency, precedence, requirement, dependency and, obligatority.

1 Constituency: Our constituency results show that in Yorùbá, nouns, pronouns, proper-nouns, adjectives, quantifiers and, determiners are allowable categories in noun phrases. These constituents are featured in noun phrases described by Ogunbowale (1970), Awóbùlúyí (1978) and Bámgbósé (1966) as well as in our CFG of Yorùbá.
2 Precedence: We induce two conditional precedence properties (although conditional precedence(adjective, noun) has two different conditions), and, nine precedence properties. Our conditional precedences imply that there are two orderings, which support different orderings in pronouns and, nouns, and, in adjectives and nouns, in for instance, in rules 10 and 11, and, 14 and 15 of our CFG, and, also in literature (Ogunbowale 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 of Awóbùlúyí (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 with determiners, as well as rules where nouns occur without determiners. This conclusion is also presented in research (Bámgbósé 1966; Awóbùlúyí 1978; Ajiboye 2006).

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 Bámgbósé (1966), but not explicit in our CFG of Yorùbá.
5 Obligatority: Obligatority succeeded in all input phrases, showing that at least one of noun, pronoun and, proper-nouns is a compulsory constituent of Yorùbá NPs. Our CFG also shows these three constituents occur at least once in all rules for NPs.

10 Possible Extensions

We covered a useful subset of Yorùbá noun phrases, but there are some phenomena not covered here. For instance, there are some words that are noun phrases with a determiner interfixed between two occurrences of the same words. For example, ọmọkọ́mọ (child any child), ilékílé (house any house), asọkásọ (dress any dress), ìwàkuwà (behaviour any behaviour), ìgbàkigbà (time any time) etc. In certain contexts, these words have different meaning, as in, ọmọkòmọ (child bad child), ilèkìlè (house bad house), ìwàkuwà (behaviour bad behaviour) etc and, these contexts, determine the categories that can occur around them. It would be interesting to develop a parser that accomodates these features. Extending our program to identify and, parse features such as vowel elision, assimilation of tones, compensatory lengthening and, other phonological processes that have grammatical functions in Yorùbá would also be helpful.

Further interesting ramifications of our work would also include: testing our system for a larger coverage of syntax (we have addressed noun phrases so far), inducing other constraints (besides precedence and, constituency constraints which we induced) present in the source language but not in the target language, testing our model on other tonal-sensitive languages, studying how assimilation influences language structure especially in tone languages, studying whether any further constraints or approach extensions would be needed to accommodate families of languages not easily describable within our approach (e.g. languages which have different categories from the source language, those who have different inflectional paradigms from the source language, those that exhibit constraints not among our allowable set of constraints).

11 Conclusions

We have shown the simplicity with which Womb Grammar automatically induces the grammar of Yorùbá from that of English despite the peculiarities in the grammar of Yorùbá and, English that can make this very difficult. This makes our model very useful in language development and, language documentation.

We have also presented a concrete system, for inducing Yoruba grammar that constitutes a proof of concept for the Womb Grammar model (Dahl & Miralles 2012b). WGallows users to input English grammar description in modular and, declarative fashion, e.g, in terms of the linguistic constraints that relate to the subset of noun phrases we consider (constituency, precedence, dependency, obligatority, requirement). Since such descriptions also stand alone as a purely linguistic model of a language's syntax, our system can cater even for users who are not conversant with coding, such as pure linguists wanting to perform syntactic model transformation experiments aided by computers. Their purely (constraint-based) linguistic descriptions of syntax automatically turn into executable code when appended to our code. We also used a subset of noun phrases which can accept any number of adjectives making our model extensible to induce even finer details such as precedence order of adjectives if we so desire.

As we have seen, our system automatically transforms a user's syntactic description of a source language into that of a target language, i.e Yorùbá, 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 descriptions can be done in terms of the supported constraints.

We have thus adapted a flexible modelling framework solidly grounded in linguistic knowledge representation which, through the magic of constraint solving, turns into an efficiently executable problem solving tool. We maintain the modularity feature of the first implementation in our adaptation, therefore, our model can support revision and, adaptation quite straightforwardly: new constraints between immediate daughters of any phrase can be modularly added and, implemented, without touching the rest of the system.

Our model allows for accommodating solution revisions at execution time by interacting with a linguist whenever constraints can not be induced consistently. Visual representations of the output that would be helpful for linguists in this respect have been explored in Adebara & Dahl (2015).

Appendix 1: Results

Due to space constraints, the failed-properties and, succeeded-properties list as well as the phrases parsed have been removed from this results. We present only a subset of the induced grammar.

```
Induced Property Grammar rules for Yoruba Noun phrases::-
-conditional precedence(pronoun,noun):-
pronoun preceeds noun if  pronoun is  plural, personal,
and can be used for both animate and inanimate nouns,
and third-person.
-conditional precedence(adjective,noun):-
adjective preceeds noun if adjective is plural,
quantity-uncountable,  an can be used for both animate and
inanimate nouns.
-obligatority((noun;proper-noun;pronoun)):-
noun;proper-noun;pronoun are obligatority.
It succeeds in all 47 input phrases.
noun succeeds in 39 relevant phrases;
pronoun succeeds in 4 relevant phrases;
proper-noun succeeds in 4 relevant phrases;
-precedence(quantifier,pronoun):-
quantifier preceeds pronoun in all 4 relevant phrases
-precedence(quantifier,adjective):-
quantifier preceeds adjective in all 3 relevant phrases
-precedence(pronoun,adjective):-
pronoun preceeds adjective in all 3 relevant phrases
-precedence(pronoun,determiner):-
pronoun preceeds determiner in all 7 relevant phrases
```

Acknowledgments

This research was supported by V. Dahl's NSERC Discovery grant 31611024.

References

1. Adebara, I. & Dahl, V. (2015). Domes as a Prodigal Shape in Synthesis-Enhanced Parsers. In Kutz, O., Borgo, S. & Bhatt, M. (eds.), *Proceedings of the Third Interdisciplinary Workshop SHAPES 3.0* (pp. 23–33). Larnaca: CEUR Workshop Proceedings 1616.
2. Adebara, I., Dahl, V. & Tessaris, S. (2015). Completing mixed language grammars through womb grammars plus ontologies. In Loiseau, S., Filipe, J., Duval, B. & van den Herik, J. (eds.), *ICAART 2015. Proceedings of the International Conference on Agents and Artificial Intelligence*, Volume 1 (pp. 292–297). Lisbon: SciTePress.

3. Adebara, I. & Dahl, V. (2015). Shape Analysis as an Aid for Grammar Inductions. In Kutz, O., Borgo, S. & Bhatt, M. (eds.), *Proceedings of the Third Interdisciplinary Workshop SHAPES 3.0* (pp. 55–57). Larnaca: CEUR Workshop Proceedings 1616.
4. Adegbija, E. (2004). Language Policy and Planning in Nigeria. *Current Issues in Language Planning*, 5(3): 181–246.
5. Adeyanju, S., Adegbola, T. & Fakinlede, O. (2015). A Supervised Phrase Selection Strategy for Phonetically Balanced Standard Yorùbá Corpus. In Gelbukh, A. (ed.), *Proceedings of the Computational Linguistics and Intelligent Text Processing 16th International Conference, CICLing 2015* Part II (pp. 565–582). Switzerland: Springer.
6. Ajiboye, O. (2006). Interpreting Yorùbá Bare Nouns as Generic. In *Berkeley Linguistic Society Proceedings of the Annual Meeting*, BLS 31.
7. Awóbùlúyí, O. (1978). *Essentials of Yorùbá Grammar*. Oxford: Oxford University Press.
8. Awoyale, Y. (1988). *Complex predicates and verb serialization*. Lexicon Project Working Papers 28. Cambridge: MIT.
9. Awoyale Y. (2008). *Global Yorùbá Lexical Database v. 1.0 Linguistic Data Consortium*. Philadelphia. Retrieved on 28/10/2015 from Natural Language Processing Lab, School of Computing Science, Simon Fraser University, British Columbia.
10. Bámgbósé, A. (1966). *A Grammar of Yorùbá*. Cambridge: Cambridge University Press.
11. Bámgbósé, A. (1993). Deprived, Endangered and Dying Languages. *Diogenes*, 41(161): 19–25.
12. Blache, P. (2000). Property grammars and the problem of constraint satisfaction. In *Proceedings of ESSLLI 2000 workshop on Linguistic Theory and Grammar Implementation* (pp. 47–56).
13. Blache, P. (2004). Property Grammars: A Fully Constraint-Based Theory. In Christiansen, H., Rossen Skadhauge, P. & Villadsen, J. (eds.), *Constraint Solving and Language Processing* (pp. 1–16), Lecture Notes in Computer Science 3438. Berlin: Springer.
14. Blache, P. & Rauzy, S. (2012). Hybridization and Treebank Enrichment with Constraint-Based Representations. In *Proceedings of LREC-2012* (pp. 6–13). Istanbul.
15. Bowen, T.J. (1858). *Grammar and Dictionary of the Yorùbá Language*. Smithsonian Institution.
16. Brill, E. & Magerman, D., Marcus, M. & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 275–282). Hidden Valley: Addison Wesley Longman.
17. Burkett D.& Klein D. (2008). Two Languages are Better than One (for Syntactic Parsing). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 877–886). Honolulu, Hawaii: Association for Computational Linguistics
18. Chelliah, S.L & de Reuse, W.J. (2010). *Handbook of Descriptive Linguistic Fieldwork*. Berling: Springer.
19. Chomsky, N. (1957). *Syntactic Structures*. Berlin: Walter de Gruyter.
20. Christiansen, H. (2005). CHR grammars. *Theory and Practice of Logic Programming*, 5(4-5): 467–501.
21. Cohen, S.B. & Smith, N.A. (2010). Covariance in Unsupervised Learning of Probabilistic Grammars. *Journal of Machine Learning Research*, 11: 3017–3051.

22. Dahl, V. & Blache, P. (2004). Directly Executable Constraint Based Grammars. In *Proceedings of the Journees Francophones de Programmation en Logique avec Contraintes* (pp.149–166). Angers.
23. Dahl, V. & Miralles, J. (2012a). Womb grammars: Constraint solving for grammar induction. In Sneyers, J. & Frühwirth, T. (eds.) *Proceedings of the 9th Workshop on Constraint Handling Rules* (pp. 32–40). Technical Report CW 624. Leuven: Department of Computer Science, K.U. Leuven.
24. Dahl, V. & Miralles, E. (2012b). Womb Parsing. In Sneyers, J. & Frühwirth, T. (eds.) *Proceedings of the 9th Workshop on Constraint Handling Rules* (pp. 32–40). Technical Report CW 624. Leuven: Department of Computer Science, K.U. Leuven.
25. Duchier, D., Dao, T.B.H & Parmentier,Y. (2014). Model-Theory and Implementation of Property Grammars with Features. *Journal of Logic and Computation*, 24(2): 491–509.
26. Fabunmi, F. & Salawu, A. (2005). Is Yorùbá an Endangered Language? *Journal of African Studies*, 14(3): 391–408.
27. Fagborun, O. (1994). *The Yorùbá Koine - its history and linguistic innovations*. Munchen: Lincom Europa.
28. Fagborun, J. G. (1989). Some Practical Problems in Yorùbá Orthography. *The Journal of West African Languages*, 19(2): 74–92
29. Frühwirth, T.W. (1998). Theory and practice of constraint handling rules. *The Journal of Logic Programming* 37(1-3), 95–138.
30. Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2): 155–171.
31. Nicolas L., Molinero, M.A., Sagot, B., Trigo, E., de La Clergerie, E., Pardo, M.A., Farré, J. & Miquel Vergés, J. (2009). Towards efficient production of linguistic resources: the Victoria Project. In Angelova, G. & Mitkov, R. (eds.), *Proceedings of the International Conference RANLP-2009* (pp. 318–323). Borovets: Association for Computational Linguistics.
32. Ogunbowale, T. (1970). *Essentials of the Yorùbá language*. London: University of London Press.
33. Olumuyiwa, T. (2013). Yorùbá Writing: Standards and Trends. *Journal of Arts and Humanities*, 2(1): 40–51.
34. Pollard, C. & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
35. Romaine, S. (2002). The Impact of Language Policy on Endangered Languages. *IJMS. International Journal on Multicultural Societies*, 4(2): 194–212.
36. Sun, M. & Bellegarda, J.R. (2011). Improved Pos Tagging For Text-to-Speech Synthesis. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5384–5387). Prague.
37. Trudgill, P. (1986). *Dialects in Contact*. Oxford: Blackwell.
38. Voutilainen, A. (2012). Part-of-Speech Tagging. In Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press
39. Wang, Y. & Li, T. (eds.) (2011). *Knowledge Engineering and Management*. Berlin: Springer.

40. Wurm, S. (1998). Methods of Language Maintenance and Revival, with Selected Cases of Language Endangerment in the World. In Matsumura, K. (ed.), *Studies in Endangered Languages* (pp. 191-211). Tokyo: Hituzi Syobo.

Author's Biodata

Ife Adebara is a research masters student at the School of Computing, Simon Fraser University, British Columbia, Canada under the supervision of Veronica Dahl. Ife's research interests are in computational linguistics, translation, constraint programming, constraint handling rule grammars, inducing the grammar of under-resource languages using womb grammars.