Proceedings of SIGUL2022 @LREC2022, pages 146–154
Marseille, 24-25 June 2022

Automatic Detection of Morphological Processes in the Yorùbá Language

Tunde Adegbola
African Languages Technology Initiative
11 Oluyole Way, New Bodija, Ibadan, Nigeria
taintransit@hotmail.com

Abstract

Automatic morphology induction is important for computational processing of natural language. In resource-scarce languages in particular, it offers the possibility of supplementing data-driven strategies of Natural Language Processing with morphological rules that may cater for out-of-vocabulary words. Unfortunately, popular approaches to unsupervised morphology induction do not work for some of the most productive morphological processes of the Yorùbá language. To the best of our knowledge, the automatic induction of such morphological processes as full and partial reduplication, infixation, interfixation, compounding and other morphological processes, particularly those based on the affixation of stem-derived morphemes have not been adequately addressed in the literature. This study proposes a method for the automatic detection of stem-derived morphemes in Yorùbá. Words in a Yorùbá lexicon of 14,670 word-tokens were clustered around "word-labels". A word-label is a textual proxy of the patterns imposed on words by the morphological processes through which they were formed. Results confirm a conjectured significant difference between the predicted and observed probabilities of word-labels motivated by stem-derived morphemes. This difference was used as basis for automatic identification of words formed by the affixation of stem-derived morphemes.

Keywords: Unsupervised Morphology Induction, Recurrent Partials, Recurrent Patterns, Stem-derived Morphemes, Word-labels.

1. Introduction

The automatic detection of morphological influences in words found on a simple list obtained from a reasonably sized corpus of unannotated written texts in natural language is an important problem in computational linguistics. There are widely varying morphological strategies for the formation of words from morphemes as sub-word elements in various natural languages. This presents a computational problem that needs to be addressed. There is a need to develop efficient algorithms that can be used to automatically identify morphemes as well as morphemic boundaries effectively in most, if not all of the languages spoken worldwide. As in all data-driven approaches to the processing of natural language, resource-scarcity poses a problem in the automatic induction of morphology.

Valuable work has been done in the unsupervised automatic induction of the morphology of some languages. Examples include Déjean (1998); Goldsmith (2000); Creutz and Lagus (2002); Creutz (2003); Creutz and Lagus (2004); Monson et al. (2007) as well as Hammarström (2009). Some of these studies have motivated the production of useful open-source application packages such as Linguistica, Morfessor and Paramor. However, it has been observed that the methods adopted in these efforts may not always scale-up to accommodate many more languages than the ones for which they were originally developed. In this regard, De Pauw and Wagacha (2007) noted the limitations of the popular methods that have been used effectively for some European languages when applied to Bantu languages of Africa. They observed in particular, that the established AutoMorphology method such as applied by Goldsmith (2000) is biased towards Indo-European languages and therefore puts it at a disadvantage when applied to a Bantu language such as Gĩkũyũ. Also, Adegbola (2016) highlighted the limitations of these methods in addressing the morphology of some other African languages. He made particular reference to the automatic induction of morphological processes such as full and partial reduplication, interfixation, compounding and others that are productively employed in Igbo, Yorùbá and some other Nigerian languages.

These methods, having been originally developed to address the morphology of a relatively few languages of Europe and Asia, essentially assume simple concatenative morphology which, even though employed in Igbo and Yorùbá, has been found to be less productively engaged in these languages than other morphological processes. Morphological processes that employ stem-derived morphemes in which affixes are dependent on and are therefore a reflection of stems cannot be automatically induced through computational methods that seek to identify recurrent partials as is used in applications such as Linguistica (Goldsmith, 2000); Morfessor (Creutz, Lagus and Virpioja, 2005) and Paramor (Monson et al,, 2007).

Hammarström and Borin (2011) prepared a comprehensive survey report on the unsupervised learning of Morphology. None of the studies in the survey addressed the unsupervised induction of partial or full reduplication, infixation, interfixation, compounding or any other morphological processes based on the affixation of stem-derived morphemes. Can and Manandhar (2014) also undertook a panoramic view of methods and algorithms used in unsupervised learning of morphology and yet strategies for addressing stem-derived morphemes did

not reflect. Marelli (2021) engaged the general subject of quantitative morphology and still yet no methods that address stem-derived morphemes featured. This study therefore addresses this important but yet outstanding problem of the automatic detection of morphological processes that employ stem-derived morphemes.

2. Recurrent Partials and Recurrent Patterns

Adegbola (2016), demonstrated that the automatic induction of Yorùbá morphology depends to a large extent on the identification of recurrent patterns rather than the identification of recurrent partials, just as Iheanetu (2015) demonstrated for Igbo. For instance, inflection of, as well as the derivation of gerunds from English verbs may be achieved by the simple suffixation of the recurrent partial *'ing'*.

In Yorùbá, however, similar derivations of nouns from verbs are achieved by prefixation, not of recurrent partials, but through a process of partial reduplication in which a consonant-vowel (CV) template is prefixed to a stem. The C being a copy of the first consonant of the stem while the V is the high tone vowel *'í'* (Oyebade, 2007a), Yorùbá, being a tone language. This implies that the CV template that is prefixed to the stem is in itself derived from the stem. Hence the idea of a stem-derived morpheme.

Table 1 shows examples of the production of nouns from verbs through the process of partial reduplication by use of stem-derived affixes in Yorùbá:

Verb	**Gloss**	**Derived Noun**	**Gloss**
Ṣe	Do	Ṣíṣe	Doing (N)
Lọ	Go	Lílọ	Going (N)
Pè	Call	Pípè	Calling (N)
Gbà	Accept	Gbígbà	Acceptance

Table 1: Yorùbá examples of partial reduplication

Other common and highly productive processes of Yorùbá morphology such as full reduplication and interfixation also conform to this approach of affixation in which affixes are derived from the stems. Tables 2 and 3 show examples of these morphological processes and the resulting words, showing clearly identifiable word patterns:

Verb	**Gloss**	**Derived Noun**	**Gloss**
Pa iná	Put out fire	Panápaná	Fire fighter
Tú ilé	Undo household	Túlétúlé	Disruptive person
Gbé ọmọ	Steal child	Gbọ́mọgbọ́mọ	Kidnaper
Wo ìran	View scene	Wòranwòran	Spectator

Table 2: Yorùbá examples of full reduplication

Based on the assumption of morphemes as recurrent partials in English, for example, Goldsmith (2001), Creutz and Lagus (2004) as well as Hammarström (2009) have developed algorithms that use probability to differentiate between random sub-word elements and recurrent partials which are valid morphemic units.

Noun	**Gloss**	**Derived Form**	**Gloss**
Ọmọ	Child	Ọmọkọ́mọ	Any child/bad child
Iye	Value	Iyebíye	Invaluable
Àgbà	Adult	Àgbàlagbà	Old/matured person
Aṣe	Doer	Aṣemáṣe	Inappropriate behaviour

Table 3: Yorùbá examples of interfixation

However, the widely used affixation of stem-derived morphemes rather than recurrent partials in Yorùbá poses a problem in the fact of the dependence of an affix on its stem. This obviates the expected relatively high frequency of such affixes to enable their classification into one of two classes of "random sub-word segments" or "significant morphemic units" based on their probabilities of occurrence.

In a bid to cluster words produced through morphological processes based on stem-derived morphemes, Iheanetu (2015) used the idea of "word-labels" derived from the patterns of arrangements of consonants and vowels in the Igbo language to cluster words according to the morphological processes through which they were formed.

3. Word-labels

As proposed by Iheanetu (2015), a word-label can be described as a textual proxy of the pattern of arrangements of consonants and vowels in a word. It provides basis for clustering words of identical patterns, in a process of unsupervised learning, thereby identifying them as derived through identical morphological processes.

Word-labels are derived from words by assigning a sequence of symbols CX or VX representing consonants (C) or vowels (V) accompanied by a numerical index (X) indicating the occurrence or reoccurrences of specific consonants or vowels in the words, from left to right (Adegbola, 2016). Table 4 shows examples of a few English words and their derived word-labels.

Word	**Word-label**
Deal	C0V0V1C1
Said	C0V0V1C1
Deed	C0V0V0C0
Seek	C0V0V0C1

Table 4: Some English words and their word-labels

The word "deal", for example, takes the word-label C0V0V1C1 because the first character, 'd' is assigned the symbol C0 and the first vowel 'e' is assigned the symbol V0. Succeeding characters 'a' and 'l' are assigned the symbols V1 and C1 respectively because

they are the second occurring vowel and consonant respectively. Using a zero-based indexing, the first occurrence of a consonant or vowel is assigned the index 0. Freshly used succeeding consonants or vowels are assigned succeeding numbers as indexes, while the reoccurrence of a consonant or vowel is reassigned the already assigned index. Based on this scheme, the word 'deed' takes the word-label C0V0V0C0 because the first and second occurring consonant as well as the first and second occurring vowel are the same. The facilitation of word-labels for the unsupervised induction of English morphology is yet to be investigated. In a Yorùbá lexicon, however, the patterns of morphological processes are clearly reflected in word-labels and it therefore becomes possible to cluster or classify words according to their morphological process, based on the manifest patterns in the words as reflected in their word-labels. The use of word-labels to cluster or classify words according to the morphological processes through which they are derived is justified by the fact that affixes derived from templates based on stems impose patterns on the produced words. These patterns are therefore reflected in the words so formed to the extent that commonality in morphological processes is reflected in a commonality in word patterns and therefore word-labels. The following examples of Yorùbá words of common morphological derivation clustered around the word-label C0V0C0V1 demonstrate this fact in Table 5. It should be noted that vowels with differing tone marks are regarded as different and that Yorùbá orthography uses the character *'n'* in three distinct ways. It is used in certain instances as a consonant, in some other instances as a syllabic nasal and in yet other instances as a nasalization indicator for a preceding vowel.

Stem	Gloss	Derived Word	Gloss	Common W. Label
Lọ	Go	Lílọ	Going (N)	C0V0C0V1
Wá	Come	Wíwá	Coming (N)	C0V0C0V1
Ṣe	Do	Ṣíṣe	Doing (N)	C0V0C0V1
Kọ	Write	Kíkọ	Writing (N)	C0V0C0V1
Ké	Cry	Kíké	Crying (N)	C0V0C0V1
Sè	Cook	Sísè	Cooking (N)	C0V0C0V1

Table 5: Some Yorùbá words (nouns) derived by partial reduplication to produce a common word-label

4. Identifying Morphologically Motivated Patterns

Every word has a word pattern. There is a need therefore to differentiate between random patterns and morphologically motivated patterns in order to be able to automatically identify words that are products of given morphological processes. The main objective of this study is to devise a means that can be used to automatically recognize pattern-inducing morphological processes in a language, using Yorùbá as an example, towards exploring the possibility of generalisation for other languages in future. Hence, we here present a scheme for automatically recognizing pattern-inducing morphological processes in Yorùbá.

To distinguish between word-labels that manifest by chance as against word-labels motivated by pattern-inducing morphological processes, it would be instructive to compute two probability measures for each word-label. The first is a predicted probability of a word-label based on an assumption of random choice of consonants and vowels in the words that produce the word-label and the second is the observed probability of the word-label in a sizable corpus of written texts in Yorùbá. These predicted and observed probabilities of word-labels may then be compared. A significant difference in the predicted and the observed probabilities of a word-label will be usable as basis for classifying word-labels as either resulting from random choice of consonants and vowels in the words that produce them or word-labels that result from significant patterns induced by morphological processes.

5. Predicted Probability of a Word-label

The predicted probability of the manifestation of a word-label is based on the assumption that all allowable consonants and vowels of the language in question occur equiprobably in words that produce such a word-label. In addition, it assumes independence between the individual consonants and vowels that make up the word. These assumptions would be valid only if there are no external influences on the choices of these consonants and vowels.

To compute this predicted probability, we consider a word-label as consisting of symbols A_iX_i where:

$$A_i \in \{C, V\}$$

$$X_i \in \{0, 1, 2, \ldots . n\}$$

In this light, the word-label C0V0C1V1 for the English word "make" for example, can be thought of as containing symbols $A_1X_1A_2X_2A_3X_3A_4X_4$, where $A_1 = C, X_1 = 0, A_2 = V, X_2 = 0, A_3 = C, X_3 = 1, A_4 = V \text{ and } X_4 = 1$.

Given a language of c consonants and v vowels, the probability of obtaining a symbol C0 for the first occurring consonant is $^{c}/_{c}$ as any of the c consonants can be chosen. This is equal to 1, implying certainty. The probability of obtaining another symbol C0 after the first consonant has taken the symbol C0 is $^{1}/_{c}$ because the only one consonant that caused the first consonant symbol to be C0 must have reoccurred. By the same token, the probability of any other consonant taking the symbol C1 is $^{(c-1)}/_{c}$, having excluded the consonant that produced C0. We can thus generalise the probability of any symbol CX as $^{(c-X)}/_{c}$. In the same vein, the probability of obtaining a symbol V0 for the first occurring vowel is $^{v}/_{v}$ and the probability of any symbol VX can be generalised as $^{(v-X)}/_{v}$ as argued above.

By virtue of the assumption of independence in the predicted probabilities of each of the symbols that make up a word-label, the likelihood $L(A_1X_1A_2X_2 \ldots A_nX_n)$ of a word-label can be

computed as the naive product of the individual probabilities of each of the symbols thus:

$$L(A_1X_1A_2X_2 \dots A_nX_n) = \prod_{i=1}^{n} P(A_iX_i) \quad (1)$$

The product of two or more probabilities may not necessarily yield a probability. Hence, to normalise the likelihood in formular (1) above into a probability, we shall multiply it by the reciprocal of the cumulative likelihoods of all conceivable word-labels in a group as shown in formular (2). This will guarantee that the probabilities of all conceivable word-labels in each group sums up to unity in accordance with probability theory.

$$P(A_1X_1A_2X_2 \dots A_nX_n) = {}^{1}/_{S} \prod_{i=1}^{n} P(A_iX_i) \quad (2)$$

$$S = \sum_{j=1}^{m} L_j(A_1X_1A_2X_2 \dots A_nX_n) \quad (3)$$

Where m is the total number of conceivable word-labels in each group and S is the cumulative likelihood of all the m conceivable word-labels in a group of equal lengths and common structure.

We define the structure of a word-label as the sequence of consonants and vowels without considering the indexes. For example, the word-labels C0V0V0 and C0V0V1 both have the same structure because they both consist of the same consonant and vowel sequence of CVV, differing only in their indexes.

6. Observed Probability of a Word-label

In considering the observed probability of a word-label, the manifestation of a given word-label is taken as a single event with a single outcome while the manifestations of all word-labels in a group is the total number of possible outcomes. Hence, the observed probability of a given word-label can be calculated as the frequency of occurrence of the word-label, based on the number of words that produced it and are thereby clustered around it, divided by the total number of occurrences of all word-labels in the same group, based on the total number of words that produced them.

To compute the observed probability of a word-label i that manifests in a given group of identical length and structure, having a cumulative total of n word-tokens, we observe the number of word-tokens n_i that produced the given word-label i. Each of the n word-tokens in the group will produce one word-label each. Hence, the probability $P(i)$ of the word-label i would be the number of word-tokens n_i that produced the word-label i divided by the total number of word-tokens n_i in the group, computed as:

$$P(i) = {}^{n_i}/_{n} \quad (3)$$

7. Automatic Detection of Morphological Processes

As already explained, the predicted probability of a word-label as computed above assumes equiprobability in the occurrences of the individual consonants and vowels combined to form the word that produced the word-label. In addition, independence between the occurrences of the consonants and the vowels is assumed. However, if the formation of a word is motivated by a morphological process, these assumptions become invalidated. For example, as Oyebade (2007a) noted, in the morphological process of partial reduplication in Yorùbá, a consonant and vowel (CV) template is prefixed to a stem, the C being a copy of the first consonant of the stem while the V is the high tone vowel *'i'*. The fact that the C is a copy of the first consonant of the stem violates the assumption of independence in the choice of that consonant. As for the assumption of equiprobability, the fact that the V in the prefix template is unconditionally the high tone vowel *'i'* violates the assumption of equiprobability. Hence, we hypothesize that the contribution of a morphological process in the formation of a word will bring about a significant difference in the predicted and observed probabilities of its word label. Word-labels derived from such a word whose formation is motivated by a morphological process that employs a stem-derived morpheme will surely feature a sufficiently significant difference to signal the involvement of such a process.

To automatically detect the morphological processes used in word formation in a language, we may therefore compare the observed and predicted probabilities of word-labels encountered in a lexicon obtained from a sizable corpus of texts in the language. It is hypothesised that in the absence of any morphological influences, we expect no significant differences in the observed and predicted probabilities of a word-label. We can therefore conclude that any significant differences between the predicted and the observed probabilities of a word-label would have been brought about by morphological influences.

The predicted probability of a word-label as described in Section 5 is a normalised product over the set of probabilities of the individual symbols that make up the word-label. The product of two proper fractions will always produce a lower value than both. Hence, the predicted probability of a word-label will depend on its length. For this reason, we opted to group together word-labels of the same lengths and structures as defined in section 5 together for consistency in the comparison of word-labels.

8. Tests and Results

To explore the difference between the predicted and observed probabilities of a word-label produced by word-tokens whose formation is motivated by a morphological process based on the affixation of stem-derived morphemes, we extracted a lexicon of 14,670 word-tokens from a Yorùbá corpus. The 14,670 tokens produced 1,282 distinct word-labels. The word-labels were grouped according to their lengths and structures and both their predicted and observed probabilities were computed, all as described in sections 5 and 6. The computed predicted probability was based on 18 consonants and 12 vowels as specified in the literature for the number of consonants and vowels of the Yorùbá

language (Oyebade, 2007b). Comparison of the predicted and observed probabilities were made and the following results were obtained.

The most productive word-label was C0V0C1V1, with a cluster of 2,716 word-tokens. This represents 18.51% of the 14,670 word-tokens in the lexicon. Examples of word-tokens that produced this word-label include *balẹ, dewé, fìjó, gbaṣọ* and *jíṣẹ́*. The overwhelming majority of these are formed by the morphological process of compounding, suggesting that the word-label C0V0C1V1 clusters Yorùbá words formed mainly by compounding. A predicted probability of 0.8657 was computed for this word-label, while the computed observed probability was 0.7690.

W. Label	P. Prob.	O. Prob.	Cardinality
C0V0C0V0	0.0046	0.0416	147
C0V0C0V1	0.0509	0.0994	351
C0V0C1V1	0.8657	0.7690	2716
C0V0C1V0	0.0787	0.0900	318
Cumulative	1.0000	1.0000	3532

Table 6: Predicted and observed probabilities of word-labels of the CVCV group

Table 6 shows all conceivable word-labels in the group (CVCV) to which it belongs. Word-labels (W. Label) are shown in the first column, while the predicted and observed probabilities (P. Prob. and O. Prob.) are shown in columns two and three respectively. The fourth column shows the number of word-tokens (Cardinality) clustered around each word-label. The fact that the cumulative predicted probability adds up to unity indicates that word-tokens producing all conceivable word-labels in this group were encountered in the corpus.

The word-label C0V0C0V0, which is a member of the CVCV group suggests a cluster of words produced by the morphological process of full reduplication. The glaring difference in its predicted and observed probabilities bears eloquent testimony to its easily perceptible symmetry.

The second most productive word-label is V0C0V1, with a cluster of 1,446 word-tokens, representing 9.86% of the 14,670 word-tokens in the lexicon. Examples of words that produced this word-label include *abi, abẹ, egbé, idán, àgbo* and *ẹ̀rọ*, all derived through the nominalisation of single syllable Yorùbá verbs by the concatenative morphological process of vowel prefixation. While we acknowledge seeming exceptions such as *abẹ́*, which, though a noun is not easily associated with a one-syllable verb with related meaning, we can say generally that this word-label clusters words formed mainly through the morphological process of concatenation by vowel prefixation. As can be observed from the sample of words shown here from this cluster, various vowels featured as the prefixed morphemes. This is consistent with Awobuluyi's (2001) observation that all Yorùbá vowels apart from *'u'* and the nasal vowels are used freely as prefixes. It stands to reason however, that these prefixes may not occur sufficiently frequently to be easily detectable automatically as recurrent partials, based solely on frequency in a process of unsupervised induction of the morphological process. The proposed approach of clustering relevant words around word-labels, however, makes it easy to perceive the prefixes, generally as vowels rather than a particular individual vowel.

The only other word-label in the group VCV is V0C0V0. Table 7 shows the predicted and observed probabilities of these two word-labels of this group.

W. Label	P. Prob.	O. Prob.	Cardinality
V0C0V1	0.9167	0.9335	1446
V0C0V0	0.0833	0.0665	103
Cumulative	1.0000	1.0000	1549

Table 7: Predicted and observed probabilities of word-labels of the VCV group

The third most productive word-label is V0C0V1C1V2, with a cluster of 1,417 word-tokens, representing 9.66% of the 14,670 word-tokens in the lexicon. Examples of words that produced this word-label include *abetí, ìbínú, ojúgbó, àbùṣán, èlùbọ́, ìbùkún, òkúta, ẹlẹ́wù* and *ọ̀mùtí*. Apart from *èlùbọ́* and *òkúta* in which the word formation processes may not be glaring to this investigator, the other words in this small sample and most of the others in the cluster feature mainly concatenation by vowel prefixation as well as compounding. For example, *ìbùkún* meaning "blessing" is a noun formed by the compounding of two words *bù* (take) and *kún* (fill) to form the verb "*increase*" followed by nominalisation of the verb *bùkún* by vowel prefixation to form the noun *ìbùkún.*

W. Label	P. Prob.	O. Prob.	Cardinality
V0C0V0C0V0	0.0004	0.0039	8
V0C0V0C1V1	0.0675	0.0640	132
V0C0V0C1V0	0.0061	0.0194	40
V0C0V1C1V1	0.0675	0.1024	211
V0C0V1C1V2	0.6745	0.6870	1417
V0C0V1C0V1	0.0040	0.0388	80
V0C0V1C1V0	0.0675	0.0344	71
V0C0V1C0V0	0.0040	0.0040	9
V0C0V1C0V2	0.0397	0.0432	89
V0C0V0C0V1	0.0040	0.0029	6
V0C0V0C0V2	0.0036	0.0000	0
V0C0V0C1V2	0.0612	0.0000	0
Cumulative	1.0000	1.0000	2063

Table 8: Predicted and observed probabilities of word-labels of the VCVCV group

The other word-labels of the VCVCV group to which V0C0V1C1V2 belongs are shown in Table 8. As would be noticed, no words that could have produced two valid word labels; V0C0V0C0V2 and V0C0V0C1V2 featured in the corpus used for this study. Particularly curious is V0C0V0C1V2 with a predicted probability of 0.0612, being the 5th highest probability in the group. A few other relatively productive word-labels are shown in Table 9.

W. Label	P. Prob	O. Prob	Cardinality
C0V0C1V1C2V2	0.6413	0.4711	620
V0C0V1C1V2C2V3	0.4810	0.5060	506
C0V0	1.0000	1.0000	430
C0V0C0V1	0.0509	0.0994	351
C0V0C1V0	0.0787	0.0900	318
C0V0V1C1V2	0.7215	0.5333	272
C0V0V1	0.9167	0.7220	226

Table 9: Predicted and observed probabilities as well as cardinality of the 4th to the 10th most productive word-labels

The core concern of this study is to automatically identify words that feature morphological processes based on stem-derived morphemes by comparing the predicted and observed probabilities of their word-labels.

Figure 1 and Figure 2 show that comparison of predicted and observed probabilities of word-labels is capable of making this important distinction. As can be observed in the charts, the observed probabilities of word-labels that incorporate stem-based morphemes are generally higher than their predicted probabilities, while the contrary holds in the case of word-labels without stem-based morphemes.

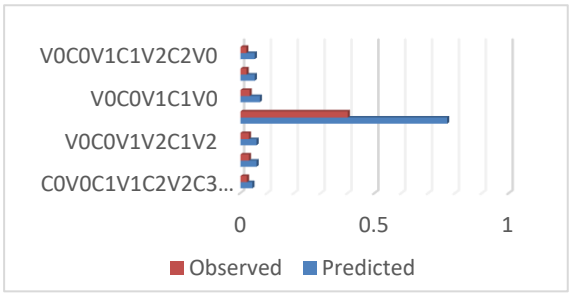

Figure 1: Observed and Predicted Probabilities of Word-labels with Stem-derived Words

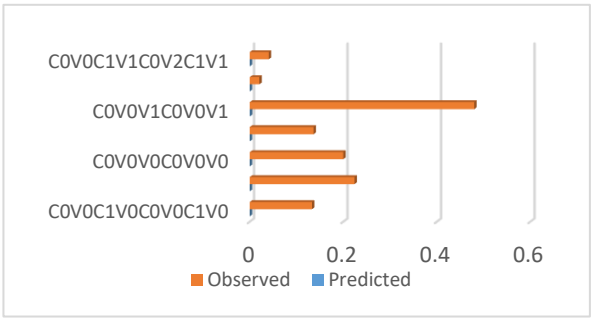

Figure 2: Observed and Predicted Probabilities of Word-labels without Stem-derived Words

We recognise the ratio of the observed and predicted probabilities of word-labels as a convenient indicator of the involvement of stem-derived morphemes.

$$ratio = {}^{O\ Prob}/_{P\ Prob}$$

Where $O\ Prob$ and $P\ Prob$ are the predicted and probabilities.

We also acknowledge two factors that would have effects on the predicted and observed probabilities of word-labels. On the one hand, as discussed in section 7, the longer a word-label, the lower its probability. Hence, the length of a word-label will affect its predicted probability to the extent that short word-labels may tend to have high probabilities while long word-labels may tend to have low probabilities. On the other hand, sampling error owing to resource-scarcity may affect the observed probability of word-labels in a group in which the word-labels cluster few words, causing them to have high observed probabilities.

The predicted probability of a word-label is computed as explained in section 5 by obtaining the normalised product of the probabilities of the individual symbols that make up the word-label, while the observed probability is computed as explained in section 6 by normalising the cardinality of a word-label with the overall cardinality of its group.

Table 10 shows a selection of word-labels, their ratios of observed and predicted probabilities and some sample word-tokens each. It can be observed from the table that the more the repetition of characters in a word-token as reflected in the word-label, the greater the ratio of the observed and predicted probabilities. Obviously, the word-labels, C0V0C1V0C0V0C1V0 and C0V0C1V1C0V0C1V1 are motivated by the morphological process of full reduplication as can be observed from the symmetry brought about by the duplication of C0V0C1V0 and C0V0C1V1 respectively. This is reflected in their relatively high ratios of 127145.48 and 19452.41 and the sample words; *biribiri* and *bọ̀lọ̀bọ̀lọ̀* as well as *bojúbojú* and *bàmùbàmù* respectively. The succeeding word-label, C0V0V0C1V0C0V0 must have been motivated by the morphological process of partial reduplication and this is reflected in the ratio of 6174.55 and the sample words: *fẹ́ẹ́rẹ́fẹ́* and *gbuurugbu*.

The word-label, C0V0C1V1C0V2 with a ratio of 0.85 and sample words of *jogòjì* and *kàgbákò* as well as the word-label C0V0C1V1C2V1C3V2 with its ratio of 0.53 and sample words of *kòbọmọjẹ́* and *mójúkúrò* provide convincing evidence that the ratio of the observed and predicted probabilities is a faithful indicator of the involvement of stem-based morphemes in certain word-labels and their absence in some others.

Word-label	Ratio	Morphological Process	Sample words
C0V0C1V0C0V0C1V0	127145.48	Full Reduplication	biribiri, bọ̀lọ̀bọ̀lọ̀, fírífírí, gbẹ̀jẹ̀gbẹ̀jẹ̀
C0V0C1V1C0V0C1V1	19452.41	Full Reduplication	bojúbojú, bàmùbàmù, fọ́rífọ́rí, jayéjayé
C0V0V0C1V0C0V0	6174.55	Partial Reduplication	fẹ́ẹ́rẹ́fẹ́, gbuurugbu, tààràtà, pẹẹrẹpẹ
C0V0C0V0C0V0	1074.26	Full Reduplication	dandandan, gangangan, jẹ́jẹ́jẹ́, tantantan
C0V0C1V1C0V2C1V1	352.40	Full Reduplication	fálafàla, jágbajàgba, kóbokòbo, pálapàla
V0C0V0C1V1C0V0	20.58	Interfixation	àgbàlágbà, ọmọkọ́mọ, òpọ̀lopọ̀
V0C0V1C1V2C2V2	1.88	Prefixation	alágídí, alákàrà, ọlọ́gẹ̀dẹ̀, ónígbèsè
V0V1C0V2C1V3	1.30	Prefix+Compounding	àìdúpẹ́, àìlera, àìmọ̀kan, àìrójú, àìgbọ̀ràn
C0V0C0V1C1V2	1.30	Partial Reduplication	dídọ́gba, jíjóná, kíkorò, lílépa, pípadà
C0V0C1V1C0V2	0.85	Compounding	jogóji, kàgbákò, láyọ̀lé, pawọ́pọ̀, ṣojúṣe
C0V0C1V1C2V1C3V2	0.53	Desentencialisation	kòbọmọjẹ́, mójúkúrò, yírapadà, ṣàfarawé

Table 10: Word-label, Ratio, Morphological Process and Sample Words

It can be inferred from the fore-going that while the predicted probability of a word-label is affected by its length, the observed probability is totally insulated from this factor. Conversely, while the observed probability may be affected by sampling error occasioned by resource-scarcity, the predicted probability is totally insulated from this effect. The sampling error is reflected in the cardinality of each word-label, which in turn reflects on the cumulative cardinality of each group.

To address the problem of disparate lengths and their effects on probabilities, word-labels of identical lengths were considered together, regardless of their structures. However, it was noticed that the effect of disparity in lengths tended to reduce as the lengths of the word-labels increased. It was also noticed that the cardinality of word-labels correlates negatively with their lengths. By treating word-labels of disparate lengths separately, it was possible to localise the effects of sampling error to extremely long word-labels that featured very low cardinality. Hence, the effect of sampling error appears to be localised to each length-based cluster of word-labels.

The 1,282 word-labels generated from the lexicon of 14,670 word-tokens extracted from the Yorùbá corpus were grouped into 24 sets of word-labels based on their lengths. The word-labels in each set were sorted according to the values of their *O Prob* and *P Prob* ratios. It was observed that in each of the 24 sets of same-length word-labels as sorted based on their ratios, the first bunch of word-labels to be encountered are those that cluster words formed by full reduplication. Followed by this first bunch of word-labels, come word-labels that cluster words formed by either partial reduplication or interfixation. We then encounter word-labels that cluster words formed by simple affixation of recurrent partials as well as compounding and then word-labels that cluster all other types of words. This is observable in Table 10 in which C0V0C1V0C0V0C1V0 and C0V0C1V1C0V0C1V1 are examples of word-labels that conform to full reduplication, C0V0V0C1V0C0V0 conforms to partial reduplication while V0C0V0C1V1C0V0 conforms to interfixation. The word-labels C0V0C1V1C0V2 and C0V0C1V1C2V1C3V2 conform to compounding and desentencialisation respectively. The successive and consistent reduction in the values of the *O Prob*/*P Prob* ratios is instructive.

All morphological processes reported in the literature of Yorùbá morphology were observed and words formed by each process were found to cluster around specific word-labels. As noted in Adegbola (2016), some word-labels clustered word-tokens formed by more than one morphological process and in some cases, a single morphological process was found to produce word-tokens that clustered around more than one word-label. In the final analysis however, the word-labels showed themselves creditably as effective purveyors of the patterns imposed on words by stem-derived morphemes and therefore an effective and efficient means of identifying the morphological processes featured in a language.

9. Conclusion

It is apparent from the results obtained in this study that the ratio of the predicted and observed probabilities of word-labels is a valuable metric for the identification of word-labels that incorporate stem-derived morphemes. This is to be expected because the involvement of morphological processes in the formation of words that produce such word-labels contradict the assumptions of equiprobability and independence in the choice of characters for the affected word-tokens. This is a radically new approach to the unsupervised induction of morphology. It should become a valuable supplement to the approach proposed by Harris (1955), which has continued to guide the approaches used in more recent studies undertaken by investigators such as Déjean (1998); Goldsmith (2000); Creutz and Lagus (2002); Creutz (2003); Creutz and Lagus (2004) as well as Hammarström (2009) as was earlier discussed.

The resource-scarce status of the Yorùbá language played out significantly in this study. The lexicon used contained only 14,670 word-tokens, which certainly left many Yorùbá words unaccounted for. Many word-labels that obviously feature stem-derived morphemes had low cardinality, some of them as low as one. The incidence of a stem-derived morpheme in a word-label is indicative of a morphological process. A morphological process is not likely to produce only one word for a language and so, these word-labels with low

cardinality are expected to have more than a few word-tokens each in their clusters. Towards addressing the resource-scarcity of Yorùbá, it should be possible to use such low cardinality word-labels to project and thereby validate or even generate out-of-vocabulary words in certain Natural Language Processing (NLP) circumstances. This is a worthy endeavour for a future study.

The cumulative values of the observed probabilities of each group of word-labels added up to unity in conformity with probability theory. This is ensured by the fact that these probabilities were computed, based solely on word-tokens that featured in the corpus that produced the study lexicon. However, the cumulative values of the predicted probabilities of some groups of word-labels did not add up to unity. This implies that certain word-labels in such groups did not feature in the modest corpus that produced the study lexicon.

As noted in Section 5, the predicted probability for each of all conceivable and valid word-labels in each group is guaranteed to sum up to unity. All word-labels that did not feature in their appropriate groups as a result of sampling error occasioned by the resource-scarcity are computationally derivable and their individual predicted probabilities can be computed with consistency. Hence, the cumulative predicted probability of all conceivable word-labels in a group is guaranteed to add up to unity. The corresponding observed probabilities of each of the unencountered word-labels due to sampling error will logically take a value of zero each, thereby ensuring that the predicted and observed probabilities for each group sums up to unity in conformity with probability theory. This was the case with the VCVCV group of word-labels as shown in Table 8, where the word-labels V0C0V0C0V2 and V0C0V0C1V2 had predicted probability values of 0.0036 and 0.0612 respectively but a cardinality of zero each and therefore observed probability values of zero each.

Unencountered word-labels, being validly computationally derivable may be useful in projecting and validating or even generating out-of-vocabulary words. The fact that the predicted probabilities of such unencountered word-labels can be calculated is of high value. Such probability values offers an important metric for assessing the coverage of available corpora in a language. The systematic use of such a metric to assess the level of coverage of available corpora of resource-scarce languages is yet another worthy issue for future study.

One interesting surprise encountered within the C0V0C1V1 cluster is the word *benson*, a foreign proper noun. Though a foreign word, it is understandable that it found its way into a Yorùbá corpus, being proper noun. It found its way into the C0V0C1V1 cluster in particular by virtue of the fact that the character *"n"* is used as the indicator for nasalisation of a preceding vowel in Yorùbá orthography as explained in Section 3. For this reason, the *"en"* and *"on"* in the word *benson* were erroneously construed as Yorùbá nasal vowels.

Many non-Yorùbá words in the corpus clustered around word-labels that admit consonant clustering which happens not to be a feature of Yorùbá syllable structure. An example of such a word-label is C0C1V0C2 under which the English words *show*, *this* and *what* were found. The words were traced to certain lines of a Yorùbá play in the corpus, in which one of the characters was showing off ability to speak the English language. In this light, it is interesting that word-labels may be usable as a means of identifying and extracting foreign words in a corpus.

Another interesting cluster is the cluster designated as "XXX", which was deliberately used to cluster words that incorporate consonants such as X, C, V and Q, which are not used in the Yorùbá language. Some of the words that found their ways into this cluster consisting of 40 words include *academic*, *achaempong*, *african* and *america*.

The problem of recognizing the presence of stem-derived morphemes in words is yet to be effectively addressed in the literature of computational morphology. Results obtained from this study show the potentials of word-labels as an effective and efficient tool for addressing this problem. A number of other important applications of word-labels have also been suggested. Locating the word-label as a proxy of words within the Chomsky hierarchy and the possible use of automata to parse and recognise valid word-labels of the Yorùbá language or any other languages for that matter are not only desirable but also pertinent. All these need to be actively engaged and further investigated in future studies.

10. Acknowledgments

We acknowledge the valuable suggestions in the excellent reviews offered by members of SIGUL2022 program committee. The Yorùbá lexicon used in this study was obtained from materials collected for the project; Development of a Yorùbá Speech Synthesizer funded by the Lagos State Research Development Council (LRDC).

11. Bibliographical References

Adegbola, T. (2016, April). Pattern-based unsupervised induction of Yorùbá Morphology. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 599-604).

Awobuluyi, O. (2001). Mọfọ́lọ́jì Èdè Yorùbá, in Ajayi B. (ed) Ẹ̀kọ́ Ìjìnlẹ̀ Yorùbá: Ẹ̀dá Èdè, Lítírẹ́ṣọ̀, àti Aṣà. Ìjẹ̀bú-òde: Shebiotimo Publications. pp47-70

Can, B., & Manandhar, S. (2014, April). Methods and algorithms for unsupervised learning of morphology. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 177-205). Springer, Berlin, Heidelberg.

Creutz, M. & Lagus, K. (2002). Unsupervised discovery of morphemes. In Proceedings of the Workshop on Morphological and Phonological Learning of ACL. Philadelphia, PA. 21–30.

Creutz, M. (2003). Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics* (pp. 280-287).

Creutz, M., & Lagus, K. (2004). Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th meeting of the ACL special interest group in computational phonology: Current themes in computational phonology and morphology* (pp. 43-51).

Creutz, M., Lagus, K., & Virpioja, S. (2005, September). Unsupervised morphology induction using morfessor. In International Workshop on Finite-State Methods and Natural Language Processing (pp. 300-301). Springer, Berlin, Heidelberg.

De Pauw, G., & Wagacha, P. W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In In Proceedings of the eighth INTERSPEECH conference.

Déjean, H. (1998). Morphemes as necessary concept for structures discovery from untagged corpora. In *New Methods in Language Processing and Computational Natural Language Learning*.

Goldsmith, J. (2000) Linguistica: An automatic morphological analyzer. In Proceedings from the Main Session of the Chicago Linguistic Society's 36th Meeting, pages 125–139, Chicago, IL.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153-198.

Hammarström, H. (2009). Unsupervised learning of morphology and the languages of the world. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg

Hammarström, H., & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, *37*(2), 309-350.

Harris, Z. (1955). From phoneme to morpheme. Language 31(2). 190–222.

Iheanetu, O. U. (2015). A data-driven model of Igbo morphology. *University of Ibadan, Nigeria (Unpublished Ph. D. Thesis)*.

Marelli, M. (2021). Quantitative Methods in Morphology: Corpora and Other "Big Data" Approaches. In Oxford Research Encyclopaedia of Morphology.

Monson, C. Carbonell, J., Lavie, A., & Levin, L. (2007). ParaMor: Finding paradigms across morphology. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 900–907). Berlin, Germany: Springer.

Oyebade, F.O. (2007a) Yorùbá Morphology. In Orẹ Yusuf (ed.) Basic linguistics for Nigerian languages teachers. Port Harcourt: M & J Grand Orbit Communications Ltd. and Emhai Press. 241-255.

Oyebade, F.O. (2007b) Yorùbá Phonology. In Orẹ Yusuf (ed.) Basic linguistics for Nigerian languages teachers. Port Harcourt: M & J Grand Orbit Communications Ltd. and Emhai Press. 221-239.