Pattern-based Unsupervised Induction of Yorùbá Morphology

Tunde Adegbola
African Languages Technology Initiative, Ibadan, Nigeria
+234 8034019398
taintransit@hotmail.com

ABSTRACT
The Unsupervised induction of morphological rules from a simple list of words in a language of interest is a productive approach to Computational Morphology. The most popular algorithms used for this purpose in the literature are based on the assumption that the relatively high occurrence frequencies of certain word segments described as recurrent partials in a lexicon suggests the existence of morpheme boundaries around such high frequency word segments. Even though this word-segment-frequency approach works well for concatenative morphology, it does not cater for some of the most productive morphological processes in Yorùbá and some other African languages. In this paper, unsupervised induction of the morphological rules of Yorùbá was achieved based on a word-pattern-frequency rather than a word-segment-frequency approach. Words in a Yorùbá lexicon were clustered according to the morphological processes on which their formation are based, producing results that hitherto were achievable only by painstaking rule-based manual classification.

Keywords
Machine Learning of Morphology; Morphological Segmentation; Probabilistic Models; Word Labels; Frequent Segments Frequent Patterns.

1. INTRODUCTION
Studies in computational morphology used Finite State Automata (FSA) to manually code the morphological rules of a language for efficient storage in computers and effective use in Natural Language Processing (NLP). The works of [1] and [2] were some of the pioneering efforts in this direction. However, with advances in Machine Learning, the focus of Computational Morphology is now directed more towards inductive data-driven approaches rather than deductive Rule-Based approaches. To this end, the early works of [3] [4] on the idea of successor frequencies as basis for determining morpheme boundaries found a new lease of life in the work of [5]. Similarly, [6] [7] developed the ideas of Arbitrary Character Assumption (ACA) and Frequent Flyer Assumption (FFA) as metaphors for distinguishing between word segments that could be regarded as stems and those that could be regarded as bound morphemes in English and some other languages. Based on these and similar ideas, other authors including but not limited to [8] [9] have developed software tools for the unsupervised induction of morphological rules with high optimism that these tools will work for any language. Indeed, these tools work well for the induction of simple concatenative morphological processes employed in many languages around the world but they do not produce the desired results when faced with serial concatenation in highly agglutinative languages. In this regard, [10] highlighted the problem of last suffix extraction in Malayalam - A morphologically rich Dravidian language. Even more serious however is the problem of inducing morphological processes such as reduplication, compounding, infixation, suprafixation, circumfixation and interfixation that do not necessarily feature frequent word segments but rather observable patterns. In this light, [11] pointed out that these segment-frequency-based approaches may not work for Bantu languages. This has been found to be true for some other African languages [12], as well as Semitic languages in which morphology is not based on concatenation of recurrent partials.

In morphological processes such as reduplication, compounding, infixation, suprafixation, circumfixation and interfixation, the characters that feature in the morphemes do not necessarily occur frequently within the vocabulary of the language. Because these morphemes usually depend on the stems with which they are associated they tend to occur almost as frequently as these associated stems. Hence, there may not be any significant differences between the frequencies of occurrence of the stems and the frequencies of occurrence of the bound morphemes, thereby compromising the high successor frequency or the frequent flyer expectations put on such morphemes.

Even though these morphemes do not manifest as frequent segments, they feature patterns which occur regularly within the lexicon. Hence, the problem of inducing the morphological rules that guide these morphological processes reduces to that of extracting these patterns. They may then be used as the relevant features on which to base the clustering of words according to the morphological processes through which they were formed.

This study analyzed this problem for Yorùbá and proposed an algorithm by which the patterns can be identified and used to cluster words produced by various morphological processes.

2. YORÙBÁ MORPHOLOGY
Yorùbá is a West African language that is widely used as a language of everyday communication in Nigeria, Benin Republic and The Republic of Togo as well as a language of *Òrìsà* religious worship in Cuba, Brazil and a number of Caribbean countries. According to [13], prefixation, reduplication and compounding constitute the main processes that characterize Yorùbá morphology. [14] in discussing the essentials of Yorùbá morphology included interfixation and desententialisation as significant morphological processes used in Yorùbá word formation. Later on, [15] observed that prefixation features prominently in Yorùbá morphology

WWW 2016 Companion, April 11-15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890563

showing copious examples of various monosyllabic and disyllabic prefixes. [14] supported these observations and further noted that these prefixes can be affixed to verbs, verb phrases, nouns, adverbs and adverbial phrases to fulfil various morphological functions. [14] also demonstrated that Yorùbá morphology features both partial and full reduplication. In the partial reduplication strategy a consonant and vowel (CV) prefix template is affixed to a stem, the C being a copy of the first consonant of the stem and the V being the high tone *'í'*. Table 1 shows examples of nouns derived from monosyllabic verbs, while Table 2, Table 3 and Table 4 show examples of the processes of full reduplication, interfixation and compounding respectively.

Table 1. Yorùbá examples of partial reduplication

Derived Noun	Gloss	Verb	Gloss
lílọ	going (N)	lọ	go (V)
rírìn	walk (N)	rìn	walk (V)
rírà	buying (N)	rà	buy (V)

Table 2. Yorùbá examples of full reduplication

Derived Noun	Gloss	Noun Phrase	Gloss
panápaná	fire fighter	pa iná	put out fire
túlétúlé	disruptive person	tú ilé	undo household
gbọ́mọgbọ́mọ	kidnapper	gbé ọmọ	steal child

Table 3. Yorùbá examples of interfixation

Derived Form	Gloss	Noun	Gloss
ọmọkọ́mọ	any child/bad child	ọmọ	child
iyebíye	invaluable	iye	value
àgbàlagbà	old/matured person	àgbà	adult

Table 4. Yorùbá examples of compounding

Derived Form	Gloss	Source Words	Gloss
etídò	river bank	etí odò	near river
ìdíkọ̀	motor park	ìdí ọkọ̀	place for vehicles
gbaná	catch fire	gba iná	catch fire

[15] as well as [14] also observed that desententialisation, in which a sentence is fused to become a noun is another important Yorùbá morphological process. Examples of such include a word like *kòtẹ́milọ́rùn* (appeal) derived from the sentence *kò tẹ́ mi lọ́rùn* (I am not satisfied).

3. WORD-SEGMENT FREQUENCY vs WORD-PATTERN FREQUENCY

In the morphology induction methods presented in the works of [3], [5], [9], [8] and a number of other scholars, the frequency of occurrence of certain word segments described as recurrent partials is the main feature. Given a balanced English lexicon for example, the word-segment 'ing' is bound to occur with relatively high frequency, thereby indicating appropriate morpheme boundaries in words such as going (go-ing), walking (walk-ing), eating (eat-ing) and teaching (teach-ing). Such a feature belongs to a mertizable space and so clustering can be undertaken using parameter-based approaches such as k-means.

The frequency of recurrent partials fail to identify morphological processes such as reduplication, interfixation and desententialisation which feature frequent patterns rather that frequent segments. Smooth functions that map these patterns to real numbers may not be available and so clustering by k-means or any other parameter-based approaches may not be possible.

To induce the morphology of Igbo; another widely spoken Nigerian language that features similar morphological processes, [16] proposed the use of 'word labels' as a textual proxy of the word patterns. These word labels are derived from the words by assigning a sequence of symbols C or V representing consonants or vowels accompanied by numerical indexes indicating the occurrence or reoccurrences of specific consonants or vowels in the words from left to right.

Table 5 shows word labels for; *'deal', 'said', deed'* and *'seek'*. The word label for *'deal'* is *C0V0V1C1* because the first character; *'d'* is assigned the symbol *C0* and the first vowel *'e'* is assigned the symbol *V0*. The succeeding characters *'a'* and *'l'* are assigned the symbols *V1* and *C1* respectively because they are the second occurring vowel and consonant respectively.

Table 5 Words and word labels

Word	Word Label
deal	C0V0V1C1
said	C0V0V1C1
deed	C0V0V0C0
seek	C0V0V0C1

The same process produced the same label for *'said'*. In the case of *'deed'* however, the first and second consonants as well as the first and second vowels are the same so they get assigned the same symbols hence the label *C0V0V0C0*. Thus, *'deal'* and *'said'* will both be clustered arond the label *C0V0V1C1*, while *'deed'* and *'seek'* will be clustered around the labels *C0V0V0C0* and *C0V0V0C1* respectively.

Subjecting an Igbo lexicon of about 30,000 word tokens to this treatment, about 2,300 word labels that clustered words around morphologically significant clusters were derived [16]. The labels were rather 'brittle' though as both deliberate and chance repetition of a consonant or vowel changed the word label around which a word was clustered. The word labels being categorical rather than rational variables, it was not possible to use them as basis of parameter-based clustering such as k-means. The word labels were therefore manually classified and segmented according to morphological structures. The study reported that a morphological analyzer based on the manual classification and segmentation of the

word labels recorded up to 88% accuracy in the determination of morpheme boundaries. [16].

Even though the manual classification of only 2,300 word labels as against the manual classification of 30,000 words represents a huge reduction in time, labor and financial costs, it would still be necessary to explore ways by which words formed by the same morphological processes can be automatically clustered. This forms the main quest of the present study.

4. YORÙBÁ WORD LABELS

Data for the present study was obtained by manually typing six published Yorùbá texts. These are *Ààrò Méta* by Oladele Sangotoye, Abdullahì Awọ̀lúmátẹ̀ẹ́ and Adesoye Ọmọlasoye, *Ẹkùn ń bímọ* by Adeboye Babalola, *Gbọ́baníyì* by Oladipo Yemitan and *Ogún Ọmọdé* by Akinwumi Isola. In addition *Àyájọ́* by Ayọ̀ Ọ̀pẹ́fèyítìmí, and *Orin Ọdẹ fún Àṣeyẹ* by Adeboye Babalola were also used, producing approximately 400 pages of text, resulting in an approximately 210,853 word document.

The 210,853 word corpus produced 14,670 Yorùbá word tokens, which clustered around 1,282 word labels based on the technique proposed in [16] as explained in Section 3.

Word labels corresponding to all the morphological processes discussed in the literature of Yorùbá morphology were found among the 1,282 cluster of words. Table 6 – Table 8 show some of these labels, some of the words clustered around them and the morphological processes through which these words were formed.

Table 6 Words clustered around prefixation labels

Label	Word	Word Gloss	Prefix-Stem
V0C0V1	àṣẹ	command	à-ṣẹ
V0C0V1	ìṣe	behaviour	ì-ṣe
V0C0V1	ìfẹ́	love (N)	ì-fẹ́
V0C0V1	àṣà	culture	à-ṣà
V0C0V1C0V2	àṣìṣe	mistake	à-ṣìṣe
V0C0V1C0V2	ìgbàgbọ́	belief	ì-gbàgbọ́

Table 7 Words clustered around partial reduplication labels

Label	Word	Word Gloss (N)	PrefixTmplt-Stem
C0V0C0V1	lílọ	going	lí-lọ
C0V0C0V1	rírìn	walk	rí-rìn
C0V0C0V1	bàbá	father	N/A
C0V0C0V0	ṣíṣí	opening	ṣí-ṣí
C0V0C0V1C1V2	kíkígbe	shouting	kí-kígbe

Table 8 Words clustered around full reduplication labels

Label	Word	Word Gloss	Stem
C0V0C1V1C0V0C1V1	panápaná	fire fighter	paná
C0V0C1V1C0V0C1V1	túlétúlé	disorganizer	túlé
C0V0C1V0C0V0C1V1	gbọ́mọgbọ́mọ	kidnaper	gbọ́mọ

As can be observed from these tables, both prefixation and partial reduplication featured more than one word labels each. For example the word label *C0V0C0V0* for the word *ṣíṣí* is different from *C0V0C0V1* for *lílọ* mainly due to the chance occurrence of the high tone *'í'* as the vowel of the stem *ṣí* in *ṣíṣí*. To be noted also is the fact that the word label *C0V0C0V1* which clusters words formed by the partial reduplication process also accommodated the word *bàbá* which is a stem and hence could not be a product of the partial reduplication process.

The main challenge of this study is the further refinement of the clustering by word labels so that all words formed by a given morphological process are identified and uniquely clustered with little or no human intervention, as an improvement over [16].

4.1 Refining word clusters

The objective of refining the word clusters is to ensure that each cluster of words represents a unique morphological process as well as differentiates words formed by different morphological processes.

In cases where a single morphological process is represented by several word labels, the inherent patterns conferred on the words by the relevant morphological processes remain apparent in the various word labels. This offers the possibility of and need for a higher ordered clustering of the word labels rather than just the words. Hence the problem of identifying all labels representing a single morphological process can be solved by a further clustering of the word labels based on their underlying patterns.

As for cases in which more than one morphological process produced words clustered around a single word label, it was observed that in such cases the mostly affected words were formed by a mixture of more than one morphological process. In such cases, there was a mixture of word-segment and word-pattern motivated morphological processes.

The problem of separating words from different morphological processes clustered around a single word label can therefore be approached by analyzing and isolating the influences of the contributing morphological processes. It is particularly necessary to separate the influences of morphological processes based on word segments from the influences of morphological processes based on word patterns in each word clustered around such a single word label.

4.2 Clustering labels with common patterns

The morphological processes described in the literature of Yorùbá morphology are prefixation, partial and full reduplication, compounding, interfixation and desententialisation. Some of these morphological processes manifest clearly recognizable patterns. For example partial and full reduplication as well as interfixation exhibit obvious patterns that are replicated both in the words formed through these processes and the word labels around which these words cluster. These obvious patterns provide basis on which the word labels they generate can be easily clustered, thereby implicitly clustering the words already clustered around these word labels hierarchically around relevant morphological process.

It can be concluded that word labels beginning with the *C0V0C0* pattern will cluster all words formed by the partial reduplication process. These include word labels such as *C0V0C0V0, C0V0C0V1, C0V0C0V1C0V2* and others which clustered such words as *títí* (spent make-up), *pípé* (complete) and *gbígbàgbé* (forgetting (N)) respectively.

Even though each of these labels may also cluster words not formed by the partial reduplication process, the issue of separating such words would be dealt with in section 5.2.

Full reduplication features obvious symmetry in the word labels. This is manifested in the possibility of splitting any word label representing full reduplication into two identical parts. Examples are labels such as *C0V0C0V0C0V0C0V0, C0V0C1V0C0V0C1V0, C0V0C1V1C0V0C1V1* clustering words like *títítítí* (continuously), *pátápátá* (completely) and *lemọ́lemọ́* (persistently) respectively.

Interfixation features two incidences of a stem sandwiching a bound morpheme that is usually based on the stem. Examples of such word labels include *V0C0V1C0V0C0V1* and *V0C0V1C1V0C0V1* which accommodate *ayérayé* (whole world) and *ọwọ́wọwọ́* (collaborative) respectively.

Due to some phonological and stylistic features of the Yorùbá language and literature, tonal variations are sometimes applied to some of the vowels in words formed by the reduplication and interfixation processes, sometimes rendering the two incidences of the stem in the resulting words slightly differently. Examples include words like *gbangbagbàngbà* (abundantly clear) producing the word label *C0V0V1C0V0C0V2V3C0V2* rather than the otherwise expected *C0V0V1C0V0C0V0V1C0V0* which could have been easily identified as a word label for reduplication. Also *ọmọkọ́mọ* (any child) which for phonological reasons is not rendered as ọmọkọmọ produces *V0C0V0C1V1C0V0* rather than the expected *V0C0V0C1V0C0V0*. There is a need to take account of these tonal variations in clustering word labels that code these morphological processes.

Prefixation, compounding and desententialisation do not exhibit word patterns that can be easily recognized from their surface structures. For this reason, it may not be possible to identify word labels that represent these morphological procceses and cluster them automatically based on the patterns they manifest.

In the case of prefixation, it was noted in [15] that Yorùbá features various monosyllabic and disyllabic prefixes, which are recurrent partials that manifest as frequent word segments. Hence, prefixation can be identified by the presence of frequent word segments which shall be addressed in section 5.2.

As for compounding and desententialisation, we do know from literature that words that result from these processes are formed by the combination of stems which are by definition free morphemes. This implies that these stems, being free morphemes could be found in the study lexicon. Hence, long word labels that can be decomposed into recognizable shorter word labels are likely to cluster words that are products of compounding and desententialisation.

To convert the foregoing arguments into actionable tasks, there is a need to implement some commonly known algorithms that can be used to:

- Collate all word labels that begin with the *C0V0C0* pattern as candidates for partial reduplication
- Collate all word labels that can be split into two identical segments as candidates for full reduplication
- Collate all word labels in which two identical segments sandwich a middle segment as interfixation
- Collate long word labels that can be decomposed into recognized shorter word labels.

4.3 Separating words from disparate morphological processes in common clusters

The clustering of words from disparate morphological processes around a single word label is observed to be majorly due to a combination of both word-segment motivated and word-pattern motivated morphology strategies in the morphological processes that produce such words. This information can be used to separate words formed by these different morphological strategies by identifying the specific processes involved.

Word labels cluster words formed by word-pattern motivated morphological strategies. Hence the fact that a given word is clustered around a given word label already identifies the word pattern that motivated the morphological structure involved. In order to identify the accompanying word-segment motivated morphological processes, it would be necessary to identify the recurrent partials in such words. The recurrent partials found in words of a given word label are bound to align within the same word positions and this should aid their easy identification.

The partial reduplication process of the Yorùbá language is based on the affixation of a consonant and vowel (CV) prefix template to a stem, the C being a copy of the first consonant of the stem and the V being the high tone *'í'* [14]. This is a clear case of a mixture of the word-pattern-based and word-segment-based morphology strategies in a single morphological process. The consistent use of the high tone *'í'* as the vowel in the second position of words formed by partial reduplication indicates that words with the character *'í'* in the second position can be separated from all other words under word labels beginning with the *C0V0C0* pattern.

Hence, the occurrence of the high tone *'í'* in the second position of words such as *lílọ*, *ríríìn* distinguishes them from words such as *bàbá* in which the consonants in the first and third positions are identical but the vowel in the second position is not *'í'*.

5. TESTS AND RESULTS

The following tests were carried out to investigate the validity of the various suggestions for the further refinement of word clusters.

5.1 Partial Reduplication

80 word labels were found to begin with the *C0V0C0* pattern and these 80 word labels had 779 words clustered around them.

The word label C0V0C0V1C0V0C0V1 which clusters words such as *kíkankíkan* are products of both partial and full reduplication processes. Also noteworthy is the word *bíbélì* which is the Yorùbá loan word for bible which is not a product of partial reduplication but could be misconstrued as a product of partial reduplication cluster around the word label *C0V0C0V1C1V2*.

5.2 Full Reduplication

Nineteen word labels were found to manifest perfect symmetry and therefore can be split into two identical segments indicating that they are word labels for full reduplication. These nineteen word labels have a total of 432 words clustered around them and the word label with the largest cluster is *C0V0C0V0* with 147 words. As was predicted in the discussions in section 5.1, the word label C0V0C0V1C0V0C0V1, which accommodates the word *kíkankíkan* formed both by the partial and full reduplication processes was also included in this collation of full reduplication word labels.

5.3 Interfixation

Only three word labels; C0V0C1V0V1C0V0C1V0, V0C0V1C1V0C0V1, and V0C0V0C1V0C0V0 manifested the type

of symmetry consistent with interfixation. However when the necessary phonological adjustment anticipated in section 4.2 was made, six more word labels were identified making a total of nine word labels representing interfixation. Typical words clustered around these labels include *iyebíye* (invaluable), *ìwàkíwà* (undesirable/random behavior) and *ìrandíran* (from generations to generation).

An observed exception is the word label V0C0V1C1V1C0V1 which featured the word *afàyàfà* (crawler). Even though the word manifests a pattern that is consistent with interfixation, it is formed by the morphological processes of prefixation and compounding, with certain vowels elided in the compounding process.

5.4 Compounding and desententialisation

Compounding and desententialisation do not manifest patterns that are easily observable at the surface structure. However, because they use free morphemes they can be identifying as word labels that can be broken into known shorter sub-labels.

5.5 Disparate morphological processes in a common cluster

Information theory dictates that a word position with recurrent partials would manifest relatively lower entropy. The noticeably lower entropy of Position 2 in the first row of Table 9 is indicative of the presence of a recurrent partial in the word label *C0V0C0V1*.

Table 9 Entropy of the four positions in two word labels

	Position 1	Position 2	Position 3	Position 4
C0V0C0V1	4.01	3.49	4.01	4.85
V0V1C0V2	3.33	3.83	4.88	4.71

Chart 1 Occurrence probabilities of consonants in *C0V0C0V1*

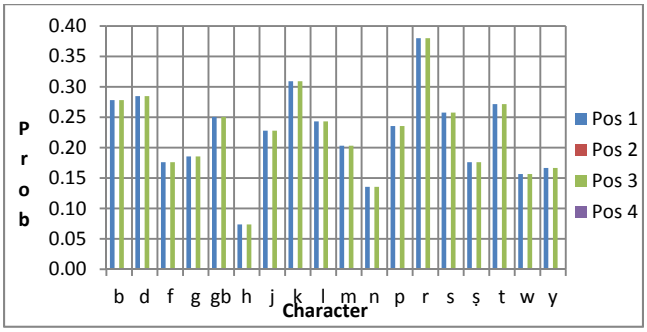


Chart 2 Occurrence probabilities of vowels in *C0V0C0V1*

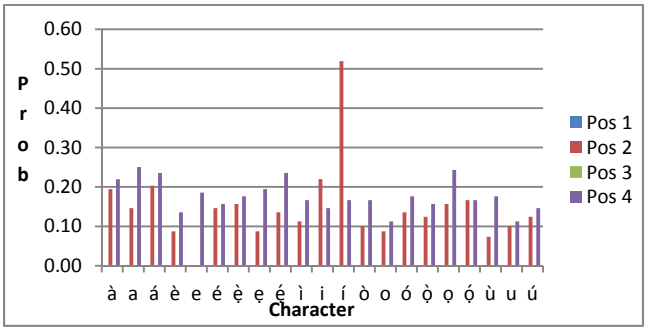
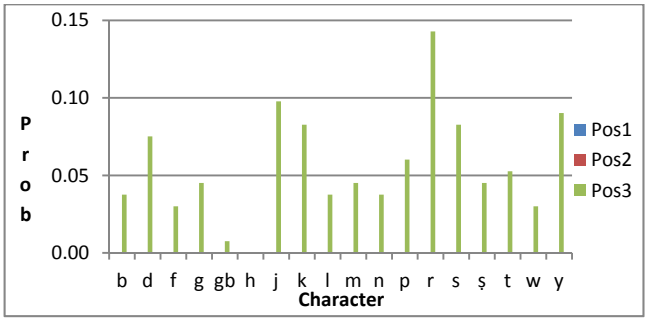


Chart 1 and Chart 2 offer graphical presentations of the occurrence probabilities of the consonants and vowels in the in the four positions of words clustered around this word label. As expected, *'í'* stands out as the vowel with the highest occurrence probability in Position 2, indicating that it is a recurrent partial in the words clustered around this word label.

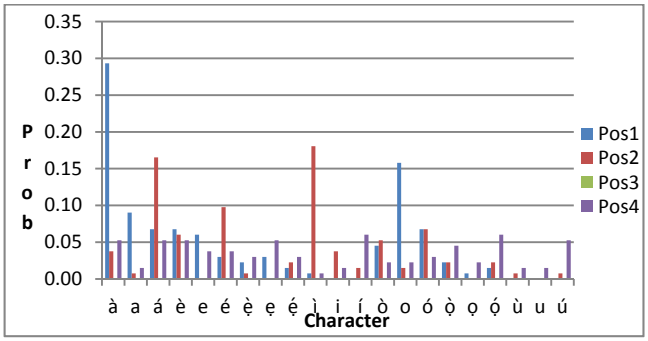
The word label *V0V1C0V2* contains recurrent partials in multiple contiguous positions. This can be inferred from the second row of Table 9 which shows the entropy values of the four positions in this word label. The entropy values of the first and second positions are noticeably lower than those of the third and fourth positions, indicating the presence of recurrent partials in the first two positions.

Chart 3 Occurrence probabilities of consonants in *V0V1C0V2*



Correspondingly, Chart 4 shows a noticeable gap in the occurrence probabilities of two sets of characters. In Position 1, the characters *à* and *o* have occurrence probabilities of 0.29 and 0.16 respectively and in position 2, characters *á, é* and *ì* feature occurrence probabilities of 0.17, 0.10 and 0.18 respectively. The occurrence probabilities of this set of characters are noticeably higher than those of the set containing all other characters. The observation that the characters *à* and *ì* have the highest occurrence probabilities in Positions 1 and 2 respectively is consistent with the observation in [15] that the recurrent partial *àì* is frequently used in Yorùbá as a negation prefix.

Chart 4 Occurrence probabilities of vowels in *V0V1C0V2*



In a similar fashion to the observation of the word *bíbélì* is an exception being a loan word in section 5.1, the word *àìsáyà* which is the Yorùbá rendition of the Hebrew name Isaiah was also observed as an exception to the use of the recurrent partial *àì* as a negation prefix in Yorùbá.

6. Conclusion

These results bear compelling evidence of the productivity of word labels as a potentially useful feature for the unsupervised induction of Yorùbá morphology.

The exceptions in the misallocation of certain loan words such as *bíbélì* the Yorùbá loan for bible and *àìsáyà* the Yorùbá rendition of the Hebrew name Isaiah further strengthens the proposition that such exceptions may point to the foreign origins of such exceptional words. However, to detect such exceptions automatically may require much larger corpora and semantic analysis.

A major attraction to unsupervised induction of morphology is its scalability to other languages. Hence dependence on foreknowledge and use of Yorùbá morphology in this study may be regarded as a limitation of the use of word labels as a feature on which the induction of the morphology of many languages could be based. The algorithm for the separation of words formed by disparate morphological process clustered around a single word label can be easily up-scaled for use in any language. However, the methods used for collating various word labels that represent the same morphological process need to be generalized beyond the known constraints of Yorùbá morphology that were applied to collate them in this study.

Morphology is foundational to many other levels of linguistic analysis and so, capacity for the automatic induction of morphology is bound to have a positive impact on higher levels of linguistic analysis. The development of NLP capacity for the 2,000 odd African languages that are yet to be subjected to NLP will receive a big boost with unsupervised induction of Morphology of these languages. It would be desirable therefore to experiment with this pattern-based morphology induction for other African languages in order to determine how well the method can up-scale to African languages other than Igbo and Yorùbá that have been so far engaged. To achieve the desired up-scaling, it would be necessary to modify the various methods that were based on foreknowledge of Yorùbá morphology in this study. This will enable the development of computational tools that can automatically induce the morphology of these many Africa languages with little or no human intervention.

The present study has concentrated mainly on establishing the viability of word labels as a productive feature for the computational induction of Yorùbá morphological rules and to devise means by which these word labels can be clustered around appropriate morphological processes automatically. Having established the viability of the use of word labels in this manner, it would be necessary in future studies to investigate the level of accuracy offered by methods based on word patterns rather that word segments. Also of interest would be the comparison of the quality of morphological analyzers based on the word-pattern approach to those based on the word-segment approach for languages such as English and the many other languages around which the word-segment approach was developed.

It was observed that certain word labels clustered spelling errors of Yorùbá words while some other word labels clustered foreign words. Future studies on the effectiveness of this approach in the automatic editing of written texts and the identification of foreign words in texts will also be desirable.

ACKNOWLEDGMENTS

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).

8. REFERENCES

[1] K. Koskienniemi, "Two Level Morphology: A General Computational Model-Form Recognition and Production," University of Helsinki, 1983.

[2] K. R. Beesley and L. Kartunen, Finite State Morphology, California: CSLI Publications, 2003.

[3] Z. Haris, "From Phoneme to Morpheme," *Language,* pp. 192-222, 1955.

[4] Z. Haris, "Morpheme Boundsries Between Words," University of Pensylvania, 1967.

[5] J. Goldsmith, "Unsupervised Learning of Morphology of a Natrural Language," *MIT Press Journal,* pp. 153-198, 2001.

[6] H. Hammarström, "Unsupervised Learning of Morphology and the Languages of the Wolrd," Gothenbiurg, 2009.

[7] H. Hammarström and L. Borin, "Unsupervised Learning of Morphology," *Computational Linguistics,* vol. 37, no. 2, pp. 309-350, 2011.

[8] M. Creutz and K. Lagus, "Unsupervised Models for Mopheme Segmentation and Morphology Learning," *ACM Transactions on Speech and Language Processing Vol 4, No. 1 Article 3,* 2007.

[9] M. Creutz, "Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition," Helsinki, 2006.

[10] N. Vasudevan and P. Bhattacharyya, "Probabilistic Approach for Automatic Last Suffix Extraction," in *Proceedings of 8th Inernational Conference on Natural Labguage Processing*, Kharagpur, 2010.

[11] G. De Pauw and P. W. Wagacha, "Bootstrapping Morphological Analysis of Gıkuyu Using Unsupervised Maximum Entropy Learning," 2007. [Online]. Available: http://aflat.org/files/0663anav.pdf. [Accessed 23 July 2013].

[12] G. De Pauw and G. M. de Scryver, "African Language Technology: The Data-Driven Perspective," in *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics*, 2009.

[13] K. Owolabi, "More of Yorùbá Prefixing Morphology," in *Language in Nigeria*, Ibadan, Group Publishers, 1995.

[14] O. Yusuf, "Yorùbá Morphology," in *Basic Linguistics for Nigerian Language Teachers*, Port Hacourt, LAN GRand Orbit Communications, 2007.

[15] O. Awobuluyi, Ẹ̀kọ́ Ìṣẹ̀dá-Ọ̀rọ̀ Yorùbá, Akure: Montem Paperbacks, 2007.

[16] O. U. Iheanetu, "A DATA-DRIVEN MODEL OF IGBO MORPHOLOGY (Unpublished PhD Thesis)," Ibadan, 2015.