Tone Mark Restoration in Standard Yorùbá Text: A Proposal

Asahiah Franklin Ọládiípọ̀[1]
Ọdẹ́jọbí Ọdẹ́túnjí Àjàdí[1]
Adagunodo Emmanuel Rotimi[1]
Olúbòdé-Sàwẹ́ Fúnmi O.[2]

Obafemi Awolowo University
Ile-Ife, Nigeria

[2]General Studies Unit,
Federal University of Technology,
Akure, Nigeria
[1]sobusola,oodejobi,eadagun@oauife.edu.ng
[2]saweff@yahoo.co.uk

Abstract. Restoring diacritics has for the most part relied either on the letter (grapheme) or the space-delineated linguistic block (word) as the lexical focus item. The usage of letter for Yorùbá text was often adduced to resource scarcity and the underlying model being language independent. On the other hand, insufficient contextual information for tone mark restoration using letters was cited for the limited performance of letter-based models. Thus, another research proposed the usage of the word as lexical token for restoration of tone marks in Yorùbá text. The result of existing word-based tone-mark restoration approaches did not indicate any improvement over the letter-based approach despite a larger training data. This situation might be due to the resource scarcity problem. In this paper, we therefore propose an alternative approach that is expected to address the twin challenges of resource scarcity and contextual insufficiency for tone mark restoration in Yorùbá text in particular and resource-scarce tone languages in general. This approach is also expected to be linguistically sensible as it tries to relate the tone mark restoration task to orthographic function of tone marks in the text to the positioning of tone within the linguistic units of the language. We propose tone mark restoration for Yorùbá text using syllables as the lexical focus, that is,syllable-based tone mark restoration for Yorùbá text.

Keywords: syllable, tone mark, restore

(Received June 5th, 2016 / Accepted March 10th, 2017)

1 Introduction

According to Encyclopedia Britannica, language is a system for communication, expression of identity and emotions. Language is effected or realized via a set of symbols which may be vocal, written or gestural. While all languages use modification in pitch for different functions, tone languages use changes in pitch patterns for lexical and/or grammatical forms of lexical items differentiation [14]. At least 60% of all languages worldwide are tone languages [29] and in Africa, the proportion increases to about 80% [4].

The term "writing" has various shades of meanings which includes the activity of forming visible or tactile marks (or letters) for language expression or the outcome of that activity or exercise. A writing system refers to the way sounds or words of human languages are written as well as the particular way of writing a specific language [5]. The alphabetic writing system makes use of distinct symbols for consonant and vowel sounds to form words. The Latin script, a subset of the alphabetic writing system, is more widely used than any other script (whether alphabetic or not) [27].

Many tone languages are written with modified Latin scripts since the basic Latin script was not designed to represent the tones in these languages where tones are integral parts of the sound systems. There are four common systems for representing tone in the orthography of languages: using diacritics, punctuation marks, numbers and unused consonant letters [12]. The modifications to the basic Latin script to enable it represent tone were accomplished mainly by the introduction of diacritics [27], symbols "placed above, through or below a letter, in order to indicate a sound different from that indicated by the letter without the diacritic" or "marks added to glyphs to change their meaning or pronunciation" [27, 8].

There are several languages all over the world and many in Africa that have modified Latin script-based orthography. In many of these orthographies, the non-standard forms, that is, with the diacritics not consistently used, are the commonest. The Yorùbá language is one of such languages whose orthography utilises the modified Latin script and tones are indicated by diacritic marks. However, these diacritic marks are not always used in many Yorùbá documents.

1.1 Classification and Degree of Use of Yorùbá

Yorùbá language is a member of the Benue-Congo subclass of the Niger-Congo class of languages [2]. It is spoken in West Africa, mainly within Nigeria where it has the status of a major language by more than 35 million people and also by a sizable number of speakers in Republic of Benin. A few speakers can also be found in Ghana, Sudan, Sierra-Leone and Côte D'Ivoire. Outside Africa, the language is used for religious purposes in Brazil, Cuba, as well as Trinidad and Tobago [7]. In Nigeria, it is a de facto provincial language in South-West, and is a language of education, both as a medium of teaching in early primary education and as a school subject in primary, secondary and tertiary institutions, up to the post-graduate level. Yorùbá is also a language of the mass media in Nigeria, being used by both private and governmental organizations. It is also taught in some colleges in the United States. Yorùbá is actually a dialect continuum, with the estimated number of dialects ranging between twelve and twenty-six [26]. These dialects have been grouped together in a way to suggest that geographical constraints have some effect on the pattern of distribution. According to Akinlabi [3], these dialect sub-groups are: North Western, South Eastern, Central, North Eastern and South Western. Significant linguistic variations in properties of the dialects are shown by these five areas [20]. There is, however, one dialect which is not tied to any geographical area and it is referred to as Standard Yorùbá (which we will henceforth be simply referred to as Yorùbá). This dialect is intelligible to the speakers of the geographical dialects and is the form taught in educational institutions and used in mass media publications, advertisements and by government agencies for dissemination of government information.

1.2 Description of Yorùbá sound system and orthography

YorùbÃ¡ has three phonological components namely, consonants, vowels and tones. There are eighteen consonants and twelve vowels; seven oral and five nasal. In addition, there are two syllabic nasals which can function as syllable nuclei but do not combine with consonants to form syllables. Finally, YorùbÃ¡ has three contrasting tones. The details of the phonological components are as follows:
The eighteen consonants: b d f g gb h j k l m n p r s á^{1}£ t w y are illustrated with verbs as follows in Table 1 The oral vowels a e e. i o ọ u are

Table 1: The eighteen consonant illustrated with verbs

S.N	**consonant**	*Word*	*Gloss*
1	**b**	**bá**	*'meet'*
2	**d**	**dá**	*'cut'*
3	**f**	**fá**	*'shave'*
4	**g**	**ga**	*'be tall'*
5	**gb**	**gbá**	*'sweep', 'hit'*
6	**h**	**há**	*'be narrow'*
7	**j**	**já**	*'cut'*
8	**k**	**ká**	*'pluck'*
9	**l**	**lá**	*'lick'*
10	**m**	**má**	*'don't*
11	**n**	**ná**	*'spend'*
12	**p**	**pa**	*'kill'*
13	**r**	**rá**	*'disappear'*
14	**s**	**sá**	*'run'*
15	**ṣ**	**ṣá**	*'be faded '*
16	**t**	**ta**	*'be spicy'*
17	**w**	**wá**	*'come'*
18	**y**	**yá**	*'loan'*

illustrated, also with monosyllabic verbs, in Table 2. Where nasal vowels occur after an oral sound, they are orthographically indicated by digraphs of an oral vowel and 'n': an, ẹn, in, ọn and un as shown in the words in Table 3.

Table 2: The oral vowels illustrated with verbs

S.N	**Oral Vowel**	*Word*	*Gloss*
1	**a**	**rà**	*'buy'*
2	**e**	**rè**	*'go to'*
3	**ẹ**	**rẹ́**	*'cut'*
4	**i**	**rì**	*'sink'*
5	**o**	**rò**	*'think'*
6	**ọ**	**rọ̀**	*'be soft'*
7	**u**	**rù**	*'lose weight'*

Table 3: The nasal vowels illustrated with words

S.N	**nasal vowel**	*Word*	*Gloss*
1	**an**	**ìran**	*'vision'*
2	**ẹn**	**ìyẹn**	*'that one'*
3	**in**	**ìrìn**	*'walk'*
4	**ọn**	**ọgbọ́n**	*'wisdom'*
5	**un**	**òórùn**	*'smell'*

The nasals ***an*** and ***ọn*** are usually in free variation except after labial consonants. So ìtàn 'story' may be written as ìtọ̀n, but ìbọn 'gun' is not *iban and ọgbọ́n 'wisdom' is not *ọgbán [19]

Syllabic nasals can occur within words (as shown in items no 1-7 of Table 4) or in phrases, where it marks the progressive (continuous) aspect (items 8 -12 of Table 4).

Table 4: Syllabic nasals in words and as progressive marker

S.N	*Word*	*Gloss*
1	**òroǹbó**	*'orange'*
2	**Oǹdó**	*name of town*
3	**àǹfààní**	*'benefit, opportunity'*
4	**gban̄gba**	*'plain view'*
5	**èròǹgbà**	*'objective'*
6	**ońjẹ**	*'food'*
7	**pan̄la**	*'stockfish'*
8	**ó ń bọ̀**	*'he is coming'*
9	**ó ń hó**	*'it is boiling'*
10	**ó ń jò**	*'it is leaking'*
11	**ó ń ká**	*'he is curling up'*
12	**ó ń sín**	*'he is sneezing'*

Before vowels, the syllabic nasal occurs only in clauses as an alternant of the first person pronoun clitic (mi) before the negative particle. Though pronounced as a velar nasal, it is orthographically represented as 'n'. See Table 5.

Table 5: Syllabic nasals as alternant of the first person pronoun clitic (mi)

S.N	*Word*	*Gloss*
1	**n ò mọ̀/n̄ ò mọ̀**	*'I do not/ did not know'*
2	**n kò fẹ́/n̄ kò fẹ́**	*'I do not want(it)'*

Yorùbá is a tone language, that is, an indication of pitch enters into the lexical realization of morphemes. Five tones are attested but only three are distinctive: high [H], mid [M] and low [L]. The other two can be argued as phonetic realizations of the others, depending on the phonological environment. [H] is orthographically represented by the acute mark (´), [M] is usually unmarked except on syllabic nasals where it is indicated orthographically by a macron (¯) and [L] is orthographically represented by grave mark (`) [21]. The tones are phonemic and are used for lexical contrast; thus minimal pairs can be created by tone variation as in Example 1 below.

Example 1: Minimal pairs created by tone variation

word	***gloss***
lọ	*'go'*
lọ́	*'be gnarled'*
lọ̀	*'grind'*

Tones in Yorùbá occur relatively indiscriminately [19]. The three tones can occur on monosyllabic verbs as shown in Example 1. In addition, in longer words, variation of tone can occur on one syllable or more to create minimal pairs in disyllabic words as shown in Example 2 or on longer words on syllables as shown in Example 3.

Example 2: Tone variation on disyllabic words

word	tone	*gloss*
ọkọ	(MM)	*'husband'*
ọkọ́	(MH)	*'hoe'*
ọkọ̀	(ML)	*'vehicle'*
ọ̀kọ̀	(LL)	*'spear'*
igba	(MM)	*'two hundred'*
igbá	(MH)	*'calabash'*
igbà	(ML)	*'climbing rope'*
ìgbá	(LH)	*'garden egg'*
ìgbà	(LL)	*'time', 'period'*

Example 3: Tone variation on longer words

word	tone	*gloss*
akòko	(MLM)	*type of tree*
àkókò	(LML)	*'time'*
àkókó	(LHH)	*'wood pecker'*
oòrè	(MLL)	*traditional title*
oore	(MMM)	*'an act of kindness'*
akọ́rọ́	(MHH)	*'billhook',*
àkọ́rọ̀	(LHL)	*'first rains'*

Each word in Example 2 contains the same sequence of vowels and consonants as another word in the list (ọ+k+ọ or i + gb + a) but may have different tones on the first or second syllable to produce different Yorùbá words. Example 3 shows similar variations in longer words. This underscores the importance of tone marking in Yorùbá texts.

Structurally, Yorùbá words are composed of one or more open syllables. The Yorùbá syllable may have one of these three forms: consonant plus vowel, vowel only or syllabic nasal. The syllable structure may be represented as $[\tau/CV]$ combination or $[\tau/V]$ or $[\tau/S]$; where τ stand for the tone, C stands for consonant, V stands for vowel and $[S]$ stands for the tone-bearing syllabic nasal. Closed syllables and consonant clusters are not permitted [3] and the total number of all kinds of syllables in Yorùbá is 690 [18].

Standard Yorùbá orthography is composed of consonants, vowels and tones. Syllables each carry a single tone which is indicated in the orthography as a tone mark over either the oral vowel or syllabic nasal. Subdots are employed on certain characters to indicate different phonemic qualities

Example 4: Diacritics marking tone and phonemic quality

word	**IPA**	**tone**	***gloss***
oko	/oko/	(MM)	*'farm'*
òkò	/òkò/	(LL)	*'stone'*
ọkọ	/ɔkɔ/	(MM)	*'husband'*
okó	/okó/	(MH)	*'penis'*
ọkọ̀	/ɔkɔ̀/	(ML)	*'vehicle'*
ọkọ́	/ɔkɔ́/	(ML)	*'hoe'*
ọ̀kọ̀	/ɔ̀kɔ̀/	(LL)	*'spear'*
ejò	/eʤò/	(ML)	*'snake'*
ẹjọ́	/ɛʤɔ́/	(MH)	*'law suit'*
èso	/èso/	(LM)	*'fruit'*
ẹ̀ṣọ́	/ɛ̀ʃɔ́/	(LH)	*'ornament'*

These items in Example 4 demonstrate the importance of diacritics in Yorùbá orthography.

1.3 Diacritic Use and Level of Usage in Yorùbá Orthography

The sub-dots modifying three letters of the Yorùbá alphabet (two vowels and one consonant) and the tone marks placed on vowels (and syllabic nasals) to indicate tones alter such letters and, in accordance with the above definitions, are diacritics. Thus, Yorùbá orthography has two kinds of diacritics: phonemic (marking vowel quality) and tonemic (marking tone quality).
The diacritically marked graphemes in Yorùbá orthography are listed as follows:

i **With dot-below only:** ẹ, ọ, ṣ

ii **With only tone marks:** á, é, í, ó, ú, à, è, ì, ò, ù, m̀, ǹ, m̄, n̄, ḿ, ń

iii **With dot-below + tone marks:** ẹ́, ọ́, ẹ̀, ọ̀,

A well written Yorùbá textual document is expected to indicate the phonemic diacritics at all times and the tone diacritics sufficiently enough to minimize ambiguity for readers. This definition of a well written text has however scarcely been adhered to [1] except in educational textbooks. The tone diacritics are the most violated, being either totally ignored, randomly used or wrongly used in many written texts. The absence of tone marking may be a minimal problem for human readers of the text who rely on context and diverse domains knowledge to disambiguate in real-time.
However, the absence of these diacritics from Yorùbá text poses a significant challenge for natural language processing systems where it may either

lead to additional processing overhead or be an outright stumbling block to the text for tasks like Text-to-Speech processing and machine translation. All the graphemes with single diacritics exist as precomposed characters with unique code-points in Unicode version 6.0 but the four graphemes with both the dot-below and tone mark have to be created using combining characters.
A simple frequency distribution of characters and diacritical marks used in a fully tone-marked text from a word count of 129317 is as shown in Table 6.

Table 6: Vowels (Oral and Nasal) and Consonant in Yorùbá Text

Oral Vowel (OV)	OV Count	Nasal Vowel (NV)	N V Count	Consonant	Consonant Count
a	13546	an	2138	b	10324
à	22287	àn	2868	d	7046
á	15542	án	517	f	4951
e	5908	ẹn	45	g	8785
è	6845	ẹ̀n	29	gb	6234
é	8932	ẹ́n	5	h	2107
ẹ	4412	in	1077	j	7098
ẹ̀	7328	ìn	1130	k	12331
ẹ́	7499	ín	632	l	16212
i	16748	ọn	4193	m	8279
ì	17473	ọ̀n	311	n	16104
í	28725	ọ́n	2402	p	6542
o	11431	un	1956	r	15843
ò	7480	ùn	1007	s	6444
ó	9973	ún	2764	ṣ	8404
ọ	9692			t	18139
ọ̀	6924			w	13829
ọ́	4909			y	9979
u	1501				
ù	3687				
ú	7924				

Total No. of Words: 132550 Total No. of (Yorùbá) Syllables: 239840
Total No. of Yorùbá Words: 129317 Average No. of Syllables per Word: 1.8546

Out of grapheme count of 418491, vowel (oral and nasal) count was 239840 while consonants count was 178651. In Yorùbá, tones were used on each vowels and thus, tones and vowels each account for 36.43% of total phonetic features while consonants accounted for 27.14%. Therefore, a little more than one-third of orthographic information is lost when tone diacritics are absent from Yorùbá text. This information goes to show the crucial need for tone mark restoration in Yorùbá text and most importantly, restoration of tone diacritics. Therefore, we will be focusing only on the graphemes that bear tone marks (with the exception of the syllabic nasals because of unreliable data on them) and tone mark plus dot-below. We shall hence assume that the text was created with the dot below diacritic.

2 Existing Approaches to Diacritic restoration for Yorùbá

The two approaches that have been applied to restoration of diacritical marks in Yorùbá are also the commonly used approaches for almost all languages in which diacritic restoration (DR) has been performed. These are namely: word-level and letter-level DR respectively.

2.1 General Review of Approaches

2.1.1 Word-level Restoration

Diacritic restoration is often performed to distinguish one word from another when without the diacritics, the sequence of letters forming the word could have multiple meanings, could have a unique meaning and the real meaning being communicated is lost or when it would have no meaning at all [25]. Thus, the space delimited item, often used to approximate a word was initially the proposed unit for DR. However, word-based DR is often knowledge intensive and relies on existence of dictionaries, statistical language models and other language processing resources like Part of Speech (POS) tagger [22]. In addition, word-based models may not be suitable for all languages. Tufiş and Ceauşu [24]claim that the word-based model is often more appropriate for languages ”where the change of diacritics has a grammatical or semantic role”. Nevertheless, its major challenge is in handling Out of Vocabulary (OOV) or unknown words due to data sparsity [9]. The often adopted solution to handle this challenge is backing-off to letter-level restoration thus yielding a hybrid solution.

2.1.2 Letter-level Diacritic Restoration

According to Mihalcea [15] and De Pauw, Wagacha and de Schryver [6], the letter constitutes ’the smallest possible level of granularity in language analysis’ and hence should ’have the highest potential for generalization’ and also that ’the local graphemic context encodes sufficient information to solve the disambiguation problem’ of DR. Letter level features are extracted from training data from which statistical model is learnt via machine learning algorithm such as Decision Trees, Instance-based algorithm and Bayesian Network. Results on various languages have shown that the letter-level DR model has wide applicability especially for resource scarce languages. Nevertheless, Šantić, Šnajder and Bašić [22] claim that letter-level DR can only be expected to yield high accuracy only 'in languages where the diacritics can be restored without examining the context'.

2.2 Existing Works on Yorùbá Diacritic Restoration

To the best of our knowledge, the existing published works on DR in Yorùbá digital text are De Pauw, Wagacha and de Schryver [6] Scannell [23] and Adegbola and Odilinye [1]. The first two are basically letter-based approaches while the last one is a word-based approach. All of them were implemented with data-driven techniques.

2.2.1 De Pauw et al.'s resource-scarce model

De Pauw, Wagacha and de Schryver's model, [6] was proposed as a data-driven technique for restoration of diacritic to some "resource-scarce" languages using letters (graphemes) as the basic unit. The languages of interest were some African languages but as a control, some European languages were included. This was done with a view to contrast the performance of the letter-based approach across language groups. Yorùbá was one of the African languages that were studied. Tilburg Memory Based Learner, a memory-based learning (a form of k-nearest neighbor approach to machine learning) implementation was used with K-value of 3 to create the restoration model from the training set for each language. The training set was built from instances extracted with the feature vector created from a sliding window of five characters immediately to the left and the right of the focus letter and the focus letter itself. The class for each focus letter was therefore determined based on the context of five previous letters and five subsequent letters and the letter itself. In addition, a metric, Lexical diffusion (LexDiff), which measures the expected difficulty in restoring diacritic for the text of a given language was also developed by the authors [6] . The reported LexDiff for Yorùbá was 1.26. Compared with other tonal languages, the LexDiff was a fair estimate of the reported performance of 40.6% accuracy on out of vocabulary words (OOV) for Yorùbá. On plain text (not filtered to focus on OOV), the letter-based DR model combined with lexicon lookup had 68.5% accuracy while pure letter-based DR yielded 76.8% accuracy for Yorùbá. On the suitability of letters for restoration of tone diacritics, De Pauw, Wagacha and de Schryver [6] reached the following conclusion:

> "While the results for Cilubá and Yorùbá have improved significantly, the diacritic restoration problem is still far from solved for these languages. The trailing results compared to the other African languages, are caused by the tonal markings present in these languages. Tonal diacritics can simply not be solved on the level of the grapheme."

Results for Chinese which marks only tone diacritics and Vietnamese which marks both phonemic and tone diacritics seems to corroborate this declaration.

2.2.2 Scannell's unicodification model

Scannell [23] investigated a range of options including the letter and group of letters as the basic unit for diacritization using the NaÃ¯ve Bayesian classifier. Yorùbá was among the languages covered in the study. Several experimental configurations were considered. Lexicon lookup (which depends on the existence of one or more lists of words in a lexicon) was considered as the baseline. The lexicon was layered such that the first layer contained words with verified diacritic form, second layer contained words with alternate diacritic form while the third layer consisted of words from training data. When a particular diacritic-less form yielded multiple diacritic forms, frequency or a word bigram was used to select the most probable. The unicodification model configurations were tied to the feature vector applied to create statistical models for DR using Naïve Bayesian classification.

FS1 was the feature vector of three single letters on both sides of focus letter while FS2 considered five single letters on the left and right of the focus letter. These were the letter-based models. FS3 and FS4 were feature vectors for models based on groups of letters. A group was made up of three consecutive 'letters' at various positions relative to the focus letter. FS3 had seven such trigrams: first starting at fourth letters preceding the focus and the next at the third letter preceding the focus and subsequently until the last trigram that was started at the second letter succeeding the focus. In this way, the sequent trigram only dropped the starting letter of the previous one and appended the next letter within the string. FS4 had three trigrams: the first started at the third letter before the focus and the second trigram was centered on the focus letter while the last trigram started from the letter immediately after the focus letter. Scannell [23] combined lexical lookup with the best statistical model (for each language) to give a hybrid model and found that ' the 3-gram models performed consistently better than the 1-gram models" [23].

On the other hand, while FS4 was considered to be generally the best feature vector across languages, the result showed that FS3 was better for Yorùbá. It can thus be safely concluded that using multi-gram as lexical unit for diacritization in Yorùbá is better than using letters (unigrams). Furthermore, for Yorùbá, the better performance of FS3 compared to FS4 suggests that a wider window was better than smaller window.

2.2.3 Adegbola and Odilinye's word trigram model

This last model relied on exclusively on the word as the basic unit for DR with the NaÃ¯ve Bayesian classifier for training the model developed. The model was developed principally to evaluate the effect of corpus size on accuracy of automatic diacritization for Yorùbá. The model was built from linearly smoothed word trigram probabilities. The model achieved a best result of 70.5% accuracy with 100 000 words with an outside test setup. An inside test setup result of 95.9% accuracy was said to indicate a likely upper bound on the DR accuracy for Yorùbá language.

Further observations made were that the monosyllabic words represented the highest source of errors for inside test while disyllabic and trisyllabic words were the most prominent sources of error in the outside test setup [1].

2.3 Comparison of models

A comparison of the outside test result of De Pauw, Wagacha and de Schryver [6], Scannell [23] and Adegbola and Odilinye [1] demonstrated that the use of the word as the basic unit for diacritization without sufficiently large training data would not yield optimal results. This is despite the fact that it provided more context than letter-based units. This comparison was based on the fact that last two works used similar data samples. It is obvious that neither the pure letter-based approach nor the pure word-based approach is an effective solution to DR challenges in Yorùbá text. There is therefore a need to investigate the use of a different lexical unit for DR for Yorùbá text. We therefore propose the use of a novel basic unit for restoration of tone marks for Yorùbá text.

3 Proposed Model

The model proposed focused on the restoration of tone-marks in Yorùbá text using syllables as basic lexical unit. The reason for restricting the study to tone marks (a subset of the full diacritic marks) in Yorùbá is twofold. Firstly, preliminary investigation showed that many typists are familiar with methods for entering Yorùbá letters with dot-below diacritics. Secondly, while there are seven characters with dot-below in the Yorùbá orthography, there are twenty letters with tone mark diacritics (with overlap of four letters with both dots-below and tone marks). Moreover, characters with dots-below account for only 20% of the character count in Yorùbá (preliminary studies). This suggests that 80% of Yorùbá words do not contain dot-below and restoring tone marks alone may be sufficient to completely disambiguate them lexically.

3.1 Corpus description and Data

The corpus from which the training and testing data is drawn from is an adhoc plain text gathered from diverse sources and on many subject matters. The sources include web documents written in Yorùbá, either as publications by non-governmental organizations for educational purposes, Yorùbá translation of multi-lingual documents by international organizations like World Health Organization, Food and Agricultural Organization, excerpts from primary school materials for teaching Yorùbá. Other sources include archives of university graduate projects written in Yorùbá and text of social media discussion and communication written in Yorùbá. The subject matters or topics covered by the 'corpus' include education with educational level ranging from basic to tertiary levels, science (mostly basic and elementary), health, politics, social life (current affairs). Other topics are the customs of the Yorùbá people, religion (covering materials extracted from the Yorùbá Bible, Quran, traditional religion (Ifa) poems) and agriculture.

However, despite the diversity in the composition of the 'corpus', close to 70% of its materials are educational in nature or origin followed by religion which accounts for more than 10% of its total size (word count) while the remaining percentage was distributed among the remaining subject matters. The ‘corpus’ comprises documents written in formal and informal styles although some normalization was carried out to ensure that conformity to Standard Yorùbá orthography. The total word count of the ‘corpus’ was about one hundred and thirty thousand (130,000). We propose the division of the corpus into 70% for training data and 30% for testing. The division will be along sentence boundaries and as such, words from different genres may aggregate unevenly between training and testing data. Further information on the corpus is shown in section 1.3.

3.2 Conceptual Description of Model

The proposed diacritic restoration technique has two stages and is described as follows:

3.2.1 Model Creation

i Training data created from Yorùbá sentences that are correctly tone-marked, designated as Y, are syllabicated into strings of syllables S_i using a syllabication tool;

ii An off-line data-driven, statistical diacritic restoration model, M, is created from the training data composed of S_i in (i) above using supervised machine learning;

3.2.2 Model Utilization

i Fresh Yorùbá text, T, without tone marks is syllabicated as in section 3.2.1 (i) to generate S_k and the off-line model M labels S_k with tags L_k, indicating the tone for each syllable. S_k and L_k are combined deterministically to yield string of syllables $\hat{S}_k$, with tone marks;

ii $\hat{S}_k$, is combined using ”*syllable aggregator*” back to get required the output text, modified with tone marks, T̂.

The process flow for the diacritic restoration with approach described above is shown in Figure 1. The diacritic restoration proper takes place in the Stage 2 of Figure 1. Stage 1 is pre-processing the input into strings of syllables while stage 3 is post-processing the tone-marked syllable strings back into words.

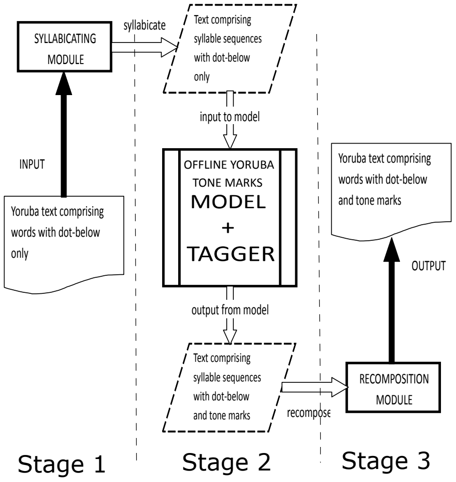
SYLLABICATING MODULE
syllabicate
Text comprising syllable sequences with dot-below only
input to model
INPUT
OFFLINE YORUBA TONE MARKS MODEL + TAGGER
Yoruba text comprising words with dot-below and tone marks
Yoruba text comprising words with dot-below only
OUTPUT
output from model
Text comprising syllable sequences with dot-below and tone marks
recompose
RECOMPOSITION MODULE
Stage 1
Stage 2
Stage 3

Figure 1: Process Flow for syllable-based diacritic restoration system for Yorùbá text

3.3 Model Assumptions

The proposed techniques made the following assumptions about the text in both stages:

1. The training data used in creating the off-line restoration model is correctly marked with dot-below and all tone marks;

2. The input text to be labeled is also correctly marked with the dot-below on appropriate letters;

3. That the text, which comprises only Yorùbá words, was created in the Standard Yorùbá dialect also known as literary Yorùbá;

4. That the both the text used as training data and as input to be labeled were created using the standard Yorùbá orthography [16] (or a very close variant) as stipulated in that report.

5. That there exists (or there will be created) a tool for syllabicating Yorùbá words.

Apart from the syllabicating tool which might be language dependent, other processes involved in the proposed technique are language independent. A rule-based syllabicating software tool, SY-syllabicator, with a reported performance of 99.99% accuracy already exists for Yorùbá [11]. The off-line, data-drive statistical DR model can be created using any appropriate supervised machine learning algorithm. The algorithm could be any of these classifiers: Memory-based learners, Naïve

Bayesian classifier, Support vector machines, Hidden Markov Model and Conditional Random Field.

3.4 Proposed model evaluation

We will be carrying out a ten-fold evaluation of the model. This will be done by developing a software prototype system in which the created model will be embedded. The proposed evaluation parameters are: syllable and word error rates. Syllable error rate is the total number of syllables that were wrongly tone-marked out of all syllables that were tone marked at testing time. Syllables will be combined into words and a word is in error if at least one of its syllables is wrongly tone-marked. Word error rate is the number of words in error relative to the total number of words restored by the system. The word error rate can be used as a basis of comparison with existing models.

4 Model Justification

Given the current state of DR in Yorùbá, a new model that can improve diacritization accuracy is necessary. Tufiş and Ceauşu [24] reports that in languages where the change of diacritics has a grammatical or semantic role, word-based DR systems are much more reliable. This claim is based on the need to include more context than can be provided by letter-based DR systems that work for languages where the diacritics can be restored without examining the context. The proposed approach based on using syllables as basic units for tone restoration is justified on the following grounds:

4.1 Failure of basic restoration units in existing works

The review in section 2 has shown beyond reasonable doubt the failure of DR for Yorùbá using the mainstream basic lexical units of letters and words. Thus, empirical evidence indicates the need to look beyond letter-based and word-based models for DR in Yorùbá text. In fact, De Pauw et al. (2007) had claimed that letter-level DR is inappropriate for tonal languages and Yorùbá is a language with a three-way contrasting tone system. De Pauw, Wagacha and de Schryver [6] seems to be further corroborated by Scannell's [23] results for tone languages. The failure of letter-based DR models for Yorùbá seems to be further aggravated by the language's 'resource scarcity'. As for word-based approaches to DR for Yorùbá, satisfactory performance can only be achieved with large corpora. As Adegbola and Odilinye [1] show, the n-gram word-based DR for Yorùbá would not be able to yield up to 95.9% accuracy even with a corpus of up to three million words. This requirement for large corpora is a significant challenge for word-based approaches to Yorùbá DR.

Scannell's results [23] indicate that sub-word units are better than letters as basic unit for DR. For about five languages that are covered in both [23] and [6], on the average, sub-word modelling out-performed letter-level modelling by 35.56%.In addition, for a resource scarce language like Yoruba, DR activities may have to make do with small corpora. With training data of approximately 5000 words, Scannell's [23] sub-word-based DR model better than Adegbola and Odilinye's [1] word-based DR model of 10,000 words. The better performance of sub-word modeling than letter-based modeling in general, and the better performance than word-based modeling with smaller training data, point to the likely superiority of sub-word models to word-based and letter-based models.

4.2 Linguistic Justification for Syllable as Basic Restoration Unit

The current sub-word model did not take advantage of the relationship between tone and syllable to optimize the performance of the model. According to Yang [28] "the domain of the tone is over the entire voiced portion of the syllable." and that "it is preferable to formalize the tone feature $\cdots$ regard them as features of individual syllables. Since phonemically, "tones are associated with the individual syllables in an utterance" [18], the orthographic corollary is that tone marks (written symbols for indicating tones) should be associated in text with syllables. This corollary is true for languages where the domain of tone is the syllable. There are a few exceptions like Sherpa where the domain of tone is the word [17, 10]. A Recognition of the relationship of tone to its prosodic domain should be an important consideration for tone mark restoration. Scannell's sub-word-based DR model [23] could be considered a language-independent approximation to syllable-based DR that gained independence at the expense of performance.

As shown in Table 6 above, the average syllable length (measured as the number of characters in a syllable) calculated from a collection of approximately 240000 syllables is 1.86 characters, or approximately, two characters per syllable. Accounting for tone marks, there will, on the average, be three symbols (two characters and one tone mark (with the null symbol to represented mid tone)). Tone mark restoration can then be postulated to be the recovery of the third part of a syllable, the remaining two parts being the characters already present. It is thus easier to recover the one-third of total information per token than an uncertain part. This is one advantage of the syllable based approach to tone marks restoration in Yorùbá. the syllable-based model. Furthermore, only one symbol needed to be appended per syllable.

4.3 Contextual Sufficiency

Another advantage of the syllable-based approach is that it enables more context to be included in the training data for the model's machine learning. Compared to the letter-based, a syllable-based approach should double the context used for each length of the token sequences. While this might not double the accuracy, it should significantly improve it. There are also a fixed number of syllables in the Yorùbá text whether taken with or without tone marks. Thus, unlike a word-based approach, where the number of tokens to be dealt with is infinite, a syllable-based approach deals with a finite number of distinct lexical tokens or items. Words which may not have appeared in the training data may contain substrings which did appear; the use of sub-word units like the syllable thus helps to reduce the incidence of unknown (out of vocabulary) words. The following examples illustrate the case we are making here: suppose a sentence in the training data contains the focus word: *Ajakaja* (M-H-H-H) but the training includes no sentence containing *Aja* (M-H). However, *Aja* (M-H) is found in a text to be restored and within a similar context, as shown below.

- TRAINING DATA: Ajakaja(M-H-H-H) ti a ba gba mu ni adugbo di pipa.
- TEXT TO BE RESTORED: Aja ti mo ba mu ni adugbo di temi.
- Ajakaja(M-H-H-H) translates as ANY DOG while Aja is simply DOG.

Although the word '*Aja*' never appeared in training data, it does appear as a sub-string, and will not be treated as OOV item. In other words, the syllable-based approach to DR, like the letter-based approach, is able to overcome the limitation of OOV in word-based approaches. However, unlike the letter-based approach in which encoding sufficient context is a challenge, syllable-based approaches for tone mark restoration can at least double the amount of context encoded for disambiguation. The model should ameliorate the contextual insufficiency that bedeviled the application of letter-based approach where, according to Luu and Yamamoto [13], "diacritics signal grammatical or semantic roles", as in case of tone marks.

4.4 Tolerance for Data Scarcity

This characteristic of a syllable-based approach reduces the challenge that a word-based approach has with data scarcity that characterizes most African languages are subjected to. This is because for the same word counts in a corpus, the number of syllable tokens are in multiple folds.

5 Conclusion

In this proposal for a new approach to tone marks restoration in Yorùbá text. A general background to problem was given by highlighting its importance and the size of problem in Yorùbá, the case study language. This was followed by a review of the existing approaches. We then presented the proposed data-driven, syllable-based approach for tone marks restoration in Yorùbá text. The key stages in the development of the proposed model are the off-line training of the tone marks model from data using supervised learning. The second stage is the tone mark restoration system where the process starts with text to be tone marked are passed to the syllabification module. The flow diagram of the proposed system was also given. In adopting an approach to diacritic restoration, several issues have to be considered: the role of diacritics in the language, availability of adequate training data, required processing speed, and users' requests and needs. The syllable-based approach has been proposed here because one would expect that it would be easier, when a portion of a linguistic token is lost, to recover $\frac{1}{3}$ of an object than $\frac{1}{2}$ of the same object. The proposed approach will be implemented and the performance of the model will be evaluated based on accuracy expected when compared to texts which to which tone marks are manually restored.

References

[1] Adegbola, T. and Odilinye, L. U. Quantifying the effect of corpus size on the qual-

ity of automatic diacritization of Yorùbá texts. In *Proceedings of 3rd international Workshop on Spoken Languages Technologies for Under-resourced Languages*, Cape Town, South Africa, 2012. online, Retrieved August 12, 2012 from http://www.mica.edu.vn/sltu2012/files/proceedings/10.pdf.

[2] Adeniyi, H. R. A comparative study of reduplication in Edo and Yorùbá. *MorphOn: e-journal of morphology*, pages 1–23, 2007. 2 April 2007.

[3] Akinlabi, A. Yorùbá sound system. *Understanding Yoruba Life and Culture*, pages 453–468, 2004.

[4] Clements, G. and Rialland, A. Africa as a phonological area. *A Linguistic Geography of Africa*, pages 36–85, 2008.

[5] Coulmas, F. *Writing systems: An introduction to their linguistic analysis.* Cambridge University Press, 2003.

[6] De Pauw, G., Wagacha, P. W., and de Schryver, G. Automatic diacritic restoration for resource–scarce languages. In Matousek V., M. P. ., editor, *Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007, Proceedings Lecture Notes in Artificial Intelligence LNAI, subseries of Lecture Notes in Computer Science LNCS*, volume 4629, page 170–179, Berlin, 2007. Springer–Verlag.

[7] Fabunmi, F. A. and Salawu, A. S. Is Yorùbá an endangered language? *Nordic Journal of African Studies*, 14(3):391–408, 2005.

[8] Gaultney, J. V. Problems of diacritic design for latin script text faces, 2008.

[9] Haertel, R. A., McClanahan, P., and Ringger, E. R. Automatic diacritization for low–resource languages using a hybrid word and consonant CMM. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, June 2010*, page 519–527, Los Angeles, California, 2010.

[10] Hildebrandt, K. A. Phonology and fieldwork in Nepal: Problems and potentials. In *Proceedings of the conference on language documentation and linguistic theory*, pages 33–44, 2007.

[11] Kumolalo, F. O., Adagunodo, E. R., and Odejobi, O. A. Development of a syllabicator for yorùbá language. In *Proceedings of OAU TekConf, September 5-8, 2010*, pages 47–51, OAU, Ile-Ife, Nigeria, 2010.

[12] Lojenga, C. K. Orthography and tone: Tone system typology and its implication for orthography development. In *Linguistic Society of America Annual Meeting*, pages 1–12, Pittsburg, US, January 2011.

[13] Luu, T. A. and Yamamoto, K. A pointwise approach for Vietnamese diacritic restoration. 2012.

[14] Maddieson, I. *Tone.* Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. Retrieved February 21, 2011 from http://wals.info/feature/13.

[15] Mihalcea, R. Diacritic restoration: Learning from letters versus learning from words. In *Proceedings of Computational Linguistics and Intelligent Text Processing, 3rd International Conference, CICLing 2002, Mexico City*, Volume 2276, pages 339–438. Springer, 2002.

[16] Nigeria. Joint Consultative Committee on Education. *1974 Revised Official Orthography for the Yoruba Language.* The Committee, 1974.

[17] Noonan, M. Recent adaptions of the devanagari script for the Tibetoburman languages of Nepal. *Indic Scripts: Past and Future*, 2005.

[18] Ọdẹ́jọbí , O. A. Recognition of tones in Yorùbá speech: Experiments with Artificial Neural Networks. In Prasad, B. and Prasanna, S., editors, *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks (Studies in Computational Intelligence)*, volume 83. Springer Science & Business Media, Berlin Heidelberg, 2008.

[19] Oyebade, F. Yoruba phonology. In Yusuf, O., editor, *Basic Linguistics for Nigerian Languages Teachers (A Publication of the Linguistic Association of Nigeria)*, pages 221 –240. M & J Grand Orbit Communications.

[20] Oyetade, S. O. A sociolinguistic analysis of address forms in Yoruba. *Language in society*, 24(04):515–535, 1995.

[21] Pulleyblank, D. G. *Tone in lexical phonology.* D. Reidel Publishing Company,, Dordrecht, 1986.

[22] Šantić, N., Šnajder, J., and Bašić, B. D. Automatic diacritics restoration in Croatian texts. In *INFuture2009: Digital Resources and Knowledge Sharing*, pages 309–318, 2009.

[23] Scannell, K. P. Statistical unicodification of African languages. *Language Resources and Evaluation*, pages 1–12, 2011. Retrieved July 20, 2011 from `http://borel.slu.edu/pub/lre.pdf`.

[24] Tufiş, D. and Ceauşu, A. Diac: A professional diacritics recovering system. In *Proceedings of the Sixth International Language Resources and Evaluation*, 2008. paper 54 on Conference CD.

[25] Tufiş, D. and Chiţu, A. Automatic diacritic insertion in Romanian texts. In *Proceedings of the International Conference on Computational Lexicography COMPLEX'99. Pecs, Hungary*, pages 185–194, 1999.

[26] UCLA. UCLA Language Materials Project: Yoruba, n.d. Retrieved October 14, 2010,from `http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=22`.

[27] Wells, J. C. Orthographic diacritics and multilingual computing. *Language problems & language planning*, 24(3):249–272, 2000. Retrieved July 12, 2010 from `http://www.phon.ucl.ac.uk/home/wells/dia/diacritics-revised.htm`.

[28] Yang, W. S.-Y. Phonological features of tones. *International Journal of American Linguistics*, 33(2):93–105, 1967. `http://www.jstor.org/stable/1263953` Accessed Sept. 07, 2011.

[29] Yip, M. *Tone.* Cambridge University Press, 2002.