http://dx.doi.org/10.7494/csci.2017.18.3.2128

Franklin Ọládiípọ̀ Asahiah
Ọdẹ́túnjí Àjàdí Ọdẹ́jọbí
Emmanuel Rotimi Adagunodo

RESTORING TONE-MARKS IN STANDARD YORÙBÁ ELECTRONIC TEXT: IMPROVED MODEL

Abstract *Diacritic Restoration is a necessity in the processing of languages with Latin-based scripts that utilizes characters outside the basic Latin alphabet used by the English language. Yorùbá is one such language, marking an underdot (dot-below) on three characters and tone marks on all seven vowels and two syllabic nasals. The problem of restoring underdotted characters has been fairly addressed using characters as linguistic units for restoration. However, the existing character-based approaches and word-based approach has not been able to sufficiently address the restoration of tone marks in Yorùbá. In this study, we address tone-mark restoration as a subset of diacritic restoration. We proposed using syllables derived from words as linguistic tokens for tone-mark restoration. In our experimental setup, we used Yorùbá text collected from various sources as data with a total word count of 250,336 words. These words, on syllabification, yielded 464,274 syllables. The syllables were divided into training and testing data in different proportions, ranging from 99% used for training and 1% used for testing to 70% used for training and 30% used for testing. The aim of evaluating the different proportions was to determine how the ratio of training-to-test data affected the variations that may occur in the result. We applied memory-based learning to train the models. We also set up a similar experiment using a character token to be able to compare the performance. The result showed that ,by using syllables, we were able to increase the word-level accuracy to 96.23% (an average of almost 15% over using characters). We also found that using 75% of the data for training and the remaining 25% for testing gives results with the least variation in a ten-fold cross validation test. Hybridizing this method that uses syllabless as processing linguistic units with other methods like lexicon lookup might likely lead to improvement over the current result.*

Keywords diacritic restoration, syllables, characters, word-level accuracy

Citation Computer Science 18(3) 2017: 301–315

1. Introduction

The Yorùbá language is spoken by more than 30 million people in places like the United States, United Kingdom, Benin Republic, and (principally) in southwestern Nigeria. The current orthography in use is based upon a report from the 1974 Joint Consultative Committee on Education, and it retained the use of subdots to distinguish some characters and diacritics to mark tones. In this orthography, tones are marked on vowels and syllabic nasals using acute accent for high tones, grave accent for low tones, and the absence of accent for middle tones (except on syllabic nasals, where it is marked with a macron). The character set for the standardized orthography of Standard Yorùbá comprises of 25 characters of the Yorùbá alphabet called Aábídí and three tone marks as shown in Table 1.

Table 1

Yorùbá alphabet and tone marks

Characters	A B D E Ẹ F G GB I H J K L M N O Ọ P R S Ṣ T U W Y
Low Tone:	grave accent mark on vowels and syllabic nasals: (\`) à è ẹ̀ ì ò ọ̀ ù ǹ m̀
Mid Tone:	indicated only on syllabic nasals as macron (¯) n̄m̄
High Tone:	acute accent mark on vowels and syllabic nasals: (′) á é ẹ́ í ó ọ́ ú ń ḿ

2. Significance of tone marks in Yorùbá text

Tones mark a lexical contrast within words with the same sequence or the order of consonant and vowels in those words. Tones are marked on vowels, and the tone-bearing unit is the syllable. In Yorùbá, a syllable does not exist without an attendant tone, and the functional load of a tone can be illustrated with the following examples:

Word	Tone pattern	Gloss
igba	mid, mid	two hundred
igbà	mid, low	climber’s belt
igbá	mid, high	calabash bowl
ìgbà	low, low	period of time
ìgbá	low, high	garden egg (a vegetable)

All of the five different Yorùbá words above were contrasted by the different tone sequence attached to the vowels. The only restriction on the distribution of tones within Yorùbá words is that a high tone shall not be the first tone in those words whose first syllable is a vowel. When the first syllable is not a vowel, this restriction does not apply. We collected a body of Yorùbá texts from several sources, such as internet archives, social media pages, webpages, OCRed pages of hardcopy Yorùbá textbooks, and storybooks. These were manually corrected. The word count was 129,317, and the number of syllables generated with an automatic syllabification program was 239,840. Frequency analysis of these syllables showed that 89,824 (37.35%) of the syllables carried the acute accent for a high tone, while 77,369 (32.26%) carried the grave accent for a low tone and 72,647 (30.29%) were mid tone with no mark or with a macron on a sylable. The frequency distribution indicated an almost even distribution of the tones, as the difference may not be considered statistically significant at a 90% confidence value. However, the dictate of standard orthography Yorùbá seems to be really different from the current writing practice. This hypothesis is born out of information from different authors about the state of documents written in the Yorùbá language. Fagbọrun [7] showed that tone-mark faithfulness among students enrolled in Yorùbá studies in tertiary institutions is low. We summarized the data taken from [7] in Table 2. The tertiary students formed the sample population, and the data size (N) was 28.

Table 2

Degree of Tone Marking (data from [7])

Parameter	Value [%]
Mean	48.888
SD	13.646
N	26
95% CI	43.377 to 54.400
Minimum	13.2
Maximum	72.4

Table 2 showed a poor average performance. One was marked as low as 13.2% of the text, while the highest tone-marking was 72.4% and mean was 48.89%. These values did not indicate whether the tone marks were correctly used but only showed efforts at the indicating tones. According to [15], tone marking was taken as optional in bibles and other religious publications, even when full use would have been more appropriate. Olumuyiwa [15] further stated that most Yorùbá print newspapers and magazines rarely mark the tones. Odejobi [16] also found out that, apart from textbooks published as materials for use in educational institutions, most of the printed materials in Yorùbá lack tone marks, while Asahiah [3] found that, except for a few digital archives and newer versions of the Yorùbá Bible in digital publications, most digital Yorùbá publications are unmarked or minimally marked for tones. Yorùbá text documents can be further complicated when ẹ, ọ and ṣ are substituted with characters not in the orthography and also when obsolete tone marks and accents like tildes, circumflexes, and breves are used.

3. Categorization of diacritic restorations

The challenge of writing without tone marks was first pointed out by Bishop Samuel Ajayi Crowther, one of the major players in the development of Yorùbá orthography. Crowther was quoted in [2] as saying that "the absence of tone" will lead to confusion with the meaning of words. Tone marks are a subset of diacritical marks applied in the text of several languages. For the most part, restoring diacritics have relied either on the character (grapheme) or the space-delineated linguistic block (word) as the lexical focus item. Tone marks are a subset of diacritical marks, and the two broad categories of approaches to diacritic restoration are the rule-based and data-driven approaches.

The rule-based approach relies on the manual crafting of linguistic, grammatical, or some other relevant rules to assign diacritical marks to a token. According to [4], "these rule-based systems achieve good performance while learning a small list of simple rules." However, when the number of rules increased beyond a point, it becomes difficult to maintain the rule-base. According to [21], linguistic rules to accomplish diacritic restoration may not even be easily discernible. Data-driven approaches shift the cost of developing rules to obtaining a useful and sufficiently large volume of data from which machine-learning algorithms can abstract parametric or non-parametric relationships. Data-driven approaches have become more common, while rule-based approaches are now incorporated either for pre-processing or post-processing. For the most part, data-driven approaches to restoring diacritics have relied on either the character or the word as the lexical focus item.

3.1. Word-level restoration

Diacritic restoration is often performed to distinguish one word from another. Without the diacritics, the sequence of characters forming the word could have multiple meanings, have a unique meaning and the real meaning being communicated is lost, or it has no meaning at all [24]. Thus, the space delimited item, often used to approximate a word, was initially the proposed unit for diacritic restoration until [13] proposed the "letter-based" approach. According to [17], word-based diacritic restorations are often knowledge intensive, relying on the existence of linguistic resources like dictionaries, part-of-speech taggers, and morphological analyzers [24]. According to [23], 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 [8, 9]. The oft-adopted solution in handling this challenge is backing-off to character-level restoration (thus, yielding a hybrid solution).

3.2. Letter-level restoration

According to [13], a letter (character) constitutes "the smallest possible level of granularity in language analysis" and, hence, should "have the highest potential for generalization." The authors in [6] stated that character-level diacritic restoration is premised on the hypothesis that "the local graphemic context encodes sufficient information to solve the disambiguation problem" of diacritic restoration. They are much simpler, faster, easier to implement, and do not require language-specific resources [17]. Character-level features are extracted from the training data from which models are learned via machine-learning algorithms, such as Decision Trees, Instance-based algorithms, and Bayesian classifiers [6, 13, 18]. Studies on various languages have shown the wide applicability of the character-level model (especially for resource-scarce languages). In fact, [17] suggested that character-level restoration can be expected to yield high accuracy only in languages where context is not a big factor.

4. Existing works on Yorùbá diacritic restoration

To the best of our knowledge, the existing published works on diacritic restoration for Yorùbá digital text are [1, 6, 18].

4.1. Diacritic restoration for resource-scarce languages

Authors of [6] proposed a data-driven technique for the restoration of diacritics to some languages using characters (graphemes) as the basic unit for several languages (one of which was Yorùbá). A memory-based learning technique was used with a K-value of 3 to create a restoration model from the training set. The class for each focus character was determined based on the context of the five previous and five subsequent characters as well as the character itself. For Yorùbá, [6] reported a performance of 40.6% word-level accuracy on out-of-vocabulary words (those not in the training set). On plain text (where words in a test set may also be in the training set), character-based restoration yielded a 76.8% word-level accuracy, while combining this approach with a lexicon lookup reduced the accuracy to 68.5% However, [6] gave the following comments on the use of characters as basic units for the restoration of tone diacritics for Yorùbá and other tone 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.”

4.2. Statistical unicodification

Scannell (2011) applied Bayesian classification with uniform smoothing to n-gram models for diacritic restoration to several languages (including Yorùbá). In the study, [18] evaluated the following algorithms: lexicon lookup (unigram) and word-bigram lookup LL2 (for basic form words [BF] with multiple diacritical forms [DF]). Others are different configurations of n-gram character features, FS1 to FS4, and a combination (CMB) of LL2 and the best of the FS series. The LL2 and LL accuracy for Yorùbá were both at 75%. The best of the FS series for Yorùbá (FS3 a seven-trigram sequence starting at offset position 4 before the target character and ending at offset position 2 afterwards) has an accuracy of 61.9%, and CMB attained 75.3%. Although [6] and [18] used different datasets, their results on plain text were comparable (76.8% and 75.3%, respectively).

4.3. Quality of diacritization of Yorùbá text

Adegbola and Odilinye [1] also used the Bayesian classifier, but with the word as the lexical token for diacritic restoration in Yorùbá. The model, developed mainly to evaluate the effect of corpus size on the accuracy of automatic diacritization for Yorùbá, was trained from different percentages of a 100,000-word corpus. The Naïve Bayesian classifier built the model from linearly smoothed word trigram probabilities. The model achieved a best result of 70.5% diacritic restoration 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 boundary on diacritic restoration accuracy for the Yorùbá language.

Comparing the outside test against both [6] and [18] showed that the use of words as a basic unit for diacritization is not optimal for resource-scarce languages, even though it provided more context than character-based units. We therefore propose the use of syllables as the basic unit for the restoration of tone marks for the Yorùbá text.

5. Syllable-based tone diacritic restoration

A syllable is defined in the Concise Oxford English Dictionary (eleventh edition) as a ”unit of pronunciation having one vowel sound with or without surrounding consonants, and forming all or part of a word.” Nearly all languages that use the Latin script mark their word boundaries with white space. Exceptions to this rule are Chinese Pinyin and Vietnamese, which both mark a syllable's boundary in written text with white space. Nevertheless, syllables have been applied in building language models for speech processing [5, 11, 12], document retrieval [10, 20], and other language-processing tasks. However, syllable-based models to diacritic restoration have only been applied to Vietnamese [14, 22], and this is most likely due to the fact that it is the syllables and not the words that are space-delimited orthographic units in Vietnamese. The results from these studies encourage the exploration of the syllable as a lexical unit for diacritic restoration for the Yorùbá language. This consideration has a linguistic justification: both languages uses diacritics for the same purposes; namely, to create phonemic variation and mark tones. While Vietnamese has six tones and uses diacritics to indicate five of them (the last tone is unmarked), Yorùbá has three lexical tones and uses diacritics to mark two (leaving the third unmarked). Seven characters in Vietnamese are modifications of basic Latin characters by adding diacritics, while only three characters in Yorùbá alphabet are modified by the addition of diacritics.

In this study, our interest is on the restoration of tone marks within the Yorùbá electronic text. This is a subset of the full set of diacritics. We limit ourselves to this subset because the degree of lexical ambiguity caused by the absence of tone marks is extremely high, and its influence is more widespread within the text. From a simple frequency count of a Yorùbá raw text of characters, it was found that orthographic consonants have a frequency of 178,651, which is comparable to the tone marks (low and high tone marks alone) with a frequency of 167,193. If the orthographic syllabics were efficiently counted, it is likely that tone marks may have a higher frequency than consonants. This is an indication of the functional load carried by tone marks. This syllable-based approach is patterned after the character-based approach as described in [6] and [13] except that syllables replaced characters in the setup. This is a semi-independent approach, since the process of generating syllables varies according to the morphological structure. A snapshot of a typical representation of the feature vectors is shown in Table 3.

Table 3
Syllable-based training features of a Yorùbá clause (tone-mark restoration)

FSy	-4Sy	-3Sy	-2Sy	-1Sy	+1Sy	+2Sy	+3Sy	+4Sy	flag	Class
a	<	<	<	<	SP	di	SP	ga	+ve	M
SP	<	<	<	a	di	SP	ga	a	-ve	N
di	<	<	a	SP	SP	ga	a	ri	+ve	M
SP	<	a	SP	di	ga	a	ri	SP	-ve	N
ga	a	SP	di	SP	a	ri	SP	si	+ve	M
a	SP	di	SP	ga	ri	SP	si	lẹ	+ve	L
ri	di	SP	ga	a	SP	si	lẹ	SP	+ve	H
SP	SP	ga	a	ri	si	lẹ	SP	e	-ve	N
si	ga	a	ri	SP	lẹ	SP	e	wu	+ve	H
lẹ	a	ri	SP	si	SP	e	wu	rẹ	+ve	L

In Table 3, the FSy column is the focus syllable, and the tone for the syllable is designated by the $Class$ column. For $1 \leq i \leq 4$, $-iSy$ is the i^{th} syllable to the left of FSy, and $+iSy$ is the i^{th} syllable to the right of FSy. Since the text for training and testing were divided into sentences, the first syllable in a sentence will four of special symbol "<" to be found in syllable positions within the feature vector is used to designate a sentence boundary symbol before the start of a word, and the first syllable in each sentence has four such symbols. The second syllable has three such sentence boundary symbols, and so on. "SP" is also a special symbol that denote the whitespace word delimiter in sentences. In the class labels, "L" stands for low tone, "M" for mid tone, "H" for high tone, and "N" for tokens *not* to be tone-marked (toneless). The training segment whose feature vectors are shown in Figure 3 is a Yorùbá clause: `a di gàárí sílẹ̀`.

6. Problem definition

In this study, tone-mark restoration for Yorùbá text is formulated as a classification problem similar to Part-of-Speech tagging. A text is comprised of a set of sentences. A particular sentence is symbolized as z, which in turn is comprised of words z_j separated by whitespace. Given sentence z, sequence of words, z_j, written without tone marks:

$$z = \{z_1, z_2, \cdots, z_{m-1}, z_m\} : 1 <= j <= m; z_j \in Z$$

Z the is set of all Yorùbá words where tone marks are absent, z is transformed by syllabification to sequence of syllables without tone marks x:

$$x = \{x_1, x_2, x_3, \cdots, x_{n-2}, x_{n-1}, x_n\} : 1 <= i <= n$$

where $x_i \in X$ and X is the set of all Yorùbá syllables in the absence of tone marks and $n(X)$ is 230.

In Yorùbá, each syllable has only one tone: either low, middle, or high. Let t be defined as the set of class tags to which x is assigned:

$$t = \{t_1, t_2, t_3, \cdots, t_{n-2}, tn - 1, t_n\} : 1 <= i <= n$$

where $t_i \in T = \{L(lowtone), M(midtone), H(hightone), N(toneless)\}$.

The fourth class is for foreign words and, therefore, not syllabicated. Thus, our problem is to find function (f) that maps a tone-mark (t) to a syllable (x); given the context of the syllable:

$$f : x \longrightarrow t$$

where f is determined via machine learning, $t|x = \underset{t\in T}{argmaxP}(t|x)$, $s_i \in S$, the set of all Yorùbá syllables written with a tone mark, $n(S)$ is 690.

$$s = \{s_1, s_2, s_3, \cdots, s_{n-2}, s_{n-1}, s_n\}$$

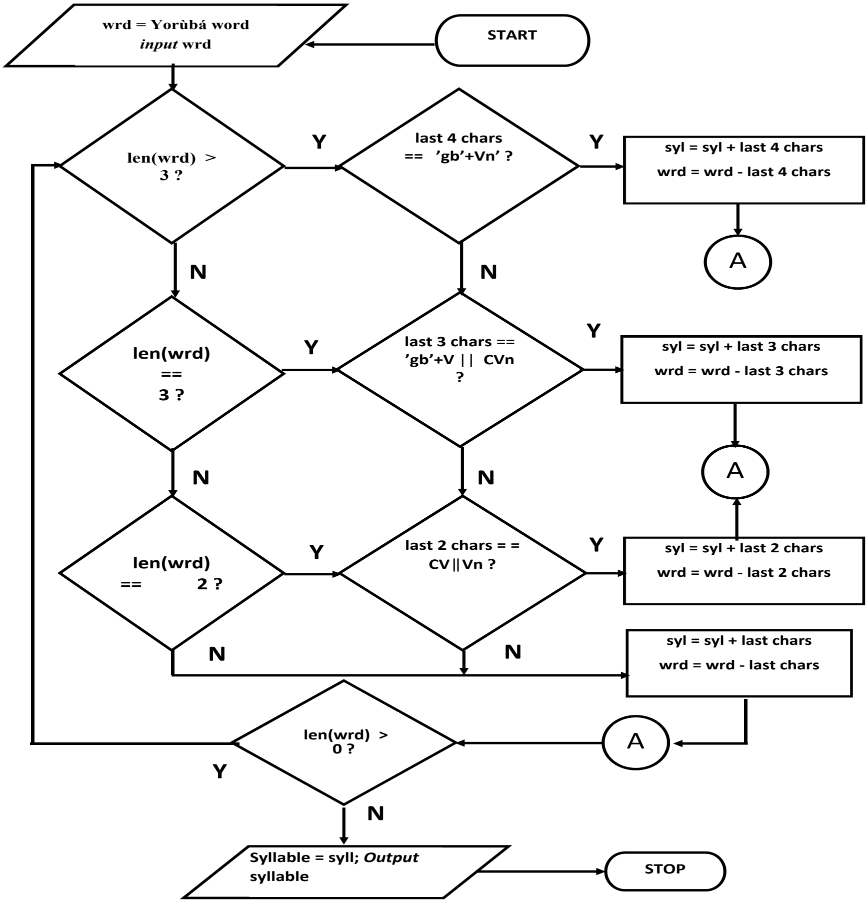

Figure 1. Flowchart for rule-based Yorùbá syllabification algorithm.

Deterministic function g: $x_i \times t_i \longrightarrow s_i$. By aggregation, s is transformed to w: $w = w_1, w_2, \cdots, w_{m-1}, w_m$; $w_j : 1 <= j <= m$; $\in W$, the set of all Yorùbá words written with tone marks.

The transformation of z to x is a form of tokenization and is accomplished in this study with the aid of a language-dependent Yorùbá syllabificator whose algorithm is presented as a flowchart in Figure 1. The main model is $f{:}x \longrightarrow t$, and we applied the memory-based learner used in [6]. This will allow us to compare the model with that of [6] (although the data set used is different). To ameliorate this, we also created a character-based model in a way similar to [6] to compare it with the syllable-based model.

7. Data set

Yorùbá language includes the following twenty three (23) diacritically marked characters: *ẹ, ọ, ṣ, à, è, ẹ̀, ì, ò, ọ̀, ù, ǹ, m̀, á, é, ẹ́, í, ó, ọ́, ú, ń, ḿ, n̄, m̄*. However, since we are concerned with tone-marked characters, we will be focusing on the twenty that bears tone marks. In this study, we have 250,336 words in our corpus. Syllabification of the words yielded 464,274 syllables for an average of approximately 1.8546 syllables per word. The total number of types (unique word forms) was approximately 14,000 compared to the 4200 used in [6]. However, to make the restoration realistic in view of the fact that it is not unusual (or rather it is usual) to find English words in Yorùbá documents (especially as names), we try to normalize English words that have known Yorùbá transliteration or leave the English word as is in the document.

The corpus is comprised of collections of digital newspaper articles as well as digital reports from several non-governmental and advocacy organizations (NGOs). Many of these NGO's digital publications are translations and advertorials from various subject areas. It also consist of a large number of academic project reports and some publications from Christian, Islamic, and Yorùbá traditional religious groups. The corpus also comprises proverbs and old school books that were scanned, OCR-ed, and manually corrected. The corpus, however, has only a few articles (mostly medical) with a science orientation.

8. Experimental setup

In this study, we assumed that the data was already comprised of the dot-below diacritic used to indicate the voiceless postalveolar fricative s (ṣ), open-mid front rounded oral vowel ɛ(ẹ) and open-mid back rounded oral ɔ(ọ).

The corpus was first divided into training and test sets. Our experimental setup was such that we evaluated the system with a different training set to test set ratios: We have the following: 0.99:0.01; 0.9:0.1; 0.85:0.15; 0.80:20; 0.75:0.25; and 0.70:0.30. The word tokens were then transformed to sets of syllable tokens through syllabification. The syllable instances and their feature vectors were then extracted from the now-transformed training set and test set, respectively. The sample is as shown in Fig. 3. For each training-test ratio configuration, a 10-fold cross validation was carried out. The training corpus was divided into 100 portions. To ensure that the 10-fold cross validation did not repeat a set used in a previous iteration, we implemented an algorithm that treated the portions as a circular-linked list. The fixed starting points for the 10-fold cross validation were as follows: {5, 90, 10, 80, 20, 70, 30, 60, 40, 50}.

As such, when the starting point plus the size of the test set selected to be used for a particular iteration of the 10-fold cross validation exceeds 100, Portion 101 was 1, and 120 was Portion 20. For example, with a test set size of 30% and training set size of 70%, the second iteration (which starts at 91 and goes virtually to 120) will actually select 91–100 and then 1–20, while the training set will be 21–90.

To be able to relatively compare this study with [6] (which we consider the current best-performing model in the literature), we also used a similar setup where a grapheme was used as a linguistic token for the tone restoration. The setup differed slightly in that, similar to what was used for syllable-based restoration, the dot-below was left in the text used for training and testing. The performance was, therefore, only tested on the restoration of tone marks.

9. Results and discussion

In this section, we present the results of the experiments carried out for both the syllable-based and grapheme-based tone-mark restorations.

Table 4

Mean accuracy of 10-fold cross validation of tone-mark restoration

Train [%]	Syllable level statistics				Word level statistics			
	Mean [%]	Min. [%]	Max. [%]	SD	Mean [%]	Min. [%]	Max. [%]	SD
70/30	97.69	97.25	98.28	0.35	92.53	91.17	94.40	1.14
75/25	97.75	97.25	98.28	0.35	92.69	91.17	94.40	1.12
80/20	97.79	97.25	98.29	0.38	92.82	91.17	94.40	1.19
85/15	97.86	97.11	98.29	0.46	92.99	90.79	94.40	1.47
90/10	97.90	96.92	98.53	0.53	93.10	89.46	95.15	1.83
95/5	98.09	96.90	98.84	0.75	93.75	89.93	96.10	2.41
99/1	97.83	95.59	98.99	1.10	92.84	85.85	96.23	3.40

Table 4 presented the results for the use of different training-test ratios. It is comprised of information on the mean of each 10-fold cross-validation as well as the best and worst performance. The performance was measured at both the syllable and word levels. A syllable with a wrong tone mark is erroneous while a word was adjudged wrongly tone marked if one or more syllables has a wrong tone mark.

The general trend as indicated by Table 4 is as follows: if the proportion of the corpus being used for training increases, the accuracy of the model generated increased. While at syllable level, the 75% training data model had better accuracy than the 70% training data model; however, both had the same accuracy at the word level. The model trained on 95% had an overall average syllable and word level. This is not surprising, since when more data is used to build a model, it is expected to better represent a language. The exception was at the point when almost all of the data (99%) was used to build the model. The model was less accurate than the model built with 85% of corpus data. One likely explanation was that the testing data was so small that it was not captured as well by the remainder of the data. A closer examination showed that when individual iterations were considered, the lowest accuracy was 95.59% and 85.85% at the syllable and word levels, respectively, and the highest accuracy (98.99% and 96.23% at the syllable and word levels, respectively) were in the 99% training data model.

However, the standard deviation columns (indicated as SD on the table) showed that the 75% training data model gave the most-reliable results since it has the set of least variations. We considered this important in further research into measuring the accuracy of models learned from data.

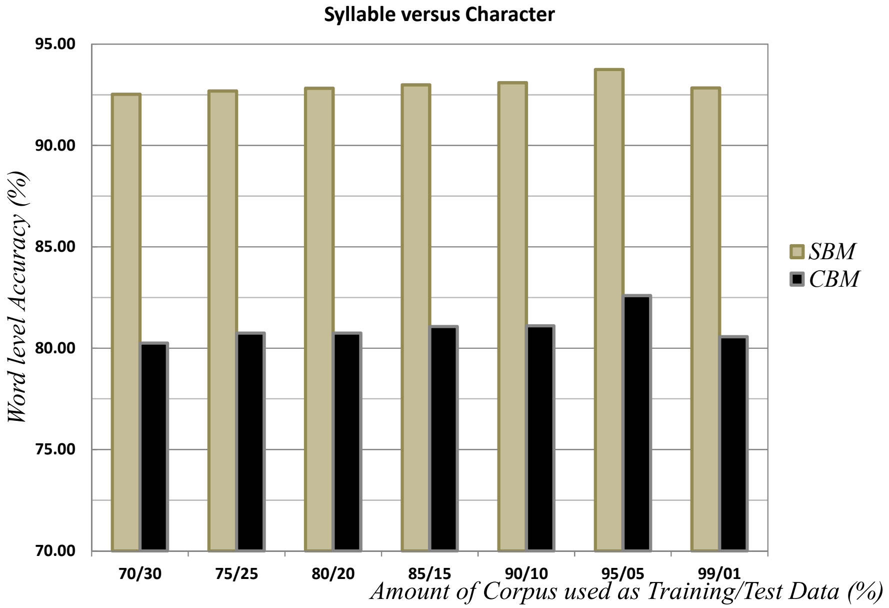

Figure 2. 10-fold cross validation mean accuracies for various training/test ratio in corpus

The chart shown in Figure 2 presents the performance of two types of models. The first is the syllable-based tone restoration model (SBM) that this study proposed, and the second is the character-based (or grapheme-based) tone restoration model (CBM). We examine the performance of the two models at the word level for the different training-to-test ratio configurations. For all of the different proportions of training data, the syllable-based models outperformed the character-based models. Using the 90:10 configuration as an example, the syllable-based model had a mean accuracy of 93.10%, while the character-based model had a mean accuracy of 81.11%. The relatively better performance for the character-based model than that reported in [6] was likely due to the fact that the focus was only on tone marks.

10. Conclusion and further work

In this study, we have presented an alternative model to the two common approaches of character-based and word-based diacritic restoration. This approch used syllable for tone-mark restoration in Yorùbá text. In previous works, no report existed on what a reliable training-to-test ratio was for learning models from data. We were able to show that a 75% training data to 25% test data ratio offers reliable measurement.

The results showed that a syllable-based tone-restoration model was more accurate than a character-based model. We can also extrapolate that, given the pattern of results from [6], [18], and [1], it will also be more-accurate than the existing word-based model. However, this conclusion is made on the basis of the current limitation of data paucity. When the available data runs into millions of words, this assertion may need revision. We will quickly point out that the maximum accuracy obtained (96.23%) could still be improved upon by combining the approach with other approaches, like lexical lookup or incorporating part of speech in the feature vector (when there are available resources).

One of the issues noticed was that some of the syllables and words that were labeled as wrong based on the data were actually the correct tonal representation when examined manually. It was the original data that contained some incorrectly labeled data. For a particular word ("oye"), the incorrect labeling occurred so often that we manually inspected the corpus and found that the word had a wrong tonal marking (Low-High). However, occurrences of the correct pattern (High-Low) were however found as substring of other words, and this correct marking was returned even when the word occurred independently. The tone marks returned by the model were not consistent with those on test text data thus considered wrong although it was the test data that had the wrong tone-marks. We also found cases in which a set of syllables forming a word were assigned a valid set of marks for a Yorùbá word. However, the set was contextually wrong and was so returned. Thus, we expect improved performance in the syllable-based model (and, of course, other models) when we have “cleaner” training data.

Further work will include a comparison of the performance of memory-based learning used in this current study, with statistical learning models like the Conditional Random Fields used in [19] and [22].This will help us examine the effect of learning algorithms on the accuracy of a model. In addition, the syllable-based model will be extended to such Nigerian languages like Igbo and Ukwuani to determine the applicability of the model across different languages.

References

[1] Adegbola T., Odilinye L.U.: Quantifying the effect of corpus size on the quality 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] Alake C.A.: Early Descriptions of the Yoruba Language: The Work of Samuel Ajayi Crowther. In: Schmitter P., Jooken L., Desmet P., Swiggers P. (eds.), *The History of Linguistic and Grammatic Praxis. Proceedings of the XIth International Colloquium of the Studienkris "Geschichte der Sprachwissenschaft", Leuven, 2nd–4th July 1998*, pp. 427–443, Peeters Publishers, 2000.

[3] Asahiah F.O.: *Development of a Standard Yorùbá Text Automatic Diacritic Restoration System.* Phd thesis, Obafemi Awolowo University, Ile-Ife, Nigeria, 2014.

[4] Brill E., Ngai G.: Man vs. machine: a case study in base noun phrase learning. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 65–72. Association for Computational Linguistics, 1999.

[5] De Palma P.A.: *Syllables and concepts in large vocabulary speech recognition.* Phd thesis, The University of New Mexico, New Mexico, The United States of America, 2010.

[6] De Pauw G., Wagacha P.W., de Schryver G.: Automatic Diacritic Restoration for Resource-Scarce Languages. In: Matoušek V., Mautner P. (eds.), *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*, vol. 4629, pp. 170–179, Springer-Verlag, Berlin, 2007.

[7] Fagborun J.G.: Disparities in Tonal and Vowel Representation: Some Practical Problems in Yoruba Orthography, *Journal of West African Languages*, vol. 19(2), 1989.

[8] Habash N., Rambow O.: Arabic Diacritization through Full Morphological Tagging. In: *Proceedings of NAACL HLT 2007*, pp. 53–56, Association for Computational Linguistics, Rochester, NY, 2007.

[9] Haertel R.A., McClanahan P., Ringger E.K.: Automatic Diacritization for Low-Resource Languages Using a Hybrid Word and Consonant CMM. In: *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, June 2010*, pp. 519–527, 2010.

[10] Larson M., Eickeler S.: Using Syllable-based Indexing Features and Language Models to improve German Spoken Document Retrieval. In: *Proceedings of Eurospeech. 8th European Conference on Speech Communication and Technology*, 2003.

[11] Liu X., Hieronymus J.L., Gales M.J., Woodland P.C.: Syllable language models for Mandarin speech recognition: Exploiting character language models. In: *The Journal of the Acoustical Society of America*, vol. 133(1), pp. 519–528, 2013.

[12] Majewski P.: Syllable based language model for large vocabulary continuous speech recognition of Polish. In: *International Conference on Text, Speech and Dialogue*, pp. 397–401, Springer, 2008.

[13] 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*, vol. 2276, pp. 339–438, Springer, 2002.

[14] Nguyen K.H., Ock C.Y.: Diacritics restoration in vietnamese: letter based vs. syllable based model. In: *PRICAI 2010: Trends in Artificial Intelligence*, pp. 631–636, Springer, 2010.

[15] Olúmúyìwá T.: Yoruba Writing: Standards and Trends, *Journal of Arts and Humanities*, vol. 2(1), p. 40, 2013.

[16] Ọdẹ́jọbí O.A.: *A Computational Model of Prosody for Yorùbá Text-to-Speech Synthesis*. Phd thesis, Aston University, Aston, 2005.

[17] Šantić N., Šnajder J., Bašić B.D.: Automatic Diacritics Restoration in Croatian Texts. In: *INFuture2009: Digital Resources and Knowledge Sharing*, pp. 309–318, 2009.

[18] Scannell K.P.: Statistical Unicodification of African Languages. In: *Language Resources and Evaluation*, pp. 1–12, 2011. Retrieved July 20, 2011 from http://borel.slu.edu/pub/lre.pdf.

[19] Schlippe T., Nguyen T., Vogel S.: Diacritization as a Machine Translation Problem and as a Sequence Labeling Problem. In: *AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pp. 270–278, Waikiki, Hawai'i, 2008.

[20] Schrumpf C., Larson M., Eickeler S.: Syllable-based language models in speech recognition for English spoken document retrieval. In: *Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib, Cortona, Italy*, pp. 196–205, 2005.

[21] Surmei M., Burileanu D., Negrescu C., Pîrvu R., Ungurean C., Derviş A.: Text-to-Speech Engines as Telecom Service Enablers. In: *Advances in Spoken Language Technology, Publishing House of the Romanian Academy, Bucharest*, pp. 89–98, 2007.

[22] Truyen T.T., Phung D.Q., Venkatesh S.: Constrained Sequence Classification for Lexical Disambiguation. In: *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 5351, pp. 430–441, Springer, 2008. Retrieved from http://www.computing.edu.au/~trantt2/pubs/pricai08.pdf.

[23] Tufiş D., Ceauşu A.: DIAC: A Professional Diacritics Recovering System. In: *Proceedings of the Sixth International Language Resources and Evaluation*, 2008, Paper 54 on Conference CD.

[24] Tufiş D., Chiţu A.: Automatic Diacritic Insertion in Romanian Texts. In: *Proceedings of the International Conference on Computational Lexicography COMPLEX'99. Pecs, Hungary*, pp. 185–194, 1999.

Affiliations

Franklin Ọládiípọ̀ Asahiah
Obafemi Awolowo University, Ile-Ife, Nigeria, Department of Computer Science and Engineering, sobusola@oauife.edu.ng

Ọdẹ́túnjí Àjàdí Ọdẹ́jọbí
Obafemi Awolowo University, Ile-Ife, Nigeria, Department of Computer Science and Engineering, oodejobi@oauife.edu.ng

Emmanuel Rotimi Adagunodo
Obafemi Awolowo University, Ile-Ife, Nigeria, Department of Computer Science and Engineering, eadagun@oauife.edu.ng

Received: 28.06.2016
Revised: 10.06.2017
Accepted: 15.06.2017