I.J. Information Engineering and Electronic Business, 2020, 6, 28-38
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2020.06.03

Spelling Error Patterns in Typed Yorùbá Text Documents

Asahiah Franklin Oladiipo
Obafemi Awolowo University, Ile-Ife, Nigeria
Email: sobusola@oauife.edu.ng

Onifade Mary Taiwo
Obafemi Awolowo University, Ile-Ife, Nigeria
Email: taiwoonifade17@gmail.com

Adegunlehin Abayomi Emmanuel
Obafemi Awolowo University, Ile-Ife, Nigeria
Email: adegunlehinabayomi@gmail.com

Received: 27 March 2020; Accepted: 24 June 2020; Published: 08 December 2020

Abstract: While writing in most of the world's major languages have a long history, Yorùbá is a relatively young language as far as writing it down is concerned. It is therefore an under-resourced language as far as tools for processing it in digital format is concerned. Spell checking is one of these tools. An analysis of the spelling error pattern is fundamental to the task of producing a good spell checker. We addressed this challenge in this article and our findings showed that spelling error pattern in Yorùbá followed that of other languages in general. There were, however, obvious departure from the norms in the specific. Diacritic-related misspelling accounted for more than 80% of all errors and words with single edit error were less than the generally expected minimum threshold of 80%. In addition, most of the errors were vowel-related with consonants accounting for less than 15% of all errors. Word-length does not seem to have any direct bearing on number of errors in a word. The research showed that the impact of diacritics on spelling error is more in Yorùbá where diacritics are majorly used for tone marking where it accounts for more than 80% of spelling errors than in languages like Brazilian Portuguese and Spanish where diacritics are used for differentiating characters where spelling error due to diacritics covered less than 60% of all errors. We thus conclude that while, to a significant extent, the character set used in a language determines distribution of spelling error, the purpose to which diacritics is employed in language also affect the distribution of spelling error in a language.

Index Terms: Yorùbá, diacritics, misspellings, patterns, spellchecking

1. Introduction

Yorùbá, a major language spoken in Nigeria, a former British colony, is used in many different contexts in several countries around the world. These countries include Benin Republic, where it is also indigenous, Ghana, Sierra Leone, Ivory Coast, the United Kingdom (UK) and United States (US), where it is spoken by immigrant communities. In some of these countries, for example the UK and US, Yorùbá language is an object in research in tertiary institutions. In the Caribbeans, Yorùbá goes by names like Lucumi, Nago and Anago where it is used as liturgical language.

Yorùbá has a written history of close to two hundred years. This might be short when compared to some European languages but is quite considerable when compared with many other African languages as there are still several African languages that are still unwritten. Yorùbá writings in Roman scripts started from the humble beginning of collections of numeral one to ten by Bodwich and few vocabularies by Kilham in 1819 and 1828 respectively [10]. A translation of some portions of The Holy Bible to Yorùbá language was first published in 1844 while the first Yorùbá newspaper: *Iwé Ìròhìn fún Àwọn Arábìgbá Àti Yorùbá*, came out in 1859 [19,17,21].

The first Yorùbá dictionary by Crowther, a clergy of the Church Missionary Society of Yorùbá descent was published in 1843 [7] and elaborate efforts at standardizing the orthography started in 1875 with a version approved by the Nigerian government committee in 1974 [18]. *Itan igbesi aiye emi "Ṣegilọla Ẹlẹyinju-Ẹgẹ", Ẹlẹgbẹrun ọkọ l'aiye*, the first Yorùbá novel, was published in 1829 [1] and since then the tradition of written literary works and studies has been ongoing with Yorùbá being an examinable subject at secondary schools in 1931 and university degree in 1974.

Despite this history of literary development in Yorùbá, and the volume of current publications in educational materials, newsprint, news broadcasting, web contents and various outlets, Yorùbá is still an under-resourced language

as far as language processing is concerned. One will be surprised that Yorùbá language spell checking facility is mostly unavailable in most text processing software packages. Microsoft Office document processing package, Microsoft Word incorporated spell checking for Yorùbá language in its 2016 version update pack available in 2018.

However, we are not aware of any existing publication on spell checking in Yorùbá. Much more important than just the development of the software, however, is the study of the spelling error pattern in Yorùbá. Existing studies on spelling error patterns have looked at some characteristics of language under investigation but we are not aware of any looking at the challenge that tone marking in addition to use of character diacritics may pose to spelling.

This study focused on the investigation of the spelling problem encountered in existing Yorùbá documents as a case study of tone languages and conclusions that can be drawn from such investigation that will better inform the development of such spell checker for Yorùbá language. The remainder of the article comprises review of relevant literature, the data used and methodology used for error detection and its analysis. Subsequently, a discussion on the implications of the analysis is given and a conclusion drawn summing the whole article.

2. Existing studies

[9] described three kinds of errors attached to text production processes: verbalization error that occurs when ideas crystallize incorrectly to thought word, spelling error that occurs in the process of converting thought word to spelled word and finally, typing/writing error that occurs in the course of physical production of letters. In [5] the first two kinds of errors above are grouped together as orthographic that spelling errors can thus be correctly modeled if the error pattern is well understood. The first two stages of errors (verbalization and spelling) in [9] were compounded together into a single type known as orthographic error in [5] which they considered to be cognitive in its origin. [11] made yet another categorization of spelling errors: sound-based error; rule-based error, writing errors; multiple errors and the error set (replacement, deletion, insertion and transposition) of [8]. The above categorizations of error types/sources are descriptions that used English as the testbed language. Researches have also been done on spelling error patterns on several other languages like Japanese, Spanish, Arabic, Portuguese and so on.

[6] analyzed the pattern of spelling errors in Spanish text and concluded that some of the pattern of spelling errors found in English might not carry over to other languages especially ones in which the letters of their alphabets are not composed of pure Basic Latin characters. Pattern for Spanish showed a significant variance with the pattern reported for English in [14] to the extent that only two out of the six pattern findings reported for English overlapped with that of Spanish [6]. Of interest in their finding was that error with diacritics formed a significant portion (> 50%) of the spelling error and errors with first letter was the third highest type of error.

[15], in a work on Urdu that utilizes the Abjad writing system also supported the observation in [6] that word-initial errors are common spelling errors. In addition, [14], there is a type of error called shape-similarity based errors of which wrong diacritical marking (presence, absence, numbers of, and placement) is also reported to be a high contributor in spelling errors found in Urdu text. Findings on Portuguese [2,13] followed a similar trend of errors in diacritics contributing significantly to spelling errors. One of the conclusions in [13] was that the distribution of errors is affected by the character set being used in language than maybe other factors. We sought to investigate this assertion for Yorùbá language. This is because like Portuguese and Spanish and even languages like Urdu (that use different script), Yorùbá character set utilizes diacritical marks. However, unlike these languages here-mentioned, Yorùbá majorly uses diacritics to mark tone, in addition to letter differentiation within the alphabet. We therefore seek to find if the additional function of diacritics (as tone marker) will have any effect on misspelling pattern.

3. Methodology

The approach applied in this research involves the aggregation of data from different sources to be used as the data for analysis. The text gathered was pruned and selected on the basis of classification into broad categories. The selected material formed the data for the research. The data was examined for spelling errors by human subjects with literacy in the language and minimum of a university degree. The outcome formed the basis for the analysis carried out in the research.

3.1 Data Collection

For this pilot research, our data is limited to those text created by entering text via keyboard. Other sources of digital text excluded from our data source are handwritten text and text created by optical character recognition of scanned hard copy of books/documents. The restriction was so informed since the spell-checker that the outcome of this research should inform is for text entered from keyboard.

The corpus of data used for this experiment were drawn from three documents which were purposefully selected to reflect different writing styles: formal, semiformal and informal styles. The documents were also of different sizes and obviously different levels of editing. The first document was a public address note by a professor of Yorùbá language at the launching of a version of the Holy Bible in Yorùbá language while the second document was drawn from a series of blog articles written in Yorùbá with coverage ranging from travels, foods, political affairs and agriculture. The last

document is a university undergraduate final year project report. These documents were selected to reflect various skillset of Yorùbá writers without the assistance of professional editors associated with book publishing.

3.2 Sources of Data

Selection of documents to be used as data source for this research followed a purposive data sampling approach. Purposive data sampling became necessary to reflect the categories of subjects that write extensively in Yorùbá language. While Yorùbá language is taught from basic educational level to graduate education level. The writer base is limited especially subjects who write in the digital medium. Secondary school candidates are taught and examined on text handwritten on paper. However, undergraduates are expected to create their text in digital media before printing them out for submission. In additional, only highly proficient Yorùbá language users create documents and blogs in Yorùbá language.

The only other large group of users that use Yorùbá language extensively that we left out intentionally are Yorùbá newspapers. The reason for their removal from the documents used for this research is their well-known non-adherence to tone-marking [18]. Using their data would have seriously biased the outcome of such data collection. Another segment that was not considered in this research are documents generated via optical character recognition of text from printed material because the characteristics of such documents are different from documents created by direct entry of text via the keyboard. Thus, the documents were purposively selected to represent as accurately as possible the expected user-base of a spellchecking system that will be developed based on the analysis carried out in this research.

After the documents were carefully selected to represent the expected/target user base, data was extracted from each document by randomly selecting several paragraphs that would contain sufficient data for analysis.

3.3 Size of Data

This study is a pilot study that hopes to be a guide to further research in this area and as such, the size of data used was restricted by easily available genres and documents. The total word count in the three documents was 34,298. Text from within a portion of each document were selected for analysis. The size of analyzed text for each document was chosen to be able to get a fair and sizable errors words to subject to analysis. Document1 (D1) was written in semi-formal style and the analyzed portion consist of 1000 words out of 3807 words, Document2 (D2), written in informal style has 2444 words in the portion analyzed out of the total word count of 11322 and Document3 (D3) consist of 19169 out which a portion made up of 1004 words were analyzed.

At the time of the analysis, even though a language pack existed in Microsoft Word package, it was not considered of sufficient use since it does not flag words written without tone-marks and the dot-below diacritics as long as the base Latin letter sequence are correct. The standard orthography requires that text be written with correct tone marks on every vowel and nasal consonant. Therefore, errors were manually identified. The errors were manually identified by two subjects who are first language speakers and writers of Yorùbá who cross-checked their results and also helped in mapping the erroneous to their corrections given their sentence contexts. A sample of the identified misspellings is shown in Table 1. Table 1 includes information at word-level, character-level, tone and dot-below diacritics, the edit operation(s) [8] that may be involved in correcting the error and finally, whether error is a real word or non-word error.

3.4 Error Analysis

Using the standard categorization of spelling error as belonging to the four possible classes of deletion, substitution, insertion and transposition, the error in each word was characterized. Additional information on some factors related to the errors were also highlighted. These include: role of tone in the misspelling, role of dot-below diacritic in the misspelling, whether the misspelling resulted in real word or non-word and also the number of errors committed per single word. It was noted that these errors can be due to one or some combinations of any two of the following error types: deletion, substitution or insertion error as far as dealing with characters were concerned.

The following additional error classifications were identified: wrong removal or insertion of dot-below diacritic, substitution of tone mark on a character with another one, phonemic substitution. However, it is worthy of note that misspelling error of the type "transposition" was not detected at all within the analyzed text. The proportion of error in each document varied greatly. The total number of misspelling in each of the three documents are 144 (14.4%), 30 (1.23%) and 153 (15.24%) while total number of unique misspellings were 90, 22 and 77 for D1, D2 and D3 respectively. The total number of misspellings and unique misspellings analyzed were 327 and 189 respectively giving the following ratio: for every seven misspelling, four of these misspelling will be unique. It might suffice to say that the low error rate of D2 is likely due to serious and rigorous effort at good presentation since the audience is wider and the impact of misspelling will affect the ability of the blog to attract readers.

Based on the analyzed text within these three documents the average misspelling error rate was 7.35% with a standard deviation of 7.86. The values for D1 and D3 are closer to those reported for English (11.8%) in [4]

Table 1. A Sample of extracted Yorùbá spelling error

Error	Error	T1	T2	dot	del	sub	ins	**Real**
lọ	ọ	m	m	wi		o		**true**
t'ọ́n	', ọ́	h	h	na		á	'	**false**
bẹ́ẹ̀n áàni	conc			na	2 sp			**false**
mọ́nra	n			na			n	**false**
en í	e, í	m,	m,	wr		ẹ, i		**true**
ọ jó	ó	h	h	wr		ọ́		**false**
ìdákọ́lẹ	ẹ	m	l	na		ẹ̀		**false**
àṣemọn	n			na			n	**false**
man	a, n			na		ọ	n	**false**
mú nwa	conc, ins			na	1 sp		n	**false**
àl áf íà	Mschar			na	à			**false**
arawa	Conc			na	1 sp			**false**
èn ìà	Mschar			na	y, n			**false**
àṣe	e	m	m	wr		ẹ		**false**
l átààn á	à			na			à	**false**
mọ̀n	n			na			n	**false**
mọ́n	n			na			n	**false**
tọ́n	ọ́	h	h	na		á		**false**
nkan	mschar			na				**false**
àd ídù	mschar			na	n			**false**
o újẹ	mschar			na	n			**false**

Keys:
T1 = tone seen in text T2 = correct tone expected
dot=dot-below; del=deletion; sub=substitution; ins=insertion
l=low-tone mark m = mid-tone mark h=high tone mark conc=two or more words concatenated
mschar=missing character(s) wi=wrongly inserted
wr=wrongly removed na=not applicable

4. Error Typology and Error Patterns

The following categorization was used in analyzing the error pattern found within the documents:

- Single-error versus multi-error in a word
- error due to diacritic misapplication or unrelated to diacritic
- error due typo or phonetic substitution
- real word errors versus non-word errors
- error due to substitution, deletion, insertion or transposition of character(s) in a word

4.1 Single vs. multiple errors at word level

An aggregation of the analysis of the three documents indicated that 145 words have single character error while the remaining 44 words have misspelling error in more than one character. One word had error in all the five vowels involved while eight other words had either three or four characters in error per word out of the total 189 words. The pie chart showing the distribution of number of characters in error as a percentage of total is shown in Figure 1 which indicates that the percentage of single error misspellings in the analyzed data is less than the generally accepted threshold of above 80%.

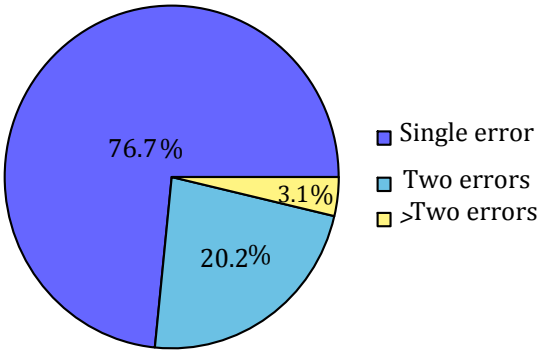

Fig. 1. Percentage of character errors per word

4.2 Error due to diacritic use or other sources

In Yorùbá orthography, diacritics are used for two purposes: to mark tone and the dot accent below is also used so as to distinguish between the characters. The dot‑below diacritical mark is used to differentiate the open form of vowel "e" and "o" from the closed vowel letters "ẹ" and "ọ" respectively. The consonant letter "s" have a similar version with a dot-below diacritic yielding consonant ṣ. Tones are marked with acute, grave, and macron for high, low and mid tone respectively. The convention is to leave out the macron from vowels and indicate it only on syllabic nasals. An analysis of the documents at both word and character levels indicated that more than 86% of the words have their errors originating from wrong or misapplication of diacritics while the remaining less than 14% are due to other reasons like typos, run-ons or phonetic substitution. At character level, 220 characters have diacritic errors out of 266 characters errors. These occurred by either leaving it out where it should on vowels or substituting the correct one with another. The confusion matrix for use of tone marks is shown in Table 2 with the variance as per weighted average of each document reflected in the table as values inside the brackets. Further details are included in Table 4. However, it is clear from Table 2 that most of the error due to tone are caused when a character (mostly vowels that should be indicated as having either high tone or low tone are indicated with mid-tone. This is not surprising since mid-tone is indicated by the absence of diacritical mark over the character. It is well known that indicating the high or low tone diacritic adds at least two extra key press (in for example Microsoft Word) and as such it is easier to make a mistake by not adding it than adding a different diacritical mark. Table 5 further down in section 4.6 give an additional support to this assertion as vowels with dot-below diacritics (ẹ and ọ) whose entry via the keyboard is further compounded to at least five keypresses when carrying tone mark diacritics have more than double the share of their open counterparts (e and o) in contribution to misspelling errors. In all, mid-tone has the highest accurate representation in text of 84.69% followed by low tone at 70.87%. High tone was least accurately represented at 63.91% accuracy.

Table 2. Aggregate Tone-marks Error Confusion Matrix

Diacritic Observed	Diacritic Expected			
		Low tone	Mid tone	High tone
	Low tone	90(0.017)	08(0.001)	03(0.0)
	Mid tone	33(0.005)	83().005)	45(0.029)
	High tone	04(0.0)	07(0.00)	80(0.022)

4.3 Error due typo or phonetic substitution

In Yorùbá there are few orthographic conventions that stipulate that some particular letters are combinable to realize some syllable sounds. The effect of not following this convention will lead to few cognitive spelling problems in which a word that is phonetically correct is orthographically wrong. All the instances of phonetic errors in Yorùbá are associated with either the substitution of ***"ọn"*** for ***"an"*** (and vice-versa) or the addition (insertion) of **"n"** to a vowel to create the nasal vowel variant when the vowel is attached to one of consonants **"m"** and **"n"** but which should not be done when using standard Yorùbá orthographic conventions. All the instances of phonetic error were found in a single document which might indicate that it is not a general problem but an individual idiosyncrasy. The whole instances presented in Table 3 showed that only eight (8) representing 4.32% were spelling errors of phonetic origin.

Table 3. Phonetic misspelling and orthographic correction

Phonetically wrong	Orthographically correct
t'ọ́n	tán
mọ́nra	mọ́ra
àṣemọn	àṣemọ
man	mọ
mu ńwa	mu ńwa
mọ̀n	mọ̀
mọ́n	mọ́
tọ́n	tán

4.4 Real word errors versus non-word errors

Real word errors are misspellings that results in another dictionary word being generated while non-word errors are misspelling whose outcome has no dictionary meaning. In Yorùbá text, the most common ground upon which misspelling often result in real word is incorrect diacritic application to a word. Some Yorùbá words without diacritic like "*igba*" has up to five variants when the diacritics have been correctly added. Most monosyllabic words have three (or at least two diacritic variants). This situation often leads to generation of real words when misspelling is due to diacritics being misplaced. Our analysis showed that most of the real word errors in text are in fact words with the sequence of tone marks or dot-below differently placed. In our text collection, 32.28% of all misspelled words were real words while the remaining 67.72% were non-words. The detail is found in Table 4.

4.5 Error due to substitution, deletion, insertion or transposition

Irrespective of the source of error, we also considered the cause of the error found in the misspelled word. A spelling error could be due to several reasons. The incorrect spelling was caused by a different letter being replaced for the correct one (substitution), or a character that should have been present not being found in the word (deletion) or an additional letter that is not need being introduced (insertion) or finally it could be due to position of two adjacent characters being interchanged transposition. We found that substitution was the most prevalent problem in the Yorùbá text followed by deletion and lastly, insertion. Transposition as a kind of spelling error was not found in the text collection analyzed. At word level, substitution accounted for 82.44% of all kinds of error, deletion accounted for 10.73% while insertion was responsible for 6.83% of all errors found in the collection. At character level, substitution all took the lion share of all kinds of error being responsible for 84.62%. The prevalence of substitution as the greatest source of misspelling is obviously directly related to prevalence of diacritic-related misspellings as indicated in section 4.2. Although other kinds of substitution misspelling occurred, it is obvious that substitution error and diacritic error are highly correlated in Yorùbá text. Further details are found in Table 4.

Table 4. Categories of Errors

Categories	**Words**	**Characters**
Multiple Error	44	
Single Error	145	
only Tone errors (TE)	94	113
only Dot-below errors (DBE)	85	105
both TE and DBE on same character		23
Character error due to diacritic		220
All character error		266
Real word error	61	
Non-word errors	128	
Deletions	22	26
Substitution	169	226
Insertion	14	14
Transposition	0	0

4.6 Error distribution by characters

The errors, by collapsing all variant diacritic forms base form (without any diacritics), the distribution of characters in errors either by deletion, substitution or insertion is shown in Table 5 from which it is obvious that all the vowels are involved in spelling errors while only two of consonants are involved misspelling in the analyzed Yorùbá text.

Table 5. Error distribution by letters

Character	Count
a (a, à, á)	48
e (e, è, é)	14
ẹ (ẹ, ẹ̀, ẹ́)	32
i (i, ì, í)	35
o (o, ò, ó)	28
ọ (ọ, ọ̀, ọ́)	72
n (n, ń, ǹ, ń)	12
s (s, ṣ)	16
u (u, ù̩, ú̩)	8
Y	1
punctuation (')	1
space character	10
Total	**277**

The letter 'n', although numbered amongst the consonant has additional functions in the Yorùbá orthography. 'n' acts anas a syllable nucleus being a syllabic nasal as well as an orthographic device for indicating nasal vowels by being appended to the end of vowels. In the context of occurrence of the letter 'n' either as deletion or insertion error, it did not appear as consonant letter but as either syllabic nasal or as part of nasal vowel.

4.7 Word length and number of errors in a word

There is no clear pattern of correlation between the number of characters in a word and the amount of character or diacritic misapplication in that word. For example, a single character can have up to two diacritic errors. However, the higher the number of the two vowels that can have both tone and dot-below diacritics in a word, the higher the likelihood of diacritic error being introduced in that word. Example is ***"ọ̀jọ̀gbọ́n"*** (six diacritic markings) without its diacritics is more likely to have error than ***"àgbàlá"*** (three diacritic markings) without its diacritics.

5. Performance of Existing Commercial Spell Checker

A word processing software (word processor) is an application that enables users to create, edit, print and manipulate documents comprised majorly of text and possibly a few graphic content. Microsoft Word also known as MS Word is the only word processing software to our knowledge with a working proofing tool (spell checking component within the word processor) for Yorùbá. We have therefore chosen compare the performance of MS Word spell checker for Yorùbá with manual identification of spelling errors on the collected misspellings that we had analyzed previously. A total of 185 erroneous words were subject to spell checking by the MS Word out of all analyzed words. This is because they were the set words that could be corrected either independent of context or together with the use of context when necessary. The algorithm(s) within the Microsoft Word proofing tools for Yorùbá (Microsoft Word Yorùbá Spell-checker: MYSC) is unknown but it indicated that it is not using a dictionary by stating that a dictionary for the language is not available.

5.1 Detection of Error

Out of the 185 words, 127 words were words that do not have meaning irrespective of the context they appear but correct spelling could still be generated for them; they are non-word errors which we tagged CI-words (Context-Independent misspelled words). The remaining 58 words are only wrong within the context they appeared but would have made sense in other contexts; they are context-sensitive errors (tagged CS-words: Context Sensitive misspelled words).

Out of these 127 CI-words, 83 CI-words representing 65.35% were not detected as misspellings by MYSC. Only 44 were correctly identified as misspellings. The detection accuracy of MYSC was approximately 34.65% on CI-words. On CS-words, MYSC was supposed to detect all of them as correct since it was checking independent of context and it scored 100% on the task of correctly detecting them as valid words.

5.2 Recognition of Correction to Errors

Furthermore, MYSC was applied to both CI-words and CS -words corrections. Of the 127 CI-words to which corrections were applied, 22 correction words or 17.32% of the correction were still not accepted as valid words and so were marked as misspellings. 1/3 of the CI-words corrections not accepted as valid comprise words had the original misspellings considered as correct spellings by MYSC. Only one CS-word word correction or 1.72% correction to CS-words out of 58 CS-words was considered a misspelling bringing the total number of corrections to spelling errors marked as misspelling themselves to 23 (or 12.43%).

The performance of MYSC on the CS-words was as expected. However, the performance on CI-words indicated that the problem of spell-checking for Yorùbá is far from being solved. A summary of the performance on CI-words is presented in Table 6 containing both raw number and percentage in the distribution of the ability to detect or otherwise based on the categorization of errors.

Table 6. Performance of MYSC on CI-word errors

	Count of UCI	UCI %	Count of DCI	DCI %
Dot below (DB) alone	49	59.04	9	20.45
Tone marks (TM) alone	15	18.07	17	38.64
DB & TM in same word	8	9.64	1	2.27
TM & other types	2	2.41	4	9.09
other types	9	10.84	13	29.55
Total	83	100.00	44	100

Keys:
UCI: Undetected CI-words
DIC: Detected CI-words
Other types: insertion, deletion and substitution

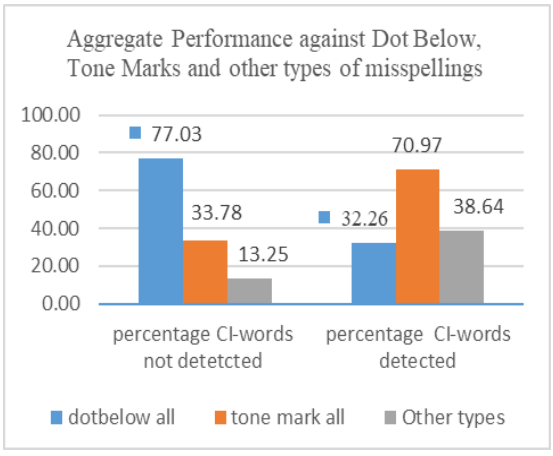

Fig. 2. Aggregate performance of spelling detection of DB, TM and Other types.

The performance in detecting CI-words as shown in Table 2 show that in words not detected as misspelling, other types of errors like insertion, deletion, substitution, splitting error and concatenation error accounted for less than 15% while the involvment of diacritic-related errors is close to 89%. On the other hand, in those CI-words whose misspelling was detected, the involvment of other types of errors was close to 39% but diacritic related error only account for 70%. Figure 2 showed that dot-below errors followed by tone-mark error were the major factors in the limitation to detecting spelling errors while tone marks followed by other types of errors were prepondrant in those spelling errors detected. By sheer number, tone marks limited the ability of MYSC to detect misspelling than assisting it to detect misspelling.

6. Discussion

6.1 Error edit

The spelling error pattern for Yorùbá followed the generally observed trends in other languages [14,6] in relation to the number of edit operations that is needed to transform the erroneous spelling to the correct one is mostly single edit misspelling. Though single error misspelling predominated (76.72%), the percentage of such is lower than the 80% benchmark [14] values for English which does not use diacritics in writing or French, Spanish and Brazilian Portuguese which employ diacritics, all of which have single error misspelling above 80% [20,13,6]. This pattern may be reflective of the fact that level of familiarity and adherence to orthographic standard is low [12].

6.2 Diacritic errors

Eighty-six percent (86%) of all spelling errors in the analyzed text were diacritic related. This situation may not be unconnected with the fact that employment of diacritics in Yorùbá language to mark tones is significantly elevated compared to other languages that employs diacritics to extend character set. According to [3], almost 70% of syllables (the tone bearing unit in tonal languages) in Yorùbá text have tone marks indicated by diacritical marks (high tone by acute accent and low tone by grave accent). In addition, three letters out of the twenty-five (25) in the Yorùbá alphabet also mark character differentiation. The pattern reflected in Table 2 is similar to Spanish [6] where diacritics were more omitted than substituted for one another. Furthermore, given that fidelity to standard orthography in relation to tone marking is reported to be averagely low [12], this pattern may not be surprising and until the correct usage of tone mark is adhered to, the pattern may continue to persist.

6.3 Phonetic, real and character errors

The phonetic errors in Yorùbá is comparable to that of Spanish with a value of 4.23% compared to 5.9%. However, the error was found in only one of the document, D2, which is also the one with the lowest error rate 1.23% (D1:14.40%, D3:15.24%) might indicate that it is due to a personal lack of understanding of orthographic convention and phonetic errors therefore might not be a common problem for Yorùbá spell checking. Similarly, misspelling that results in real words accounted for 32.28% of all errors. This value fell within the range of 30% – 40% reported [14]. The authors in [16] was of the opinion that the average rate of occurrence of non-word error could be put at 67.5%. Our result on ratio of real word to non-word therefore agree with general trends found in other languages. It is also interesting to find that errors in vowels accounted for over 85% of all errors. This disproportionate distribution is due to the preponderance of diacritic-related error since diacritics are mostly placed on vowels. It is therefore likely that any effort to address vowel correction will address diacritic-related errors and vice-versa. Deletion of the space between two or more words was also a common phenomenon as was also recorded for other languages but punctuation error occurred only once in all the analyzed text collection.

7. Conclusion

In this paper we have examined various error classification for the Yorùbá language text categorizing errors by edit operations required to transform misspelling to correct word, error on diacritic placement, amount of real errors versus non-words, phonetic errors and distribution by characters. We established, in agreement with [6] that character set affects the pattern of distribution of spelling error in a text. We were also able to show that pattern for text of a language that uses diacritics to mark tones is different to those that either do not use diacritics at all or that use diacritics for other purposes like differentiating characters. We found that of the four error categories by [7], unlike what was reported for most other languages, substitution was the largest problem being responsible for more than 80% misspelling problem. In addition, any spell checking and correction system that will be effective for a word-processing Yorùbá text will need to make sufficient provision for up to two (2) edit operations as single edit operation will cover less than the 80% threshold found in other languages. Effort should also be made to design software that will cover context-sensitive spell checking as real word misspelling is a significant problem in the language, most especially due to ability of different diacritic placement on characters to yield different valid words.

Further research that we are already working on is the implementation of the information yielded by this analysis to produce a robust spell checking and correcting software for the Yorùbá language. Our first prototype is focusing on spell-checking and correction of non-word misspelling in Yorùbá text while subsequent version will include context-sensitive real-word errors that also happen to be significant in proportion of total error trend.

References

[1] Adejumo, A.: A postcolonial analysis of the literary and cultural consequences of the abolition of the 18th century transatlantic slave trade on the Yorùbá of south western nigeria. Lumina, 2010, 21(2), 1–1

[2] Andrade, G., Teixeira, F., Xavier, C.R., Oliveira, R.S., Rocha, L., Evsukoff, A.G.: Hasch: high performance automatic spell checker for Portuguese texts from the web. Procedia Computer Science, 2012, 9, 403–411
[3] Asahiah, F.O., Odejobi, O.A., Adagunodo, E.R.: Restoring tone-marks in standard Yorùbá electronic text: improved model. Computer Science, 2017, 18(3,) AGH University of Science and Technology Press. DOI: https://doi.org/10.7494/csci.2017.18.3.2128
[4] Bebout, L. An error analysis of misspellings made by learners of English as a first and as a second language. Journal of Psycholinguistic Research, 1985, 14(6), 569–593.
[5] van Berkelt, B., Smedt, K.D.: Triphone analysis: A combined method for the correction of orthographical and typographical errors. In: Second Conference on Applied Natural Language Processing. Association for Computational Linguistics, Austin, Texas, USA 1988). https://doi.org/10.3115/974235.974250 1988, 77–83.
[6] Bustamante, F.R., Díaz, E. L.: Spelling Error Patterns in Spanish for Word Processing Applications. In: LREC, 2006, 93–98
[7] Church Missionary Society: Dictionary of Yorùbá Language Church Missionary Society, Lagos, 1913.
[8] Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of the ACM, 1964, **7**(3), 171 – 176
[9] Deorowicz, S., Ciura, M.G.: Correcting spelling errors by modelling their causes. Int. J. Appl. Math. Comput. Sci., 2005, 15(2), 275 – 285
[10] Desmet, P., Jooken, L., Schmitter, P., Swiggers, P. (eds.): The History of Linguistic and Grammatical Praxis. Leuven/Paris/Sterling: Peeters, 2000
[11] Elliott, G., Johnson, N.: All the right letters–just not necessarily in the right order. spelling errors in a sample of gcse english scripts. In: Paper presented at the British Educational Research Association Annual Conference, Edinburgh, UK. 2008
[12] Fagborun, J.G.: Disparities in tonal and vowel representation: Some practical problems in Yorùbá orthography. Journal of West African Languages **19**(2) 1989,
[13] Gimenes, P.A., Roman, N.T., Carvalho, A.M.B.: Spelling error patterns in Brazilian Portuguese Computational Linguistics, 2015, 41(1), 175–183
[14] Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR), 1992, 24(4), 377–439
[15] Naseem, T., Hussain, S.: A novel approach for ranking spelling error corrections for urdu. Language Resources and Evaluation, 2007, 41(2), 117–128.
[16] Nguyen, T.T.H., Jatowt, A., Coustaty, M., Nguyen, N.V., Doucet, A.: Deep statistical analysis of OCR errors for effective post-ocr processing. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL),IEEE, 2019,. 29–38.
[17] Ogunbiyi, I.A.: The search for a Yorùbá orthography since the 1840s: Obstacles to the choice of the Arabic script. Sudanic Africa: A Journal of Historical Sources, 2003, 14, 77–102
[18] Olúmúyìwá, T.: Yorùbá writing: Standards and trends. Journal of Arts and Humanities, 2013, 2(1), 40
[19] Omu, F.I.A.: The 'ìwe irohin', 1859-1867. Journal of the Historical Society of Nigeria, 1967, **4**(1), 35–44, http://www.jstor.org/stable/41971199
[20] Ren, X., Perrault, F.: The typology of unknown words: an experimental study of two corpora. In: COLING 1992 Volume 1: The 15th International Conference on Computational Linguistics, 1992.
[21] Salawu, A.: The Yorùbá and their language newspapers: Origin, nature, problems and prospects. Studies of Tribes and Tribals, 2004, **2**(2), 97–104. https://doi.org/10.1080/0972639X.2004.11886508.

Authors' Profiles

Asahiah F. O. was born in 1972 and earned his B.Sc., M.Sc. and PhD from the Obafemi Awolowo University, Ile-Ife, Nigeria in 1997, 2005 and 2014 respectively. He joined the Department of Computer Science and Engineering of Obafemi Awolowo University, Ile-Ife as a Graduate Assistant and has since risen to the position of Senior Lecturer. He has several publications to his credit including "The development of a syllabicator for Yorùbá language (Proceedings of OAUTekConf, 2010, Nigeria)", "Restoring Tone-Marks in Standard Yorùbá Electronic Text: Improved Model (Computer Science, 18(3), Poland)" "Survey of Diacritic Restoration in Abjad and Alphabet Writing Systems (Journal of Natural Language Engineering, 24(1), 2018, UK)" and "Computational Modelling of an Optical Character Recognition System for Yorùbá Printed Text Images (Scientific Africa, Vol 9, 2020)". His research interest is in Human language processing especially in developing resources for low-resourced languages and application of machine learning to text processing.

Onifade M.T is a Nigeria born in 1986 and earned her M.Sc in 2018 through the Department of Computer Science and Engineering of Obafemi Awolowo University, Ile-Ife.

She has four years' experience in the application of artificial intelligent to the area of computational linguistics especially text processing in Nigeria indigenous languages. She is currently a doctoral student working on development of a grammar and context-sensitive spell checking for Yorùbá Language. She has a publication: "Investigation of feature characteristics for Yorùbá named entity recognition system (Proceedings of AICTTRA, 2019, Nigeria)" to her credit

Adegunlehin A. E. was born in 1990 and earned his Master's degree from the Obafemi Awolowo University, Ile-Ife, Nigeria in 2019. He is currently a PhD student in the Department of Computer Science and Engineering of Obafemi Awolowo University, Ile-Ife. He has a publication to his credit "Investigation of feature characteristics for Yorùbá named entity recognition system" (Proceedings of AICTTRA, 2019, Nigeria)". His research interest is in the application of machine learning techniques to solving Human language processing tasks.

How to cite this paper: Asahiah Franklin Oladiipo, Onifade Mary Taiwo, Adegunlehin Abayomi Emmanuel, " Spelling Error Patterns in Typed Yorùbá Text Documents", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.6, pp. 28-38, 2020. DOI: 10.5815/ijieeb.2020.06.03