



**ADDRESSING TONE-MARKING CHALLENGES IN DIGITISED YORUBA  
TEXTS FOR WEB COMMUNICATIONS**

**JAPHET, Akintoye Samson**

Department of Linguistics and Nigerian Languages

Osun State University, Osogbo

[akintoye.japhet@uniosun.edu.ng](mailto:akintoye.japhet@uniosun.edu.ng) | [japhetakintoye@gmail.com](mailto:japhetakintoye@gmail.com)

Tel: 08059159778 | <https://orcid.org/0000-0003-3904-9103>

**AWONIYI, Folorunso Emmanuel**

Department of Linguistics and Nigerian Languages

Osun State University, Osogbo

[folorunso.awoniyi@uniosun.edu.ng](mailto:folorunso.awoniyi@uniosun.edu.ng) | Tel: 08035216039

<https://orcid.org/0000-0003-3595-8684>

&

**ADEYEMI Aderogba**

Department of Communication Studies

Osun State University, Osogbo

**Email:** [aderogba.adeyemi@uniosun.edu.ng](mailto:aderogba.adeyemi@uniosun.edu.ng) | Tel: 07069550894

**ORCID:** <https://orcid.org/0000-0002-0357-1891>

DOI : <https://doi.org/10.5281/zenodo.16281733>

### Abstract

This study is driven by the following questions: Do Yoruba born-digital datasets consistently observe tone orthography? What are the primary implications of neglecting tone-marking? How can tone orthography be effectively implemented in Yoruba born-digital texts? The study aims to examine the extent to which tone-marking is applied in born-digital Yoruba datasets, identify the key implications of tone-marking neglect, and explore practical strategies for ensuring the proper implementation of tone orthography in Yoruba digital texts. Utilizing the Yorùbá Web 2015 corpus (yorubawac15), data were sourced from websites such as *wikipedia.org*, *alaroyeonline.com*, *nairaland.com*, and others. A toneless search query, *aja*, was used via the Sketch Engine tool to generate a concordance, from which 50 lines were manually extracted and analyzed. The findings reveal that tone orthography is poorly implemented in the selected born-digital texts and underscore the linguistic and communicative consequences of this neglect, including ambiguity and loss of meaning. Two practical solutions are proposed: the adoption of predictive typing tools for general users and the use of Unicode keyboards for proficient writers. Due to the limited scope of the study, the analysis was confined to 50 concordance lines. This study ultimately advocates for the integration of tone orthography in Yoruba digital writing and emphasizes the importance of linguistic accuracy in the era of African digital humanities.

**Keywords:** African digital humanities, tone orthography, digital orthography, tone diacritics, born-digital data

Word count: 150

### Introduction

One of the main problems confronting digitized texts in Yorùbá is the tone-marking orthography. Tone marking is facing a growing neglect in Yorùbá textual data which include religious publications, Yorùbá print newspapers, magazines and students' writings (Fagborun, 1989; Odejobi, 2005; Olumuyiwa, 2013). This study is focusing on the extension of such tone marking neglect in digital archives of the language (Asahiah, 2014). There is a serious concern on the poor condition of the Yorùbá born-digital textual data currently available on the web (Japhet,

Afolabi & Olatuji, 2024). The study reveals a crucial problem that can hinder the on-going digitisation projects in the language.

Born-digital data are the data that are originally codified in digital forms. They differ from analogue data. Analogue data has the following characteristics. As analogue data, audio data exist as natural voices from speakers or singers; pictures data are captured on hard surfaces: stones, wood, walls and paper; moving image data are viewed as actions. These analogue data can later be digitised by converting them to the desired digital formats. Born-digital data, on the other hand, are not just digital by conversion. They are generated as digital products from the outset. They occur in soft copies produced by software. Sound data occur as audio files (e.g. .mp3, .wav, .acc, .midi, .wma); pictures and still images occur as image files (e.g. .jpg, .png, .tif, .bmp, .gif, .svg); moving images are created as video files (e.g. .mp4, .mov, .avi, .wmv); 3D objects are created in soft copies with the 3D perspective (e.g. .fbx, .obj, .3ds, .skp, .stl, .blend).

The study reveals a crucial problem that can hinder the on-going digitisation projects in the language. For a tone language like Yoruba which has a lot of minimal pairs and minimal sets where tones contrast, tone-marking problems can adversely affect comprehension. Where the context fails, digital texts often lack enough information to sustain comprehension where ambiguity surfaces from tone-marking neglects, a problem which humans may cope with in real life situations. It is therefore important to ensure tones are properly marked rather than depending on the context that is not digitally available. Due to the current rise in ICT, new language data on the web are usually born-digital. It is cheaper to type directly into word processor pages or web pages than to write on paper before scanning and digitising the script using OCR tools. This new drive for digital textual input is yet to be matched with the users competence in the orthography. The scope of this study is therefore limited to the role of tone in those Yoruba born-digital textual data.

### **Methodology**

The study was conducted using the Yorùbá corpus called Yorùbá Web 2015 also known as *yorubawac15*. The corpus was accessed using Sketchengine's search on Yorùbá Web 2015 from <https://auth.sketchengine.eu/> (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, Rychlý, & Suchomel, 2014). The data wrangling process in the study focuses on cleaning, structuring and preparing the raw unprocessed Yorùbá texts for proper tone-marking. The data were enriched with tone-marking to make them usable and meaningful.

The data used for the study were collected from 13 diverse websites. These are *wikipedia.org*, *gospelgo.com*, *alaroyeonline.com*, *nairaland.com*, *olayemioniroyin.com*,  
p g . 8 9

*slideshare.net, blogspot.cz, anti-el7ad.com, vegetable-gardens.biz, tripod.com, loveandtips.com, loveandtips.com, and themediaonline.co.za.* They are selected with the following conditions: first, they are digitised Yorùbá texts; second, they have not been properly edited for thorough tone-marking. The data cleaning operation was done manually by reading through the lines of concordance to remove unnecessary duplicates and making the data representational of its source. Data analysis was planned to uncover patterns, themes, and insights recoverable from within the Yorùbá textual data on the selected websites.

*Aja* was selected because it is a common Yoruba word that can display tonal minimal pairs. It is preferred to other sets like *igba* and *ila*, which have been widely used in Yoruba tone illustrations. If these ones were used, most of the data mined will be from language learning corpora, and they will be too narrow to demonstrate the problem the study wanted to solve.

The data wrangling process includes selecting the right term to investigate in the corpora through concordancing, cleaning the data through deduplication, structuring the data sources and contextual relations, and preparing the raw. data for analysis through proper tone-marking. Due to these processes, only 50 lines of the concordance were available for analysis in the study. Inconsistency in tone marking across Yorùbá websites can be problematic when such websites form part of a corpus.

## Data

The data in the study comprises several websites from which the toneless word: *aja* was put in concordance from the websites. The concordance is provided in table 1 below.

| N | Left  | wic | Right  |              |
|---|---|-----|--|--------------|
|   | Kebu, Animere.<br>Gbe :- Lábé èyí ni a ti rí: Ewe<br>ati Gen/ | ja  | , Fon-Phla-Phera BENUE<br>CONGO                  | Adja         |
|   | Suru, Yauri, Zuru.<br>KOGI Kogi Adavi,                        | ja  | okuta, Ankpa, bassa,<br>Dekina, Ibaji, Idah,     | Ajaoku<br>ta |
|   | lati ko lo si Abomey,<br>larin won ni a ti ri                 | ja  | , awon omo eya Gbe ti<br>awon akotan nigbagbo pe | Adja         |

|   |   |    |  |                 |
|---|---|----|--|-----------------|
|   | ni won se idasile ilu na. Ibasepo larin awon eya                    | ja | ati eya Fon to fa idasile eya titun ti a mo si "Dah  | Adja            |
|   | agbegbe Abomey ni Arin Gusu; ati Mina, Xueda, ati                   | ja | (ti won wa lati Togo) ni ekun odo. Fon ni  | Adja            |
|   | o n so ede Fon, Yoruba tele won pelu 1.2 legbegberun,               | ja | (600,000), Bariba (460,000), Ayizo (330,000), the  | Adja            |
|   | eya eniyan 99% je àwon omo Afrika (Yoruba, Fon,                     | ja | , Bariba at.bb.lo), awon omo Europe ko ju 10,000   | Adja            |
|   | to gajulo laye pelu igasoke 4,900 m, won n pe bi                    | ja | aye. Himalaya Àsiá ilẹ́ Índià je asia orile-edo  | loft            |
|   | ati Ebali, Sefo, ati Onamu. 24 Wọnyi si li awon omọ Sibeoni; ati    | ja | on Ana: eyi ni Ana ti o ri awon isun omi gbigbona ni ijù, bi o ti mbọ awon ketekete Sibeoni baba | Ajah Gen.36: 24 |
| 0 | o si se arewa enia. 43 Filistini na si wi fun Dafidi pe, Emi ha nse | ja | bi, ti iwọ fi mu opá tò mi wá? Filistini na si fi Dafidi re nipa awon olorun re.                 | dog             |
| 1 | re. 14 Nitori tani o ba Israeli fi jade? tani iwọ nlepa? Okú        | ja | , tabi esinshin? 15 Ki Oluwa ki o se onidajo, ki o si dajo larin emi ati                         | dog             |
| 2 | nitori orọ wọnyi ti Işboşeti sọ fun u, o si wipe, Emi ise ori       | ja | bi? emi ti mo mba Juda ja, ti mo si şanu loni fun idile Saulu baba re,                           | dog             |
| 3 | teriba, o si wipe, Kini iranşe re jasi, ti iwọ o fi ma wo okú       | ja | bi emi? 9 Oba si pe Siba iranşe Saulu, o si wi fun u pe, Gbogbo                                  | dog             |

|   |   |    |  |     |
|---|---|----|--|-----|
| 4 | won. 13 Hasaeli si wipe, Sugbon kinla? Iranşe rẹ iṣe          | ja | , ti yio fi ṣe nkan nla yi? Eliṣa si dahùn wipe, Oluwa ti fi hàn mi pe,                          | dog |
| 5 | omọ Nebati, ati bi ile Baasa omọ Ahijah; 10 Awon              | ja | yio si jẹ Jesebeli ni oko Jesreeli, ki yio si eniti yio sinkú rẹ. O si ṣi                        | dog |
| 6 | owọ Elijah iranşe rẹ ara Tiṣbi wipe, Ni oko Jesreeli li awon  | ja | yio jẹ eran-ara Jesebeli: 37 Okú Jesebeli yio si dàbi imí ni igbé, ni                            | dog |
| 7 | li ẹrẹkẹ; iwọ o mu mi dubulẹ ninu erupe ikú. 16 Nitoriti awon | ja | yi mi ka: ijo awon enia buburu ti ká mi mó: nwon lu mi li owọ, nwon si lu mi li ẹṣẹ. 17 Mo le ka | dog |
| 8 | lowo. 20 Gbà okàn mi lowo idà; eni mi kanna lowo agbara       | ja | nì. 21 Gbà mi kuro li enu kiniun ni; ki iwọ ki o si gbohùn mi lati ibi                           | dog |
| 9 | nwon ndubulẹ, nwon fẹ ma tōgbé. 11 Nitōto ojeun               | ja | ni nwon ti kì iyó, ati oluṣọ agutan ti kò moye ni nwon:  | dog |
| 0 | 23 Ki ẹṣẹ rẹ ki o le pón ninu ẹjẹ awon ota rẹ, ati ahon awon  | ja | rẹ ninu rẹ na. 24 Nwon ti ri irin rẹ, Olorun; ani irin Olorun mi, Oba                            | dog |
| 1 | aṣiwèrè si iṣẹ, ti o si gba awon olurekoja si iṣe-owo. 11 Bi  | ja | ti ipada sinu èbì rẹ, bẹ̀lì aṣiwèrè itun pada sinu wèrè rẹ.                                      | dog |
| 2 | fun awon eeyan re, sugbon kokoro kan wa nibe to ba eyin       | ja | je. ... Ija n bo! Awon gomina ko fee sanwo osu tuntun mo o                                       | dog |

|   |   |    |  |                    |
|---|---|----|--|--------------------|
| 3 | Okan ninu awon ohun<br>ti Yoruba n pe ni a-<br>teyinrogbon ree o. A-teyin-<br>rogbon, a ge eti      | ja | , o wa n loo fi obe pamo,<br>se iyen tun ran oro re ni. Jonathan<br>yoo tun fowo                                   | dog                |
| 4 | bi nnkan se n lo ninu<br>egbe oselu ACN ipinle Ondo<br>bayii, o see se ki ija ajadiju,<br>ija       | ja | pin yeleyele sele ninu egbe<br>naa laipe nitori bi opo awon omo<br>egbe naa se n mura lati di gomina<br>lodun      | à-jà:<br>nom-fight |
| 5 | Nibo leyin n gbe,<br>Nibo leyin wa O doluwooro<br>jin wooro, o dolu-wooroo jin<br>wooro, oku        | ja | ki i gbo, oku agbo ki i<br>kan... Bi Haruna isola se mo orin i<br>ko to, ti okiki re si gbale kankan,<br>kinni kan | dog                |
| 6 | ni kan to mo-on-mo<br>ran ara e lo sorun apapandodo<br>laduugbo Idi-Oro                             | ja | to ba fee sonu, ko ni i gbo<br>fere olode Mo ti soro yii leekan o,<br>sugbon n ko so o ni ekun-un-re               | dog                |
| 7 | Ile n jeeyan, awon eni<br>nla ti re koja lo Akintola lo si<br>Kano o bo,                            | ja | to rele ekun to bo ni, ba a<br>ba ki aja ku ewu, ka si ki ekun naa<br>pe o ku ewu                                  | dog                |
| 8 | awon to je olugbe<br>adugbo kan ti won n pe ni<br>Peace land, ... Read More<br>Alabi to ji          | ja | alaja gbe ateni to ta a fun<br>n'Iganmu ti dero ile-ejo Leyin to<br>gbe aja meta to je ti Ogbeni                   | dog                |
| 9 | a mi o, oko mi ti gba<br>ile ti mo fowo mi ra Ojukwu<br>n sepade, Obasanjo naa n<br>sepade, afi bii | ja | to ri ikooko to n gbo, ninu<br>aja pelu ikooko, eni kan naa lesu<br>yoo se Ni asiko ti awon omo-<br>ogun Biafra    | dog                |
| 0 | ti mi o mobi ti mo wa,<br>ti mi o si mo ohun ti mo n se<br>rara                                     | ja | to ba rele ekun to bo,<br>keeyan maa ki i ku ewu ni, bee<br>gege loro omo  | dog                |

|   |   |    |  |                    |
|---|---|----|--|--------------------|
| 1 | E mura si i daadaa<br>Eepa n pa ara re, o ni oun n<br>pa                                    | ja | . Awon ti won n pese<br>telifoonu fun gbog   | Dog                |
| 2 | ko e l'obe je tan So<br>fun bobo Ifeleke yen pe ere ti                                      | ja | ba f'ogun odun sa, ko ni<br>gba esin ni ojo kan.   | Dog                |
| 3 | padanu eto re ni ile<br>Nairaland. Mo lero wipe bi  | ja | ba n siwin o ye ki o mo oju<br>olowo re o. Abata loro eleyi  | Dog                |
| 4 | Ma se be mo oooo, is<br>not good at all.  | ja | la jenu iwo omugo yi,kete<br>kete le si ri,kata kata nbo lona,                                       | Dog                |
| 5 | Sebi awon agba loni<br>ilu ki i wa laini oba alade.<br>Leyin ti Oba Adebiyi Adesida<br>goke | ja | , awon afobaje ilu Akure ti<br>pe birikoto lati fi Omooba 'birin<br>Adetutu Adesida Ojei to je beere | Loft               |
| 6 | Sebi awon Yoruba<br>loni, are-maja kan ko si,   | ja | -ma-pari-e lo buru jojo.<br>Bi o tile je wi pe iyapa D'banj ati<br>Don                               | a-jà:<br>nom-fight |
| 7 | si lule eleyii to mu ki<br>koto ti won wa ni inu re ja<br>sinu                              | ja | ile sohun raurau. Kelly<br>Hansome ti gbe orin jade lati fi  | cavity,<br>hole    |
| 8 | ko le gbagbo wipe<br>Jesu le sami awon Keferi bi  | ja | ati elode ati ki o si ma ba<br>iya re soro pe Obirin yi,   | dog                |
| 9 | olori- buruku.” Matteu<br>7:6: “E mase fi ohun mimo<br>fun                                  | ja | , ki e ma si se so peril nyin<br>siwaju elode...”  | dog                |
| 0 | ohun ti a maa n<br>bukaata sinu re, tori<br>naasanma ni                                     | ja | re ti a gbega lori re, ati pe<br>ile ni ibusun atiite ati  | loft               |



|   |   |    |  |                                    |
|---|---|----|--|------------------------------------|
| 1 | ee lo, a gbo pe leyin ti<br>o pa Lekham tan o gun oke o<br>si wo inu    | ja | ile lo. Ibeere niyi: Ti o ba<br>je pe moleka / angeeli lo pa<br>Lekham bi Ahmadi se fe ki a<br>gbagbo,     | depth                              |
| 2 | oro etan tikarare. A<br>gba itoni emi ati iru<br>omoleyin, ti o buru ju | ja | ati igbe aye aimo awon<br>maa so asotele nile, ati wipe pelu<br>owo ara won ati pelu ogbon<br>alumo-koroyi | dog                                |
| 3 | itijú bo Àjàpá.   | JA | ILE NI ABEOKUTA<br>Olorun ti se tan lati gbe ija awa<br>omo naijiria                                       | cavity                             |
| 4 | mu ki o baa le ya oorun<br>ni ile rẹ. Inu                               | ja | naa dun lati kuro ninu iji.<br>Ni alẹ ojọ naa, ile gba ina.  | dog                                |
| 5 | mimu iderun ba awon<br>omo ile Naijiria 2. gbigbe<br>ogun               | ja | aiwehin ti iwa ibaje 3. sise<br>ajinde igbagbo   | dog                                |
| 6 | mehe pupo ni agbegbe<br>wa nibi Ogbontagi                               | ja | -fun-eto omoniyan kan,<br>Ogbeni (Komureedi) Mashood<br>Erubami  | a-jà:<br>NOM-fighter<br>“activist” |
| 7 | Ni ibere lati ọririn ara<br>jẹ ti o ni inira Oníwúrà,<br>ewúre, awon    | ja | di aso ti awo-ara, o dara<br>fun awon manufacture ti yangan<br>alawo                                       | dog                                |
| 8 | Ninu awon etutu ti<br>won ma nfi sami si odun naa<br>ni ki won pa       | ja | ati Agbo. Odokunrin ati<br>odobinrin ti won tagun yoo da eje<br>won si ara, leyin naa won yoo wa<br>fi     | dog                                |
| 9 | Boya wa ti jẹ ẹya<br>atijo-asa yinyin ipara firisa<br>ninu rẹ oke       | ja | . O le jẹ Fun lati gba diẹ<br>ninu awon yinyi  | loft                               |

|   |   |    |   |     |
|---|---|----|---|-----|
| 0 | o ba ohun ini nipasẹ re<br>ebi . 201 . O ti je nipasẹ<br>awọn | ja | , jackals , wolves ati aran,<br>Crows ati hawks ju je | dog |
|---|---|----|---|-----|

Table 1. “Aja” in yorubawac15 corpus concordance

### Discussion of the analysis

Many of the born-digital Yorùbá textual data are rendered in forms that violate the standard orthography, especially in tone-marking. In table 2, Olumuyiwa (2013:47) cites how Abẹ̀òkúta community web page violated tone-marking amidst other orthography violations.

| Web<br>form   | Clean<br>form | gloss           |
|---------------|---------------|-----------------|
| E<br>kaaro-ro | ẹ<br>káàárọ̀  | good<br>morning |
| Okuuri<br>n   | ọkùnrí<br>n   | male            |
| Obirin        | obìnrin       | female          |
| Bee ni        | bẹ́ ẹ̀ ní     | yes             |
| Odaa          | ó dáa         | it is good      |
| Ejo           | ẹ jọ ọ́       | please          |
| Se e ni?      | şé ẹ ni?      | do you<br>have? |
| E se          | Ẹ şé          | thank you       |
| Ta-ni?        | Ta ni?        | Who is it       |

|        |       |            |
|--------|-------|------------|
| Kí-ni? | Kí ni | what is it |
|--------|-------|------------|

Table 2. Orthography cleaning (adapted from Olumuyiwa, 2013:47)

The concordance in table 1 is compressed to align with the source website in Figure 1. There are 13 sources. The frequency of the keyword-in-context is placed after each of the websites. *wikipedia.org* has 8 mentions (concordance lines: 1-8) The highest occurrence of the KWIC is in *gospelgo.com* with 13 mentions (concordance lines: 22-31). At *alaroyeonline.com*, there are 10 mentions (concordance lines: 9-21). *Nairaland.com*, *olayemioniroyin.com* and *slideshare.net* have 3 mentions each. *Blogspot.cz*, *anti-el7ad.com* and *vegetable-gardens.biz* have 3 mentions each. *tripod.com*, *loveandtips.com*, *loveandtips.com*, and *thediaonline.co.za* have a mention each.

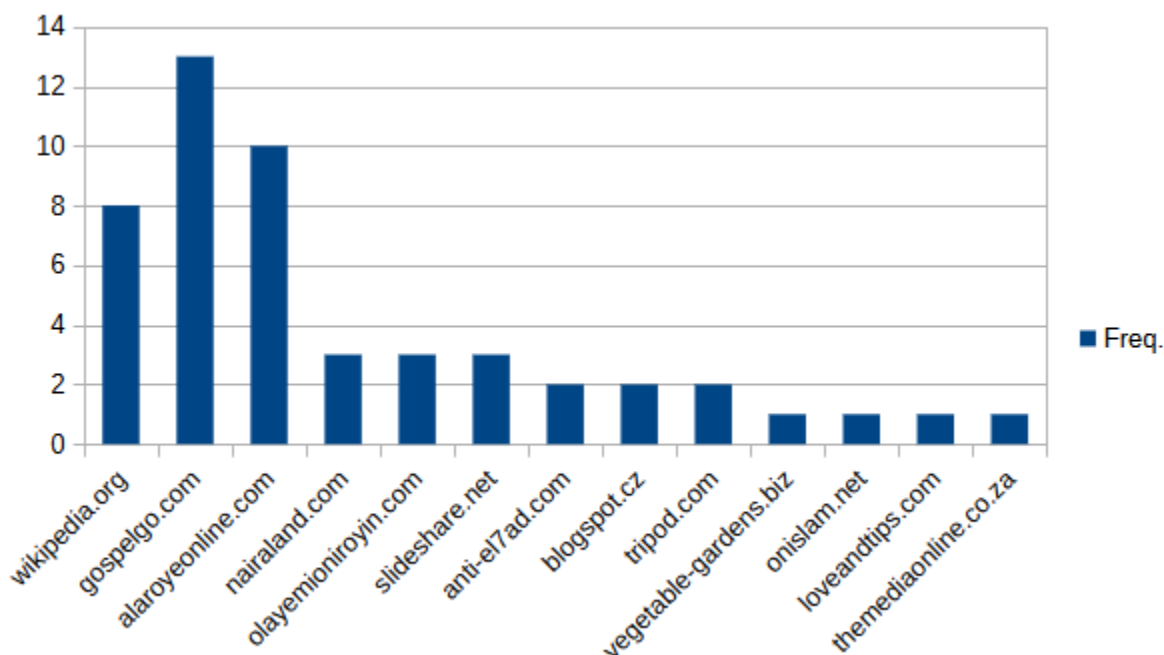


Figure 1. Distribution of “aja” per listed website

The concordance entries are categorised by their sources respectively in Figure 1. The first group comprises data from *Wikipedia* documents. The second group is from a religious website with the domain name: *gospelgo.com*. The third group is the online edition of the Yorùbá newspaper, *Alaroye*, available at *alaroyeonline.com*. Data from *Nairaland*, a very popular Nigerian social website, comes next. In short, the grouping follows the order in which the sources are listed above.

This study used *aja* as the *key word in context* (kwic) to construct the concordance. *Aja* output of the 50-line concordance has the following outcomes: *ajá* “dog” (31 times); *àjà* “loft/cavity” (9 times); *Adja* “a people” (6 times); *A-jà* a- NOM+VP (4 times); *Ajah* “name in the bible” (1 time); and *Aja-* as part of *Ajaokuta* (1 time).

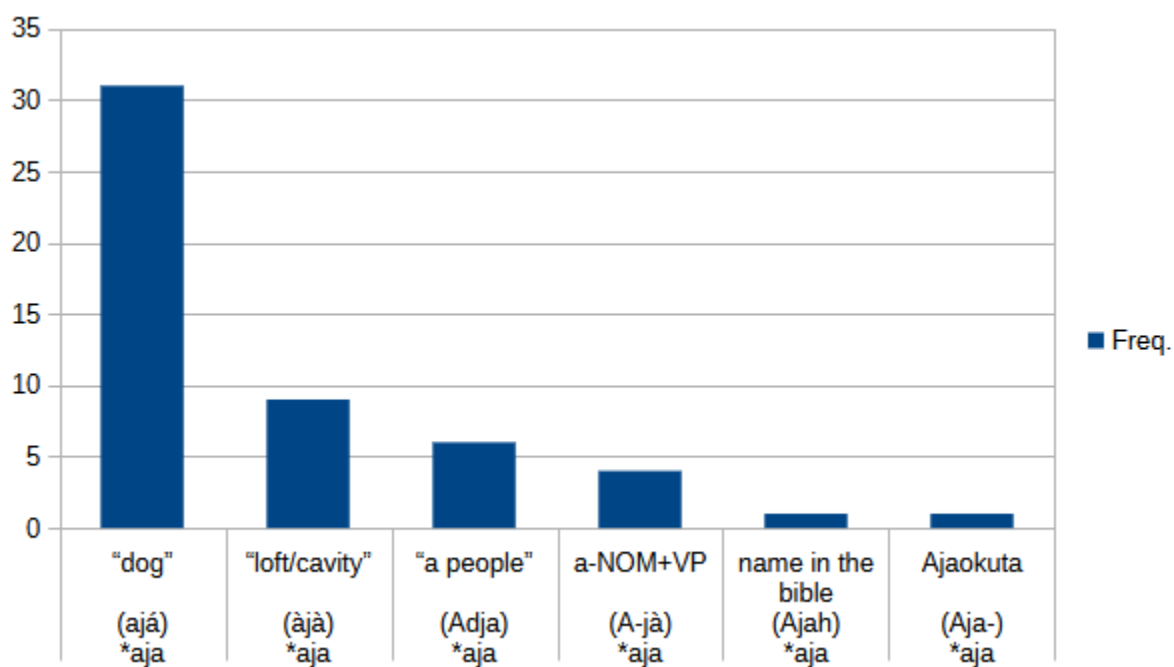


Figure 2. Presentation of the data cleaning on “aja” \*raw data; ( ) cleaned data.

With the tone clearly marked, the words have their tags interpretable even without formal lemmatisation. The most frequent item is *ajà* “dog”; it occurs 31 times. Next to this is *àjà* “loft/roof/hole” which occurs 9 times. There are two cases where *aja* came from non-Nigerian languages as in Figure 2. The first is *Adja* written as *Aja* in the Yorùbá textual source where the crawler mined it. The word represents an ethnic group in Benin and Togo Republics. The second one is *Ajah*, also written as *Aja* in its source document, the Yorùbá bible. This is a Hebrew name found in Yorùbá version of the bible in Genesis 36: 24. Both words are pronounced with their first syllables stressed. This could have been rendered in Yorùbá with a high tone on the stressed syllable as reproduced in Table 3.

| N | raw form | clean form from the source | Clean form for Yorùbá | gloss             |
|---|----------|----------------------------|-----------------------|-------------------|
|   | aja      | Ajah                       | Ájà                   | “a people”        |
|   | aja      | Aja                        | Ájà                   | name in the bible |

Table 3. Presentation of the data cleaning on “aja” for “Adja” and “Ajah”

Having the data in Table 3 bearing the high tone on the first syllable clearly shows that those items are not Yorùbá words. The clue provided through the high tone mark is enough to show they are borrowed.

Another issue of ambiguity caused by poor tone marking results in the possibly mistaking the bound morpheme, *A-jà*, for the free morpheme, *aja*. *A-jà* forms part of a verbal noun which is derived by combining the nominalising prefix *a-* with the verb *jà* “fight”. The three cases are analysed in Table 4.

| N | OM<br>prefix | STEM with clean<br>Yorùbá data                 | Gloss   |
|---|--------------|--|---|
|   | -            | jà pín yéḗyèḗ<br>fight split scatter-about     | “a fight that breaks into several<br>factions and fronts” |
|   | -            | jà-fún-ètó ọmọniyàn<br>fight-for-right-human’s | “human right activist”                                    |
|   | -            | jà-má-parí-ẹ<br>fight-NOT-finish-it            | “a monger of unendless strife”                            |

*Table 4. Presentation of the data cleaning on aja for “àjà pín yéḗyèḗ”, “ajà-fún-ètó ọmọniyàn” and “ajà-má-parí-ẹ”.*

In Table 4, there are two types of nominalisation prefixes. The first one has a low-toned prefix (marked with a grave accent). This prefix carries thematic imports. It simply assigns a name to the action of the verb (See the first item in Table 4). The second and the third item in Table 4 have the mid-toned agentive nominalisation prefix. Mid-tone is unmarked in Yorùbá texts. The tone mark information on the nominalisation prefixes is part of grammar. Disregarding such vital rules of the grammar will make digitised Yorùbá texts difficult for accurate digitisation projects.

### **Discussion of the findings**

The study has revealed the importance of tone marking in digitised Yorùbá texts. Due to poor use of tone diacritics on large data sets of digitised Yorùbá texts, it has become difficult to have good results with properly tone-marked Yorùbá texts while using search engines on the web (Asubiaro, 2014). The search engines are simply confused when the tone-marked search words do not match the desired results because of their incompatibility in form due to unmatching tone diacritics.

Rather than dwelling on the problem caused by neglect of tone marks on Yorùbá digitised text, more efforts have been placed on the solution to restore tone in the language (Asahiah, 2014;

Asahiah, Odejobi, and Adagunodo, 2017). This study provides two main patterns for general users of the language in providing born-digital textual data that will contribute to the ongoing digitisation of the language. The first option is the use of unicode Yorùbá keyboard. The second option is the use of predictive tone-marking input tools. The predictive input tools differ from the typical input keyboards, because it offers language users the opportunity to make the right choice from a list of options by merely clicking on it.

*Yorùbá unicode keyboard applications:* There are several keyboards software that can be installed on the personal computer which can be used to type Yorùbá tones. Ajao, Babatunde, Asapetu and Yusuf's (2020) Yorùbá Character Keyboard proposes a combination of ASCII and Unicode to generate unicode for Yorùbá characters with tone marks and subdot diacritics. See screenshot in Figure 3.

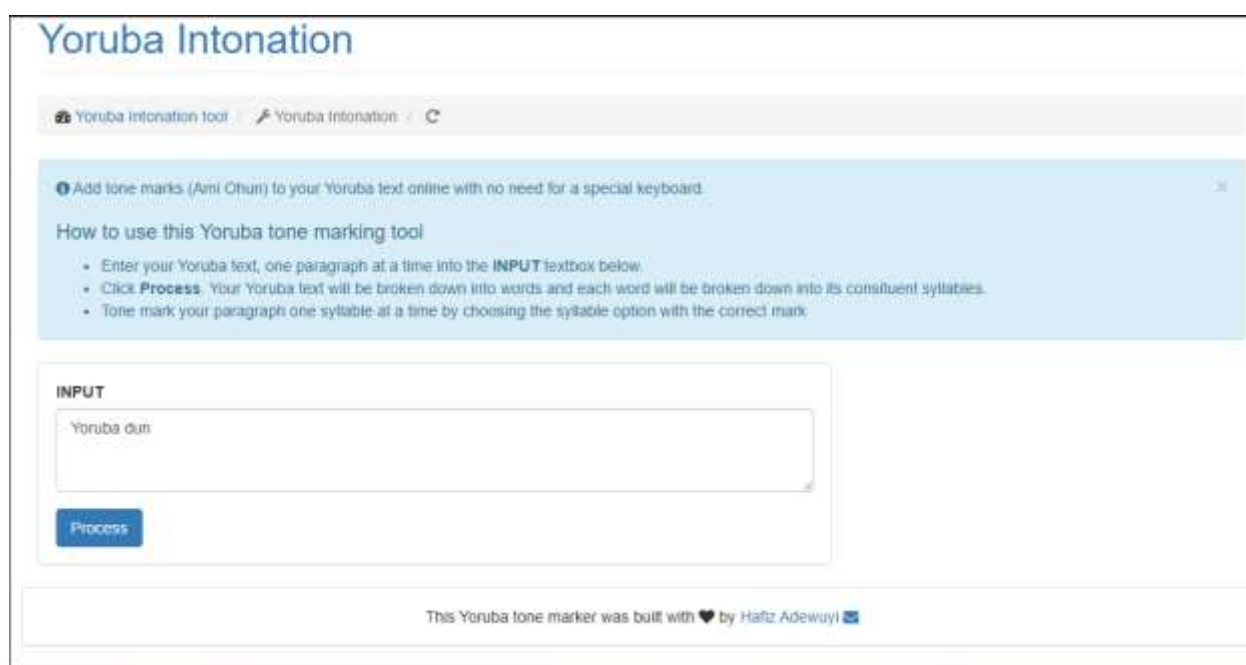


*Figure 3. Tone-marking approach in Yorùbá character keyboard by reading tone mark and letter as a single character on the keyboard*

The goal of this Yorùbá Character Keyboard is to generate a single code point for Yorùbá characters instead of combining two or more keys on the keyboard to type a single character. The

keyboard is so detailed that users can type the high-toned vowel *á* and the low-toned vowel *à* from different keys labeled after them. A combination of different keys is not required. Availability and the use of input tools like this will reduce tone-marking problem in Yorùbá born-digital textual data sets.

*Predictive applications:* Some predictive applications are also available. Users just type the words in plain Latin letters, then some tone-marked options will appear from which the intended word can be selected (Adewuyi, 2017; Adebara and Adelani, 2022).



**Yoruba Intonation**

Yoruba Intonation tool / Yoruba Intonation

Add tone marks (Ami Oṣun) to your Yoruba text online with no need for a special keyboard.

**How to use this Yoruba tone marking tool**

- Enter your Yoruba text, one paragraph at a time into the **INPUT** textbox below.
- Click **Process**. Your Yoruba text will be broken down into words and each word will be broken down into its constituent syllables.
- Tone mark your paragraph one syllable at a time by choosing the syllable option with the correct mark.

**INPUT**

Yoruba dun

Process

This Yoruba tone marker was built with ❤️ by Hafiz Adewuyi

*Figure 4. Typing in the toneless Yorùbá data “Yoruba dun” into the input box.*





*Figure 5. Tone-marking the first syllable “yo” by clicking on the right option “Yo”.*



*Figure 6. Tone-marking the second syllable “ru” by clicking on the right option “ru”.*



*Figure 7. Tone-marking the third syllable “ba” by clicking on the right option “ba”*



*Figure 8. Tone-marking the second word “dun” by clicking on the right option “dùn”.*

Adebara and Adelani’s (2022) tone-marking software provides a word-length tone-marking application in similar way choosing from available options provided for the writer.

## Conclusion

While improper tone-marking is generally seen as abnormal in Yorùbá orthography, its use in digital texts has a greater consequence. It can confuse digital textual processing tools in their analyses as it does in the use of *Sketch Engine's* corpora and concordance compilation in the current study. Erroneous judgments coming from tone-marking can lead to false outputs in various analyses making use of such digital tools. Improper tone-marking can also hinder the ongoing, as well as future, digitisation projects in Yorùbá. Thus, proper tone-marking is crucial to digital humanities projects in Yorùbá. awareness on the proper use of Yoruba tones in the digital space should therefore be encouraged. Software developers should carry along Yorùbá digital content creators on their current developments of various input software that can aid their efficiency. Linguists should also be considering necessary revisions in the tone orthography that will ensure born-digital data sets are properly codified to reflect the norm in the orthography. Automation of predictive typing of tone marks should be a way to go; this will reduce the time spent in typing Yorùbá texts and the mental load facing textual input of Yorùbá born-digital data.

## References

- Adebara, I. & Adelani, D.I. (2022, April 20). Digital Indaba: Yorùbá NLP and machine learning University of Kansas IDRH <https://africandh.ku.edu/yoruba-nlp-and-machine-learning> and [https://www.youtube.com/watch?v=64N8b\\_1VEKo&t=14s](https://www.youtube.com/watch?v=64N8b_1VEKo&t=14s)
- Adewuyi, H. (2017a). The Yorùbá tone marking web application <https://www.hafiz.com.ng/2017/02/the-yoruba-tone-marking-web-application.html>
- Adewuyi, H. (2017b). Yorùbá tone marker <http://yorubaintonation.somee.com/>
- Ajao, J. F., Babatunde, R.S., Asapetu, S.O., Yusuf, S.R. (2020). Implementation of Yorùbá unicode generation for an indigenous keyboard *Technoscience Journal for Community Development in Africa* 1:(1), 71–80.
- Asahiah F. O. (2014). Development of a standard Yorùbá text automatic diacritic restoration System. Phd thesis, Obafemi Awolowo University, Ile-Ife, Nigeria,.



Asahiah, F. O., Odejobi, O. A., and Adagunodo, E. R. (2017). Restoring tone-marks in standard yorùbá electronic text: improved model. *Computer Science*, 18.

Asubiaro, T. V. (2014). Effects of diacritics on web search engines' performance for retrieval of Yorùbá documents. *Journal of Library & Information Studies*, 12(1).

Fagborun, J. (1989). Disparities in tonal and vowel representation: some practical problems in Yorùbá orthography. *Journal of West African Languages* 19(2):74-92

Japhet, A.S. Afolabi, I.O. & Olatuji, R.O. (2024). Ìjáfáfá Ògbúfò Gúgù lédè Yorùbá *Journal Yorùbá: Journal of Yorùbá Studies Association of Nigeria*, 13(1), 85-94.

Japhet, A. S. (2024, October 14-18). Digitizing Yorùbá for web communication. [Paper presentation]. 35th Conference of Linguistic Association of Nigeria, Al-Qalam University, Katsina, Nigeria.

Kilgarrieff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36. Link: <https://www.sketchengine.eu/bibliography-of-sketch-engine/>

Olatunbosun. Y. This Day 1, Nov 2020. Digital revolution as hope for preserving arts, cultural heritage  
<https://www.thisdaylive.com/index.php/2020/11/01/digital-revolution-as-hope-for-preserving-arts-cultural-heritage/#>

Olumuyiwa, T (2013). Yorùbá writing: standards and trends. *Journal of Arts and Humanities*, 2, 40-51.

Odejobi, O.A. (2005). A Computational Model of Prosody for Yorùbá Text-to-Speech Synthesis. Phd thesis, Aston University, Aston.