Developing an Open-Source Corpus of Yoruba Speech

Alexander Gutkin[1], *Işın Demirşahin*[1], *Oddur Kjartansson*[1], *Clara Rivera*[1], *Kọ́lá Túbọ̀sún*[2]

[1]Google Research, London, United Kingdom
[2]British Library, London, United Kingdom
{agutkin,isin,oddur,rivera}@google.com

Abstract

This paper introduces an open-source speech dataset for Yoruba – one of the largest low-resource West African languages spoken by at least 22 million people. Yoruba is one of the official languages of Nigeria, Benin and Togo, and is spoken in other neighboring African countries and beyond. The corpus consists of over four hours of 48 kHz recordings from 36 male and female volunteers and the corresponding transcriptions that include disfluency annotation. The transcriptions have full diacritization, which is vital for pronunciation and lexical disambiguation. The annotated speech dataset described in this paper is primarily intended for use in text-to-speech systems, serve as adaptation data in automatic speech recognition and speech-to-speech translation, and provide insights in West African corpus linguistics. We demonstrate the use of this corpus in a simple statistical parametric speech synthesis (SPSS) scenario evaluating it against the related languages from the CMU Wilderness dataset and the Yoruba Lagos-NWU corpus.

Index Terms: speech corpora, open-source, West Africa

1. Introduction

The Yoruba language, together with the other 11 languages from the Ede language group, belong to the Yoruboid subbranch of the Benue-Congo branch of a Niger-Congo family [1]. The geographic continuum where Yoruba is commonly spoken (often referred to as *Yorubaland* [2]) stretches from the southeastern part of Nigeria into Benin and Togo [3] across colonially determined boundaries [4]. Yoruba is the official language of these three countries and is one of the few West African languages considered to be regional lingua franca [5]. The estimates of the Yoruba speaker population vary in the literature. According to the conservative Ethnologue estimates [6], the total population of Yoruba native (L1) speakers is above 20 million, with further 2 million people speaking it as a second (L2) language. The speaker population size is likely to be significantly higher due to population growth.

Despite being one of the largest Nigerian languages second only to Hausa [6], Yoruba, together with the majority of smaller languages of Nigeria [7], is considered by many to be endangered due to multiple confounding factors, such as dominance of English (and its perceived prestige) in education and administration, economic and religious policies [8]. This situation is reflected in the relative scarcity of resources for Yoruba noted by language and speech technologists and researchers [9]. Despite this, there is a growing awareness to the importance of Yoruba reflected in the rising number of applications in several areas of natural language processing, such as machine translation [10] and morphological analysis [11].

Although there have been multiple efforts to bridge the gap between Yoruba and higher-resource languages for both automatic speech recognition (ASR) [12] and text-to-speech (TTS) [13], the situation with speech applications is more complex because obtaining the speech corpora poses a different set of challenges than collecting text, primarily because building high-quality speech resources is expensive. More often than not, it is time consuming to set up the recording logistics, collect and analyze the data, and procure linguistic expertise and additional corpora for constructing further components for Yoruba speech application, such as diacritization [14] and grapheme-to-phoneme (G2P) conversion [15]. Moreover, for applications such as text-to-speech, further complications arise as one needs to find an adequate voice talent and a recording studio.

Our work relies on several methods to mitigate some of these data collection issues in a low-resource language setting, which served us well in the past [16], such as recording smaller amounts of high-quality audio from multiple volunteer speakers using affordable recording equipment. The main contribution of this work is the Yoruba speech corpus that is free for commercial, academic and personal use, and is of sufficiently high-quality to be used in state-of-the-art speech applications and corpus linguistics. To the best of our knowledge, based on the review of existing work provided next, our dataset is the second Yoruba speech resource freely available online.

2. Background

Related corpora. While the situation with the availability of Yoruba text corpora [17], as well as with text-based NLP models [18], has been improving in recent years, the Yoruba speech resources are still very scarce. The few corpora used for developing Yoruba speech applications have predominantly been developed in-house: In [19], the authors describe a single-speaker unit selection database and the corresponding application to TTS, but provide no details about availability of their corpora. In similar applications to TTS, such as intonation modeling or system overviews, such as [20], the corpora are also not available in public domain. This situation is mirrored in ASR, where lack of good quality Yoruba resources has been noted [21].

We are aware of only one open-source Yoruba speech corpus available online – the Lagos-NWU Yoruba Speech Corpus collected by Lagos University (Nigeria) and North West-University (Vanderbijlpark, South Africa) [22]. The corpus is distributed under Creative Commons (Attribution 2.5 South Africa) public license by the South African Centre for Digital Language Resources (SADiLaR), which is a national center supported by the Department of Science and Innovation (DSI). The corpus consists of 4,316 recordings (2.75 hours in total) and the corresponding transcriptions from 17 male and 16 female speakers recorded at 16 kHz as 16-bit PCM RIFF. The corpus is primarily intended for the fundamental frequency analyses in the study of Yoruba tone realization for TTS [23], but has also been used for other applications, such as gender recognition [24]. Compared to the Lagos-NWU dataset, the corpus presented in this work is of higher recording quality (recorded at 48 kHz) and is also

Table 1: *The SY phonemes and corresponding graphemes.*

Consonants						Vowels			
P	G	P	G	P	G	P	G	P	G
/b/	b	/k͡p/	p	/t/	t	/a/	a	/ɛ/	ẹ
/d/	d	/l/	l	/w/	w	/e/	e	/ĩ/	in
/d͡ʒ/	j	/m/	m	/g/	g	/i/	i	/ũ/	un
/h/	h	/n/	n	/g͡b/	gb	/o/	o	/ɔ̃/	ọn
/j/	y	/r/	r	/ʃ/	ṣ	/u/	u	/ɛ̃/	ẹn
/k/	k	/s/	s	/f/	f	/ɔ/	ọ		

bigger (just over four hours).

A very low-resource relatives of Yoruba from the Ede language cluster, the Ifè and Okpela languages [25], are present in the open-source CMU Wilderness Multilingual Speech Dataset containing force-aligned text and audio of Bible translations for over 700 languages [26]. The dataset currently does not include Yoruba. The original data is mined from the New Testament website (`http://bible.is`), which includes Yoruba, and it is very likely that Yoruba will be included in the CMU Wilderness Dataset in the future.

Language overview. Similar to other languages, Yoruba has many dialects exhibiting considerable linguistic variation. The speakers of diverse dialects communicate using Standard Yoruba (SY), a literary form of Yoruba used in Nigeria in education, media and governance [15], which derives from the first attempts at standardization in the 19$^{\text{th}}$ century [27].

The SY writing system consists of lower and upper-case character pairs representing 18 consonants (Bb, Dd, Ff, Gg, GBgb, Hh, Jj, Kk, Ll, Mm, Nn, Pp, Rr, Ss, Ṣṣ, Tt, Ww, Yy) and 7 vowels (Aa, Ee, Ẹẹ, Ii, Oo, Ọọ, Uu), where gb (GB) is a digraph [28]. An older orthographic convention (e.g., ạ or ṛ), which is easily normalized to modern standards, can occasionally be encountered as well [29]. The SY phoneme inventory consists of 18 consonants and seven oral vowels mapping one-to-one to the orthographic representation mentioned above. In addition, it has four nasal vowels represented in writing as ẹn, in, ọn and un [30]. The SY phonemes (P), based on the inventory in [31], and their grapheme correspondences (G) are shown in Table 1. The potential fifth nasal vowel sound [ã] (represented in writing as an) is often treated as an allophone of /ɔ̃/ [28].

Similar to many other languages from the Niger-Congo family [32], Yoruba is a tonal language distinguishing between three levels of tones LOW (/ˋ/), HIGH (/ˊ/) and MID (/¯/) on oral and nasal vowels, where the first two tones are marked in the orthography by the "̀" and "́" diacritics, while the third tone is unmarked by default. The absence of tone marks in Yoruba writing can become highly problematic for pronunciation models as tone is used to convey the differences in lexical disambiguation, pronunciation and prosody [33]. For example, the set of minimal pairs ọwọ́ ("hand"), ọ̀wọ̀ ("honour") and ọ̀wọ́ ("group") are reduced to a homograph ọwọ [34]. Several approaches to automatic diacritization have been recently proposed to tackle this problem, which still remains a non-trivial challenge [35].

Another potential complication in the processing of Yoruba text may arise due to the occasional use of apostrophe character that marks orthographic contractions that may correspond to various phonemic processes, such as vowel elision, obvious to the reader. For example, "je'ṣu" ("eat yam") expands to "jẹ iṣu" and "M'ọ́wọ́ wá" ("bring hand come/bring your hand") expands to "Mú ọwọ́ wá". The resolution of orthographic contractions is usually not covered by the existing G2P approaches [15] and needs to be implemented as a separate stage in the text normalization instead.

Table 2: *Corpus identifiers and the hosting URL.*

ISLRN	`464-321-727-765-2`
OpenSLR ID	`SLR86`
URL Link	`http://openslr.org/86/`

Table 3: *Recording script lines and the audio properties.*

Gender	Lines	Tokens					Speakers	Audio Duration	
		min	max	avg	Total	Unique		Total [h:m:s]	avg[s]
Female	1,892	2	23	8.5	15,880	4,113	19	2:03:21	3.9
Male	1,691	3	22	8.4	14,242	3,835	17	1:58:10	4.2
Total	**3,583**	–	–	–	**30,122**	–	**36**	**4:01:31**	–

3. Corpus details

Distribution and licensing. The corpus is released under Creative Commons Attribution-ShareAlike 4.0 International (`CC BY-SA 4.0`) license [36] and is made available for download from the Open Speech and Language Resources repository (OpenSLR) [37] as shown in Table 2 along with the International Standard Language Resource Number (ISLRNs) [38] and the OpenSLR Speech and Language Resource (SLR) identifier. The ISLRN is a 13-digit number that uniquely identifies the corpus and serves as official identification schema endorsed by several organizations, such as European Language Resources Association (ELRA) and Linguistic Data Consortium (LDC).

Collections of audio and the corresponding transcriptions are stored in a separate compressed archive (in `zip` format) for each gender. Transcriptions are stored in a line index files, one for each gender, which contain a tab-separated list of pairs consisting of the audio file names and the corresponding hand-curated transcriptions. The name of each utterance consists of three parts: the three-letter symbolic dataset name (`yof` for female and `yom` for male), the five-digit speaker ID and the 11-digit hash, separated by underscores. The 48 kHz single-channel audio files are provided in 16 bit linear PCM RIFF format.

Recording process. Instead of renting an expensive professional recording studio, the recordings took place in two different office locations in Lagos, Nigeria: the African Arts Foundation (AAF) in Victoria Island and the office of LifeBank Nigeria in Yaba. Neither location was unechoic and sound-isolated, but considerable effort went into reducing the echoes and the environmental noise (e.g., by recording at the weekends). The male and female participants were recruited through a Twitter feed and were all native SY speakers. The ages of participants, obtained through the voluntary form they all completed, range from 21 to 43. Participants were asked to read the script prepared by us in advance from various public domain texts, such as news, Wikipedia, religious stories and folk tales.

A regular Chromebook laptop was used for the recordings. It was not fanless but because it was placed far enough from the microphone, it did not introduce many fan-related recording artifacts. The audio was recorded using a Neuman KM184 diaphragm condenser cardioid microphone, a Blue ICICLE XLR to USB analogue to digital (A/D) converter, which also provides power to the microphone. The laptop stand was a repurposed musical sheet stand. All the recordings went through a quality control process performed by a native speaker and problematic lines that could not be re-recorded were dropped.

Some corpora properties. Even though transcriptions mostly contain sequences of natural language words, because they have not been text normalized they also contain punctuation symbols, such as commas and quotation marks. Therefore, here and be-

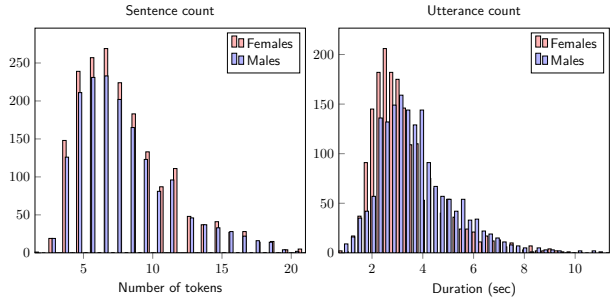

Figure 1: *Sentence lengths (left) and recording durations (right) for the females (red) and males (blue).*

Table 4: *Statistics for Yoruba tones in the corpus.*

Phoneme	Females			Males		
	LOW	MID	HIGH	LOW	MID	HIGH
/a/	2,493	2,029	2,204	2,237	1,825	1,954
/e/	547	613	1,078	491	557	945
/i/	1,384	2,373	2,597	1,268	2,123	2,319
/ĩ/	1	191	–	1	176	–
/o/	845	1,432	1,162	787	1,272	1,051
/u/	520	187	1,126	481	181	1,019
/ũ/	–	282	1	–	254	1
/ɔ/	770	1,310	859	700	1,178	758
/ɔ̃/	–	331	–	–	299	–
/ɛ/	728	605	861	663	526	780
/ɛ̃/	–	112	–	–	102	–
Total	7,288	9,465	9,888	6,628	8,493	8,827

low we refer to the constituent space-separated elements of transcriptions as "tokens" rather than words. The total number of recording script lines, the minimum, maximum, average and total (and unique) number of tokens per sentence for each gender are shown in the first six columns of Table 3. The corresponding distributions of sentence lengths per gender (in number of tokens) are represented by the histograms on the left-hand side of Figure 1. As can be seen from the figure, the shapes and the modes of the sentence length distributions for male and female datasets are very similar, even though the script content varies across genders. The last three columns of Table 3 show various properties of the recorded audio that include the total number of speakers, the total duration of each dataset and the average duration of individual utterance for each gender. The corresponding duration distributions (measured in seconds) for each gender are shown on the right-hand side of Figure 1. Due to the variety of speaking styles the distributions are different, with the male portion of the dataset containing longer utterances on average than its female counterpart.

Disfluency annotation. The recordings contain annotations in the script, which include disfluencies such as hesitations or abrupt stops. The additional annotations are denoted by square brackets. In total, the recording scripts contain 1,306 annotations, 682 for the female recordings and 625 for the male recordings. The annotations can be located at the start, the end or anywhere in the line. Some lines contain more than one annotation. Overall there are five types of annotations: the audio stops abruptly at the end of the recording (`[abrupt]`), audible breathing (`[breath]`), external noise (`[external]`), hesitation (`[hesitation]`) and a minor snap of the tongue or lip movement (`[snap]`).

Phoneme distributions. In order to compute phoneme coverage statistics for our Yoruba corpus we applied Epitran, a G2P conversion system that supports over 69 languages including Yoruba [39]. The phoneme inventory used by Epitran is identical to the phoneme inventory in Table 1. Prior to applying Epitran a very basic text normalization was performed that consisted of removal of punctuation tokens, point-fixing of the old orthography (e.g., "ẹ̀wọ̀n" → "ẹ̀wọ̀n") and the foreign word spellings (e.g., "city" → "siti"). The resulting phoneme distributions for the female (red) and male (blue) portions of our dataset are shown on the left histogram of Figure 2.

Table 5: *Objective evaluation results.*

Dataset			GR-CG		CG		RF-CG	
Name	Gen.	Lang.	MCD	F0	MCD	F0	MCD	F0
$\texttt{yof}_{16}$	♀	yor	6.45	20.21	6.21	18.70	**6.10**	17.94
$\texttt{yom}_{16}$	♂	yor	6.36	14.28	6.31	14.11	**6.19**	13.10
$\texttt{yo\{m,f\}}_{16}$	♀+♂	yor	6.52	20.80	6.39	23.66	**6.26**	22.79
$\texttt{yof}_{48}$	♀	yor	6.03	27.52	5.97	25.97	**5.77**	17.58
$\texttt{yom}_{48}$	♂	yor	6.01	14.00	5.96	13.95	**5.78**	13.38
CMU	♂	ife	7.15	–	–	–	6.80	–
CMU	♂	atg	5.55	–	–	–	5.39	–

In order to compare the Yoruba phoneme distribution to other corpora we computed phoneme distribution over the JW300 dataset, which is a parallel corpus of over 300 languages with around 100K parallel sentences per language pair on average, covering miscellaneous material crawled from the website of Jehovah's Witnesses [40]. The Yoruba portion of JW300 corpus, post-processed and described in [14, 35], is significantly larger than our data that we summarized in Table 3: it consists of over 470 thousand sentences with over 10 million tokens. The JW300 phoneme distributions are shown on the right-hand side histogram of Figure 2. As can be seen from the two figures, the shapes and the modes of the histograms are very similar, with /a/, /i/ and /n/ being among the most frequent phonemes, and /ɛ̃/, /ĩ/ and /g/ among the rarest.

The distributions of tones over the vowels in our corpus are shown separately for male and female portions of the corpus in Table 4. As can be seen from the figure, the four nasal vowels in our corpus carry default (MID) tone, with extremely rare singleton occurrences of high and low tones likely due to spelling errors. For the five oral vowels, the distribution of tones is significantly more even, which is due to the nature of Yoruba open syllables for which any tone can be preceded or followed by any vowel tone.

The utterance transcriptions in our corpus contain 240 unique instances of contractions (e.g., "r'ógun"), briefly introduced in Section 2. Since disambiguating contractions requires lexical and part-of-speech knowledge, they often cannot be resolved using pure G2P approaches, such as Epitran or the finite state transducer (FST) based Yoruba grammars [41] from Phonetisaurus framework [42].

4. Application to TTS

In order to gauge the quality of the Yoruba speech corpus we built the traditional Hidden Markov Model (HMM)-based parametric speech synthesis system [43] using CLUSTERGEN method [44] provided by the Festival Speech Synthesis System, which was also used in [26] for evaluating the CMU Wilderness Dataset. The goal of this simple experiment is not to construct the state-of-the-art system, but rather to ascertain the usefulness of our data compared to other Ede languages in [26]. Using our corpus we constructed 15 TTS voices along two dimensions. The first dimension was the training data, configured in three possible ways: multi-speaker male-only (`yom`) and female-only (`yof`) data (in 48 and 16 kHz), and combined multi-speaker multi-gender data (`yo{m,f}` in 16 kHz). The second dimension was the CLUSTERGEN recipe, where, for each data configura-

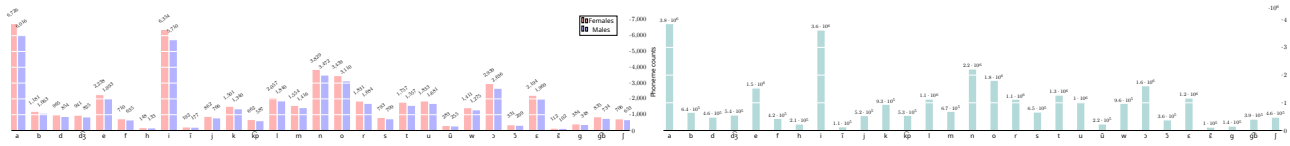

Figure 2: *Phoneme distributions for our male and female datasets (left) and Yoruba portion of JW300 corpus (right).*

Table 6: *Subjective preference test results for* RF-CG *voices.*

A	B_1	Score$_{A,B_1}$	t_{A,B_1}	B_2	Score$_{A,B_2}$	t_{A,B_2}
yof$_{16}$	lnf$_{16}$	1.022±0.065	31.148	lnf+yof$_{16}$	0.504±0.081	12.311
yom$_{16}$	lnm$_{16}$	1.065±0.078	27.206	lnm+yom$_{16}$	0.649±0.087	14.740

tion we constructed a grapheme-based system [45] (GR-CG), a regular phoneme-based system (CG) and a voice trained using random forests [46] (RF-CG). For both phoneme-based systems (CG and RF-CG), pronunciations were generated using the Epitran G2P system described in Section 3 using the phonology from Table 1. Disfluency tokens were removed. Each vowel-tone pair represented a separate phoneme. Sentences with orthographic contractions were removed from the training data resulting in 1491 and 1691 sentences for `yom` and `yof` datasets, respectively.

Objective evaluation was used for quality testing, whereby 10% of the data was held out during the training and on this data the synthetic speech was compared with the reference recorded speech using the Mean Mel-Cepstral Distortion (MCD) [47] of the predicted cepstra, measured in dB. A lower value of MCD suggests better synthesis and its typical ranges are from 4.0 (very good) to 8.0+ (not good) [26]. MCD is linked to perceptual improvement in the intelligibility of synthesis, an improvement of about 0.08 is perceptually significant [45]. Because we are building multi-speaker and multi-gender voices without performing speaker adaptation, we expect the values of MCD to be higher on average compared to a single-speaker scenario.

Results of the objective evaluation experiments are shown in Table 5, where for each of the 15 configurations the MCD is shown along with the root mean square error (RMSE), in Hz, for fundamental frequency (F0). The ISO 639-3 language code is shown in the third column. We also show the results for the two SY relatives from the CMU Wilderness dataset, Ifè (`ife`) and Okpela (`atg`), as reported in [26]. Both datasets are multi-speaker and mostly contain male recordings. We have not found any female speakers in that data. Similar to the results reported in [26], we find that the phoneme-based random forest configurations RF-CG are the best, significantly outperforming in terms of MCD other voice types for all of our datasets. The grapheme-based systems GR-CG are the worst performing of all the configurations, leading us to conclude that the use of a reasonably accurate SY G2P improves the quality of the voices. All 48 kHz configurations significantly outperform the 16 kHz voices, while combining female and male 16 kHz data did not lead to better quality. We hypothesize that the confounding factor is the increased number of diverse speaking styles and speaker characteristics due to including both genders in the training data. As can be seen from Table 5, all SY voices built using our data significantly outperform the multi-speaker Ifè CMU voices reported in [26]. However, this is not the case for multi-speaker Okpela CMU voices. We hypothesize that this is due to the nature of MCD metric that favors datasets with lower cepstral variance in the reference data. Some of the CMU Wilderness languages, such as Spanish, are provided as multiple datasets with the RF-CG MCD scores varying from 5.12 (voice `SPNBDA`, good) to 7.09 (voice `SPNNVI`, bad). It is likely that combining this data in a single voice may have the averaging effect decreasing the quality compared to the best constituent dataset while improving the quality compared to the worst performing dataset.

The subjective A/B listening tests were performed with 10 native Yoruba speakers using headphones. Each listener was presented with 100 stimulus pairs generated from out-of-domain biblical sentences. Each pair had to receive a forced-choice rating on a 7-point preference scale: B preferred over A ({-3,-2,-1}), no preference (0) and A preferred over B ({1,2,3}). Each stimulus pair had to have at least 8 distinct raters. Table 6 shows the results of subjective preference tests of several 16 kHz RF-CG voices comparing our datasets (A) against voices constructed purely from the Lagos-NWU datasets (described in Section 2, denoted `lnf` and `lnm` under B_1) and by combining Lagos-NWU dataset with our data (B_2). Same linguistic front-end was used during the training of all voices. The scores are shown along with the corresponding 95% confidence intervals. Columns four and seven show the corresponding values of t-statistic from a two-sided t-test with $\alpha = 0.01$ and degrees of freedom $df = 99$. All results are statistically significant with $p < 10^{-5}$. The significant levels of reverberation in Lagos-NWU data degrade the quality of multi-speaker voices B_1 and B_2 for both genders. As can be seen from Table 6, in both cases, the voices constructed solely from our data were consistently preferred by the listeners. It is also interesting to note that combining our data with Lagos-NWU data results in better perceived quality than using Lagos-NWU data on its own.

5. Conclusions and Future Work

This paper introduced the high quality multi-speaker Yoruba speech corpus. The corpus has been designed with speech applications in mind, such as multi-speaker TTS and ASR speaker adaptation. We described the process used to construct the corpus. The data is released with a permissive open-source license. We showed how this data can be used to construct a simple, yet viable, TTS system using open-source tools. We hope that this data will aid research and development of speech applications for this important language. In the future we plan to use this corpus as a high-quality language adaptation data in state-of-the-art multilingual TTS and ASR deep learning architectures.

6. References

[1] H. Hammarström, R. Forkel, and M. Haspelmath, "Glottolog 4.1," Max Planck Institute for the Science of Human History, Jena, 2019. [Online]. Available: http://glottolog.org

[2] B. Adediran, *The Frontier States of Western Yorubáland: 1600–1889*. IFRA-Nigeria, 1994.

[3] A. Kluge, "A synchronic lexical study of the Ede language continuum of West Africa: The effects of different similarity judgment criteria," *Afrikanistik-Aegyptologie Online*, no. 4, 2008.

[4] M. O. Kehinde, "Implications of colonially determined boundaries in (West) Africa: the Yoruba of Nigeria and Benin in perspective," Ph.D. dissertation, School of Government and International Affairs, Durham University, UK, 2010.

[5] D. Gnisci and M. Trémolières, "Atlas on Regional Integration in West Africa: Population Series, Languages," Organisation for Economic Co-operation and Development (OECD), Tech. Rep. ECOWAS/SWAC-OECD, 2006.

[6] G. F. Simons and C. D. Fennig, *Ethnologue: Languages of Africa and Europe*, 21st ed. SIL International, 2018.

[7] F. M. Ikoro, "Development and Sustenance of Indigenous Languages in Nigeria: The Role of Ninlan and Its Library," in *Proc. of International Conference on Social and Education Sciences*. Denver, CO, USA: ERIC, October 2019, pp. 45–126.

[8] K. M. Oparinde, "Language Endangerment: The English Language Creating a Concern for the Possible Extinction of Yorùbá Language," *Journal of Communication*, vol. 8, no. 2, pp. 112–119, 2017.

[9] D. Goldhahn, M. Sumalvico, and U. Quasthoff, "Corpus collection for under-resourced languages with more than one million speakers," in *Proc. of CCURL*, 2016, pp. 67–73.

[10] I. I. Ayogu, A. O. Adetunmbi, and B. A. Ojokoh, "Developing Statistical Machine Translation System for English and Nigerian Languages," *Asian Journal of Research in Computer Science*, pp. 1–8, 2018.

[11] M. Straka, J. Straková, and J. Hajič, "UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging," *arXiv preprint arXiv:1908.06931*, 2019.

[12] O. Adetunmbi, O. Obe, and J. Iyanda, "Development of Standard Yorùbá speech-to-text system using HTK," *International Journal of Speech Technology*, vol. 19, no. 4, pp. 929–944, 2016.

[13] A. R. Iyanda and O. D. Ninan, "Development of a Yorúbà Text-to-Speech System Using Festival," *Innovative Systems Design and Engineering (ISDE)*, vol. 8, no. 5, 2017.

[14] I. Orife, "Attentive Sequence-to-Sequence Learning for Diacritic Restoration of Yorùbá Language Text," in *Proc. of Interspeech*, 2018, pp. 2848–2852.

[15] A. R. Ìyàndá, O. A. Odéjobí, F. A. Soyoye, and O. O. Akinadé, "Development of Grapheme-to-Phoneme Conversion System for Yorùbá Text-to-Speech Synthesis," *INFOCOMP Journal of Computer Science*, vol. 13, no. 2, pp. 44–53, 2014.

[16] J. A. E. Wibawa, S. Sarin, C. F. Li, K. Pipatsrisawat, K. Sodimana, O. Kjartansson, A. Gutkin, M. Jansche, and L. Ha, "Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech," in *Proc. of LREC*, 2018, pp. 1610–1614.

[17] O. Fagbolu, A. Ojoawo, K. Ajibade, and B. Alese, "Digital Yorùbá Corpus," *International Journal of Innovative Science, Engineering and Technology*, pp. 2348–7968, 2015.

[18] J. O. Alabi, K. Amponsah-Kaakyire, D. I. Adelani, and C. España-Bonet, "Massive vs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yorùbá and Twi," *arXiv preprint arXiv:1912.02481*, 2019.

[19] T. K. Dagba, J. O. Aoga, and C. C. Fanou, "Design of a Yoruba Language Speech Corpus for the Purposes of Text-to-Speech (TTS) Synthesis," in *Proc. of Asian Conference on Intelligent Information and Database Systems*, 2016, pp. 161–169.

[20] J. O. Aoga, T. K. Dagba, and C. C. Fanou, "Integration of Yoruba language into MaryTTS," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 151–158, 2016.

[21] S. A. M. Yusof, A. F. Atanda, and M. Hariharan, "A review of Yorùbá Automatic Speech Recognition," in *Proc. of IEEE 3rd International Conference on System Engineering and Technology*. IEEE, 2013, pp. 242–247.

[22] D. van Niekerk, E. Barnard, O. Giwa, and A. Sosimi, "Lagos-NWU Yoruba Speech Corpus," North-West University, South Africa, 2015. [Online]. Available: https://repo.sadilar.org/handle/20.500.12185/431

[23] D. R. Van Niekerk and E. Barnard, "Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis," *Speech Communication*, vol. 56, pp. 229–242, 2014.

[24] T. J. Sefara and A. Modupe, "Yorùbá Gender Recognition from Speech Using Neural Networks," in *Proc. 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2019, pp. 50–55.

[25] R. Blench, *An Atlas of Nigerian Languages*. Cambridge, 2019.

[26] A. W. Black, "CMU Wilderness Multilingual Speech Dataset," in *Proc. of ICASSP*, 2019, pp. 5971–5975.

[27] S. Crowther, *A Grammar of the Yoruba Language*. Seeleys, 1852.

[28] A. Bamgbose, *A Grammar of Yoruba*, ser. West African Language Monograph Series. Cambridge University Press, 1966, vol. 5.

[29] T. Olúmúyìwá, "Yoruba Writing: Standards and Trends," *Journal of Arts and Humanities*, vol. 2, no. 1, pp. 40–51, 2013.

[30] O. O. Oyelaran, "Yoruba Phonology," Ph.D. dissertation, University of Michigan, Ann Arbor, 1971.

[31] C. Chanard and R. L. Hartell, *Yoruba Sound Inventory (AA)*. Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/inventories/view/848

[32] G. N. Clements and A. Rialland, "Africa as a phonological area," in *A Linguistic Geography of Africa*, B. Heine and D. Nurse, Eds. Cambridge University Press, Cambridge, 2008, pp. 36–85.

[33] E. Fajobi, "The Nature of Yoruba Intonation: A New Experimental Study," in *Yoruba Creativity: fiction, language, life and songs*, T. Falola and A. Genova, Eds. Trenton, NJ: Africa World Press, 2005, pp. 183–221.

[34] T. Adegbola and L. U. Odilinye, "Quantifying the effect of corpus size on the quality of automatic diacritization of Yorùbá texts," in *Proc. of SLTU*, 2012, pp. 48–53.

[35] I. Orife, D. I. Adelani, T. Fasubaa, V. Williamson, W. F. Oyewusi, O. Wahab, and K. Tubosun, "Improving Yorùbá Diacritic Restoration," *arXiv preprint arXiv:2003.10564*, 2020.

[36] Creative Commons, "Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)," 2019. [Online]. Available: http://creativecommons.org/licenses/by-sa/4.0/deed.en

[37] D. Povey, "Open SLR," John Hopkins University, US, 2020. [Online]. Available: http://www.openslr.org/resources.php

[38] V. Mapelli, V. Popescu, L. Liu, and K. Choukri, "Language Resource Citation: the ISLRN Dissemination and Further Developments," in *Proc. of LREC*, 2016, pp. 1610–1613.

[39] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," in *Proc. of LREC*, 2018, pp. 2710–2714.

[40] Ž. Agić and I. Vulić, "JW300: A wide-coverage parallel corpus for low-resource languages," in *Proc. of ACL*, 2019, pp. 3204–3210.

[41] K. Kirchhoff, M. Hasegawa-Johnson, P. Jyothi, and L. Rolston, "LanguageNet Grapheme-to-Phoneme Transducers," Statistical Speech Technology, University of Illinois, 2018. [Online]. Available: https://github.com/uiuc-sst/g2ps

[42] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[43] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[44] A. W. Black, "CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling," in *Proc. of Interspeech*, 2006, pp. 1762–1765.

[45] S. Sitaram, A. Parlikar, G. K. Anumanchipalli, and A. W. Black, "Universal Grapheme-based Speech Synthesis," in *Proc. of Interspeech*, 2015, pp. 3360–3364.

[46] A. W. Black and P. K. Muthukumar, "Random Forests for Statistical Speech Synthesis," in *Proc. of Interspeech*, 2015, pp. 1211–1215.

[47] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion," in *Proc. of SLTU*, 2008, pp. 63–68.