Full Paper

STATISTICAL TEXT ANALYSIS FOR YORÙBÁ SPEECH GENERATION USING ZIPF'S LAW

A.R. Iyanda
Computer Science and Engineering Department
Obafemi Awolowo University, Ile-Ife, Nigeria
abiyanda@oauife.edu.ng

ABSTRACT.

The practical challenge of creating a Yorùbá text-to-speech synthesis has initiated our work on statistical text analysis. Language and speech technology applications have gained an increasingly wide-spread use in several languages/countries, and this has necessitated the importance of examining how much difference exists between English (in most cases the first language for most technologies and applications) and tone languages, specifically, Yoruba. These differences are studied and described in detail in linguistics but they rarely quantified and used by technology developers. In this paper, Yoruba language was described using text corpora from textbooks and newspapers. Other texts from Internet sources were also used. The corpus size was 291,392 word forms and the data was analyzed using Zipf. Based on the statistical analysis, it was found that the coverage of corpora by the most frequent words follows a parallel logarithmic rule for all languages in coverage range, known as Zipf's law in linguistics.

Keywords: *Text corpora, Data analysis, Standard Yoruba, Text-to-speech synthesis*

1. INTRODUCTION

As Text-To-Speech (TTS) systems find application in varied fields such as (i) computer assisted language learning (CALL) system; (ii) text-to-speech synthesis for Linguistic and Psycholinguistic Experimentation; (iii) telecommunication services (such as mobiles, multimedia, man machine Communication); and (iv) automatic text read aloud system and so on (Ngugi *et al.*, 2005). There is a need to create an adequate language resource which requires inter-disciplinary efforts amongst phoneticians, linguists as well as the use of powerful speech processing software.

A TTS synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. A text-to-speech system consists of two main parts: The first part is the natural language processing module, which produces a phonetic transcription of the input text, together with the prosodic information, that is information about say pitch, duration and phrasing (Dutoit, 1997). The second part is the digital signal-processing module, which transforms the information given by the natural language processing model into synthesized speech. Systems that simply concatenate isolated words or parts of sentences, denoted as Voice Response Systems, are only applicable when a limited vocabulary is required (typically a few one hundreds of words), and when the sentences to be pronounced respect a very restricted structure, as is the case for the announcement of arrivals in train stations for instance. In the context of TTS synthesis, it is impossible (and useless) to record and store all the words of the language. It is thus more suitable to define Text-To-Speech as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter (Dutoit, 1997).

1.1. Standard Yorùbá

Yorùbá is classified as a Niger-Congo language of the Yoruboid branch of Defoid, Benue-Congo (Riddle and Stahlke, 1991). Although Yorùbá has several dialects with linguistic variations, all speakers can communicate effectively using the Standard Yorùbá (SY) (also known as Yorùbá Koine (Fagborun, 1994). SY is used in language education, the mass media and everyday communication (Odejobi, 2007). Yorùbá is a tone language in which a pitch of an utterance is used to express differences in meaning; or when a particular pitch or change of pitch constitutes an element in the intonation of a phrase or sentence, such as high, mid or low. The tone information in Yorùbá texts is indicated by diacritics - tone marking on top of vowels and syllabic nasals to give meaning to the contexts (Iyanda, 2014). Yorùbá alphabet comprises of 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).

2. DATA

Data can exist in two forms: primary data and secondary data. Primary data are those collected by the researcher for the purpose of the survey in mind. They are always given in the form of raw materials and are originals in character. To be able to analyse and interpret such data, there is a need for the application of statistical technique. The data can be collected in three different ways such as by: (i) experiment, (ii) survey using structured questionnaires and (iii) interview, participant observation or focus group. Secondary data, on the other hand, is information that has been collected by somebody for other purpose. These are data collected through censuses, organisational records as well as through qualitative methodologies or qualitative research. Secondary data must possess the qualities of reliability, suitability and adequacy. For the purpose of our design, secondary data was used. In designing database, there are four major procedures to consider and these include; collecting databases, labelling, extraction of features and building models from the data (Black and Lenzo, 2007).

i. Collecting databases: Getting the right type of data is important when one is going to be using its content to build models. The quality of synthesized speech of data collected, for example a database of isolated words provide clearly articulated synthesized speech that will sound like isolated words, even when used for the synthesis of continuous speech. Both the acoustic properties and the prosodic properties of the database are important. In designing a database, the specific points worth to be considered are: phonetic coverage, dialect, voice quality, prosodic coverage as well as size.

ii. Labelling databases: In order for Festival to use a database it is most useful to build utterance structures for each utterance in the database. Festival is a multi-lingual TTS engine and a general purpose concatenative TTS system that offers a general framework for building speech synthesis systems. Generally, there is need for labels for: segments, syllables, words, phrases, intonation events, pitch targets.

iii. Extracting features: The easiest way to extract features from a labelled database is by loading in each of the utterance structures and dumping the desired features for training using the Festival script 'dumpfeats'.

iv. Building models: In this stage, model is built from data extracted from databases using the CART building program, 'wagon' which is available as with the Edinburgh Speech Tools Library.

2.1. Data Collection

The domain for the speech synthesis for this research is in language education, religious and mass media from various sources such as Internet, digitized printed material and existing digital materials from non-Internet sources. Two SY newspaper (Aláròyé and Yorùbá Gbòde) and two SY textbooks (Awobuluyi, 2008, Owolabi, 2011) were selected. The two newspapers were not toned mark, and Yorùbá Gbòde is without under dots. The texts selected from the newspaper for the text corpus were fully tone marked and under dotted. To increase the quality and coverage of the acquired materials, additional Yorùbá text data were gathered from existing printed materials by scanning through a process known as optical character recognition. The volume of textual data generated using this technique was about 291,392 words.

The texts were edited using a Tákàdá text editor (www.sourceforge.net/projects/takada) and were corrected for graphemic items (tone marks and under dots) using Àkotó Yorùbá. Àkotó Yorùbá is a software developed by Asahiah (2014) for restoration of Yorùbá diacritics. In this system, database of pre-selected units (syllables) was used. The creation of this database involved selecting a single unit, for inclusion in the system representing all possible occurrences. To cater for all possible segmental (auto and supra) contexts, the unit needs to be as neutral as possible.

For the Yorùbá TTS, the database contains syllable segments and their feature descriptions were translated into a data structure. The data (that is syllable segments and their feature descriptions) were stored in a hashtable using linked lists. The selection of a set of phonemically balanced sentences that contain the units was done by comparing the text corpus and the transcribed text. The choice of this unit is critical to realizing the selection of phonemically balanced sentences. The sample data collected is shown in Table 1. The data is based on the Yorùbá syllable structure. There are about 690 syllables (V=7x3, Vn=5x3, N=2x3, CV=126x3 and CVn=90x3) collected.

The concatenation of those syllables form a word and the concatenation of words produce a sentence. It should be noted that not all the syllables presented in Tables 1 are valid Yorùbá syllables. For example, "l" cannot precede a nasal vowel (for instance, **lan, lẹn, lọn, and lun,** do not exist in the Yorùbá words).

2.2. Data Analysis

Data analysis was done using Zipf which observed a phenomenon in human languages (and some other social structures) that indicated that there is a pattern in the distribution of tokens and that a few very frequent words make up a very large portion of any text or collection of texts, while the large majority of words occur relatively rarely. Zipf's law is an empirical law formulated using mathematical statistics. It refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution, one of a family of related discrete power law probability distributions (Chen, 2012).

The law is named after the American linguist George Kingsley Zipf (1902-1950), who first proposed it.

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc (Sorell. 2012). For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus (Manning and Schutze, 1999).

Table 1: Yorùbá syllable data for mid tone

Consonant (C)	Oral Vowel (V)								Nasal Vowel (Vn)				Syllabic Nasal (N)
	a	e	ẹ	i	o	ọ	u	an	ẹn	in	ọn	un	n, m
C				CV						CVn			
b	ba	be	bẹ	bi	bo	bọ	bu	ban	bẹn	bin	bọn	bun	Syllabic nasals can stand alone as a syllable not in combination with any consonant or vowel.
d	da	de	dẹ	di	do	dọ	du	dan	dẹn	din	dọn	dun	
f	fa	fe	fẹ	fi	fo	fọ	fu	fan	fẹn	fin	fọn	fun	
g	ga	ge	gẹ	gi	go	gọ	gu	gan	gẹn	gin	gọn	gun	
gb	gba	gbe	gbẹ	gbi	gbo	gbọ	gbu	gban	gbẹn	gbin	gbọn	gbun	
h	ha	he	hẹ	hi	ho	họ	hu	han	hẹn	hin	họn	hun	
j	ja	je	jẹ	ji	jo	jọ	ju	jan	jẹn	jin	jọn	jun	
k	ka	ke	kẹ	ki	ko	kọ	ku	kan	kẹn	kin	kọn	kun	
l	la	le	lẹ	li	lo	lọ	lu	lan	lẹn	lin	lọn	lun	
m	ma	me	mẹ	mi	mo	mọ	mu	man	mẹn	min	mọn	mun	
n	na	ne	nẹ	ni	no	nọ	nu	nan	nẹn	nin	nọn	nun	
p	pa	pe	pẹ	pi	po	pọ	pu	pan	pẹn	pin	pọn	pun	
r	ra	re	rẹ	ri	ro	rọ	ru	ran	rẹn	rin	rọn	run	
s	sa	se	sẹ	si	so	sọ	su	san	sẹn	sin	sọn	sun	
ṣ	ṣa	ṣe	ṣẹ	ṣi	ṣo	ṣọ	ṣu	ṣan	ṣẹn	ṣin	ṣọn	ṣun	
t	ta	te	tẹ	ti	to	tọ	tu	tan	tẹn	tin	tọn	tun	
w	wa	we	wẹ	wi	wo	wọ	wu	wan	wẹn	win	wọn	wun	
y	ya	ye	yẹ	yi	yo	yọ	yu	yan	yẹn	yin	yọn	yun	

The analysis of the data used in this study for frequency distribution was performed using a free concordance software Simple Text Analysis Tool (TextSTAT) version 2.9.0.0 (Hüning, 2014). TextSTAT was designed to be user friendly and provide simple Internet functionality. Texts can be combined to form corpora (which can also be stored as such). The program analyses these text corpora, displays word frequency lists, concordances, and keywords in context according to search terms. TextSTAT can be used to search large amounts of text. It can also help to learn how often a certain word occurs or in what contexts it is used. Word combinations can as well be examined.

TextSTAT) version 2.9.0.0 was used to generate the frequency count for the word tokens in the text data. It was also used to detect words that have incorrect forms such as wrong diacritics (tone marks and under dots). In the analysis of the words in the main text corpus, 6857 words appeared only once in the corpus and the word with the highest appearance in the data set (*tí*) occurs 8257 times, followed by *ni* with frequency of 6358 and followed by *àwon* with frequency of 5753. This pattern is expected of natural language (Sorell, 2012). Table 2 shows the distribution of some selected words in conformity with Zipf's law for linguistic data. The word n is in the tenth position and it occurred 4365 times, *ti* is in the twentieth position and occurred 2673 times and so on.

Table 2: Ranks and frequencies of the text data

Word Type	Rank (r)	Frequency (f)	Proportion (%)
n	10	4365	1.4978
ti (has)	20	2673	0.9173
fún (give)	30	2215	0.7601
gẹ (cut)	40	1560	0.5354
èdè (language)	50	1153	0.3957
o (you)	60	947	0.3250
ìṣe (doing)	70	786	0.2697
ìtàn (story)	80	694	0.2382
pọ	90	599	0.2056
ìmọ (knowledge)	100	524	0.1798
ìpínlẹ (state)	200	189	0.0649
àtijọ (former)	300	110	0.0377
yí (turn)	400	77	0.0264
ọjà (market)	500	57	0.0196
Adéyẹmí (Yoru□ba□ name)	600	44	0.0151
Ààyè (opportunity)	700	37	0.0127
Àgbàlagbà (adult)	800	30	0.0103
gbeyẹwò (consider)	900	25	0.0086
déésì (translated name)	1000	21	0.0072
kórìira (hate)	10000	1	0.0003

Zipf's law states that there is an inverse proportional relationship between the rank and frequency of words in a text. Rank is the position of a word in a table of words ordered by frequency of occurrence (rank one being the most frequent) (Sorell, 2012).

The general nature of Zipf's distribution as it applies to human language use whether in oral or written communication or discourse (Asahiah, 2014) is as highlighted below:

i. a few word tends to score very high on the frequency table meaning that they appear regularly in discourse;
ii. a medium number of words or linguistic tokens appears with relatively medium frequency indicating that they are regularly used;
iii. a huge number of words in a document have very low level of occurrence, indicating that they are rarely used.

With the proportion of these tokens in relation to the total token count, it means that these words (*tí* and *ni*) occurred in many contexts and are good features to be used in grapheme to phoneme conversion system for Yorùbá language. This agrees with existing work (Asahiah, 2014). The Zipf curve for the words whose rank ranges from tens to thousands (the regularly occur to the rarely occur), to have a wide coverage of the database is shown in Fig. 1. The shape of the curve justifies Zipf's law for linguistic data. This means that a few words appear regularly in discourse; a medium number of words or linguistic tokens appears with relatively medium frequency indicating that they are averagely used; a huge number of words in a document have very low level of occurrence, indicating that they are rarely used.

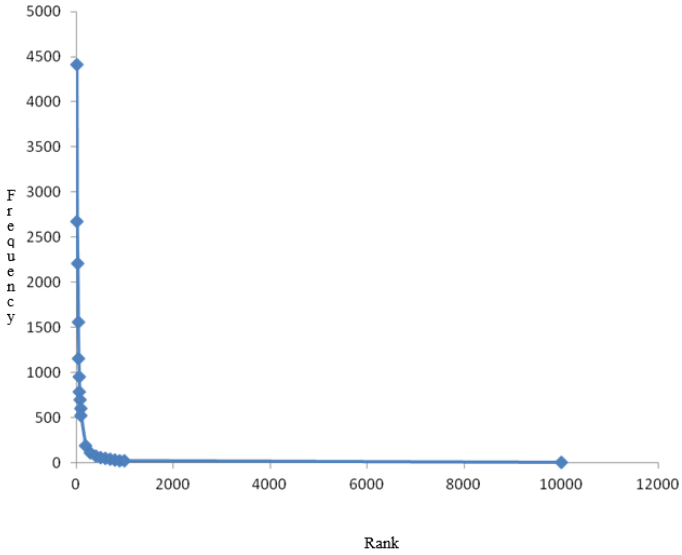

Fig. 1: Zipf frequency-rank curve for Yorùbá texts

Zipf also argued that the proportion of a text made up by words with a certain frequency should be equal (Sorell, 2012):

$$1=f\ (f+1) \qquad (1)$$

Where f is the frequency of each word in the database.

3. IMPLEMENTATION

R

Language resource such as specially prepared speech material (e.g. text and speech sound) has become a central issue in the development of a TTS for any language. The design, collection, recording and annotation of such materials depend on several factors. These factors include: the scope of the research, the linguistic features of the language, and the domain of application.

The data used in studies like this varies widely in scale and scope. The creation of an adequate language resource is a very complex task which requires inter-disciplinary efforts amongst phoneticians, linguists as well as the use of powerful speech processing software (Odejobi, 2005). There are no publicly available language resources for standard Yorùbá, so a small language resource was created for use in this research.

3.1. Speech Corpus

The analysis of the text database discussed earlier informed the selection of text for the speech corpus. Six hundred and ninety (690) syllables were used for the speech database. The following procedures are carried out:

i. Recording: 740 words and 690 syllables were read by two males and two females who are native speakers of SY. The age of the speakers ranges from 22 to 37 years old. Each speaker reads the text at his/her own pace, resulting in the average number of syllable per seconds ranging from two to three. The sounds were prepared by recording at a 44100 Hz sampling rate and 32-bit quantization. After that, they were down-sampled to 16 kHz for analyzing and use in the

implementation of the system in Festival TTS engine. All speech units were recorded in normal speaking rate.

ii. Recording equipment: The corresponding speech data for the Yorùbá syllables collected was recorded in a quiet environment with a noise cancelling microphone on a typical multimedia computer system using the Speech Filing System (SFS) software. Adobe Audition 3.0 was used to normalise and to reduce adaptive noise. SY which is the one being used in education, the mass media and everyday communication was chosen for the recording. The recorded speech data was analysed and annotated using the PRAAT speech processing software.

iii. Speech file annotation: Each file of the recording was loaded into PRAAT and annotated manually. For the annotation, TextGrid is created in PRAAT for each speech waveform file. The TextGrid and waveform files were selected for annotation and editing. There are two tiers in the annotation: word and syllable. The labelling was done to identify syllabic boundaries. In the annotation of the syllable speech files, only one tier was specified (the syllable tier). The syllable is labelled with its associated tone as shown in Fig. 2. In the annotation of the word speech files, two tiers were specified (the syllable and word tiers) as shown in Fig. 3. Both the spectrogram and the waveform were used in determining syllable and word boundaries.

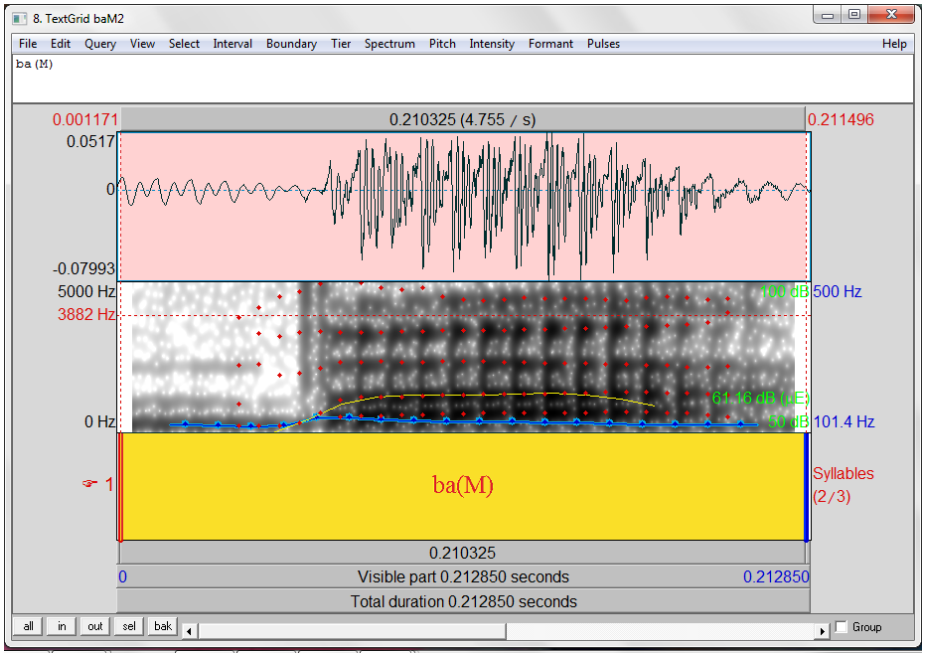

Fig. 2: Screen capture of annotation of the syllable bá

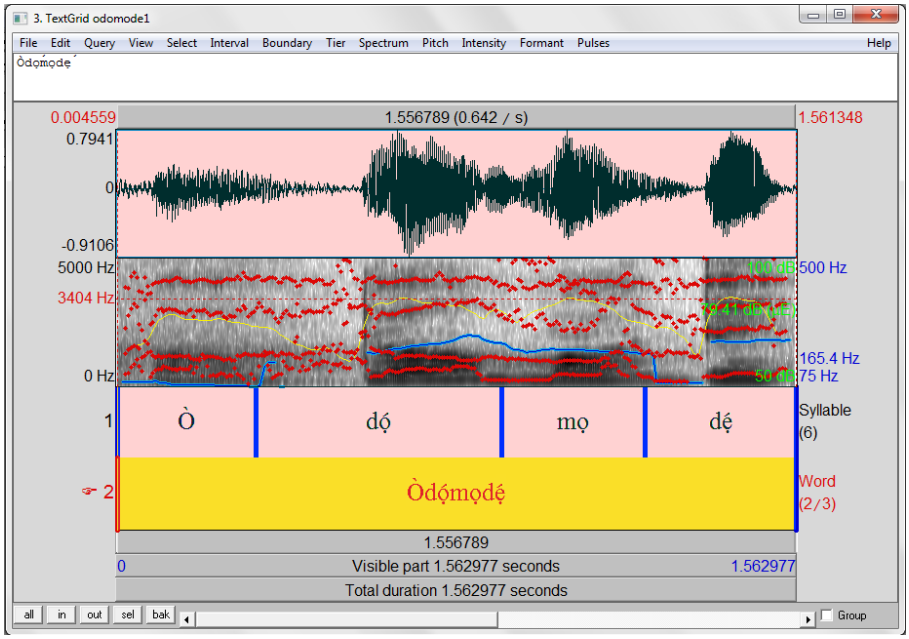

Fig. 3: Screen capture of annotation of the word Ọdọmọdé

4. CONCLUSION

The text corpus shows a very similar coverage distribution which can be well approximated by straight lines on a logarithmic scale. The shape of the curve conforms to Zipf's law for linguistic data and agrees with existing works (Lin *et al.*, 2014; Manning and Schutze, 1999). This means that a few words appear regularly in discourse; a medium number of words or linguistic tokens appears with relatively medium frequency indicating that they are averagely used; a huge number of words in a document have very low level of occurrence, indicating that they are rarely used. This work contributes significantly to the knowledge for natural language engineering for Yorùbá such as predictive text input, diacritic restoration, word hyphenation, language modelling in speech recognition, corpus-based speech synthesis, and so on.

REFERENCES

Asahiah, F. O. The Development of a Standard Yorùbá Diacritics Restoration System. PhD thesis,Obafemi Awolowo University, Ile-Ife, Nigeria, 2014.

Awobùlúyì, O. Èkó Ìsèdá-Òrò Yorùbá . Montem Paperbacks, Akure, Ondo State, 2008.

Black, A.W. and Lenzo, K. A. Building synthetic voices. available from festvox.org/bsv/bsv.pdf. 2007.

Chen, Y. Zipf's law, 1/f noise, and fractal hierarchy. Chaos, Solitons & Fractals 45.1: 63-73. 2012.

Dutoit, T. An introduction to text-to-speech speech system. isbn 0-7923-7923- 4498-7. Master's thesis, Kluwer Academic, 1997.

Fágborún, J. G. The Yorùbá Koiné: Its History and Linguistic Innovations, volume 6. LINCOM Europa, München, Newcastle, linguistics edition, 1994.

Hüning, M. Textstat - simple text analysis tool/ concordance software. available from http://neon.niederlandistik.fu-berlin.de/en/textstat/. visited: April, 2014.

Ìyàndá, A. R. Design and Implementation of a Grapheme-to-Phoneme Conversion System for Yorùbá Text-to-Speech Synthesis. PhD thesis, Obafemi Awolowo University, Ile-Ife, Nigeria, 2014.

Lin, R., Ma, Q. D. Y. and Bian, C. Scaling laws in human speech, Decreasing emergence of new words and a generalized model. arXiv preprint arXiv:1412.4846. 2014.

Manning, C. D. and Schutze, H. Foundations of Statistical Natural Language Processing, volume 999. MIT Press, 1999.

Ngugi, K., Okelo-Odongo, W., and Wagacha, P. W. Swahili Text-to-Speech System. African Journal of Science and Technology (AJST) Science and Engineering Series, 6(1):80-89. 2005.

Odéjobí, O. A. A Computational Model of Prosody for Yorùbá Text-to-Speech Synthesis. PhD thesis, Aston University, 2005.

Odéjobí, O. A. A quantitative model of yorùbá speech intonation using stem-ml. INFOCOMP Journal of Computer Science, 6(3):47–55, 2007.

Owólabí, K. Ìjìnlè Ìtúpalè Èdè Yorùbá (1): Fònétîkì ati Fonólojì. Universal Akada Books Nigeria Limited, 2011.

Riddle, E. and Stahlke, H. Linguistic typology and Sinospheric languages. First Annual Meeting of the Southeast Asian Linguistics Society. 1991.

Sorell, C. Zipf's law and vocabulary. The Encyclopedia of Applied Linguistics. Wiley Online Library 2012.