Design and Implementation of a Luganda Text Normalization Module for a Speech Synthesis Software Program

Ronald Kizito†, Wayne S. Okello‡, and Sulaiman Kagumire‡

***Abstract*—This paper describes a Luganda text normalization module, a crucial component needed for a Luganda Text to Speech system. We describe the use of a rule-based approach for detection, classification and verbalization of Luganda text. At the core of this module are the Luganda grammar rules that were hand-built to normalize Non-Standard Words (NSWs) from different semiotic and noun classes. Input text is first analyzed, matched against handcrafted patterns developed using regular expressions to detect any NSWs. Upon detection, NSWs are tokenized and classified into one of the semiotic classes and then if necessary, into one of the Luganda noun classes. These are subsequently verbalized, each according to its semiotic as well as noun class, and a new text file is produced. We tested the module with 7 datasets and achieved average detection and normalization rates of 82% and 77.7% respectively.**

***Index Terms*—Automatic Speech Recognition, Detection-conversion, Luganda, Machine Translation, NLP, Number system, Speech Synthesis, Text Normalization, Text-to-speech, TTS.**

I. Introduction

TEXT normalization is the conversion of non-standard words (NSWs) like **$4** or **Mw.** into standard words like **Ddoola nnya**, *four dollars* or **Mwami**, *Mister* respectively [1]–[4]. These standard words can then be converted into speech using a Text-to-Speech system [1], [4], [5] or translated into a different language using a Machine Translation system. A typical Luganda document may contain several NSWs that can be categorized into many semiotic classes. The most common of these are abbreviations, numbers, measurement units, ranges, dates, times, currencies, abbreviations and letter sequences [2]. A non-standard word like **4** could belong to a currency, a date, a time or another semiotic class. Therefore it can be converted into different standard words. Therefore a good text normalization system must be able to correctly detect that **4** is a NSW, classify and verbalize it into a standard word. Luganda text normalization systems must solve several other problems that are not found in languages like English which are not agglutinating and do not have as many noun classes [6]. The presence of many noun classes means that the non-standard word, **1**, can belong to anyone of 10 classes meaning that a good noun classifier must be included in any text normalization module. The agglutinating nature of Luganda means that it is impossible to create a wordlist that contains all the standard words in the language and simply assume everything else is a non-stardard word. Text normalization and text-to-speech (TTS) systems for low-resourced languages like Luganda face even bigger challenges as they cannot take advantage of the most recent advances in natural language processing (NLP) like the use of deep neural networks. Deep neural networks not only require large, high quality annotated datasets and a large commercial interest but also are prone to making unacceptable errors [7], [8]. On the other hand rule-based NLP systems have shown promise as a first step for NLP development of low-resource languages. [9], [10].

This paper describes the architecture of a Luganda Text Normalization module. We worked with linguists and native speakers of Luganda to create grammar rules for normalizing NSWs from the most common semiotic classes.

The semiotic classes used in Luganda was based on what is commonly found in English text normalization [2] and in Luganda news articles and recommendations of the language experts. Like many other authors [8], we excluded data from short message services (SMS) texts, advertisements and social media sites since they normally contain a lot of spelling errors, electronic addresses, rare abbreviations, code mixing and code-switching (English words in Luganda sentences) [11]. The rest of the paper is organized as follows: In Section II, we discuss the related work and highlight the key contributions of our work in this paper. Then in Section III, we describe the implementation process that included collecting required data about some noun classes, the Luganda number system, the pronunciation of various nouns [12] and numbers from each semiotic class used. Then in Section IV, we show evaluation of the module. Finally, in Section V, we end the paper with future work.

II. Related Work

Most of the previous work done in the field of text normalization has been focused on the large commercial languages like English. This includes seminal work presented in [2] where most of the common semiotic classes are introduced and [3], [13], [14]. Further work was done in [15] where Google's Kestrel text normalization engine is described. There are a few papers devoted to text normalization systems that deal with large but low-resourced languages like Bangla [9] and

†R. Kizito is with the Department of Electrical and Computer Engineering, Makerere University, P.O. Box 7062, Kampala, Uganda (email: ronald.kizito@cedat.mak.ac.ug).

‡W. S. Okello and S. Kagumire are with netLabs!UG, a Research Centre of Excellence in the Department of Electrical and Computer Engineering, Makerere University, P.O. Box 7062, Kampala, Uganda (email: wayneredemption@outlook.com, kagumire1840@gmail.com).

Myanmar [16]. However, there is very little published work devoted to low-resourced Bantu languages like Luganda or Kiswahili despite the development of TTS systems that support voices in these languages [17]–[19]. Therefore, most of these TTS systems generally cannot handle NSWs. Fortunately, most of the techniques and semiotic classes used in building text normalization modules for English can be transferred to Luganda. This has enabled us to build what as far as we know, is the first Luganda text normalization module.

The main contributions of this paper are as follows:

- a description of a rule-based approach to text normalization of texts written in the Luganda language, that to our knowledge is the first of its kind.
- a description of the Luganda number system.
- a Luganda dataset[1] that contains a wide range of common semiotic classes found in newspapers, the bible, and books.

We believe that these contributions could be very useful for researchers working in the field of natural language processing for low resourced languages.

III. Implementing Text Normalization

The text normalization module takes plain text or a text file as the input with words already separated by white spaces. As in [16], text normalization is divided into two phases. First, input text is analysed, tokenized, and NSWs detected and classified into semiotic classes. In this phase, some input tokens may be grouped together. Secondly, a verbalizer module for each semiotic class and noun class converts the classified NSWs into standard text. Although Luganda texts contain very many semiotic classes, we decided to limit our scope to the most common ones. These are listed in Table I.

Table II shows the noun (and adjective) classes used in Luganda to refer to countable things as well as some examples of each class. Table II shows that, like most Bantu languages, Luganda nouns have initial vowels (a, e, o) that generally act like articles [20]. This initial vowel (when it exists) and the prefix are used in the classification of a noun. Correct noun-classification is very important as it determines whether for instance, the digit **1**, is verbalized as e**mu**, o**mu**, eki**mu**, olu**mu**, ogu**mu**, aka**mu**, oku**mu**, or eri**mu**. The adjective stem **–mu** represents the value, *one*.

A. Luganda Number System and Semiotic Classes

The Luganda word order differs from the English one in that the number always comes after the noun it qualifies. For instance, **10 kg** is read as **kkirogulaamu kkumi (kg 10)** and not as *ten kilograms*. Therefore, our algorithm looks at the standard word before the digit symbol and not after it. This is an important rule to remember when doing semiotic and noun classification.

Table III illustrates some number ranges in Luganda, their parts-of-speech (PoS), the noun class to which they belong, some specific examples and that large numbers are usually expressed as a combination of multiples (of powers of ten) and additions of small numbers [6], [21]–[23]. In general, the numbers that are larger than five are invariable nouns and so always belong to the same noun class [6], [22]. The only exception to this is if the number contains a 1, 2, 3, 4, or 5 in the ones position. For instance, numbers like 6**1**, 8**3** and 2,00**4**. The numbers 1 to 5, whenever they appear in the ones position, are adjectives and therefore must belong to the same noun class as the noun they qualify. For instance, **Aba**ntu **ama**kumi **a**biri mu **ba**taano, *twenty five people* is factored as ((2*10) + 5) people. Note that the adjective **ba**taano, *five* agrees with the class of the noun, a**ba**ntu, *people*. The noun stem **-kumi** represents powers of ten and **Amakumi** specifically refers to multiples of ten that are less than 60. The adjective stem **–biri** is two and the **–taano** is five. **Ba-** and **(a)ba-** are prefixes of the plurals of noun class I and **(a-, ama-)** are prefixes of the plurals of noun class V as shown in Table II. In this context, **mu** and **na** are equivalent to "and" in English. The noun stems **-kaaga**, *six*, **-sanvu**, *seven*, **-naana**, *eight* and **-enda**, *nine* are multiplied by powers of 10 to create specific multiples of the number and specific noun class. For instance, **mukaaga** (6), **nkaaga** (60), **lukaaga** (600), **kakaaga** (6000) as shown in Table III.

*1) **Cardinals**:* The Luganda cardinals differ from the English ones and so the normalization must be done differently. For instance, the number 12 is verbalized **kkumi na bbiri** (10 with 2 ones) as opposed to a single word, *twelve* in English. Another example is the number 23 that is verbalized as **amakumi abiri mu ssatu** (2 Tens and 3 ones) as opposed to *twenty-three in English*.

*2) **Ordinals**:* The Luganda ordinals also differ from the English ones and must therefore be normalized differently. The English word, first, can be translated as an adjective stem **–beryeberye** or verb stem **–sooka** and used with a possessive and must agree with the noun class. For instance, in Luganda, **Ssekabaka Muteesa I** is verbalized as **Ssekabaka Muteesa ow'oluberyeberye** *The late king Muteesa of the first* instead of *Late king Muteesa the first*. Similarly, **Ekyasa I** is verbalized as **Ekyasa ekyasooka**, *the century that was first*.

The ordinal numbers second to fifth are created by taking the possessive, adding the prefix and then the number stem to it. For instance, **Muteesa V** is verbalized as **Muteesa ow'okutaano** *Muteesa of the five* instead of *Muteesa the fifth*.

The rest of the ordinal numbers are created by taking the possessive and adding the number stem to it. For instance, **Muteesa VII** is verbalized as **Muteesa ow'omusanvu**, *Muteesa of the seven* instead of *Muteesa the seventh*.

In each of these cases the possessive must agree with the noun class. Therefore, the algorithm must classify the noun correctly. These rules also apply to non-standard words used in dates, time, and fractions.

*3) **Measurements**:* The order of reading measurements for both Luganda and English also differ. For instance, *1 kg* in Luganda is read as **kkirogulaamu emu** as opposed to one kilogram **emu kkirogulaamu** in English.

[1]https://github.com/rkizito/luganda/blob/master/sampleForNormalization.txt

TABLE I
SEMIOTIC CLASSES COVERED BY THIS PAPER

Semiotic Class	Description	Example of input (Luganda)
ABBREVIATION	Abbreviation	Mw. (Mwami), Owek.(Oweekitiibwa)
LETTERS	Sequence of letters	UN (U N), FM (Ffa Mma)
CARDINAL	Cardinal numbers	12 (Kkumi na bbiri), 3 (Ssatu)
DATE	Date expression	22/5/2018, 2-01-2019
MEASUREMENT	Unit Symbols	3 kg (Kkirogulaamu ssatu), 5'(Yiinci ttaano)
CURRENCY	Currency symbol	$2.50, Ksh 7.38, Ugx 34,000
TELEPHONE	Telephone numbers	+941234567, 0786346968
TIME	Time expression	4:55, 11.33
ORDINAL	Ordinal numbers	Muteesa II, (Muteesa Ow'okubiri)

TABLE II
NOUN CLASSES USED IN LUGANDA COUNTABLE NOUNS, WITH EXAMPLES

class		Examples (Singular)	Examples (Plural)
Number	(Vowel) Prefix		
I	(o)mu-, (a)ba-	Omulenzi, omuntu, omusajja	Abalenzi, abantu, abasajja
Ia	Ba-	Kabaka, makanika, nnaalongo	Bakabaka, bamakanika,bannaalongo
II	(O)mu-, (E)mi-	Omucungwa, omuti, omwalo	Emicungwa, emiti, emyalo
III	(E)n/m-, (E)n/m-	Ente, embwa, embuzi	Ente, embwa, embuzi
IV	(E)ki-, (E)bi-	Ekitabo, ekintu, ekikonde	Ebitabo, ebintu, ebikonde
V	(E)ri-, (A)ma-	Erinnya, eryato, eriiso	Amannya, amaato, amaaso
VI	(A)ka-, (O)bu-	Akabwa, akantu, akaana	Obubwa, obuntu, obwana
VII	(O)lu-, (E)n-	Oluguudo, oluggi, olunaku	Enguudo, enziggi, ennaku
VIII	(O)gu-, (A)ga-	Oguntu, ogukazi, ogusajja	Agantu, agakazi, agasajja
IX	(O)ku-, (A)ma-	Okugulu, okutu	Amagulu, amatu
XI	(A)wa-, -	Awantu	-

TABLE III
NUMBER RANGES IN LUGANDA

Numbers	Part-of-Speech	Noun Class	Examples
1-5	Adjective	Noun's class	Ekintu kimu (1 thing), ente emu (1 cow), omuntu omu (1 person)
6-9	Noun	II	Ebintu musanvu (7 things), ente musanvu (7 cows)
10	Noun	V	Abantu kkumi (10 people)
11-15	Noun + Adjective	V + Noun's class	Abantu kkumi na babiri ((10 + 2) people)
16-19	Noun + Noun	V + II	Ebintu kkumi na mukaaga ((10+6) people)
20-50	Noun	V	Abantu amakumi abiri ((2*10) people)
60-70	Noun	III	Abantu nsanvu ((70) people)
80-90	Noun	IV	Abantu kinaana ((80) people)
100-500	Noun	IV	Abantu ebikumi bitaano ((5*100) people)
600-900	Noun	VII	Abantu lukaaga ((600) people)
1,000-5,000	Noun	VII	Abantu enkumi bbiri ((2*1,000) people)
6,000-9,000	Noun	VII	Abantu kakaaga ((6,000) people)
20,000-900,000	Noun	II + Noun's class	Emitwalo ebiri (2*10,000), abantu emitwalo lukaaga ((600*10,000) people)
2,000,000,000 and above	Noun	VI + Noun's class	Obuwumbi bubiri (2*1,000,000,000), Obuwubi lukumi (1000*1,000,000,000)

*4) **Currency**:* Luganda word order still applies to the way currencies are read. The name of the currency is always read first. For instance, in Luganda, £50 is read as **Pawundi amakumi ataano**, *pounds fifty* as opposed to fifty pounds in English or the way measurements such as 50 kg are read. The Luganda hand-built rules are therefore responsible for interchanging the currency symbol and value accordingly. 5 USD is read as **Ddoola z'Amerika ttaano**, *Dollars of America five* as opposed to five United States (of America) dollars (**ttaano Ddoola z'Amerika**). Note that the adjective, **ttaano** qualifies the noun, **Ddoola** and not the name, **Amerika**. Therefore, the text normalization algorithm must recognize this and put the adjective in the correct noun class. 100.50/= is read as **Ssiringi kikumi n'ennusu amakumi ataano**, *Shillings 100 and Cents fifty* and handled the same way we deal with a measurement NSW.

*5) **Abbreviations**:* The Luganda abbreviation non-standard words are constructed in similar way as the English ones. The first letter is usually capitalized, optionally followed by small letters and a full stop. For instance, **Mw.** represents the word **Mwami**, *Mister*. Some abbreviations like **kg** short for **kkirogulaamu**, do not obey Luganda morphological rules and so can also be detected that way. For instance, all Luganda words end in a vowel. So if the last character in a word is not a vowel then it is an NSW especially an abbreviation as in the example above.

*6) **Time**:* In Luganda the day starts at 7:00 am (roughly sunrise in Kampala, Uganda) and ends at 6:59 pm (roughly sunset in Kampala, Uganda). Therefore **Ssaawa emu**

ey'enkya, *hour one of the dawn* is equivalent to 7:00 am. Similarly the next 12 hours are roughly considered to be night. Therefore **Ssaawa bbiri ez'ekiro**, *hour two of the night* is 8:00 pm. Therefore Luganda is based on a 12-hour system and not a 24-hour system. So, any TTS/computer system using "English language" (e.g Coordinated Universal Time, UTC or East African Time, EAT) time must convert that time to the Luganda time system. Our text normalization module offers the option of doing that conversion. The time divisions in Luganda differ from those used in English. 7 am to 11 am is **ez'enkya**, *of morning*, 1 pm to 5 pm is **Olw'eggulo**, *of afternoon or evening*, 6 pm is **Akawungeezi**, *dusk*, 7 pm to 5 am is **ez'ekiro**, *of night*, 6 am is **ku makya**, *at dawn*. The phrase "**ssaawa 5 ez'ekiro**" is normalized as **ssaawa ttaano ez'ekiro**, *hour five of the night* i.e 11 pm. The phrase "**5:26 ez'enkya**" is normalized as **eddakiika abiri mu mukaaga eziyise ku ssaawa ttaano ez'enkya**, *twenty six minutes have passed hour five of the morning* i.e 11:26 am.

*7) **Dates**:* Dates in Luganda texts are generally written as sequences of numbers separated by hyphens or slashes. The day, month and year are read as ordinal numbers. For instance, 1-6-1999 or 1/6/1999 is verbalized as **olunaku olusooka, omwezi ogw'omukaaga, omwaka ogwa lukumi mu lwenda mu kyenda mu mwenda**, *the first day, the sixth month, the year of one thousand and nine hundred and ninety-nine*. The hyphens and slashes are silent.

*8) **Telephone numbers**:* In general, the telephone numbers that appear in Luganda text are not written with hyphens, brackets or plus signs but have a relatively standard number of digits. For instance, 070278319 is verbalized as a counting sequence of cardinal digits **zeero, musanvu, zeero, bbiri, munaana, ssatu, emu, mwenda**. The adjectives 1 to 5 belong to the noun class III. Every now and then command symbols like * and # are used and they are also read as words.

*9) **Letters**:* As is the case with English, a sequence of capital letters can be read in a number of ways which creates ambiguity. FM (short for Frequency modulation) is read one letter at a time as **ffa mma** (i.e the Luganda letter pronunciation) but DP (short for Democratic party) is read one letter at a time as **ddi ppi** (i.e the English letter pronunciation). ISO can be read as a sequence of letters **ayi essi o** (i.e English letter pronunciation short for International Organization for Standardization) or as the word **iso** (i.e as a normal Luganda word and short for the Internal Security Organization). Normalizing this class of text will be done in the future.

B. Detection and Classification

The detection and semiotic classification of non-standard words in Luganda text was achieved using Python regular expressions [3]. The detected NSW is tokenized as a single unit and also uses whitespaces. For instance, "10 kg" is treated as a single token not as two tokens. Further splitting is done within the verbalization module. Note that some authors [13] use a list of words from an English dictionary to find NSWs. This cannot work for agglutinating languages like Luganda because sentences can be combined to form one "word". For instance, **Ndikimuguliranga**, *i will always buy it for him* is one "word/sentence" but would not appear in a dictionary. NSWs that belong to the time, cardinal, ordinal, currency, date, telephone and measurement classes are then classified into the correct noun class. For cardinals numbers that have 1 to 5 in the ones position, NSW noun classification is based on the initial vowel (if it exists) and the prefix of previous token. For instance **Emifaliso 2**, the initial vowel **E** and the prefix **mi** help to determine that the number **2** falls in the noun class II (see Table II). The rest of the numbers are invariant nouns and are handled as described in subsection III-A. In general, currencies and SI units like kg, ft, mm, V, A, and mL that were borrowed from English words belong to noun class III. This simplifies noun classification.

Table IV illustrates some of the python regular expression patterns we created for detection and classification [3].

TABLE IV
SOME OF THE PATTERNS USED FOR DETECTION AND CLASSIFICATION

Pattern used to detect and classify	Semiotic class after detection and classification
$r'\backslash d\{1,2\}:\backslash d\ \{2\}$'	Time
$r'(?[A-Z]\ \{3\}\ \backslash s)\ (?\ \backslash d\ (?\ [.,]?\ \backslash d^*)\ *)$'	Currency
$r'\backslash b$ (? [a-z] *[A-Z] [a-z] *) $\{2,\}$'	Abbreviations

C. Verbalization and Replacement

Verbalization was achieved by developing handcrafted ad-hoc rules that are specifically for the described semiotic classes. It is these rules that are embedded into the sub-modules that handle verbalization after classifying the semiotic and noun class. The NSW in the semiotic class is then replaced with defined standard Luganda words. For instance, 8:27, would be detected as a non-standard word, then classified to fall in the time semiotic class. It's now the time verbalization sub-module that expands it into its defined standard Luganda words i.e. **ssaawa munaana n'eddakiika abiri mu musanvu**. On the other hand, abbreviations and letter sequences are verbalized using a look-up table.

IV. EVALUATION

As the first stage of the normalization, the accuracy and efficiency of the detection was analyzed using seven different test samples of Luganda text. For each test sample, we give the number of NSWs, correctly detected NSWs, undetected NSWs, the falsely detected NSWs, and the detection percentage accuracy. Table V shows the NSWs detection accuracy and efficiency, using the seven test samples with the system, depends on the number of correctly detected NSWs and the total NSWs in each sample. The percentage of the accuracy and efficiency of detection decreases with increase in the number of non-standard words. A simple precision metric was used for evaluation of detection.

$$\text{Precision} = \frac{\text{Correctly detected NSWs}}{\text{Correctly detected NSWs + undetected NSWs}}$$

The evaluation for the detection yields a 91.2% accuracy for the first sample. However, as the number of text samples increases, the accuracy percentage decreases up to 75.1%, with increase in false positives, increase in true positives and increase in the falsely detected NSWs from each sample tested.

As an example, the first bold text block below shows some sample input text that was tested. The NSWs detected are highlighted in italics.

Musoke, ye muddusi omusajja eggwanga gwe litunuulidde okulikiikirira mu *24 m*, kyoka naye obudde bukyamulemye okutuusa. Abantu *24* ku buli *100* be twayogedde nabo baagambye nti ebipipa omusuulwa kasasiro babirina naye tebimala. Nze Kaggwa. Mbeera Jinja. Nnina emyaka *24*.

(Musoke, is the runner that the country is looking to in the 24 m, but he has failed to meet the time. 24 out of the 100 people that we talked to said that they have dust bins but they are not enough. I am Kaggwa. I live in Jinja. I have 24 years).

When the block of Luganda text above is fed into our Luganda text normalization module, the output below is generated.

Musoke, ye muddusi omusajja eggwanga gwe litunuulidde okulikiikirira mu *mmita abiri mu nnya*, kyoka naye obudde bukyamulemye okutuusa. Abantu *abiri mu bana* ku buli *kikumi* be twayogedde nabo baagambye nti ebipipa omusuulwa kasasiro babirina naye tebimala. Nze Kaggwa. Mbeera Jinja. Nnina emyaka *abiri mu ena*.

The second text block shows that although the input text contains three examples of the same number 24, the normalized output gives three different verbalizations. Each verbalization is dependent on the context in which the 24 is used as well as the noun class of the word before it. As expected the adjective for cardinal number 4, **-na**, agrees with the noun it qualifies. So we have 3 different noun classes and three different adjective classes (giving **nnya, bana** and **ena**)

The system was tested a number of times with Luganda data from various sources. The same precision metric used in detection, was also used to evaluate the verbalization. From Table VI, it is seen that as the number of samples increases with increase in correctly detected NSWs, the accuracy drops from 90.8% to 69.6% because the number of incorrectly detected NSWs also increased.

V. Future work

Luganda is a complex language with many noun classes and some non-standard Luganda words falling into more than one semiotic class. For example, radio frequencies such as *93.3 KFM* can be read as **kyenda mu ssatu katonyeze ssatu**, *Ninety three point three* not as a decimal **kyenda mu ssatu n'obutundu busatu**, *Ninety three and three parts*. As a result, these tokens decrease the accuracy and efficiency of the system quite significantly. Some Luganda text comprises of many different letter sequences from other languages especially English. Some of these abbreviations do not have corresponding Luganda standard text. For example, IEEE, UIPE, H.E, NRM, and DP. Some people read them as English letters while others read them as Luganda letters. Lastly, compound tokens such as Bible verses e.g. **Mat. 5:21-48** short for **Matayo**, *Matthew*, are misinterpreted. This is because the text normalization module is unable to correctly classify such tokens. For instance, **Mat. 5:21–48** can be incorrectly classified as **time 5:21, *ssawa ttaano n'eddakiika abiri mu emu*** and at the same as **subtraction, 21-48, *abiri mu emu yawulako ana mu munana***. These kinds of errors reduce the system's accuracy.

For those reasons, we plan to add more rules to cater for NSWs that fall under more than one semiotic class and to add other semiotic classes found in Luganda language. Luganda requires significantly more time to develop than, say, English, due to the complexity of morphology. Future work will include the building of annotated datasets that can be used in training machine learning models in order to perform noun classification, abbreviation and cardinal numbers normalization.

Acknowledgment

We would like to acknowledge the contributions of the following to the development of the Luganda text normalization module.

First and foremost, We would like to express our deepest thanks to the people below for their vital input along the way: Mr. William Kibirango, Dr. Deo Kawalya, Mr. Solomon Elton Muwebwa, Ms. Phoebe Katwesigye, Mr. Samuel Olwe, Mr. Methodius Uwizera, Dr. Richard Sproat and Dr. Fridah Katushemererwe.

We would also like to express our indebtedness appreciation to the management and members of netLabs!UG for providing necessary resources and guidance during the course of this research.

Lastly, the staff of the Department of Electrical and Computer Engineering Makerere University for their management and academic support.

References

[1] P. Taylor. *Text-To-Speech Synthesis*. Cambridge University Press, Cambridge, UK, 2009.

[2] R. Sproat et al. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333, 2001.

[3] D. Jurafsky and H. J. Martin. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2009.

[4] X. Huang et al. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, Inc, 2001.

[5] I. Nwakanma et al. Text–To–Speech Synthesis (TTS). *International Journal of Research in Information Technology*, 2(5):154–163, May 2014.

[6] J. D. Chesswas. *The Essentials of Luganda*. Oxford University Press, London, UK, 3 edition, 1963.

[7] Navdeep Jaitly and Richard Sproat. An rnn model of text normalization. https://arxiv.org/pdf/1611.00068.pdf, 2017.

[8] R. Sproat. Lightly supervised learning of text normalization: Russian number name. In *IEEE Workshop on Spoken Language Technology*, pages 436–441, Berkeley, California, 2010.

[9] K. Sodimana et al. Text Normalization for Bangla, Khmer, Nepali, Javanese,Sinhala and Sundanese Text-to-Speech Systems. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 147–151, Gurugram, India, August 2018.

TABLE V
DETECTION ACCURACY AND EFFICIENCY

Test data	NSWs in data	Correctly detected NSWs	Undetected NSWs	Falsely detected NSWs	Detection accuracy (%)
1	239	218	21	31	91.2
2	586	523	63	57	89.2
3	1001	844	157	64	84.3
4	1402	1109	293	111	79.1
5	1697	1332	365	125	78.5
6	2101	1614	487	176	76.8
7	2684	2017	667	291	75.1

TABLE VI
ACCURACY AND EFFICIENCY ANALYSIS FOR VERBALIZATION

Test data	Correctly detected NSWs	Properly normalized NSWs	Improperly normalized NSWs	Normalization accuracy (%)
1	218	198	20	90.8
2	523	432	91	82.6
3	844	662	182	78.4
4	1109	853	256	76.9
5	1332	997	335	74.8
6	1614	1146	468	71.0
7	2017	1404	613	69.6

[10] B. Plank. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, volume 16, pages 13–20, Bochum, Germany, 2016.

[11] D. Supranovich and V. Patsepnia. IHS_RD: Lexical normalization for English tweets. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 78–81, Beijing, China, jul 2015.

[12] J. D. Murphy. *Luganda-English dictionary*. Catholic University of America Press, Washington, D.C., 1972.

[13] E. Flint et al. A Text Normalisation System for Non-Standard English Words. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 107–115, Copenhagen, Denmark, Sep 2017.

[14] D. van Esch and R. Sproat. An expanded taxonomy of semiotic classes for text normalization. In *Proceedings of Interspeech 2017*, pages 4016–4020, 2017.

[15] P. Ebden and R. Sproat. The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):333–353, May 2015.

[16] A.M. Hlaing et al. Myanmar Number Normalization for Text-to-Speech. In *15th International Conference of the Pacific Association for Computational Linguistics, PACLING*, volume 781, pages 263–274. Springer, August 2017.

[17] I. Nandutu. Luganda Text-to-Speech Machine. Unpublished master's thesis, Uganda Technology And Management University, Kampala, Uganda, 2017.

[18] A. M. Oirere et al. Swahili text and speech corpus: a review. *Asian Journal of Computer Science and Information*, 2(11):286–290, November 2012.

[19] M. Gakuru et al. Development of a kiswahili text to speech system. In *Interspeech 2005*, pages 1481–1484, Lisbon, Portugal, September 2005.

[20] F. Katamba. A Non-Linear Analysis of Vowel Harmony in Luganda. *Journal of Linguistics*, 20(2):257–275, September 1984.

[21] R. J. Hurford. *The Linguistic theory of numerals*. Cambridge Studies in Linguistics (Book 16). Cambridge University Press, Cambridge, UK, 2010.

[22] E. O. Ashton et al. *A Luganda Grammar*. Longmans, Green and Co., London, 1954.

[23] S. Ritchie et al. Unified verbalization for speech recognition & synthesis across languages. In *Proceedings of Interspeech 2019*, pages 3530–3534, 2019.

Ronald Kizito received the B.Sc., M. Sc., and Ph.D degrees in electrical engineering from Michigan Technological University, Houghton, MI, USA in 1997, 1999 and 2005 and the B. S. degree in business administration from Michigan Technological University, Houghton, MI, USA in 1999. From 1997 to 2004, he was a research assistant at Michigan Technological University. He is currently a Lecturer in the Department of Electrical and Computer Engineering, Makerere University. His research interests include digital speech processing, Luganda natural language processing, spoken dialog systems, and electronic systems.

Wayne Steven Okello received the B.Sc. degree in Computer Engineering from Makerere University, Uganda in 2020. He is currently a Graduate Research Assistant at netLabs!UG a research Centre of Excellence in the Department of Electrical and Computer Engineering, Makerere University. His research interests include Luganda natural language processing, distributed information systems, data mining and database systems, cloud computing and cyber security. He was a winner of the student award at the Uganda Institute of Professional Engineers Annual Dinner Awards 2019.

Sulaiman Kagumire received the BSc degree in Computer Engineering from Makerere University, Uganda in 2020, and the certificate (with honours) in Machine Learning and Statistical Analysis(Applied Data Science) from WorldQuant University, New Orleans, LA, USA in 2019. He is currently a Graduate research assistant at netLabs!UG, Makerere University. He won the best Student Project award at the Uganda Institute of Professional Engineers Annual Dinner Awards 2019. His research interests include Machine Learning and Data mining, Luganda Natural Language Processing, and Communications, Digital signal processing and control.