18

Chapter 1

Toward the development of a crowd-sourced lexical database for Lubukusu: A focus on nouns & noun classes

Zuzanna Fuchs
University of Southern California

Travis Major
University of Southern California

Justine Mukhwana Sikuku
Moi University

Catherine Pan
University of Southern California

The present paper introduces a preliminary lexical database for Lubukusu (JE31c). The database presents information on a large number of nouns to enable future psycholinguistic and formal research on nouns and noun classes. We outline the crowd-sourcing approach used for data collection, wherein many participants each provided relatively small amounts of data. The resulting database has written forms of words that are representative of how native speakers are likely to use or encounter them in daily life. This resource will enable experimental research involving written stimuli, and it will support efforts to create a standardized orthography for the language.
Keywords: database, lexicon, crowd-sourcing, Lubukusu, noun classes

Change author. Change title. To appear in: Change Volume Editor. Change volume title. Berlin: Language Science Press. [preliminary page numbering]

1 Introduction

Access to comprehensive language resources is critical for language research. In particular, experimental and psycholinguistic research requires language resources that enable the researcher to control for experimental factors such as word length or frequency. Although experimental work is still uncommon in African linguistics, recent work in this domain (e.g. Kgolo & Eisenbeiss 2015, Ciaccio et al. 2020, Lawyer et al. 2024, Kanampiu, Martin & Culbertson 2025, Kanampiu, Martin, Culbertson & Kanampiu 2025) demonstrates the significance of experimental work on African languages to theories of language and language processing.

We highlight one challenge in creating language resources that might support such experimental work on more African languages: some languages do not have a standardized orthography. Notation in existing databases for such languages is typically done in the International Phonetic Alphabet or in a notational system designed by a linguist or lexicographer. This is useful for certain areas of formal linguistic research, but successful experimental work requires presenting native speakers with forms they recognize as words in their language.

The present manuscript presents one approach to tackling this challenge. Using a crowd-sourcing approach, we construct a lexical database for Lubukusu, a Bantu language spoken in Kenya that does not have a standardized orthography. Speakers of Lubukusu are nevertheless typically literate in Swahili and/or English and use the Roman alphabet to represent Lubukusu words in informal contexts such as social media interactions, as discussed in Section 2.1. By crowd-sourcing data on how words are spelled by native speakers themselves, we are able to observe converging population-level trends in spelling and identify the written forms of words that are most likely to be encountered by native speakers in daily life.

We present here a preliminary phase of this database, currently constrained to nouns, given our concurrent experimental investigations of noun classes (e.g. Fuchs et al. under review). Section 2 provides background on Lubukusu, its noun classes, and existing resources. Section 3 outlines the crowd-sourcing approach, and Section 4 details how data from 32 speakers was processed, yielding a database of nearly 700 nouns, with corresponding noun class information and other measures useful for controlling various experimental factors. Section 5 discusses the opportunities and challenges of this approach, including the potential for this resource to support work on the development of a speaker-based standardized

orthography.

2 Background

2.1 Lubukusu and its speakers

Lubukusu (JE31c) belongs to the Luhya subgroup of Bantu languages. Lewis et al. (2016) list the population of Lubukusu speakers at 1,433,000 based on the 2009 census. The 2019 census lists only slightly over 1 million speakers. The disparity is perhaps related to the fact that in the latter census, people had an option not to disclose their linguistic identity. Lubukusu speakers predominantly occupy Bungoma and Trans Nzoia counties in Kenya's western region.

The status of Lubukusu speakers' exposure to written Kiswahili, English, and Lubukusu is characterized by a hierarchical dominance, with English and Kiswahili generally having greater written presence and utility compared to Lubukusu. English and Kiswahili hold a significant position in Kenya as official languages utilized in education, government, and formal business sectors. In addition, Kiswahili is the major national lingua franca and is taught as a compulsory subject in schools. Consequently, Lubukusu speakers, like other Kenyans, are exposed to written English and Kiswahili through various channels, including educational materials from primary school through university, where English is the medium of instruction for many subjects, media (newspapers, magazines, online content), official documents, legal texts, and international communication. As a local language, Lubukusu is primarily used for oral communication within the Bukusu community. In contrast to English and Kiswahili, the exposure of Lubukusu speakers to written Lubukusu is considerably more limited to religious texts such as the Bible, educational material in a few schools with a strong Lubukusu presence, and social media interactions.

2.2 Nouns & noun classes

Because one of the goals of the database is to present information regarding noun classes, we briefly introduce the noun class system of Lubukusu here. It is well known that most Bantu languages exhibit a rich system of noun classes, the number of which varies across languages in the family, but most languages have between 12 and 21 classes (Nurse & Philippson 2003). Lubukusu has 19 noun classes, as exemplified in Table 1.

Table 1: Noun classes in Lubukusu; example words are presented in modified orthography

Class	Gender	Number	Pre-prefix	Prefix	Example	Translation
1	A	SG	o-	mu-	omuundu	'person'
2		PL	ba-	ba-	babaandu	'people'
3	B	SG	ku-	mu-	kumukhono	'hand'
4		PL	ki-	mi-	kimikhono	'hands'
5	C	SG	li-	li-	liliino	'tooth'
6		PL	ka-	ma-	kameeno	'teeth'
7	D	SG	si-	si-	sisiindu	'thing'
8		PL	bi-	bi-	bibiindu	'things'
9	E	SG	e-	n-	enyungu	'pot'
10		PL	chi-	n-	chinyuungu	'pots'
11			lu-	lu-	luluuchi	'river'
12			kha-	kha-	khakhaandu	'small thing'
14			bu-	bu-	bubuukhi	'honey'
15			khu-	khu-	khukhwaasaaka	'to split'
16a			a-	-	aanju	'at/by the house'
16b			sya-	-	syaanju	'toward the house'
17			khu-	-	khuunju	'on the house'
18			mu-	-	muunju	'in the house'
20			ku-	ku-	kukuundu	'big thing'
23			e-		ebung'oma	'at Bungoma'

Note that Lubukusu nouns mark noun class with both a pre-prefix and a prefix. The project that the present database feeds into is focused on abstract gender that is responsible for linking singular and plural within each pair of classes Carstens (1991). For this reason, our database primarily targets nouns that fall within the paired classes (1-10). As with most other Bantu languages, most noun classes have a semantic core of nouns that are consistent with certain semantic properties, but there are many exceptions within each class, ultimately making class membership an idiosyncratic property of nominals.

2.3 Existing Lubukusu resources

While efforts exist to promote local languages in Kenya, written materials in Lubukusu are generally scarce. These might include some religious texts (e.g., portions of the Bible), cultural publications, or educational materials for early primary grades in specific regions where mother tongue instruction is implemented.

Though there is no written grammar of Lubukusu, a number of papers have made important contributions regarding structural properties of the language, such as phonology and morphology (Mutonyi 2000), wh-movement (Wasike 2007), locatives (Diercks 2011), anaphoric constructions (Sikuku 2011), adjectival constructions (Wasike 2018), and object marking (Sikuku & Diercks in press). Also from the structural perspective, The Aphranaph Project[1] has publicly available data on various anaphoric constructions in Lubukusu as well as in other African languages.

There are a few existing resources for the lexicon of Lubukusu. The first is the Bukusu-English dictionary (Marlo et al. ms). This database was generated by a native-speaker co-author (Sifuna) translating the word list from the Luhya-English Vocabulary resource (Appleby ms), and combining the outcome with a number of other existing smaller resources, including Mutonyi (2000)'s dissertation and De Blois (1975)'s Bukusu wordlist, as well as the Bukusu lexicon from the Comparative Bantu Online Dictionary. All words were then revised to be presented using a single Roman-alphabet-based transcription system that marks tones and vowel length. Parts of speech and definitions are also provided. However, the authors note that transcriptions and/or accuracy of some lexical items were not verified with other native speakers of the language.

A more recent lexicon has been compiled (Marlo et al. n.d.) for the purposes of phonological research and is publicly available on Marlo's academic website. Transcription was agreed upon by the co-authors and spellings were determined by the two co-authors who are native speakers of the language. The database tracks many properties that researchers in phonology may find useful, such as tonal marking, syllable count, syllable length, fine-grained tonal properties, and the presence of enclitics, among others.

2.4 Research goals & approach

While the resources listed above constitute valuable resources for some domains of linguistic research, they remain insufficient for experimental work on Lubukusu. We discuss here important desiderata for a database that would enable such experimental work and thus what the goals are for the database presented in the current project.

The first desideratum is to compile a database that presents words in the written form that native speakers are likely to encounter in daily life. The ex-

[1]Available at https://afranaphproject.afranaphdatabase.com/

isting resources for Lubukusu are compiled by and oriented toward other researchers, encoding in their transcriptions many properties useful for linguistic research, but not necessarily reflecting how a native speaker would spell those words or see them written. Even when one or two native speakers (often themselves trained linguists) have been involved in the preparation of the database, that may not be representative of how larger portions of the population spell those same words. Meanwhile, having recognizable written words is critical for linguistic studies, since for the study to work, when participants see a string of letters they must reliably recognize it, i.e., be able to map that onto a lexical item in their mental lexicon and subsequently access connected information such as the word's meaning or grammatical properties. Our primary goal, then, is to identify the most common spelling of Lubukusu words, as used by native speakers of Lubukusu.

The second desideratum for a lexical database supporting future experimental work is some measure of relative frequency. The existing dictionaries and lexicons for Lubukusu contain large numbers of words, but they do not provide information regarding which words occur more or less commonly in daily use. It is well established that participants in language experiments react differently to words that they are more familiar with than those they are less familiar with, in a way that may interfere with the validity of experimental results. Existing studies of real-time processing on languages with no standardized orthography like Setswana (e.g. Kgolo & Eisenbeiss 2015, Ciaccio et al. 2020), demonstrate the success and impact of using proxy measures of frequency to inform experimental design. Our secondary goal, then, is to obtain some proxy measure(s) of frequency for the words in our Lubukusu database.

Our approach to achieving these goals is to employ crowd-sourcing. In traditional fieldwork, large amounts of language data and judgments are collected from a small number of speakers; crowd-sourcing collects small amounts of data from a large number of people. Specifically in our case, each participant provides a relatively small number of nouns and their spellings. One advantage of this approach is that we can compare the spellings of the same word provided by multiple speakers, thus allowing us to identify population-level patterns and trends. This ensures that the database is representative of how a larger proportion of the population represents their language in written form. A second advantage is that this provides a natural, albeit coarse, measure of relative frequency: by calculating which words are provided more often, we can obtain a preliminary frequency measure. An additional benefit of this approach is that it makes large-

scale data collection more feasible, because no individual is being given a large, time-consuming task; rather, each individual who participates in data collection completes a short task in one sitting.

In the current phase of the database, we constrain the domain to nouns. This ensures feasibility and allows us to test the effectiveness of our data collection tool and procedure prior to large-scale data collection on other lexical categories, while still creating a resource that already enables experimental work pertaining to nominals. Consistent with our team's current and planned experimental work on noun classes in Lubukusu (Fuchs et al. under review), we prioritize nouns that occur in paired singular/noun classes (cf. Section 2.2), and nouns that are likely to be imageable, i.e. easy to represent as a single image, to support a variety of experimental methodologies.

3 Data collection methodology

3.1 Design

The structure of the data collection tool was modeled on (semantic) verbal fluency tasks, which are commonly used in the psycholinguistic literature to assess vocabulary knowledge (Hall et al. 2010). In these tasks, participants are shown a semantic category (ex. "fruit", "clothing") and given a set amount of time (typically 30 or 60 seconds) to verbally name as many words within that category as they can. This is repeated for some number of categories. Research on verbal fluency tasks suggests that highly frequent nouns are more likely to be provided more often than nouns that are infrequent (Taler et al. 2020).

Our task makes a number of modifications to the traditional verbal fluency task. Our data collection tool was untimed, because our priority was collecting more tokens rather than testing participants' ability to provide tokens quickly. The task was also written rather than oral, because this aligned with our goals of identifying most common written forms of words. Finally, because our research interests pertain to noun classes, the task involved providing not only the singular form of a noun, but also its plural. Participants were also asked to provide an English translation or description; this was deliberately broad, under the assumption that not all Lubukusu words have a direct translational equivalent in English, or that some English translational equivalents may not be known to a participant.

The categories used in the data collection tool are presented in Table 2.

Selection of semantic categories was guided by experimental goals, prioritizing categories that were likely to yield concrete and picturable nouns that occur in the paired noun classes. We also aimed to make the tool flexible to accommodate varying enthusiasm and time commitments from participants. We therefore divided the categories into three bins (see Table 2), where Bin 1 contained categories that would be obligatory for participants, and Bins 2 and 3 contained categories that would be optional, completed only if participants opted to continue the task (see below).

3.2 Procedure

The study was implemented in Qualtrics. Participants viewed one category at a time and were provided with twenty rows of blanks, allowing them to provide up to 20 responses (Figure 1); they were encouraged but not required to provide at least 10. Each row contained three blanks: one for the Lubukusu word in the singular, one for the corresponding plural word in Lubukusu, and one for the English translation or description. Participants were allowed to leave blanks within a given row, as no fields were required.

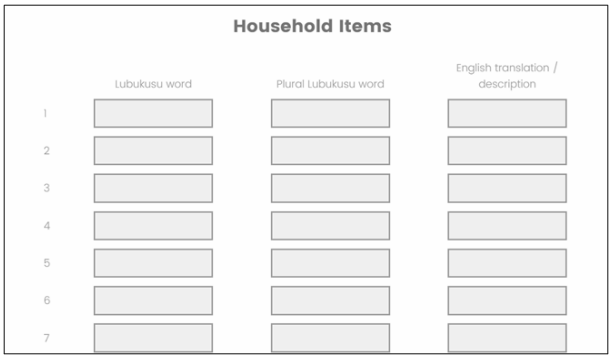

Figure 1: Sample survey screen (partial view; screen extends to include 20 rows)

Each participant viewed four categories from Bin 1 – these four categories were randomly selected for each participant. They were then informed that they could end their participation in the study, or they could proceed to see four more categories. These next four categories were randomly selected for each participant from Bin 2. If they opted to continue, after the second round of four categories, they were once again given the choice to end the study or complete four more categories. This final set of categories were randomly selected for each par-

ticipant from Bin 3. The study ended after the third round. Participants were then presented with space to optionally provide anonymous feedback regarding their experience.

3.3 Participants

Participants were recruited via word of mouth in Uasin Gishu County and Bungoma County in Kenya. Prior to data collection, the authors obtained a research permit from Kenya's National Commission for Science, Technology and Innovation (NACOSTI) for the legal and ethical conduct of studies involving human subjects in Kenya. Participants were compensated $10 USD for their participation.

3.4 Results

We included in analysis 32 survey responses with at least one word each. From those responses, 1172 response rows were recorded. Two response rows contained only an English translation and no Lubukusu word, so they were excluded from further processing.

4 Data processing & construction of database

4.1 Structure of database entries

The intended structure of the current form of the database is for each entry to represent one noun in Lubukusu and contain several pieces of information about it that are straightforwardly computed from participant responses: the most commonly provided spelling(s) (of its singular and plural forms), alternate spellings, its most commonly provided English translation, its semantic category/ies, and a raw count of the number of times it was provided by participants (referred to as "occurrences"). Additionally, each entry contained three derived measures: the lengths of the most commonly provided spellings of the singular and plural forms, the noun classes of the singular and plural forms, and an approximate measure of frequency. Figure 2 illustrates this structure.

Table 2: Number of respondents and response rows per semantic category

Bin	Category	# Respondents	# Total Response Rows
1	Clothing & jewelry	19	194
	Cooked foods	14	149
	Fruits	11	87
	Furniture	8	46
	Household items	15	196
	Human body parts	19	298
	Vegetables	17	164
2	Animal products	5	59
	Domestic animals	5	48
	Insects	3	33
	Plants & flowers	6	62
	Things to eat	5	80
	Tools of trade	2	29
	Water creatures	6	39
	Wild animals	6	86
3	Bathroom objects	0	0
	Classroom objects	2	19
	Colors	3	30
	Drinks	1	10
	Emotions	3	41
	Family relations	2	33
	Geological features	3	38
	Musical instruments	2	31
		Total	**1172**

4.2 Tools

The database was constructed using mySQL, a free and open-source database management system. mySQL provides the ability to create structured tables that are easily queryable and combinable based on a multitude of conditions, and it has the additional benefit of being well-established and widely used in the technology industry, with extensive documentation available online.

To process and insert the collected data into the database, as well as to au-

Lubukusu (singular)	Lubukusu (plural)	English translation	Singular noun class	Plural noun class	Category	Frequency	Occurrences
kumurwe *(alternate spelling: kumurue)*	kimirwe *(alternate spelling: kimirue)*	head	3	4	Animal products, Human body parts	0.64	16
emoni	chimoni	eyes	9	10	Human body parts	0.80	16
litore ~ *(alternate spelling: litoore)*	kamatore *(alternate spelling: kamatoore)*	bananas	5	6	Cooked foods, Fruits, Things to eat	0.53	16
busuma *(alternate spelling: buusuma)*	busuma *(alternate spellings: buusuma, nusuma)*	ugali	14	14	Cooked foods, Things to eat	0.79	15
liru *(alternate spelling: liruu)*	kamaru *(alternate spelling: kamaruu)*	ear	5	6	Animal products, Human body parts	0.60	15

Figure 2: Screenshot of selected database entries

tomatically construct queries to retrieve data based on certain criteria, a server-side scripting language known as PHP was used. PHP is commonly used with mySQL to create dynamic websites whose content is drawn from a database. PHP is widely used, free, and open-source, and like mySQL, it has extensive documentation and enjoys widespread support among website hosting services. Finally, the user-visible interface for the database, being a website, was constructed with HTML, CSS, and Javascript.

4.3 Processing

In this section, we present the data processing procedure used to combine response rows pertaining to the same word into a single database entry. A four-step process was followed for each response row provided by each participant: (1) the response row was fed into a parser that added noun class information; (2) the parser attempted to group the row with previously inserted words in the database; (3) the parsed rows were inserted into the database; (4) additional modifications on the new database entries were performed as necessary.

In Step (1), the noun class for each singular and plural word in a given response row was identified. To identify noun class information, the parser matched the first 1-3 letters of each word to a pre-existing list of singular or plural noun class prefixes (based on Table 1). For singular words, the parser searched a list containing only singular noun class prefixes, and for plural words, the parser

searched a list containing only plural noun class prefixes. If a matching prefix was found, the word was assigned the corresponding noun class. If no match could be found, the noun class was left blank. If a match could not be found for the singular Lubukusu word but could be found for the plural word, or vice versa, the missing noun class was determined by recourse to common noun class pairings (Table 1).

In Step (2), the parser tried to group response rows with existing database entries. For a given response row, the algorithm first attempted to find an existing database entry with similar spellings and translations. If a candidate entry exactly matched the current response row in two or more of (i) the singular Lubukusu word, (ii) the plural Lubukusu word, or (iii) the translation or description, they were grouped together. If not, candidate entries that matched only one of (i)-(iii) were considered. Any non-matching fields were compared with corresponding ones in the response row to ensure similarity. If the entry matched the singular Lubukusu word exactly, grouping required that the plural spelling in the response row and the entry have a Levenshtein distance (Levenshtein 1966) of at most 0.35 * L, where L is the maximum length of the singular or plural word. Similarly, entries and response rows matching exactly in the plural word must have singular spellings within 0.35 * L of each other to be grouped together. Entries matching only the English translation must have a singular or plural spelling within 0.25 * L of the singular or plural words in the response row.

In Step (3), if an existing matching entry was identified, it was updated with any new spellings and/or semantic category/ies provided in the response row, and the number of occurrences of the entry was incremented by one. Because the semantic categories in Table 2 are not mutually exclusive, some words were provided in multiple categories. For instance, the word for "chicken" was provided for both "Domestic animals" and "Things to eat." If no existing matching entry was found, the response row was treated as a new word, and a new database entry with a unique numeric identifier was created.

As an example of how the grouping algorithm works, consider a response with singular word *lipwoni*, plural *kamapwoni*, and English translation "sweet potatoes." A search for similar entries in the database might return an entry with singular *lipwondi*, plural *kamapwondi*, and English translation "sweet potatoes." Because only the English translations match, the response and the entry are not immediately grouped together. Instead, the algorithm computes the Levenshtein distance between *lipwoni* and *lipwondi* as well as the distance between *kamapwoni* and *kamapwondi*, and sees that the minimum distance—in this case, 1—is

within the threshold to be considered similar enough. The response and the entry would then be grouped together, and the entry would be updated with the new spellings *lipwoni* (sg.) and *kamapwoni* (pl.).

The algorithm outlined above was cautious, designed to err on the side of keeping responses and entries separate rather than erroneously grouping them together. In particular, the algorithm was able to avoid erroneously grouping two different words together if they only shared an English description (ex. 1). This was a regular occurrence, especially for words described in general terms. The trade-off in having a cautious algorithm is that sometimes words that should be grouped together were not (ex. 2).

(1) Two responses **correctly not grouped together**

Singular	Plural	English translation / description
namasaka	namasaka	a type of bitter leaf
esaaka	esaaka	a type of bitter leaf

(2) Two responses **incorrectly not grouped together**

Singular	Plural	English translation / description
enduyu	chinduyu	rabbit
enduuyu	chinduuyu	hare

Step (4) involved manual modifications to the database. Combining entries that were not automatically grouped together but did indeed correspond to the same word (ex. 2) was the most common form of manual modification, with about 100 combinations performed. In contrast, separating erroneously grouped entries was one of the least common forms of manual modification and was performed less than 10 times. All post-insertion modifications were recorded for transparency, as illustrated in Figure 3.

After completing all the processing steps, additional measures and information were calculated or derived for each entry. The most frequent spellings of the singular and plural were identified straightforwardly. In the event of two spellings being equally frequent, both were marked as such. The length of these was also calculated straightforwardly. A measure of relative frequency was calculated by dividing the number of occurrences for an entry by the number of

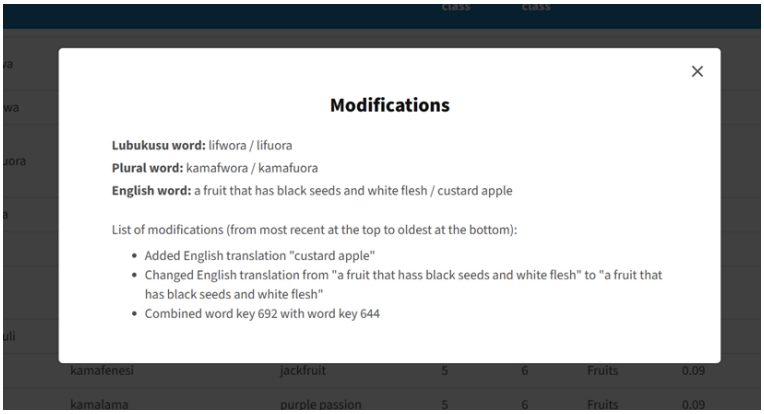

Figure 3: Sample reporting of manual modifications for a database entry

respondents to that category, following the assumption that participants would be more likely to provide more frequently used words over less frequent ones (cf. Section 3.1).

4.4 Database outcomes

The database is publicly accessible via an online interface[2] . Each entry is displayed with all its associated information as a row, as in Figure 2 in Section 4.1. The database also has a robust, flexible search and filtering system, intended to facilitate the selection of words based on a range of common criteria. Users can search by the singular word, the plural, or the English translation or description. They can additionally filter results based on noun class, semantic category, and/or the number of characters in the most common spelling of the singular or plural Lubukusu word. Users can also sort the returned results alphabetically, by number of occurrences of a given word, by frequency, or by the numeric identifier of the entry. Which information per entry is returned in a search can also be adjusted by users to suit their needs.

The current version of the database has 696 entries. Database entries span semantic categories and noun classes. Categories that occurred in Bin 1 had expectedly more responses than categories in Bins 2 and 3. With the exception of "Bathroom objects", each semantic category has at least 10 entries (Table 3). The

[2]The database can be accessed at the following URL: https://lubukusu.shaoart.com/

number of entries in noun classes 1-10 ranges from 17 (noun class 2) to 284 (noun class 10) (Table 4).

Table 3: Number of entries by semantic category

Bin 1	# Entries	Bin 2	# Entries	Bin 3	# Entries
Clothing & jewelry	73	Animal products	35	Bathroom objects	0
Cooked foods	67	Domestic animals	20	Classroom objects	15
Fruits	35	Insects	24	Colors	20
Furniture	24	Plants & flowers	53	Drinks	10
Household items	82	Things to eat	44	Emotions	38
Human body parts	55	Tools of trade	26	Family relations	26
Vegetables	46	Water creatures	17	Geological features	24
		Wild animals	40	Musical instruments	26
Other / miscellaneous[a]	16				

[a]The "Other/miscellaneous" category was not a category in any of the bins in the survey, but instead was created by the co-authors as part of the manual modifications processing step, to store anomalous entries and thus maintain coherent semantic categories without losing the data.

Table 4: Number of entries by noun class

Noun class (sg.)	# Entries	Noun class (pl.)	# Entries
1	25	2	17
3	55	4	55
5	135	6	139
7	97	8	97
9	263	10	284
11	43		
12	1		
14	44		
15 or 17	12		
16a	0		
16b	0		
18	2		
20	8		
23	3		
Blank (for sg. & pl.)	10		

5 Discussion

5.1 Outlook and implications

Although the present version of the database is only the preliminary phase, it demonstrates that crowd-sourcing is a very promising approach to creating a native-speaker-based lexical database for a language without a standardized orthography. With data from only 32 speakers, each of whom was able to make their contribution within a reasonable amount of time and in one sitting, we have a database of 696 nouns spanning a range of semantic categories and noun classes.

Researchers pursuing experimental work on lexical processing and/or noun classes can make use of the current database for the purposes of designing balanced experimental stimuli. Nouns in the database span a wide range of semantic categories and noun class (cf. Tables 3 and 4), and care was taken to select semantic categories likely to contain nouns that are concrete and picturable, for use in experimental methodologies such as picture-naming or visual world eye-tracking. The most frequent spellings are likely to be useful in experimental paradigms involving written presentation of stimuli, as in lexical decision tasks. The database also allows researchers in this domain to control for a number of important factors, such as word length and frequency, though we note limitations regarding the latter in Section 5.2.

The tool may also enhance formal work on nouns and noun classes. The database provides unprecedented empirical depth regarding the morphology and semantic scope of individual noun classes in Lubukusu. Formal linguistic studies often rely on a limited corpus or elicitation from a few speakers; meanwhile, the crowd-sourced approach yields data from many more speakers, with the potential to reach vast numbers of speakers across different dialects and social strata. Researchers interested in noun classes may be able to sort the database by noun class(es) and thus identify subtle semantic distinctions within noun classes that would not be immediately apparent in smaller datasets, such as the specific types of nouns that consistently fall into certain classes, or the nuances of meaning associated with class membership beyond the typical broad categories. In addition, this research may also shed light on synchronic variation and ongoing diachronic change within the Lubukusu noun class system. Moreso, by collecting data on plural forms and their corresponding singular noun classes, the database offers detailed insights into the intricate interaction between noun class and number in Lubukusu, which is a complex area of study in Bantu linguistics, with differ-

ent theoretical analyses proposed (e.g. Fuchs & van der Wal 2021). The current research provides a robust empirical basis to test these theories, revealing which noun class pairings are most common for singular-plural formation, and identifying any irregularities or exceptions that might challenge existing analytical frameworks. Additionally, this crowd-sourced database facilitates comparative work with other Bantu languages.

Finally, this research holds promise for the development of a standardized written orthography for Lubukusu. A corpus of spoken Lubukusu or a lexical frequency database for the language, especially one that records speakers' spelling conventions rather than IPA transcriptions or lexicographers' preferred spelling, does not currently exist, to the best of our knowledge. The development of standardized written forms supported by digital resources such as databases for Bantu languages elsewhere (e.g. South African Bantu languages) highlights the feasibility of creating such resources for a language like Lubukusu. Furthermore, native-speaker-based writing systems are more likely to gain a wider acceptance than those imposed by lexicographers and linguists. Indeed, native speakers will be involved in the process of creation of the orthography and recognized for their contributions.

5.2 Additional considerations

In this subsection we present a few considerations, given challenges that arose during data processing. For one, the spellings that speakers provided did not differentiate tone marking. Lubukusu is a tonal language, and other existing resources of Lubukusu do notate tone marking, but participants do not seem to include it in their own spellings. This leads to ambiguity, wherein two words that are in oral speech distinguished by tone appear in writing as homographs (ex. 3). Our database correctly distinguishes them as different entries based on the lack of similarity in their English translation, but future work will need to determine whether and how to incorporate tone into the database.

(3) Two entries that were incorrectly grouped together by the algorithm.

Singular	Plural	English translation / description
enda or endaa	chinda	stomach
eda	chinda	lice

Another consideration pertains to mass nouns. Likely as an outcome of

the semantics of mass nouns, participants vary in their response to what the plural form should be: leave it blank, provide the same word as the singular, or provide an infrequently used plural form (ex. 4). The algorithm does group these response rows together because their singular spelling and English translation are identical, but it remains unclear how best to present the plural form, if at all.

(4) Participant responses for a mass noun.

Singular	Plural	English translation / description
echai	—	tea ($n = 1$)
echai	echai	tea ($n = 3$)
echai	chichai	tea ($n = 1$)

A final consideration concerns our current measure of relative frequency. Its reliability is naturally modulated by the number of respondents to a particular category. Frequency measures for categories with fewer respondents (ex. "Drinks", n=2 respondents) are likely to be less reliable than those for categories with several (ex. "Cooked foods", n=14 respondents). Until more data is collected, care should be taken to cross-reference the raw number of occurrences (also listed for each entry in its row, cf. Fig. 2) when making use of the frequency measure. Future work will also collect subjective frequency ratings for words in the database (Schreuder & Baayen 1997), which have been found to correlate with corpus-derived frequencies (Kgolo-Lotshwao & Otlogetswe 2023) and to be effective in designing stimuli for psycholinguistic studies, e.g. for Setswana (Kgolo & Eisenbeiss 2015, Ciaccio et al. 2020).

6 Conclusions & next steps

The present study detailed the process of data collection and data processing for the creation of a preliminary lexical database for the Bantu language Lubukusu. Using a crowd-sourcing approach, data from 32 native speakers of Lubukusu yielded a database of just under 700 entries, each representing a noun in its singular and plural form, and containing other relevant information, such as the corresponding noun classes, alternative spelling(s), the English translation, word length, raw number of occurrences, and relative frequency. The breadth and depth of this resource highlights the potential of the crowd-sourcing approach,

and we aim to make public further documentation of our process for use by other scholars who want to pursue a similar approach in other languages.

The database is still in progress, with future work intended to collect more data on nouns and to expand to other lexical categories. Still, even at this stage, the database is able to support formal and experimental work on nouns and noun classes in Lubukusu. Most importantly, the database presents native-speaker-based written forms, with clear uses both for experimental linguistics research, in which recognizability of words to native-speaker participants is critical, and for scholars and community members aiming to develop a standardized orthography.

Acknowledgements

[Redacted for peer review.]

References

Appleby, L.L. ms. *An english-luluhya vocabulary*. Maseno, Kenya: C.M.S.

Carstens, Vicki. 1991. *The morphology and syntax of determiner phrases in kiswahili*. UCLA. (Doctoral dissertation). https://ling.auf.net/lingbuzz/001568.

Ciaccio, Laura Anna, Naledi Kgolo & Harald Clahsen. 2020. Morphological decomposition in bantu: a masked priming study on setswana prefixation. *Language, Cognition and Neuroscience* 35. 1257–1271. DOI: 10.1080/23273798.2020.1722847.

De Blois, Keis. 1975. *Bukusu generative phonology and aspects of bantu structure*. koninklijk Museum voor midden Afrika, Tervuren, Belgie.

Diercks, Michael. 2011. The morphosyntax of lubukusu locative inversion and the parameterization of agree. *Lingua* 121(5). 702–720.

Fuchs, Zuzanna, Travis Major & Justine Mukhwana Sikuku. under review. A triad study of noun similiarity judgmetns in lubukusu: the role of abstract grammatical gender. In *Selected papers from the 55th annual conference on african linguistics*. Contemporary African Linguistics.

Fuchs, Zuzanna & Jenneke van der Wal. 2021. The locus of parametric variation in bantu gender and nominal derivation. *Linguistic Variation*. 1–57. DOI: 10.1075/lv.20007.fuc.

Hall, J, R.E. O'Carroll & C.D. Frith. 2010. Neuropsychology. In Eve C Johnstone, David Cunningham Owens, Stephen M Lawrie, Andrew M McIntosh & Michael D Sharpe (eds.), *Companion to psychiatric studies, 8th edition*, 121–140. Churchill Livingstone.

Kanampiu, Patrick, Alexander Martin & Jennifer Culbertson. 2025. Experimental evidence for semantic and morphophonological productivity in kîîtharaka noun classes. *Glossa Psycholinguistics* 4. 1–45. DOI: 10.5070/G601120527. https://doi.org/10.5070/G601120527.

Kanampiu, Patrick Njue, Alexander Martin, Jennifer Culbertson & Patrick Kanampiu. 2025. Semantic and morphophonological productivity in the kîîtharaka gender system: a quantitative study. *Glossa: a journal of general linguistics* 10(1).

Kgolo, Naledi & Sonja Eisenbeiss. 2015. The role of morphological structure in the processing of complex forms: evidence from setswana deverbative nouns. *Language, Cognition and Neuroscience* 30. 1116–1133. DOI: 10.1080/23273798.2015.1053813.

Kgolo-Lotshwao, Naledi & Thapelo Otlogetswe. 2023. Evaluating the representativeness of the setswana corpus using behavioural data. *Studies in African Linguistics* 52. 121–136.

Lawyer, Laurel A, Jean-Paul Ngoboka, Willem S van Boxtel & Kyle Jerro. 2024. Meaning or morphology: individual differences in the categorization of kinyarwanda nouns. *Glossa*.

Levenshtein, Vladamir I. 1966. Binary codes capable of correcting delatinos, insertions, and reversals. *Soviet Physics Doklady* (10(8)). 707–710.

Lewis, Paul M., Gary F. Simons & Charles D Fennig. 2016. *Ethnologue: languages of the world (19th edition)*. SIL International.

Marlo, Michael, Michael Diercks, Adrian Sifuna & Justine Sikuku. N.d. *Bukusu lexicon*. Google sheet spreadsheet file [database].

Marlo, Michael, Adrian Sifuna & Aggrey Wasike. ms. Bukusu-english dictionary. Indiana University.

Mutonyi, Nasiombe. 2000. *Aspects of bukusu morphology and phonology*. The Ohio State University.

Nurse, Derek & Gérard Philippson. 2003. Introduction. In *The bantu languages*, 164–181. London/New York, Routledge.

Schreuder, Robert & R Harald Baayen. 1997. How complex simplex words can be. *Journal of memory and language* 37(1). 118–139.

Sikuku, J & Michael Diercks. in press. *Open generative syntax: object marking in bukusu: at the interface of pragmatics and syntax*. Language Science Press.

Sikuku, Justine M. 2011. *Syntactic patterns of anaphoric relations in lubukusu: representation and interpretation in a minimalist perspective*. University of Nairobi, Kenya. (Doctoral dissertation).

Taler, Vanessa, Brendan T Johns & Michael N Jones. 2020. A large-scale semantic analysis of verbal fluency across the aging spectrum: data from the canadian longitudinal study on aging. *The Journals of Gerontology: Series B* 75(9). e221–e230.

Wasike, Aggrey. 2018. Adjectives in lubukusu. *African linguistics on the prairie*. 325.

Wasike, Aggrey Khaoya. 2007. *The left periphery, wh-in-situ and a-bar movement in lubukusu and other bantu languages*. Cornell University. (Doctoral dissertation).