Chapter 21

A phylogenetic classification of Luyia language varieties

Michael R. Marlo[a], Rebecca Grollemund[a], Thanh Nguyen[a], Erik Platner[a], Sarah Pribe[b] & Alexa Thein[c]

[a]University of Missouri [b]Ohio State University [c]Washington University in St. Louis

This paper presents the results of a comparative study of the Luyia cluster of Bantu languages spoken in western Kenya and eastern Uganda. We propose a new classification of Luyia and neighboring languages using phylogenetic methods. Our study is based on a 200-item wordlist of basic vocabulary, representing 33 language varieties from the Luyia cluster and its closest neighbors, including Ganda, Gwere, and Soga to the west, and Gusii and Kuria to the south. Our results are broadly consistent with past classifications by Mould (1976, 1981) and Williams (1973), but refine our understanding of the relatedness of the target languages by employing more extensive data from more languages within the Luyia cluster and others in the region.

1 Introduction

The Luyia language cluster consists of around 20 Bantu language varieties spoken in western Kenya and eastern Uganda. A map with several Luyia varieties and neighboring Bantu languages is shown in Figure 1. The languages that we refer to as part of the Luyia cluster are circled. This includes the Kenyan language varieties spoken by members of "Luyia" or "Luhya" ethnic communities that were politically united in the first half of the 20th century (see MacArthur 2016) as well as closely related linguistic varieties on the Ugandan side of the border that were not part of the ethnopolitical unification that took place in Kenya.

Michael R. Marlo et al. 2022. A phylogenetic classification of Luyia language varieties. In Galen Sibanda, Deo Ngonyani, Jonathan Choti & Ann Biersteker (eds.), *Descriptive and theoretical approaches to African linguistics: Selected papers from the 49th Annual Conference on African Linguistics*, 383–407. Berlin: Language Science Press. DOI: 10.5281/zenodo.6393807

A second map of Kenyan Luyia varieties from Heine & Möhlig (1980: 35) is given in Figure 2.

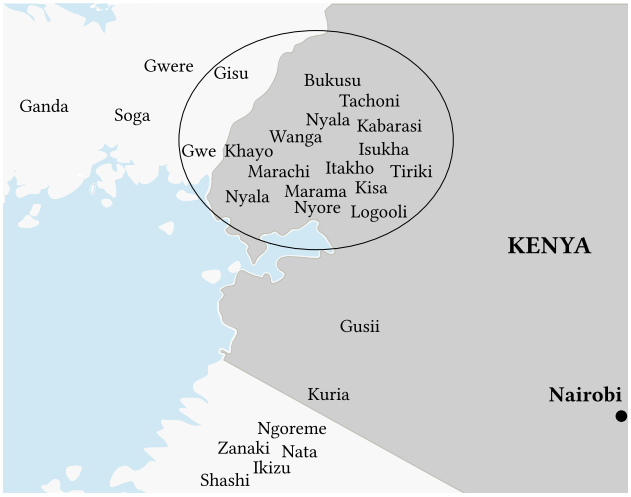

Figure 1: Map of the Luyia language cluster and its nearest neighboring Bantu languages

The overarching goals of this project are to understand the relationship between the languages of the Luyia cluster and their closest neighbors and to understand the internal structure of the Luyia cluster. To achieve this goal, we propose a new classification of Luyia and its nearest neighbors using phylogenetic methods.

2 Prior classifications of Luyia

In this section, we provide a detailed overview of prior classifications of Luyia. These include geolinguistic classifications in §2.1, genetic classifications in §2.2, a "dialectometrical" classification in §2.3, and the referential classification found in *Glottolog* in §2.4. We conclude the section by identifying some lingering questions about specific varieties of Luyia mentioned in prior classifications in §2.5.

2.1 Geolinguistic classifications (Guthrie 1948, 1967, Maho 2009, Lewis et al. 2016)

The first geolinguistic classification of Luyia was established by Guthrie (1948, 1967) and updated by linguists in Tervuren. These results, which were largely adopted by the *Ethnologue* (Lewis et al. 2016), are presented in an accessible way

Figure 2: Map of Kenyan Luyia varieties (Heine & Möhlig 1980: 35)

in Maho (2009). In these classifications, most language varieties of the Luyia cluster fall under JE30, the Masaaba-Luyia Group, shown in Table 1. Within this classification, a large number of central Kenyan Luyia varieties fall within JE32, the so-called "Lu(h)yia cluster".

Many broad aspects of this classification are uncontroversial – for instance, the fact that the language varieties of JE30 should be grouped together. One aspect of the Guthrie/Maho classification that is controversial is the placement of a few language varieties that we consider part of the Luyia cluster outside of the JE30 decimal series. As shown in Table 2, the southeastern varieties Isukha (JE412), Itakho (JE411), Logooli (JE41), and Tiriki (JE413) are in JE40, the Logooli-Kuria Group, along with Gusii (JE42) and Kuria (JE43), and several other languages of the Mara region of northwest Tanzania. Additionally, as shown in Table 3, Maho (2009) places Nyala West in JE18, part of JE10 along with Soga and Ganda.

Nyala West is sometimes referred to as "West Nyala" and Nyala East is sometimes referred to as "East Nyala" or "Nyala K", referring to Kakamega – the name of the county where Nyala East is spoken. As shown in Figure 2, Nyala East is surrounded by several other Luyia varieties: Bukusu, Tachoni, Kabarasi, Wanga, and Tsootso. Nyala West is spoken in Busia County adjacent to the southwestern variety Saamia.

Table 1: JE30: Masaaba-Luyia Group (Maho 2009)

Bantu Code	Name	ISO Code
JE31	Masaaba cluster	[myx]
JE31a	Gisu	[myx]
JE31b	Kisu	[myx]
JE31c	Bukusu	[bxk]
JE31D	Syan	
JE31E	Tachoni	[lts]
JE31F	Dadiri	[myx]
JE31G	Buya	[myx]
JE32	Lu(h)yia cluster	[luy]
JE32a	Wanga	[lwg]
JE32b	Tsotso	[lto]
JE32C	Marama	[lrm]
JE32D	Kisa	[lks]
JE32E	Kabarasi	[lkb]
JE32F	Nyala East	[nle]
JE33	Nyore	[nyd]
JE34	Saamia	[lsm]
JE341	Khayo	[lko]
JE342	Marachi	[lri]
JE343	Songa	
JE345	Nyole	[nuj]

In 2007, Luyia was introduced as a "macrolanguage" in the *Ethnologue*, and many of the Luyia varieties were recognized with distinct ISO codes. As part of this change, Nyala East was renamed "Olunyala" (i.e. "Nyala" without the cl. 11 noun class prefix), and although Nyala West was not mentioned in ISO 639-3 Change Request Number 2007-171, Nyala West was also included as part of Nyala in the changes implemented in the ISO 639-3 reclassification. Nyala West and Nyala East were thus unified as part of JE32 in the 16th edition of *Ethnologue*, but this seems to have been an accident resulting from the shared language name and the lack of mention of Nyala West in the change request. A subsequent ISO change request (2014-001) to reintroduce Nyala East and Nyala West as separate language varieties with distinct ISO codes was rejected, citing a lack of linguistic evidence. (The authors of the change request failed to cite Heine & Möhlig (1980),

Table 2: JE40: Logooli-Kuria Group (Maho 2009)

Bantu Code	Name	ISO Code
JE401	Ngoreme	[ngq]
JE402	Ikizu	[ikz]
JE403	Suba	[suh]
JE404	Shashi	[szk]
JE405	Kabwa	[cwa]
JE406	Singa†	[sgm]
JE407	Ware†	[wre]
JE41	Logooli	[rag]
JE411	Itakho	[ida]
JE412	Isukha	[ida]
JE413	Tiriki	[ida]
JE42	Gusii	[guz]
JE43	Kuria	[kuj]
JE431	Simbiti	[ssc]
JE432	Hacha	[ssc]
JE433	Surwa	[ssc]
JE434	Sweta	[ssc]
JE44	Zanaki	[zak]
JE45	Ikoma-Nata-Isenye	[ntk]

which includes Nyala East in the northern-central Bukusu-Wanga cluster and Nyala West in the southwestern Saamia-Nyala cluster.) This decision failed to recognize that the merger of these two languages 7 years earlier had also been done without any linguistic evidence or even a specific request to merge the two language varieties under one name.

2.2 Genetic classifications (Mould 1976, 1981, Williams 1973, Nurse & Philippson 1980)

The first genetic classification of Luyia language varieties was done by Williams (1973), who presents a relatively extensive internal classification of Luyia with data from 16 Luyia varieties. Williams (1973) primarily uses lexicostatistic methods and the 200-item Swadesh list.[1] Lexicostatistics – the method developed

[1]Williams (1973) also identifies some phonological correspondences across varieties and compares the noun class prefixes across Luyia varieties.

Table 3: JE10: Nyoro-Ganda Group (Maho 2009)

Bantu Code	Name	ISO Code
JE101	Gungu	[rub]
JE102	Talinga-Bwisi	[tlj]
JE103	Ruli	[ruc]
JE11	Nyoro	[nyo]
JE12	Tooro	[ttj]
JE121	Hema	[nlx]
JE13	Nkore	[nyn]
JE14	Kiga	[cgg]
JE15	Ganda	[lug]
JE16	Soga	[xog]
JE16	Kenyi	[lke]
JE17	Gwere	[gwr]
JE18	Nyala West	[nle]

by Swadesh (1952) – measures the percentages of shared cognates by comparing the similarity of words from the basic vocabulary of Swadesh between two or more related languages. Williams' (1973) approach yields a geographically-based clustering of varieties, shown in Figure 3, which has a flat structure with five branches: Western, Northern, Central, Eastern, and Southeastern. Note that in contrast with the Guthrie/Maho system, the southeastern varieties Isukha, Itakho, Logooli, and Tiriki, and the southwestern variety Nyala West are treated as part of Luyia in William's (1973) classification.

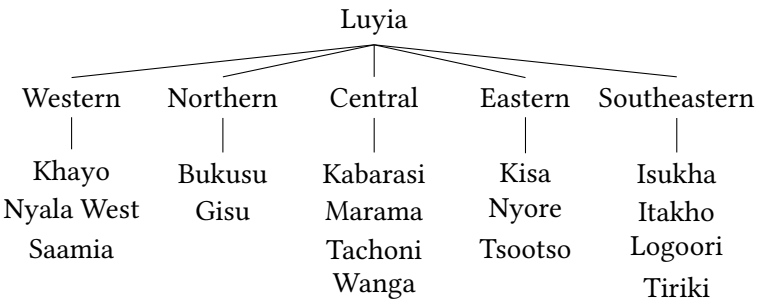

Figure 3: Williams' (1973) classification of Luyia

Mould's (1976, 1981) classifications generally accord with Williams (1973), though Mould worked with different languages. Using a 200-item wordlist, Mould (1976, 1981) carried out a lexicostatistic analysis including five Luyia varieties (Bukusu, Itakho, Logooli, Saamia, and Wanga) along with Ganda and Soga to the west and Gusii to the south. Mould's (1976, 1981) results, summarized in Figure 4, show the overall unity of Luyia, as the varieties we consider part of the Luyia cluster are more similar to one another than they are to Ganda, Soga, and Gusii. Internally within Luyia, the Southeastern varieties Itakho and Logoori branch off from Saamia, Wanga, and Bukusu, which is divided into a Northern branch with Bukusu and a Western-Central branch with Saamia and Wanga.

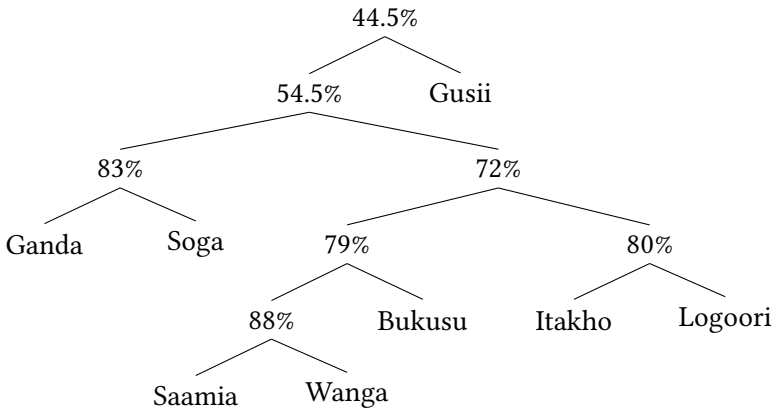

Figure 4: Mould's (1981: 185) classification of Luyia based on lexicostatistics

Mould (1981: 201) also considers sound change in creating a second tree with similar results, shown in Figure 5. There are two differences in this tree: (i) Logoori branches off from Itakho at a higher level in the tree, and (ii) Bukusu, Saamia, and Wanga are not subdivided further. Mould (1981: 201) also computes a third tree based on a comparison of tense/aspect markers; its results are identical to the tree based on sound change.

Nurse & Philippson's (1980) classification is based on a lexicostatistical study with a 400-item wordlist. It includes four Luyia language varieties and essentially gives the same results as Mould (1976, 1981) but with fewer languages. As shown in Figure 6, there is a geographical split that divides Saamia and Bukusu from Itakho and Logoori. Nurse & Philippson (1980) treat this as a North-South split, though "West" vs. "East" appears to us to be equally tenable labels for the two groups.

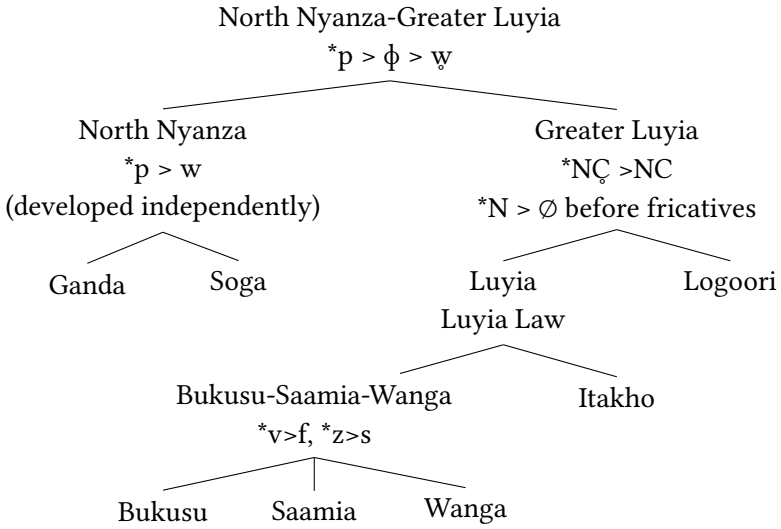

Figure 5: Mould's (1981: 201) classification of Luyia based on sound change

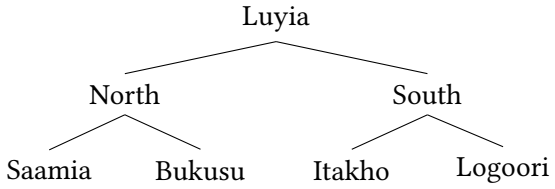

Figure 6: Nurse & Philippson's (1980) classification of Luyia

2.3 Dialectometrical classification (Heine & Möhlig 1980)

The classification of 15 varieties of Kenyan Luyia by Heine & Möhlig (1980) is part of a large-scale study of languages in Kenya, which includes a wordlist of 640 concepts, documentation of each variety's phonological system, and basic features of grammar (Heine & Möhlig 1980: 9). Their classification of Luyia is an areal grouping, based on "geographical and synchronic dialectal proximity (Heine & Möhlig 1980: 13)." Heine & Möhlig (1980: 32) state that the Luyia varieties, "neither form a single dialect cluster nor even represent dialects of variations of a single language. The term [Luyia] as such is geographical and has no further dialectological significance." Internal subgroupings are based on "dialectal proximity", which measures the degree of of linguistic similarity in linguistic features, e.g. isoglosses, across dialect clusters.

The four subgroupings established by Heine & Möhlig (1980) are shown in Figure 7. Logooli is viewed as a "separate language", and three other "cluster[s] of dialects" are identified: a Southwestern cluster, a Central-Northern cluster, and a Southeastern cluster. The separation of Logoori from the Southeastern languages and the inclusion of Central and Northern varieties in a single larger branch is similar to Mould's (1976, 1981) classifications but different from Williams (1973).

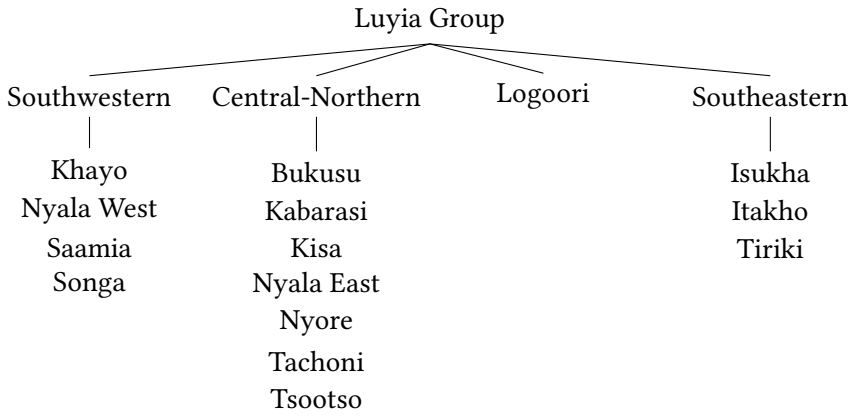

Figure 7: Heine & Möhlig's (1980) classification of Luyia

The authors indicate that the linguistic data undergirding the atlas would be published in future volumes, but no subsequent studies on Luyia languages were published from the project. See Heine (2013) for a later overview of the *Language and Dialect Atlas of Kenya* project and see Möhlig's (1985) review of Angogo Kanyoro (1983) for some additional comments on his "dialectometrical" analysis of the Luyia area, which includes the identification of three "dialect continua": Marachi-Khayo in the west, Kabarasi-Tachoni-Nyala East in the east, and Nyore in the south.

2.4 Referential classification (Hammarström et al. 2020)

A recent referential classification of "Greater Luyia", which includes both Kenyan and Ugandan language varieties, is found in the *Glottolog* (Hammarström et al. 2020). The *Glottolog* 4.3 classification, which is based on secondary materials – primarily Mould (1981) – is shown in Figure 8. As noted above, Mould (1981) deals with only a small subset of the language varieties represented in the *Glottolog* classification. It appears that the many other language varieties present in the *Glottolog* classification are populated from uncited sources, with the *Ethnologue*

database being a possible source. The *Glottolog* represents the prior classification with the most Luyia language varieties displayed in a tree format, but it is not a genetic classification based on original data, and the justification for many aspects of its structure is unclear.

Our representation of the *Glottolog* classification given here in Figure 8 differs from the original in a few ways. First, we harmonized some language names (e.g. "Idakho" → "Itakho", "Kabras" → "Kabarasi"), and we eliminated the cl. 11 *ulu-* noun class prefix from the languages under the Masaaba node.

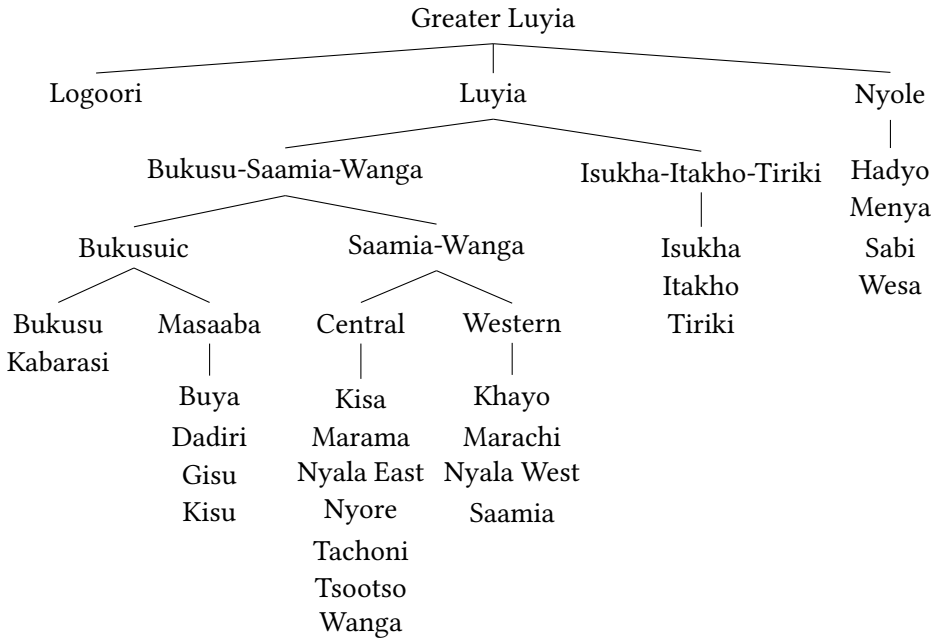

Figure 8: Glottolog 4.3 classification of Luyia (Hammarström et al. 2020)

Second, two languages are listed in multiple locations in the *Glottolog* tree, and we have retained the languages in only one position. Tachoni is given under a node for Bukusu where it is contrasted with "Nuclear Bukusu" and under the Central Luyia node. We have not seen evidence for two distinct varieties of Tachoni. Following Odden (2009: 305), who states that "Tachoni most resembles "mainstream" Luhya varieties such as Tsootso, Nyala, Wanga, Kisa, and Marachi, least resembling Bukusu, [Gisu], and Logoori," we have included Tachoni only under the Central Luyia node, and we have simplified the Bukusu node, eliminating "Nuclear Bukusu". Bukusu is also listed in two locations in the *Glottolog*

tree: under the "Bukusuic" node with Kabarasi and under Masaaba, where presumably it refers to Ugandan Bukusu. Lacking evidence for a distinct form of Ugandan Bukusu, we have not included Bukusu under Masaaba.

Third, Nyole and its subvarieties are considered "Unclassified Luyia" in the *Glottolog* and are listed as a highest level branch of Greater Luyia. We have maintained the position of Nyole within the tree but have removed the "Unclassified Luyia" label, acknowledging here that further study of Nyole and other Bantu language varieties of eastern Uganda might place those languages in a different position within the tree. Maho (2009) includes Nyole (JE345) within JE34, which makes it most similar to other southwestern varieties Saamia (JE34), Khayo (JE341), and Marachi (JE342); no other classifications indicate the position of Nyole within Greater Luyia, as far as we know.

Given that Mould (1981) is cited as the basis for the *Glottolog* classification, it is not surprising that the *Glottolog* tree reflects the results of Mould (1981) (see Figures 4–5 above). However, as Mould (1981) includes only 5 language varieties of Greater Luyia (Logoori, Itakho, Bukusu, Wanga, Saamia), many details concerning the internal structure of Luyia are underspecified in Mould (1981). These uncertain details include the grouping of other languages with those studied by Mould as well as decisions to subdivide Mould's (1981) groupings further.

Although it is unclear how the *Glottolog* arrived at its classification, several aspects of its structure are consistent with the results of other research:

- Setting off Logoori at the highest level of the tree follows Mould's (1981) trees based on sound changes tense-aspect markers. This is also consistent with Heine & Möhlig's (1980) treatment of Logoori as a separate language.
- The inclusion of Isukha and Tiriki alongside Itakho in a high-level southeastern branch of Luyia is consistent with the results of Williams (1973) and Heine & Möhlig (1980).
- The inclusion of Masaaba varieties alongside Bukusu in a northern branch of Luyia is supported by Williams (1973).
- The division of Saamia-Wanga into a Central branch and Southwestern branch is supported by Williams (1973).
- The inclusion of Marama, Tachoni, and Wanga in the Central branch is consistent with the results of Williams (1973).
- The inclusion of Khayo, Nyala West, and Saamia follows both Williams (1973) and Heine & Möhlig (1980).

A few aspects of the *Glottolog* classification differ from other prior classifications:

- The *Glottolog* classification of Kabarasi with Bukusu in a Northern branch differs from the results of Williams (1973), which classifies Kabarasi with the Central varieties.
- Williams (1973) subdivides the languages of the *Glottolog* Central group into Central and Eastern subgroups.
- Heine & Möhlig (1980) differs from Mould (1976, 1981), Williams (1973), and *Glottolog* in collapsing the Central and Northern groups.
- No prior studies include Marachi, so it is unclear on what basis Marachi is included with the Western varieties in *Glottolog*.

2.5 Issues in prior classifications

There are a few questions about some of the specific varieties listed in past classifications. Within the Masaaba cluster, the *Ethnologue* treats Gisu and Masaaba as alternative names for the same language, of which Kisu (JE31b) is another alternative name. We are doubtful that there are distinct language forms "Gisu" vs. "Kisu", but we have retained Maho's labels in Table 1 because distinct codes are given to the two language names. The *Glottolog* also includes both "Gisu" and "Kisu", but we are not aware of any discussion of this distinction in the literature.

Buya (JE31G) and Dadiri (JE31F) are treated as dialects of Masaaba in Maho (2009) and the *Glottolog*. Despite the existence of an early grammar (Purvis 1907), a robust wordlist (Siertsema 1981), and a book on Gisu/Masaaba dialect variation (Brown 1972), there is little available information concerning the classification of Gisu/Masaaba language varieties. Similarly, we are unaware of materials dealing with the classification of the Nyole language varieties identified in the *Glottolog*.

A further question concerns Syan, a variety described in Huntingford (1965) based on materials collected in 1924 from Syan migrant workers in Uasin Gishu province in Kenya. Huntingford confirmed that Syan people lived in Bulegenyi District in eastern Uganda as late as 1930, though other investigators reported not encountering a Syan ethnic group in that area. Syan is no longer listed in the most recent versions of the *Ethnologue*, and we are not aware of any additional research on the language. Citing Schoenbrun (1994), the *Glottolog* states that "Syan is a missing language of the North Nyanza subgroup of Bantu, which [is] lexicostatistically too divergent" to be [mutually] intelligible [with] any other

language in JE10 or JE20. In the *Glottolog* classification, North Nyanza includes Ganda, Gwere, Kenyi, Lamogi, and Soga. Maho (2009) however does classify Syan as part of JE31 (JE31D).

It is also unknown how Songa (JE343) fits in. The *Ethnologue* lists it as a dialect of Saamia, and Heine & Möhlig (1980: 32) include it as part of their Saamia-Nyala dialect cluster, with which it is geographically adjacent (see the map in Figure 2). Heine & Möhlig (1980: 32), who focus on languages in Kenya, list a population of 10,000 Songa speakers. This figure is consistent with the fact that the 1979 Kenyan census, cited by Were & Odak (1987: 26), identified 9,000 inhabitants of Usonga location in Siaya District. As reported in Marlo (2007: 2-3), Marlo attempted in 2006 to collect information on Songa, but it was unclear how the variety differed from Nyala West. Confusingly and possibly erroneously, various editions of the *Ethnologue* (e.g. 13th edn.) identify 10,000 speakers of Songa in Uganda, but we have been unable to find any other sources that identify a Songa language variety of Uganda. As noted by Marlo (2009), a 2004 report by the SIL Language Assessment team on Gwe and Saamia in Busia District, Uganda (Anderson et al. 2004), does not mention Songa in its results. We have not seen any Songa data in the literature.

There is also at least one Luyia variety not listed in prior classifications, Tura, which is described in Marlo (2008). Although the exact classification of this variety is unclear, it is geographically and linguistically most proximal to Bukusu, Khayo, and Wanga, and should fit in with the JE30 languages.

A few communities of Greater Luyia have offshoots in the diaspora. Marlo et al. (2017) provide a description of a variety of Nyole spoken in southern Busoga, Uganda. In addition, through various migrations and resettlement patterns, there are sizable communities of Logooli speakers in western Uganda and southern Kenya, around Migori (Chavasu 1997, Heine & Möhlig 1980: 70). As some of diasporic Luyia communities have been separated for 50-60 years or more, in different contact situations, it might be appropriate to treat some of them as distinct varieties or at least to investigate them separately, leaving open the possibility that they are distinct.

To conclude, over the past 50 years, different techniques have been applied to the study of Luyia languages in order to better understand their internal classification: referential classifications, lexicostatistics studies, "dialectometric" studies, and (rarely) classifications based on the study of linguistic innovations. Due to the selection of different type of data (geography, cognate sets or sound changes), methods (lexicostatistics or use of shared phonological innovations in order to make groupings) and the number of languages selected (from 5 to 16 languages), these studies lead sometimes to different classifications. However, we do find

some accordance in their conclusions: the Luyia group includes languages such as Bukusu, Itakho, Saamia, and Wanga, and most sources recognize the distinctness of southeastern Luyia varieties and the fact that Logoori is the most distinct within Luyia or possibly even a separate language.

3 Methodology

In light of recent developments in the field of historical linguistics which include the appearance of phylogenetic methods borrowed from the field of biology being used to classify languages, we decided to propose the first phylogenetic study of the Luyia languages. Phylogenetic methods are based on a simple principle: languages and species evolve in a similar way, by a process of descent with modification. Therefore, when similarities are observed between species or languages, they can be explained by a common ancestor from which they have descended. By extension, the evolutionary tools used to investigate biological evolution in order to classify organisms in terms of their genealogical relation to one another can also be applied to the study of languages.

In order to carry out our analysis, we compiled wordlists from several sources into a database. These include prior research by Brown (1972), Williams (1973), and Nurse & Philippson (1975), as well as more recent work on Luyia languages by members of our extended research team. This includes a series of Luyia lexical materials that Michael Marlo collected in 2006 based on Appleby (1943), Swadesh lists for several varieties collected in 2016 and 2018, and more extensive lexical research carried out on Bukusu, Tiriki, and Wanga in 2016. It also includes a Swadesh list collected by Deo Kawalya in 2016, and data extracted from lexical materials collected by Kristopher Ebarb in 2012–2013 and David Odden in 2014–2018. The 61 total datasets in our database at the time of our analysis in 2018 are listed in Table 4.

Table 4: Datasets in our database

	Language	Maho code	Source	Year
1.	Bukusu*	JE31c	Marlo	2016
2.	Bukusu	JE31c	Mould	1976
3.	Bukusu	JE31c	Williams	1973
4.	Ganda*	JE15	Nurse & Philippson	1975

Continued on next page

Table 4 – continued from previous page

	Language	Maho Code	Source	Year
5.	Ganda	JE15	Mould	1976
6.	Gisu*	JE31a	Marlo	2016
7.	Gisu (Soba)	JE31a	Brown	1972
8.	Gisu (Fumbu)	JE31a	Brown	1972
9.	Gisu (Hugu)	JE31a	Brown	1972
10.	Gusii*	JE42	Nurse & Philippson	1975
11.	Gusii	JE42	Mould	1976
12.	Gwe*	JE34	Marlo	2016
13.	Gwere*	JE17	Marlo	2016
14.	Gwere*	JE17	Nurse & Philippson	1975
15.	Ikizu*	JE402	Nurse & Philippson	1975
16.	Isukha	JE412	Marlo	2006
17.	Isukha*	JE412	Marlo	2018
18.	Itakho*	JE411	Ebarb	2013
19.	Itakho	JE411	Marlo	2006
20.	Itakho	JE411	Marlo	2006
21.	Itakho	JE411	Mould	1976
22.	Kabarasi*	JE32E	Marlo	2006
23.	Kabarasi	JE32E	Marlo	2018
24.	Kabarasi	JE32E	Marlo	2018
25.	Khayo*	JE341	Marlo	2018
26.	Kisa	JE32D	Marlo	2006
27.	Kisa*	JE32D	Williams	1973
28.	Kuria (Mago)*	JE43	Nurse & Philippson	1975
29.	Kuria (Tari)*	JE43	Nurse & Philippson	1975
30.	Logooli	JE41	Marlo	2006
31.	Logooli*	JE41	Odden	2018
32.	Logooli	JE41	Mould	1976
33.	Logooli*	JE41	Nurse & Philippson	1975
34.	Logooli*	JE41	Williams	1973
35.	Marachi*	JE342	Marlo	2009
36.	Marama*	JE32C	Marlo	2018
37.	Marama*	JE32C	Williams	1973
38.	Nata*	JE45	Nurse & Philippson	1975
39.	Ngoreme*	JE401	Nurse & Philippson	1975

Continued on next page

Table 4 – continued from previous page

	Language	Maho Code	Source	Year
40.	Nyala East*	JE32F	Marlo	2018
41.	Nyala West	JE18	Marlo	2006
42.	Nyala West	JE18	Marlo	2018
43.	Nyala West*	JE18	Marlo	2018
44.	Nyala West	JE18	Williams	1973
45.	Nyole (Lower)	JE345	Marlo	2016
46.	Nyole (Upper)	JE345	Kawalya	2016
47.	Nyore*	JE33	Ebarb	2013
48.	Nyore	JE33	Marlo	2016
49.	Nyoro*	JE11	Nurse & Philippson	1975
50.	Saamia	JE34	Marlo	2006
51.	Saamia	JE34	Mould	1976
52.	Shashi*	JE404	Nurse & Philippson	1975
53.	Soga	JE16	Mould	1976
54.	Soga*	JE16	Nurse & Philippson	1975
55.	Tachoni	JE31E	Marlo	2006
56.	Tiriki*	JE413	Marlo	2016, 2018
57.	Tooro*	JE12	Nurse & Philippson	1975
58.	Tsootso	JE32b	Marlo	2006
59.	Wanga*	JE32a	Marlo	2016, 2018
60.	Wanga	JE32a	Mould	1976
61.	Zanaki*	JE44	Nurse & Philippson	1975

For the present classification, we used a subset of the datasets in our database. We eliminated languages that had too few words represented in the 200-item Swadesh wordlist. This removed the datasets from Brown (1972) and Mould (1976), and a few others such as those on Upper Nyole and Lower Nyole, which are based on the 100-item Swadesh list. We also eliminated a handful of our research team's datasets where we felt we had a more accurate dataset for the same language. For instance, several of Marlo's 2006 wordlists are preliminary lists of translated words in a practical orthography provided by a speaker working alone that have not been vetted by a linguist. For several varieties, there are more recently developed wordlists with more reliable data (e.g. the materials by Ebarb and Odden), and in such cases we did not use the materials from 2006. We have also collected some Swadesh lists (e.g. on Kabarasi and Kuria) since we completed the analysis reported here; such data are also not included in the results reported.

Our analysis is based on 33 primary datasets, plus 4 outgroup languages: Ha (JD66), Vinza (JD67), Lega (D25), and Yaka (H31). The 33 primary datasets, which are indicated with an asterisk in Table 4, represent 29 language varieties: 16 Luyia varieties (e.g. Bukusu, Tiriki, Wanga) and 13 Bantu languages to the west (e.g. Ganda, Soga) and south (e.g. Kuria, Gusii).

We eliminated words from the 200-item Swadesh wordlist that had entries from fewer than 21 datasets. As a result, our analysis is based on 151 entries from the 200-item Swadesh wordlist.

Once the database of 35 datasets of 151 words was established, we carried out cognancy judgments for each of the 151 words. We used color-coding to form cognate sets based on predictable sound changes between the languages, as in Figure 9. We then transformed the colors into numbers to use for further analysis, as shown in Figure 10.

Next, we built two trees: a network representation shown in Figure 11 and a Bayesian tree-like representation shown in Figure 12. The network was built using a Neighbor-Net algorithm (Bryant & Moulton 2004) which uses a distance-based method that calculates the distance between pair of languages in order to produce a distance matrix. Distances between two or more languages are measured by the percentage of cognates shared. The Bayesian method allows the construction of a sample of trees. The use of the Markov chain Monte Carlo (MCMC) approach (Larget & Simon 1999, Pagel & Meade 2004) allows us to sample trees in proportion to their likelihood. In the tree presented in Figure 12, we can see numbers under the nodes. These numbers correspond to the posterior probability of each node on the tree (which is similar to the proportion of trees in the sample containing that node).

The network presented in Figure 11 displays the relationships between languages studied. If we want to measure the closeness or the distance between two languages, we have to look at the path from language X to language Y. If the path involves a great number of rectangles, it means that the languages are not closely related, e.g. Zanaki (JE44) and Logooli (JE41). But if the path between two languages is short (small number of rectangles), we will consider the two languages close to each other, e.g. Zanaki (JE44) and Shashi (JE404).

4 Results

The analysis of the network presented in Figure 11 shows three main groups: the Luyia group (pink), the Ganda group (green) composed of JE10 languages, and the Kuria group (blue) composed of JE40 languages. The Ganda group (green) is

	1 all	4 ash	10 belly	15 blood
D25_Lega	-nsé	lwìtɔ̀	ndà; kibú	mikilá
H31_Yaka	-óóso, -óoso	bóóma	vúmú	éngá
JE11_Nyoro	byona	eiju	enda	esagama
JE12_Tooro	-ona	iju	enda	esagama
JE15_Ganda	-nna	vvu	?	musaayi
JE16_Soga	bona-bona, byona-byona	eivu	enda	owusaai
JE17_Gwere_1	βónaβóna	ei-kóke	?	ómú-sááyi
JE17_Gwere_2	byonabyona	?	kidda	musaaye
JE31a_Gisu	βóóse	?	múu-n-da	mú-sááyi
JE31c_Bukusu	-osi	liikoxe	eenda	kámafuki ~ kámalasile
JE32a_Wanga	-osi	liikoʃe	inda	liiˡlásíle
JE32C_Marama_1	-osi	-koshe	-nda	-tsayi
JE32C_Marama_2	-osi	likoshe	inda	amatsayi
JE32D_Kisa	-esi	?	-nda	-tsayi
JE32E_Kabaras	ohosi	likoshe	eyinda	amabanga

Figure 9: Establishing cognate sets

	1 all	4 ash	10 belly	15 blood
D25_Lega	1	1	1	1
H31_Yaka	2	3	2	3
JE11_Nyoro	3	6	1	5
JE12_Tooro	3	6	1	5
JE15_Ganda	3	6	?	6
JE16_Soga	3	6	1	6
JE17_Gwere_1	3	2	?	6
JE17_Gwere_2	3	?	1	6
JE31a_Gisu	1	?	1	6
JE31c_Bukusu	1	2	1	7
JE32a_Wanga	1	2	1	7
JE32C_Marama_1	1	2	1	6
JE32C_Marama_2	1	2	1	6
JE32D_Kisa	1	?	1	6
JE32E_Kabaras	1	2	1	2

Figure 10: Multistate encoding

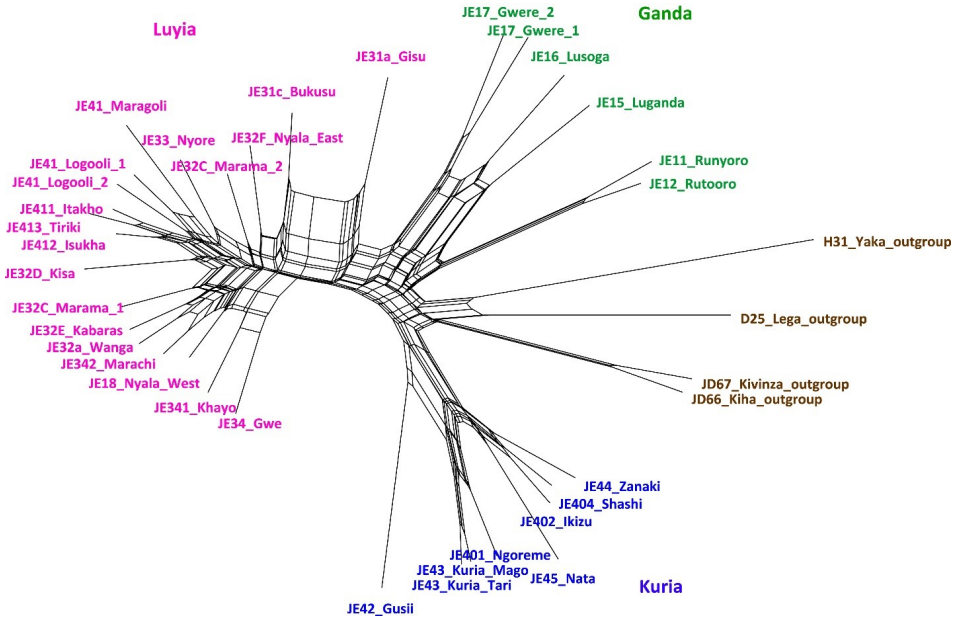

Figure 11: Unrooted Neighbor-Net network

more closely related to the Luyia group (pink) than the Kuria group (blue). According to Fitch (1997), the analysis of the webbing or "netting" in a phylogenetic network allows to visualize alternative histories because phylogenetic networks are a generalization of phylogenetic trees that display the representation of conflicting signal or alternative evolutionary histories in a single diagram (Bryant & Moulton 2004).

Within the Luyia group (pink), we can distinguish several small subgroups: (i) a Central-Western group with Marama (JE32C), Kabarasi (JE32E), Wanga (JE32a), Marachi (JE342), and Nyala West (JE18), (ii) a Southeastern group with Isukha (JE412), Tiriki (JE413), Itakho (JE411), and (iii) Logooli (JE41). The amount of webbing observed between the languages in the Luyia group (pink) suggest that these languages are similar and that they must be in a situation of contact (as opposed to the Kuria or Ganda groups where the webbing is reduced).

The analysis of the Ganda group (green) shows two subgroups: (i) Nyoro (JE11) and Tooro (JE12), have a very long common branch showing that these two languages are similar (because they are sharing a high percentage of cognate sets), and (ii) Ganda (JE15) and Lusoga (JE16) linked to Gwere (JE17).

Thanks to the network representation, we can also note that Gisu (JE31a) is situated in between the Luyia group (pink) and the Ganda group (green), showing

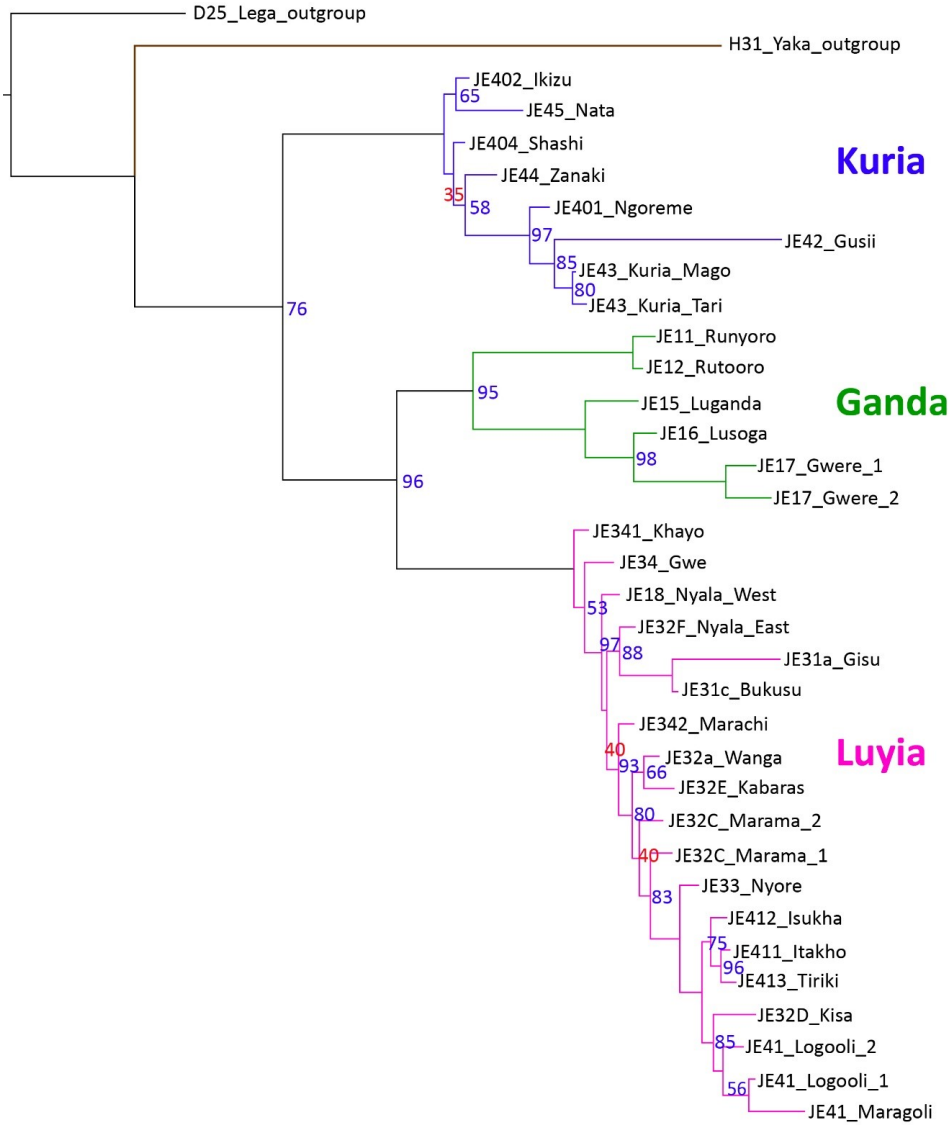

Figure 12: Bayesian consensus tree based on a 200-item Swadesh list with 151 words. The numbers show posterior probabilities (nodes with no numbers = 100%): the numbers indicate the percentage of trees in the sample containing that node.

that this language is close to Bukusu (JE31c) and Gwere (JE17). This position of Gisu (JE31a) is not surprising as it is geographically located between Bukusu (JE31c) and Gwere (JE17). Therefore, we can assume that Gisu (JE31a) has some cognate sets in common with the two languages.

The analysis of the Kuria group (blue) shows three subgroups: (i) Gusii (JE42), (ii) Tari Kuria (JE43), Mago Kuria (JE43), and Ngoreme (JE401), and (iii) Nata (JE45), Ikizu (JE402), Shashi (JE404), and Zanaki (JE44). The division into three groups can be corresponds to the geographical location of the languages. In the network, Gusii (JE42) has a separate branch from the other Kuria languages because Gusii (JE42) is geographically remote from the other Kuria languages (see the map in Figure 1). The subgroup with Kuria (JE43) varieties and Ngoreme languages (JE401) corresponds to the languages spoken near the Kenya-Tanzania border. The third subgroup includes the languages spoken in the Mara region of Tanzania: Zanaki (JE44), Shashi (JE404), and Ikuzu (JE402).

The analysis of the Bayesian tree in Figure 12 is in accordance with our analysis of the network. Indeed, the analysis of the tree suggests three groups: Kuria (blue), Ganda (green), and Luyia (pink). Each of these three groups show strong support (i.e. posterior probability: 100% for the Kuria group (blue), 95% for the Ganda group (green), and 100% for the Luyia group (pink). The tree-like representation allows us to visualize the relatedness between the languages, and it also represents the hierarchy between the groupings and the languages studied. First of all, we can notice that the Kuria group (blue) is the first one to diverge, followed by the Ganda group (green), and then the Luyia group (pink). The ordering of the nodes implies that the Luyia group (pink) is the most recent one that has diverged from the Ganda group (in the network representation, the Luyia (pink) and the Ganda (green) groups are very close).

Within the Kuria group (blue), we can observe the same subgroupings as the ones distinguished in the network: Ikizu (JE402), Nata (JE45), Shashi (JE404), and Zanaki (JE44) vs. Ngoreme (JE401), Gusii (JE42), and the Kuria (JE43) dialects.

The Ganda group (green) splits into two groups as in the network with Nyoro (JE11) and Tooro (JE12) vs. Ganda (JE15) and Soga (JE16) linked to Gwere (JE17).

Finally, the Luyia group (pink) shows a succession of branches. The first languages to branch off are the Southwestern varieties Khayo (JE341), Gwe (JE34), and Nyala West (JE18). Then, we have a subgroup composed of the three Northern languages Nyala East (JE32F), Gisu (JE31a), and Bukusu (JE31c) – these three languages are also close in the network – followed by various Central languages branching off in succession: Marachi (JE342), then Wanga (JE32a) and Kabarasi (JE32E), then Marama (JE32C) and Nyore (JE33). Finally, there is a Southeastern

group that splits into two subgroups: with Isukha (JE412), Itakho (JE411), and Tiriki (JE413) in one subgroup and Kisa (JE32D) and Logooli (JE41) in the other.

The network and the Bayesian trees show the fundamental unity of Luyia (pink group), the unity of the Bantu languages to the south of Luyia (blue group), and the unity of the Bantu languages to the west of Luyia (green group), supporting the view of Mould (1976, 1981) and arguing against the geolinguistic classifications that place southeastern Luyia within the JE40 group and Nyala West in the JE10 group. In each tree, the Ganda group of languages (green) is more closely related to Luyia than the Kuria group of languages (blue).

As for the internal structure of Luyia, we will focus our discussion on the Bayesian tree in Figure 12, which uses the 151-item wordlist and which expresses the confidence in each branch of the tree. In general, our tree expresses the unity of Southeastern Luyia with Logooli (JE41), Isukha (JE412), Itakho (JE411), and Tiriki (JE413) (and also confirms the subgrouping of Isukha-Itakho-Tiriki), but there is one surprise, which is the inclusion of Kisa (JE32D) within this cluster. Other prior studies have placed Kisa (JE32D) within a Central cluster, but our tree here unfolds more like an onion with a number of layers that are added as one moves to the west and north within Luyialand.

Several Central varieties cluster next with the Southeastern group, beginning with Nyore (JE33) and followed by the two Marama (JE32C) datasets. It is surprising that the two Marama datasets do not cluster with one another first, but the low-confidence grouping of Marama_1 before the grouping with Marama_2 may reflect the uncertainty of grouping the two Marama datasets together first. Next, there is a branching with a cluster that includes Wanga (JE32a) and Kabarasi (JE32E) (though the confidence in the Wanga-Kabarasi cluster is somewhat low at 68%). Next is a low-confidence (40%) branching with Marachi, followed by a surprising branching with Nyala West and then a fairly low confidence (53%) branching with a Northern cluster that includes a Bukusu-Gisu subgroup and Nyala East. The branching with Nyala West is surprising based on geography because prior classifications like Williams (1973) include it in a cluster with Southwestern Luyia varieties like Khayo (JE341) and Saamia/Gwe. Instead, Saamia/Gwe and Khayo (JE341) attach at the highest most levels to the Luyia cluster. The Bukusu-Gisu cluster follows prior classifications and history that connects Bukusu and Gisu communities.

5 Future research

In future research, we would like to include additional languages in our database, including the JE20 languages around Lake Victoria for better establishing the

position of Luyia with respect to its regional neighbors. Our collaborator Minah Nabirye has recently collected a 200-item Swadesh wordlist for Kenyi, a language that has not figured in prior classifications. We would like to work with Nabirye and Gilles-Maurice de Schryver to include as many Bantu languages of Uganda as possible, including those studied in Nabirye's (2016) dissertation.

As far as further studying the internal structure of Luyia is concerned, we would like to add data on Luyia varieties currently missing from our 200-word comparison, such as Saamia, Tachoni, and Tura, and we would like to incorporate data from another speaker for several of the datasets from Marlo's *Luyia Dictionary Project*, which collected preliminary dictionary materials based on Appleby's (1943) *Luluhya-English Vocabulary* in 14 varieties: Bukusu, Isukha, Itakho, Kabarasi, Kisa, Khayo, Logoori, Nyala West, Nyore, Saamia, Tachoni, Tiriki, Tsootso, and Tura. The highest priority are varieties like Kabarasi and Kisa, which have a somewhat unexpected position in the tree generated by the present classification.

Acknowledgments

We are grateful for financial support from the University of Missouri College of Arts & Science Undergraduate Research Mentorship Program, Campus Writing Program, Honors College, Office of Undergraduate Research, and Research Board, and National Science Foundation Award BCS-1355750. We would like to thank two anonymous reviewers and University of Missouri student Bobby Love for thoughtful feedback on our paper. We also thank Kristopher Ebarb, Deo Kawalya, Minah Nabirye, and David Odden for sharing data with us, Thilo Shadeberg for providing a scanned copy of Williams (1973), Kelvin Alulu and Alfred Anangwe for their assistance in data collection, and our many language consultants for sharing their languages with us and making the present analysis possible.

References

Anderson, Heidi, Leah Schreiner & Sabine Weidemann. 2004. *Lusamia-Lugwe Language Assessment report*. Tech. rep. Entebbe, Uganda: SIL International.

Angogo Kanyoro, Rachel Msimbi. 1983. *Unity in diversity: A linguistic survey of the Abaluyia in Western Kenya*. Vienna: Afro Publications.

Appleby, Leonora L. 1943. *A Luluhya-English vocabulary*. Maseno: Church Missionary Society.

Brown, Gillian. 1972. *Phonological rules and dialect variation: A study of the phonology of Lumasaaba*. Cambridge: Cambridge University Press.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–265.

Chavasu, Henry O. 1997. *British colonialism and the making of the Maragoli diaspora in Western Kenya, 1895–1963*. Eldoret: Moi University. (MA thesis).

Fitch, Walter M. 1997. Networks and viral evolution. *Journal of Molecular Evolution* 44(1). S65–S75.

Guthrie, Malcolm. 1948. *The classification of the Bantu languages*. London: Oxford University Press.

Guthrie, Malcolm. 1967. *Comparative Bantu: An introduction to the comparative linguistics and prehistory of the Bantu languages*. Farnborough: Gregg Press.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2020. *Glottolog 4.3*. Jena. DOI: 10.5281/zenodo.4061162.

Heine, Bernd. 2013. The language and dialect atlas of Kenya project 1973–1980. In Karsten Legère (ed.), *Bantu languages and linguistics: Papers in memory of Dr. Rugatiri D.K. Mekacha*, 187–197. Bayreuth: Bayreuth African Studies (BASS).

Heine, Bernd & Wilhelm J. G. Möhlig. 1980. *Language and dialect atlas of Kenya*. Vol. 1. Berlin: Deitrich Reimer Verlag.

Huntingford, G.W.B. 1965. The Orusyan language of Uganda. *Journal of African Languages* 4. 145–169.

Larget, Bret & Donald L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16. 750–759.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the World*. 19th edn. Dallas, Texas: SIL International. http://www.ethnologue.com.

MacArthur, Julie. 2016. *Cartography and the political imagination: Mapping community in Colonial Kenya*. Ohio University Press.

Maho, Jouni Filip. 2009. NUGL Online: The online version of the New Updated Guthrie List, a referential classification of the Bantu languages.

Marlo, Michael R. 2007. *The verbal tonology of Lumarachi and Lunyala-West: Two dialects of Luluyia (Bantu, J.30, Kenya)*. Ann Arbor: University of Michigan. (Doctoral dissertation).

Marlo, Michael R. 2008. Tura verbal tonology. *Studies in African Linguistics* 37. 153–243.

Marlo, Michael R. 2009. *Luyia tonal dialectology*. Paper presented at the University of Nairobi, Department of Linguistics and Languages.

Marlo, Michael R., Minah Nabirye, Deo Kawalya & Gilles-Maurice de Schryver. 2017. A sketch of Lower Nyole tone. *Africana Linguistica* 23. 215–257.

Möhlig, Wilhelm J.G. 1985. Review of Kanyoro (1983). *Journal of African Languages and Linguistics* 7(2). 203–207.

Mould, Martin J. 1976. *Comparative grammar reconstruction and language subclassification: The North Victorian Bantu languages*. Los Angeles: UCLA. (Doctoral dissertation).

Mould, Martin J. 1981. Greater Luyia. In Thomas J. Hinnebusch & Derek Nurse (eds.), *Studies in the classification of Eastern Bantu languages*. Hamburg: Helmut Buske Verlag.

Nabirye, Minah. 2016. *A corpus-based grammar of Lusoga*. Ghent, Belgium: Ghent University. (Doctoral dissertation).

Nurse, Derek & Gérard Philippson. 1975. *Tanzania language survey*. http://www.cbold.ish-lyon.cnrs.fr/.

Nurse, Derek & Gérard Philippson. 1980. The Bantu languages of East Africa: A lexicostatistical survey. In Edgar C. Polomé & C. P. Hill (eds.), *Language in Tanzania*, 26–67. London: Oxford University Press.

Odden, David. 2009. Tachoni verbal tonology. *Language Sciences* 31. 305–324.

Pagel, Mark & Andrew Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 53(4). 571–581.

Purvis, John Bremner. 1907. *A manual of Lumasaba grammar*. London: Society for promoting Christian knowledge.

Schoenbrun, David L. 1994. Great Lakes Bantu: Classification and settlement chronology. *Sprache und Geschichte in Afrika/Language and Culture in Africa* 15. 91–152.

Siertsema, Berthe. 1981. *Masaba word list: English-Masaba, Masaba-English*. Tervuren: Royal Museum for Central Africa.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4). 452–463.

Were, Gideon & Osaga Odak. 1987. Siaya district: Socio-cultural profile. A joint research and training project of The Ministry of Planning and National Development and the Institute of African Studies, University of Nairobi.

Williams, Ralph M. 1973. *A lexico-statistical look at Oluluyia*. Paper presented at the 4th Annual Conference on African Linguistics. New York, NY: Queens College.