Bantoid lexical diversity from an individual-based perspective

Jeff Good*, Nelson C. Tschonghongei*, Pierpaolo Di Carlo†, and Clayton Hamre*

*University at Buffalo

†University of Naples "L'Orientale"

Abstract

This chapter presents the results of a study of lexical data from a set of Bantoid languages spoken in the Lower Fungom region of the Cameroonian Grassfields, an area characterized by a high degree of multilingualism. Individual-based wordlists are compared with each other in a manner analogous to how wordlists representing distinct languages are compared in more typical kinds of investigation. The results reveal a higher level of individual-level variation than would be expected based on the way that wordlist data has generally been presented from the Grassfields area. This suggests that modeling patterns of language diversification in the Bantoid area may need to take into account a higher level of baseline variation among individuals than has been the norm in earlier work. This work also has implications for our understanding of lexical variation in highly multilingual societies and for the historical stability of specific lexical items, which is significant from a prehistorical perspective given that early Bantu populations may have shown similar patterns to what is presented in this study.

1 Lexical diversity in Bantoid languages

To the extent that it can be historically reconstructed, it appears that the communities of the Bantoid area (including the Bantu area) have long been characterized by widespread individual-level multilingualism (see, e.g., Schadeberg 2003: 158–159), and these historical patterns continue to the present day in many areas (Di Carlo, Good & Ojong Diba 2019).[1] This is especially true of the Cameroonian Grassfields (Warnier 1980), an important region within

[1] The authors would like to acknowledge comments made on presentations of this work by audiences at the Small-Scale Multilingualism Conference 2 at the University of Helsinki, the University of Rochester Department of Linguistics, the Linguistic Convergence Laboratory of HSE University, and the Princeton Phonology Forum, as well as feedback from anonymous reviewers. The first author had the primary responsibility for the writing of this paper, the overall framing of the presented results, and the application of the computational methods, the second author was responsible for the collection of the data, the third author contributed to the overall conceptualization of the work, and the fourth author conducted the initial data analysis that the analytical discussion of this paper is based on. This work was supported by NSF Award No. BCS-1761639 and NSF Award No. BCS-2109620.

the non-Bantu Bantoid area both due to the extent of its linguistic diversity (Watters 2003; Blench 2015) and its location in the general area that has been proposed as the Narrow Bantu homeland (Nurse & Philippson 2003: 5; Grollemund et al. 2015) (though see Idiatov & Van de Velde 2021: 98 for an alternative proposal).

Recent phylogenetic work has advanced our understanding of the internal and external relations of the Bantoid languages (Grollemund 2012; Hombert & Grollemund 2018), but, to this stage, work of this kind has not been able to directly consider how the multilingual realities of Bantoid speakers might have impacted patterns of language change. Not only might we expect greater opportunity for lexical borrowing across languages in multilingual societies, but they would also present a context where more distinctive kinds of changes, such as patterns of complexification associated with linguistic esoterogeny might be expected to occur.[2] Multilingual individuals, with knowledge of many languages spoken in a given area, would be in a position to initiate specific kinds of changes that could make their primary language more distinctive in the local linguistic space, if this was deemed useful to achieve some set of social goals (see Mve et al. 2019) for a potential case of this in the Bantoid area). Such changes could take place alongside better-studied kinds of changes such as contact-induced convergence, stability, and simplification (see Trudgill 2004; Kühl & Braunmüller 2014; and Di Carlo & Good 2023 for relevant discussion).

If patterns of multilingualism in the Bantoid area are to be more directly incorporated into historical linguistic studies, individual-level linguistic knowledge will necessarily take on an important role in such investigations. This is because, in non-urban areas characterized by extensive interaction among language communities, the individuals comprising a given language community by virtue of sharing at least one common language, will otherwise have

[2] Linguistic esoterogeny refers to changes that make a language harder for outsiders to learn (Thurston 1987, 1989, Dimmendaal 2009).

different multilingual repertoires connected to their specific life histories.[3] This further calls for methodological approaches that put individual-level data more directly in focus rather than working with data primarily at the level of “languages”. Individual-level approaches to linguistic analysis are not unusual in some areas of linguistics, such as sociolinguistics and, in particular, in so-called third-wave sociolinguistic studies that emphasize how language users employ variation to craft linguistic styles and, in turn, their social identities (see Eckert 2012). However, individual-level data in historical-comparative work is not regularly employed. For instance, it is standard to use a single wordlist to represent the lexical patterns of a specific language without detailed consideration of its provenance, such as whether it was collected from a single speaker, comprises an amalgamation of data from multiple speakers, or represents the consensus of multiple speakers.

In this paper, we present the results of comparative work based on wordlists collected at the individual-level from thirteen Bantoid linguistic varieties associated with the Lower Fungom region of Cameroon. While this work is still exploratory, we believe that it demonstrates the need to consider linguistic variation from an individual-based perspective when reconstructing the history of Bantoid, and, by extension, Bantu as well. We further hope that our work can serve as a foundation for more extensive studies of this kind that can ultimately be used to give us a clearer view of Bantoid prehistory. In §2, we describe the dataset that this study is based on. In §3, we describe the patterns of variation that have been found in our individual-based wordlist data at the present state of investigation. In §4, we place the results of this study within the wider comparative Bantoid context and consider its implications for future work on the comparative linguistics of this group with a focus on the role of multilingualism in its historical development.

[3] This sociolinguistic configuration falls under the heading of what has been referred to as small-scale multilingualism (Lüpke 2016, Pakendorf, Dobrushina & Khanina 2021)—i.e., the multilingualism of small-scale societies—in contrast to urban multilingualism.

2 A new dataset of individual-based lexical data

The work described in this paper is being undertaken in the context of the Key Pluridisciplinary Advances on African Multilingualism project (KPAAM-CAM), which is studying endangered languages of the Grassfields Region of Cameroon. This project has focused, in particular, on the languages of an area known as Lower Fungom (Good et al. 2011; Di Carlo 2011), from which the data used in this paper is drawn. The languages of Lower Fungom are all classified within Bantoid, but the precise relationships of most of the languages of the region with the rest of Bantoid remain unclear. A map of the region is provided in Figure 1, where Lower Fungom itself is encircled with a dotted line.[4] Our current understanding of the linguistic situation of the region suggests that its thirteen villages should be treated as associated with eight languages. Six of them, which are grouped under the referential label Yemne-Kimbi, do not have established close relatives outside of Lower Fungom, and they do not appear to be closely related to each other beyond the fact that the language associated with the villages of Mufu and Mundabli and the one associated with the village of Buu form a small subgroup. The Missong variety of Mungbam should, perhaps, also be considered a separate languag from the other Mungbam varieties rather than a dialect due to its distinctiveness from the other four Mungbam dialects. (A partial lexical basis for these classifications will be presented in §3.) Four of the region's language groups, Ajumbu, Fang, Koshin, and Kung, are restricted to a single village. However, Kung has been classified with the Central Ring group of languages and has close relatives outside of Lower Fungom (Akumbu & Kießling 2023). The village of Mashi is associated with a variety of a language known as Naki in reference sources, which is also associated with a number of villages outside of Lower Fungom, as indicated in Figure 1.

[4] While early classifications treated most of Lower Fungom's languages as part of the Western Beboid subgroup of Beboid (Hombert 1980), no evidence for the coherence of the Western Beboid group or a close affinity between Western Beboid and Eastern Beboid, as originally proposed, has been found. Accordingly, Good et al. (2011) propose a referential label Yemne-Kimbi for the languages formerly referred to as Western Beboid. The precise relationship between each of the Yemne-Kimbi languages and the rest of Bantoid remains an open question.

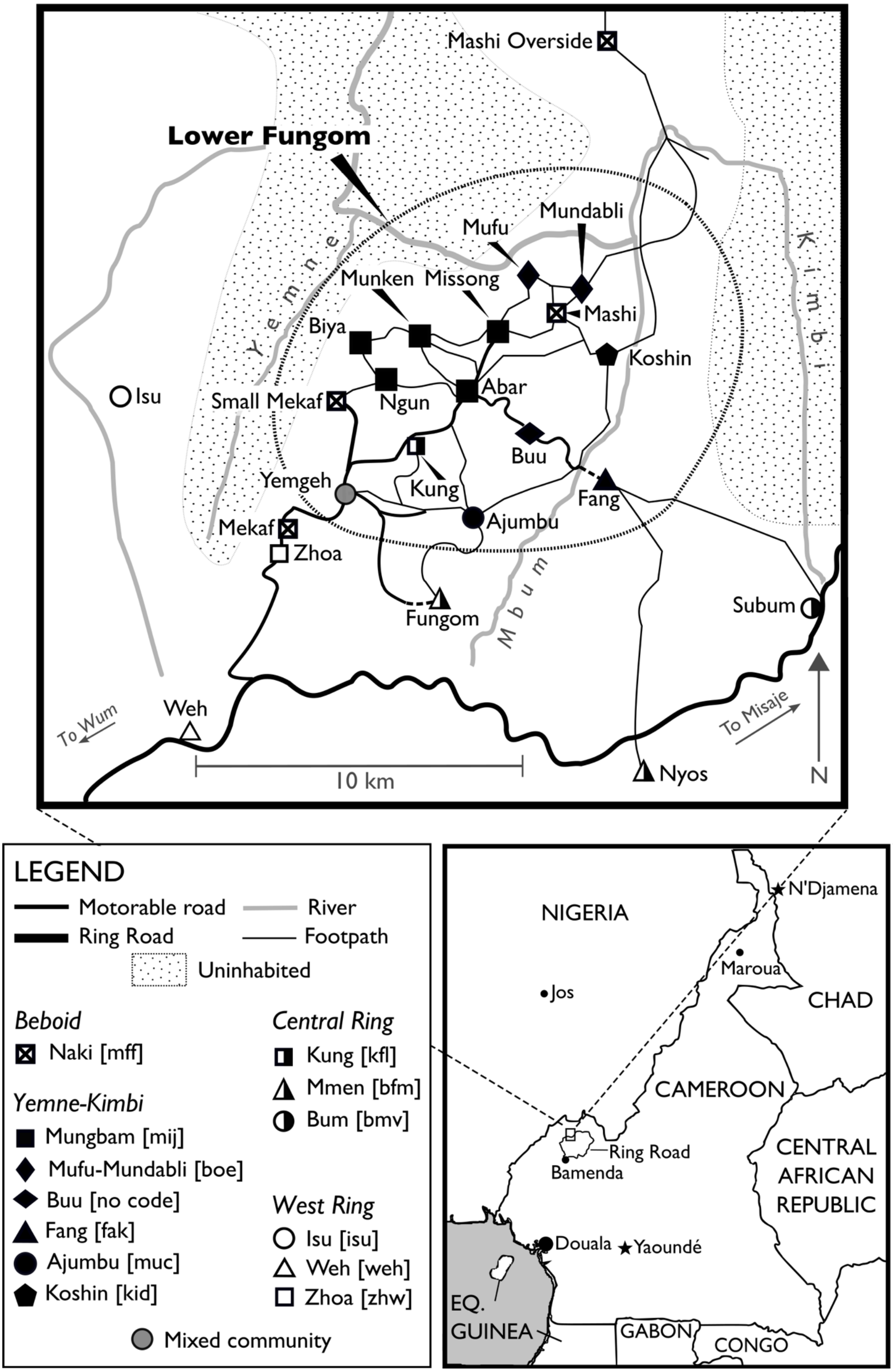

Figure 1: Lower Fungom and the surrounding region (map created by Pierpaolo Di Carlo)

In Table 1, each of the Lower Fungom villages, grouped by the languages that they are associated with are listed, with their ISO 639-3 code provided. The dotted line separating Missong from the rest of Mungbam in Table 1 is used to indicate its divergent status from the other Mungbam varieties. The variety of Buu is currently classified with Mufu and Mundabli in the ISO 639-3 system, but it is clearly a distinct language, which is why it is separated from them using a solid line.[5] The current classifications of the languages are also provided, though, as noted above, Yemne-Kimbi should be understood as a referential classification rather than a genealogical one. The final column of the table includes a list of the identifiers for each of the individual-based wordlists examined in this study, discussed further below. Data from each speaker is treated as an individual-based doculect (see Cysouw & Good 2013), which is identified by a sequence of initials for the speaker (e.g., *ECL*) followed by the name of a locally recognized variety (e.g., *Abar*) followed by a number that is used to ensure that the same identifier will not accidentally be used twice (e.g., *8*).[6] For purposes of this study, the most informative part of the identifier is the variety name, which begins with the fourth capital letter of the identifier. Each of the identifiers used in this study is included in Table 1 to show how the doculects are linked to Lower Fungom's village-level linguistic varieties.

While scholarly classifications treat Lower Fungom as associated with seven to nine languages (depending on where the line between a dialect and a language is drawn), in the local sociolinguistic space, each village is understood as having a distinct linguistic variety, and this is one of the defining features of the village as an independent political entity (see, e.g., Di Carlo & Good 2014 for relevant discussion). Scholarly assessments are in line with

[5] The current ISO 639-3 naming scheme confusingly groups all three varieties under the name *Mundabli*. This appears to reflect the fact that this village was surveyed as part of the work described in Hamm et al. (2002), while the other villagers were not. The use of the name Mundabli to cover all three varieties has no grounding in local sociolinguistic realities and is also not useful for scholarly linguistic work.

[6] In this system for identifying individual-based varieties used here, the alternate spelling Mumfu is used for the variety referred to as Mufu elsewhere in this paper.

the local understanding insofar as they have also found that each village is associated with a clearly distinct variety, even if some of these are classified as dialects of each other in reference sources. Accordingly, when considering the lexicogrammatical diversity of a region like Lower Fungom, it is important to consider data on the varieties of all of its villages in order to fully capture the linguistic situation. From the perspective of the study of Bantoid languages, Lower Fungom's geographic compactness allows its high degree of linguistic diversity to be explored in an unusual level of detail. We believe that this allows it to serve as a microcosm for the much larger Bantoid and Bantu areas and that its linguistic patterns are likely to hold lessons for the study of high-level patterns of diversification within the family, as further discussed in §4.

SUBGROUP	LANGUAGE	VILLAGE	DOCULECTS
Yemne-Kimbi	Mungbam [mij]	Abar	ECLAbar8, NACAbar2, NMAAbar1, NVBAbar7
		Munken	NEAMunken1, NGTMunken3, NUNMunken4, TNTMunken2
		Ngun	AOMNgun2, KBMNgun4, MCANgun3, WCANgun1
		Biya	ENBBiya1, FBCBiya8, ICNBiya2, NFKBiya7, NJNBiya6, NSFBiya5
		Missong	ABSMissong1, AGAMissong2, NDNMissong5, NMSMissong4
	Ji group [boe]	Mundabli	CENMundabli2, LFNMundabli1, NINMundabli4, NMNMundabli3
		Mufu	APBMumfu1, DNMMumfu2, MEAMumfu3, NCCMumfu4
		Buu	KCYBuu2, KEMBuu1, MNJBuu4, NNBBuu3
	Fang [fak]	Fang	DPNFang13, KDVFang1, KHKFang12, KJSFang2
	Koshin [kid]	Koshin	JGYKoshin3, MRYKoshin2, TELKoshin4
	Ajumbu [muc]	Ajumbu	KDCAjumbu10, KMNAjumbu2, NEMAjumbu9, NVIAjumbu1
Beboid	Naki [mff]	Mashi	BAAMashi4, BKBMashi2, KFKMashi1, NCMMashi5
Central Ring	Kung [kfl]	Kung	BNMKung2, KCSKung3, NJSKung4, ZKGKung1

Table 1: Lower Fungom's linguistic varieties and the doculects examined in this study

In order to explore this research possibility concretely, wordlists are being collected from multiple individuals across all of Lower Fungom's thirteen villages. This is an ongoing project, and we report on the current state of our results here. The work builds on methods developed by Angela Nsen Tem which are discussed in Mba & Nsen Tem (2020: 212–213). The key feature that makes this work different from more typical approaches to wordlist collection is the lack of any attempt at standardization of the wordlists across speakers in order to create a single description of the vocabulary of a "community".[7] Instead, the work with each individual is conducted without the presence of other speakers of the variety, and the words that they produce are recorded as provided. These wordlists can subsequently be compared with each other to assess how similar or different they are. For this study, data was only gathered from individuals who could be reasonably classified as first-language speakers of a variety.[8]

The dataset on which this paper is based consists of more than 18,000 individual wordlist entries across fifty-three speakers. Four wordlists are available for eleven of Lower Fungom's thirteen varieties, three wordlists used for one of the remaining two varieties (Koshin), and six for the last variety (Biya). Data collection was primarily the responsibility of the second author and work was done in two phases using two different versions of a standardized concept list. Individuals were chosen as consultants largely based on their availability. Collected forms were then entered into a database from handwritten notes by Charles Nyoh Abang and further processed using the CLDFBench framework (Forkel & List 2020) to facilitate analysis using LingPy (List & Forkel 2021). While some degree of semi-automated data cleaning was done using the tools provided by CLDFBench, individual forms have not yet been double checked by hand, and some errors almost certainly remain in the data. Nevertheless, we believe that the large size of the dataset limits the extent to which these would significantly impact the results presented here, especially since this study is considered to be exploratory rather than definitive. The dataset and methods are discussed in more detail in §3.2.

[7] We are not aware of other work that adopts the approach developed here. The work that we have found that is most similar is Slaska (2005, 2006), but it is focused on eliciting multiple wordlists to examine methodological aspects of wordlist collection rather than to study variation with the languages from which the wordlists were collected.

[8] Since high levels of individual-level multilingualism are the norm in Lower Fungom (Esene Agwara 2020), we hope to extend this work in the future to collect wordlists drawn from different varieties from the same individual as a means of gathering data on the full range of their linguistic repertoires. Among other things, extending our approach in this way will help allow us to avoid some of the problems associated with determining what counts as a "first language" in multilingual African contexts (see, e.g., Lüpke & Storch 2013:22).

Alongside the wordlist data, detailed sociolinguistic information was also collected from each speaker (see Esene Agwara (2013: 118–119) for an example of the kind of questionnaire used in this study). This information was collected on the assumption that there will be important correlations between an individual's patterns of lexical knowledge and their sociolinguistic background, though studying possible connections of this kind is outside of the scope of the present chapter.

3 Patterns of individual-based variation

3.1 Analyzing synchronic variation for historical applications

Our analysis of the wordlist data is guided by several research questions, and this has also influenced a number of our methodological choices. These are: (i) How extensive is individual-level lexical variation within the linguistic varieties of Lower Fungom? (ii) Do some varieties show more individual-level variation than others? (iii) What are the overall patterns of lexical similarity across Lower Fungom varieties, and how clear-cut are the boundaries between languages (as classified by scholars) and varieties (following the local categorization system)? And, (iv) which concepts appear to be associated with more stable patterns of expression (e.g., based on similar roots), and which appear to be less stable in their expression?

The results presented below are intended to be interpreted primarily in synchronic terms, though one of our key goals in doing this work is to develop more accurate models of language change within Bantoid. This is because we believe that this requires a better understanding of the synchronic sociolinguistic dynamics within communities whose patterns of variation are likely to be representative of the historical situation for the family. In order to undertake this synchronic analysis, we make use of tools originally designed for diachronic investigation due to the fact that these tools are designed to detect and visualize similarities in lexical data. The work described here also overlaps with work in dialectometry, i.e., the study of dialects using computational and statistical methods (Wieling & Nerbonne 2015), given that individual-based wordlists can be analogized to data collected from dialects of a single language. The fact that the dataset includes individual doculects drawn from varieties associated with different languages means that it also overlaps with areal linguistic studies (Good 2013). This methodological eclecticism is intended to lay the groundwork for more detailed work on sociolinguistic reconstruction which can, in turn, inform more traditional historical work (see, e.g., Good under review).

As discussed below, in §3.2 and §3.3, implementing this overall analytical approach requires making a number of concrete methodological choices, some of which are motivated by practical considerations and others of which are motivated by a mix of practical and conceptual considerations. While we think the results of this work demonstrate the promise of our general approach and the value of individual-based wordlists for improving our historical models, we see this work as largely exploratory and expect that significantly more methodological experimentation will be needed to take full advantage of this dataset, or any similar ones that might be collected.

3.2 Structure of the dataset

In order to use the dataset to explore the questions discussed in §3.1, we made use of methods implemented in LingPy (List & Forkel 2021) that were originally designed to facilitate the historical analysis of wordlist data (see, e.g., List et al. 2018). In particular, we made use of the features of LingPy designed to detect cognate forms in wordlists, while accounting for phonetic differences across varieties. We specifically used the work reported on in Hantgan & List (2022) as a model, due to its consideration of both genealogical and contact relations using wordlist data and the fact that it was focused on language groups of Africa. However, for this work, we are not interested in whether or not words are true historical cognates but, rather, the extent to which speakers produced relatively similar forms for a given concept.

Therefore, while many of the "cognate" sets found in the data clearly represent actual cognates, we avoid use of the word cognate below and instead refer to the groupings detected as *similarity sets* since our initial focus is not on historical relationships but synchronic similarities.

In order to detect similarity sets, the Sound-Class-Based Phonetic Alignment (SCA) method of List (2012) was employed, as implemented in LingPy.[9] This method was chosen because it is useful as a measure of the synchronic sound-based similarity holding among words rather than being designed specifically to detect older historical correspondences (as is the case for the LexStat method discussed in Hantgan & List 2022). Therefore, we view it as an appropriate method for understanding the synchronic areal linguistic situation of Lower Fungom, at least at this initial stage of investigation.[10]

Unlike work focused on genealogical relations, which might specifically choose to use a Swadesh list (see, e.g., Swadesh 1955) or the Leipzig-Jakarta list (Tadmor 2009:67), the list of concepts used as the basis of this study was developed in a more ad hoc fashion. It was first based on a reduction of longer wordlists (with basic concepts retained), such as Roberts & Snider (2006), which was then augmented with terms for salient local cultural concepts. The precise selection of these terms was the primary responsibility of the second author, who is from a region close to Lower Fungom and who is quite familiar with the area. At the present stage of this research, it is not clear to us how the specific choice of concepts used for this study may have led to significantly different results than if a more standardized list of concepts had been used, and we see this as something to explore in further research.

[9] For the results presented in this paper, the distance threshold for detecting members of similarity sets was set to 0.45. This threshold was used in Hantgan & List (2022), and it follows the recommendation of List, Greenhill & Gray (2017:9).

[10] In our view, the ideal comparison method would not be based on general linguistic principles of phonetic similarity but, rather, a metric that corresponded to local perceptions of similarity and difference by the multilingual speakers themselves. However, given that we lack the information needed to develop such a metric, a sound-based method was seen as the most practical approach here.

The presently available dataset consists of more than 18,000 individual wordlist entries across fifty-four speakers. While verbs were collected for some speakers, only data from nouns was consistently collected for all speakers, and the data reported on here, therefore, only involves nouns.[11] Four wordlists are available for twelve of Lower Fungom's thirteen varieties and six for one variety (Biya). However, one wordlist was removed for the studies reported on here, resulting in just three wordlists for Koshin, due to the fact that the forms produced by the relevant individual varied so extensively from those of all of the other speakers that it was, in effect, an outlier.[12]

While the wordlists were based on similar concept lists that allowed for comparability between them, their overall coverage differed across each variety both due to differences in responses from speakers and due to adjustments to the standard concept list made during the course of data collection. In the results reported below, only concepts for which there were entries across at least forty of the fifty-three wordlists are presented, which represents around 7,000 total words. This specific cutoff, at 75% coverage, was intended to achieve a balance between ensuring there was decent coverage across all the concepts analyzed while also working with a concept list that would be long enough for clear results to emerge. For this dataset, this resulted in a concept list consisting of 138 concepts, presented in Table 2 with an indication of the number of wordlists the concept appeared in.

3.3 Analysis of the variation in the data

In order to keep the scope of the present study manageable, various analytical choices needed to be made. In some cases, these were primarily practical in nature. For example, in the

[11] The entry for one noun in the list, 'fly', was removed from the study after it was noted that the actual forms listed appear to have been inadvertently mixed in with forms for the verb 'fly'.

[12] Due to ongoing conflict in Cameroon referred to under the heading of the Anglophone Crisis (see Pommerolle & De Marie Heungoup 2017, Anchimbe 2013), it has not been possible to locate this speaker to arrive at a clearer understanding of the source of this variation. However, an informal inspection of the data that they provided shows that, for many nouns, they produced singular–plural pairings showing different roots in the singular and the plural, and the plural forms often showed a close match with the plural forms of other speakers, even when the singular forms did not. This interesting pattern of variation clearly merits further investigation, but this is outside the scope of the present paper.

complete wordlist dataset, a given concept was often associated with more than one word either because a consultant produced variant forms or because multiple forms of a word were provided (e.g., the singular and plural forms for nouns). In this study, only the first word provided for a given concept was considered in the analysis due to the data processing challenges involved with determining which forms represented alternate forms of the same word and which represented variants of other kinds.[13] Similarly, the original transcribed data made use of a wide variety of characters which were mapped onto a standard character set, with the result that some relatively minor distinctions may have been lost in the mapping. Based on our experience working with different mappings, the precise choices can influence the similarity sets that were detected, though not at a level where we believe the overall results of the paper would change. The dataset, character mapping, analysis scripts, and related materials on which this work is based are provided in a Zenodo repository so that the overall process of analysis can be made more transparent and that all of the similarity sets that were detected can be individually examined.[14]

One important choice in the analysis of the data was made for a mix of practical and conceptual reasons. One of the most significant of these is the fact that words were not segmented morphologically (e.g., to separate noun class prefixes from roots). This would have improved the overall results of the application of the automatic alignment algorithms, for instance by minimizing the chances that a prefix sequence in one variety will be aligned with a phonetically similar portion of a root in another variety. If this study were primarily focused on

[13] We have not undertaken the kind of analysis needed to determine how the inclusion of alternate forms might impact the results presented here. There are various possibilities that need to be kept in mind. If plural forms were included, we would expect them to generally share roots with the singular forms, but variation in prefixes used to form plurals could result in them being placed in different similarity sets from corresponding singular forms. If variant forms are included, presumably those would increase similarities among some doculects, for example, where one speaker's first form matched another speaker's alternate forms, but the alternate forms may also introduce new kinds of variation that would reduce similarity scores. Finally, there is the question of how to interpret variant forms in sociolinguistic terms. Should the first form that speakers produce be given more weight than a second variant form, for example? We leave these general issues open for further research here, while acknowledging that they could impact the overall results in significant, if unknown, ways.

[14] The repository can be accessed at: https://doi.org/10.5281/zenodo.15814992.

the reconstruction of proto-forms or the establishment of historical relationships, then morphological parsing would have definitely been warranted. However, for this study the value of that is less clear since the presence or absence of a noun class prefix, or the use of a different noun class prefix, could be a significant marker of similarity or difference in the local sociolinguistic space (see Di Carlo & Good 2023:§5 for relevant discussion). How precisely to handle nominal morphology in a study like this one remains an open question, in our view.

axe	53
bird	53
breast	53
cat	53
chief	53
corn	53
devil	53
ear	53
egg	53
fish	53
forest	53
garden egg	53
grave	53
hair	53
hand	53
head	53
heart	53
house	53
jaw	53
ladder	53
name	53
nose	53
palm nut	53
palm tree	53
pepper	53
pig	53
place	53
plantain	53
pot	53
potato	53
rain	53
snake	53
song	53
stomach	53
tongue	53
yam	53
animal	52
bag	52
banana	52
basket	52
bitter leaf	52
blood	52
book	52
broom	52
caterpillar	52
chair	52
child	52
crab	52
cricket	52
cup	52
cutlass	52
deity, god, God	52
dust	52
eye	52
faeces	52
farm	52
fowl	52
grass	52
leaf	52
monkey	52
moon	52
root	52
salt	52
smoke	52
sun	52
tooth	52
trap	52
tree	52
water	52
air	51
bed	51
compound	51
cow, cattle	51
day	51
drum	51
friend	51
gong	51
gun	51
hoe	51
intestine	51
mother	51
mouth	51
owl	51
pineapple	51
sand	51
seed	51
sieve	51
soap	51
stone	51
storm (wind)	51
story	51
war	51
case (court) (n)	50
father	50
feather	50
fire	50
goat	50
grasshopper	50
hailstone	50
hill	50
horse	50
medicine	50
person	50
raffia bamboo	50
sheep	50
toilet	50
umbrella	50
xylophone	50
oil	49
rope	49
soldier ant	49
star	49
dog	48
fence	48
bat	47
dance	47
sky	47
bridge	46
cloth	46
headpad	46
knife	46
zinc	46
cap	45
fireside	45
rainbow	45
camwood	44
gizzard (fowl)	44
pap	44
road	44
work (n)	44
spider	43
belly	42
termite	42
wingless termite	42
dry season	41
elephant stalk	41
horn (head)	40
mushroom	40

Table 2: Concepts used in this study, including number of wordlists each concept appears in

Despite these limitations, where key linguistic patterns of differentiation were already known as a result of qualitative research, the results to be described below are in line with them, suggesting that the overall patterns found via automated similarity set detection are reliable for exploratory work of the sort undertaken here. In this regard, the fact that this study is based on a region where we, independently, have detailed knowledge of its linguistic and sociolinguistic situation is useful since it helps us assess the extent to which the new results are sensible in the context of what is already known. However, it would clearly be beneficial to make use of more of the data and to segment it morphologically where possible in future work, if for no other reason than to compare how this would change the overall results. This would be especially valuable if there were attempts to apply similar methods to regions that are not as well studied as Lower Fungom and where it would not be possible to do a "reality check" of how well the results compare to the linguistic patterns discovered using more traditional techniques.

In Table 3, an example is provided of the kind of data that forms the basis of this study. It provides the forms collected for the concepts of 'rain' and 'heart' across the fifty-three speakers from whom wordlists were analyzed. The transcription of the forms has been standardized to IPA for segments with superscript numbers used to represent tones. Three levels are distinguished, with a 5 representing a high tone, a 3 representing a mid tone, and a 1 representing a low tone. This was done to facilitate processing using LingPy. The data is divided into the similarity sets detected using the tool's SCA algorithm.[15]

[15] The file kplfSubset-SCA-0.45_threshold-aligned.html in the supplementary materials presents all of the similarity sets, including the analysis of segmental alignments, in a format that is relatively easy to read. These materials also include a machine-readable version of the same information in the file kplfSubset-0.45_threshold-cognates.tsv.

Forms collected for ‘rain’

DOCULECT	FORM
ECLAbar8	ɪ^{1}bʷu^{51}
NACAbar2	ɪ^{1}bwu^{51}
NMAAbar1	i^{1}bwu^{51}
NVBAbar7	ɪ^{1}bʷu^{51}
ENBBiya1	ɪ^{5}bʷu^{1}
ENBBiya1	ɪ^{5}bʷu^{1}
FBCBiya8	ɪ^{5}bʷuː^{5}
ICNBiya2	ɪ^{1}bʷu^{1}
ICNBiya2	ɪ^{5}buː^{5}
NFKBiya7	ɪ^{1}bʷu^{1}
ABSMissong1	i^{1}bwu^{1}
AGAMissong2	ɪ^{5}bwuː^{5}
NDNMissong5	ɪ^{1}bʷu^{51}
NMSMissong4	ɪ^{5}bʷuː^{51}
NEAMunken1	ɪ^{1}bwu^{51}
NGTMunken3	ɪ^{1}bo^{1}
NUNMunken4	ɪ^{5}bʷu^{51}
TNTMunken2	ɪ^{1}bu^{51}
AOMNgun2	ɪ^{5}bu^{1}
KBMNgun4	ɪ^{5}bʷu^{1}
MCANgun3	ɪ^{1}bʷuː^{5}
WCANgun1	ɪ^{1}bu^{1}
KDCAjumbu10	bʷə^{1}
KMNAjumbu2	bwoː^{1}
NEMAjumbu9	bʷɛː^{51}
NVIAjumbu1	bwə^{1}
DPNFang13	bʷə^{5}lə^{5}
KHKFang12	bʷə^{5}lə^{5}
KDVFang1	bwə^{5}lə^{5}
KJSFang2	bwə^{5}lə^{5}
KCYBuu2	dʒə^{5}ŋ
KEMBuu1	dʒə^{5}ŋ
MNJBuu4	dʒa^{5}ŋ
NNBBuu3	dʒə^{5}ŋ
JGYKoshin3	dza^{1}ŋ
MRYKoshin2	dza^{1}ŋ
TELKoshin4	za^{1}ŋ
BAAMashi4	dza^{1}ŋ
BKBMashi2	dza^{1}ŋ
KFKMashi1	dza^{1}ŋ
NCMMashi5	dza^{1}ŋ
APBMumfu1	ɠjə^{3}ŋ
DNMMumfu2	gɪː^{5}ŋ
MEAMumfu3	gjə^{5}ŋ
NCCMumfu4	gjə^{5}ŋ
CENMundabli2	dzə^{3}ŋ
LFNMundabli1	dzə^{3}ŋ
NINMundabli4	dzə^{5}ŋ
NMNMundabli3	dzə^{5}ŋ
BNMKung2	i^{1}wo^{5}l
KCSKung3	ɪ^{1}ɣo^{5}l
NJSKung4	ɪ^{1}ɣo^{5}l
ZKGKung1	i^{1}wo^{5}l

Forms collected for ‘heart’

DOCULECT	FORM
ECLAbar8	n^{1}ʃa^{3}m
NACAbar2	ɪ^{1}ʃa^{51}m
NMAAbar1	ɪ^{1}ʃa^{51}m
NVBAbar7	n^{1}ʃa^{3}m
ENBBiya1	i^{1}ʃa^{5}m
FBCBiya8	ɪ^{1}ʃa^{3}m
ICNBiya2	ɪ^{5}ʃa^{5}m
NFKBiya7	a^{5}ʃa^{5}m
NJNBiya6	ɪ^{1}ʃa^{3}m
NSFBiya5	ɪ^{1}ʃa^{15}m
MNJBuu4	ʃɪː^{15}m
NNBBuu3	ʃi^{5}m
DPNFang13	si^{3}m
KHKFang12	si^{3}m
KJSFang2	si^{3}m
JGYKoshin3	ʃə^{5}m
MRYKoshin2	ʃə^{3}m
TELKoshin4	ʃə^{5}m
BNMKung2	i^{1}tə^{5}m
KCSKung3	i^{5}ta^{5}m
NJSKung4	ta^{3}m
BAAMashi4	ʃə^{3}m
BKBMashi2	ʃə^{3}m
KFKMashi1	ʃə^{3}m
NCMMashi5	ʃə^{3}m
ABSMissong1	i^{1}ʃam
AGAMissong2	ɪ^{5}ʃa^{3}m
NDNMissong5	ɪ^{1}ʃa^{5}m
NMSMissong4	ɪ^{3}ʃa^{3}m
APBMumfu1	ʃa^{5}m
DNMMumfu2	ʃa^{3}m
MEAMumfu3	ʃa^{3}m
NCCMumfu4	ʃa^{3}m
CENMundabli2	sa^{5}m
LFNMundabli1	sa^{3}m
NINMundabli4	sa^{3}m
NMNMundabli3	sa^{5}m
NEAMunken1	ɪ^{1}ʃa^{3}m
NGTMunken3	ɪ^{5}ʃa^{5}ma^{51}
NUNMunken4	ɪ^{1}ʃa^{3}m
AOMNgun2	ɪ^{1}ʃaː^{51}m
KBMNgun4	ɪ^{1}ʃa^{3}m
MCANgun3	ɪ^{1}ʃa^{3}m
KCYBuu2	n^{1}tʃʊː^{51}
KEMBuu1	n^{1}tʃʊː^{51}
KDVFang1	n^{1}tsʊ^{1}
TNTMunken2	n^{1}tso^{1}lə^{1}
KDCAjumbu10	ʃi^{1}n
KMNAjumbu2	ʃi^{1}n
NEMAjumbu9	ʃi^{1}n
NVIAjumbu1	ʃjə^{51}
ZKGKung1	te^{5}ɪ^{5}nɛː^{15}zə^{5}
WCANgun1	ɪ^{5}ʃa^{3}mɪ^{1}tsʊ^{5}tsʊ^{15}

Table 3: Detected similarity sets for wordlist entries for ‘rain’ and ‘heart’

In the discussion below, quantitative patterns found across individual wordlists are calculated on the basis of whether or not the forms associated with a given word are treated as belonging to the same similarity set. Thus, for instance, for ‘rain’, all of the varieties in the first block would be treated as using the “same” element for this concept, which would, in turn, mean that the other varieties use a different element. As is clear from the data, there is significant phonetic variation among the forms in the similarity sets. Nevertheless, this measure provides a good indication of lexical distance among varieties in the local space. At the same time, the automated nature of the comparison does result in some cases where the grouping of forms into similarity sets may be different from what a human would produce if conducting an analysis by hand. For instance, the form from one Ajumbu speaker for the word ‘heart’, *ʃjə*^{51} from NVIAjumbu1, shows significant formal overlap with the forms for the other Ajumbu speakers, which all have the shape *ʃi*^{1}*n*. However, as seen, the algorithm separated them into two distinct sets.[16]

The forms in Table 3 also give some indication of the kinds of individual-based variation found in the wordlists that were collected. In some cases, the variation is primarily phonetic or phonological, as can be seen in some of the forms provided for the concept ‘rain’ within a given variety. This is found in the forms *i*^{1}*wo*^{5}*l* (for BNMKung2 and ZKGKung1) and *ɪ*^{1}*ɣo*^{5}*l* (for KCSKung3 and NJSKung4) provided by Kung speakers and the forms *b*^{w}*ə*^{1} (for KDCAjumbu10) and *bwoː*^{1} (for KMNAjumbu2) provided by Ajumbu speakers.

There are also cases where two clearly different roots are used within the same variety. This can be seen, for instance, in forms for the concept ‘heart’. Most of the words provided for ‘heart’ fall into a single similarity set, which is the first one in the table. However, the comparison algorithm placed some of the words into two additional small similarity sets

[16] List, Greenhill & Gray (2017) discuss the process through which forms are grouped into similarity sets using LingPy. This process involves setting a specific threshold for distances among forms that is used as part of the grouping process. Following the recommendation of List, Greenhill & Gray (2017:9), a threshold of 0.45 was used in this study, and experimentation with different thresholds is left for future work.

while also placing three words in their own separate classes—that is, they were deemed too dissimilar from the other words to be put in the same class as any of them. Looking at the overall patterns across the similarity sets, there are several cases where the entries for the individual-based wordlists drawn from the same variety are found in different similarity sets and are clearly formally divergent from each other (unlike the Ajumbu example just discussed above). This is seen, for instance, in the Buu variety, where two speakers produced forms in the first similarity set, namely *ʃi*^{5}*m* in doculect NNBBuu3 and *ʃɪː*^{15}*m* in doculect MNJBuu4, and another two speakers produced forms from the second similarity set, both with shapes *n*^{1}*tʃʊː*^{51}, as seen in doculects KEMBuu1 and KCYBuu2. A similar pattern is found for the Fang and Munken varieties, where the forms for three of the doculects are found in the first similarity set, while the form for the fourth is found in the second.

The forms for ‘heart’ for ZKGKung1 and WCANgun1 appear to be morphologically complex given their length, and, for the WCANgun1 form, *ɪ*^{5}*ʃa*^{3}*mɪ*^{1}*tsʊ*^{5}*tsʊ*^{15}, there is also overlap between its initial sequence of transcribed characters and the forms seen for the other Ngun doculects, such as *ɪ*^{1}*ʃa*^{3}*m* for MCANgun3, in the first similarity set. In a more traditional approach to wordlist collection, a form like the one for WCANgun1 would likely have been filtered out as not representing the most typical way of expressing the meaning ‘heart’ in Ngun. However, as discussed above, because we are specifically interested in individual-based variation for this research, such variants were retained. However, this again raises the issue of what kind of comparison algorithm should be used in work such as this since, while it is clear that the form found in the WCANgun1 doculect is distinctive, the current way of assembling similarity sets treats it as completely distinctive from the forms associated with the other Ngun doculects considered here, even though there is a partial overlap (see Wu & List 2023 for relevant discussion).

In the following sections, we discuss some of the broader patterns that were detected in the wordlists based on the similarity sets detected using LingPy.

3.4 Overall patterns found via wordlist comparison

The overall patterns of similarity and dissimilarity across the wordlists as detected via the overlapping similarity sets among their forms are presented in Table 4, which provides the distances presented numeric form, and Figure 2, which represents the same information via a heatmap where warmer colors (i.e., red) indicate two varieties are lexically closer and cooler colors (i.e., blue) indicate that they are more distant. In the heatmap, there is a dark red diagonal line indicating where each individual-based doculect is compared with itself. Otherwise, the most striking feature of the heatmap are the series orange-to-red squares seen along the diagonal which, for the most part, represent clusters of wordlists from the same variety.

NFKBiya7	1.00																																																				
FBCBiya8	0.85	1.00																																																		
NJNBiya6	0.80	0.85	1.00																																																
NSFBiya5	0.76	0.78	0.80	1.00																																														
ICNBiya2	0.71	0.76	0.78	0.72	1.00																																												
ENBBiya1	0.76	0.82	0.81	0.74	0.83	1.00																																										
NUNMunken4	0.66	0.68	0.66	0.68	0.68	0.69	1.00																																									
NGTMunken3	0.67	0.69	0.66	0.70	0.66	0.69	0.80	1.00																																								
TNTMunken2	0.67	0.70	0.73	0.68	0.66	0.69	0.79	0.75	1.00																																							
NEAMunken1	0.66	0.71	0.70	0.69	0.74	0.72	0.77	0.76	0.85	1.00																																						
WCANgun1	0.74	0.73	0.76	0.76	0.66	0.69	0.69	0.70	0.70	0.69	1.00																																					
KBMNgun4	0.73	0.76	0.78	0.76	0.71	0.74	0.69	0.67	0.72	0.70	0.83	1.00																																				
AOMNgun2	0.72	0.72	0.72	0.74	0.63	0.68	0.73	0.69	0.71	0.69	0.83	0.78	1.00																																			
MCANgun3	0.72	0.71	0.72	0.73	0.66	0.68	0.66	0.67	0.67	0.68	0.78	0.82	0.81	1.00																																		
NVBAbar7	0.64	0.66	0.67	0.67	0.60	0.64	0.68	0.65	0.70	0.68	0.71	0.73	0.73	0.70	1.00																																	
NACAbar2	0.65	0.68	0.68	0.69	0.67	0.66	0.70	0.69	0.71	0.71	0.73	0.76	0.77	0.73	0.90	1.00																																
ECLAbar8	0.66	0.68	0.70	0.70	0.67	0.66	0.69	0.69	0.70	0.75	0.73	0.76	0.73	0.70	0.86	0.88	1.00																															
NMAAbar1	0.62	0.67	0.66	0.67	0.62	0.63	0.70	0.69	0.69	0.68	0.69	0.75	0.70	0.70	0.85	0.86	0.83	1.00																														
NMSMissong4	0.55	0.55	0.56	0.56	0.53	0.54	0.60	0.57	0.56	0.58	0.61	0.59	0.61	0.57	0.66	0.63	0.65	0.57	1.00																													
NDNMissong5	0.55	0.59	0.59	0.57	0.56	0.59	0.59	0.59	0.59	0.61	0.62	0.64	0.60	0.58	0.68	0.68	0.67	0.61	0.88	1.00																												
AGAMissong2	0.49	0.53	0.55	0.52	0.54	0.55	0.57	0.54	0.59	0.58	0.55	0.58	0.58	0.54	0.63	0.63	0.59	0.56	0.84	0.85	1.00																											
ABSMissong1	0.50	0.53	0.55	0.51	0.53	0.53	0.57	0.54	0.54	0.58	0.57	0.56	0.59	0.53	0.65	0.66	0.63	0.60	0.83	0.84	0.83	1.00																										
ZKGKung1	0.19	0.19	0.20	0.19	0.19	0.19	0.18	0.21	0.21	0.21	0.17	0.20	0.18	0.20	0.18	0.20	0.22	0.20	0.16	0.16	0.16	0.15	1.00																									
BNMKung2	0.19	0.20	0.20	0.22	0.21	0.19	0.20	0.21	0.19	0.23	0.17	0.21	0.18	0.21	0.20	0.21	0.23	0.22	0.19	0.19	0.18	0.18	0.86	1.00																								
NJSKung4	0.19	0.21	0.20	0.20	0.19	0.21	0.19	0.21	0.20	0.22	0.18	0.20	0.19	0.21	0.22	0.21	0.23	0.23	0.21	0.21	0.20	0.20	0.71	0.80	1.00																							
KCSKung3	0.21	0.22	0.23	0.22	0.22	0.24	0.21	0.22	0.23	0.25	0.20	0.23	0.19	0.23	0.23	0.21	0.25	0.23	0.23	0.23	0.22	0.21	0.72	0.76	0.90	1.00																						
MEAMumfu3	0.09	0.10	0.10	0.12	0.13	0.11	0.09	0.10	0.10	0.12	0.09	0.09	0.10	0.10	0.09	0.11	0.10	0.11	0.15	0.14	0.15	0.15	0.12	0.14	0.14	0.12	1.00																					
DNMMumfu2	0.10	0.10	0.13	0.13	0.10	0.11	0.10	0.10	0.10	0.11	0.10	0.09	0.11	0.10	0.10	0.10	0.11	0.10	0.18	0.15	0.15	0.17	0.12	0.14	0.15	0.13	0.86	1.00																				
NCCMumfu4	0.10	0.11	0.12	0.12	0.13	0.13	0.11	0.11	0.11	0.11	0.09	0.11	0.11	0.11	0.09	0.12	0.10	0.11	0.14	0.15	0.15	0.14	0.12	0.14	0.15	0.14	0.83	0.85	1.00																			
APBMumfu1	0.11	0.11	0.12	0.12	0.12	0.11	0.09	0.12	0.10	0.11	0.11	0.10	0.10	0.11	0.09	0.12	0.10	0.11	0.15	0.16	0.14	0.14	0.13	0.14	0.15	0.14	0.77	0.81	0.81	1.00																		
NMNMundabli3	0.12	0.12	0.12	0.11	0.14	0.13	0.11	0.13	0.11	0.12	0.11	0.10	0.11	0.12	0.09	0.12	0.10	0.11	0.15	0.15	0.13	0.14	0.12	0.13	0.12	0.12	0.74	0.73	0.74	0.74	1.00																	
CENMundabli2	0.10	0.11	0.12	0.11	0.15	0.13	0.11	0.11	0.11	0.12	0.10	0.11	0.11	0.12	0.10	0.14	0.10	0.13	0.14	0.15	0.15	0.14	0.14	0.15	0.15	0.15	0.76	0.74	0.78	0.76	0.85	1.00																
NINMundabli4	0.12	0.13	0.12	0.14	0.15	0.14	0.12	0.13	0.14	0.13	0.12	0.12	0.12	0.13	0.12	0.15	0.12	0.15	0.15	0.16	0.16	0.15	0.14	0.15	0.15	0.15	0.72	0.70	0.70	0.73	0.82	0.82	1.00															
LFNMundabli1	0.11	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.12	0.09	0.10	0.10	0.10	0.10	0.13	0.10	0.12	0.14	0.14	0.15	0.14	0.13	0.14	0.15	0.15	0.76	0.76	0.72	0.77	0.80	0.82	0.82	1.00														
NNBBuu3	0.28	0.29	0.30	0.27	0.30	0.31	0.30	0.30	0.32	0.35	0.28	0.28	0.30	0.30	0.32	0.34	0.34	0.35	0.30	0.32	0.30	0.32	0.25	0.26	0.24	0.24	0.36	0.37	0.40	0.35	0.40	0.41	0.37	0.36	1.00													
MNJBuu4	0.26	0.29	0.27	0.25	0.31	0.31	0.32	0.30	0.30	0.35	0.26	0.26	0.29	0.28	0.29	0.31	0.30	0.32	0.31	0.32	0.30	0.33	0.25	0.26	0.25	0.23	0.37	0.38	0.41	0.36	0.42	0.40	0.37	0.37	0.89	1.00												
KEMBuu1	0.24	0.29	0.29	0.27	0.30	0.31	0.31	0.28	0.35	0.36	0.31	0.30	0.30	0.28	0.31	0.32	0.32	0.33	0.31	0.31	0.35	0.33	0.23	0.23	0.24	0.23	0.39	0.40	0.41	0.38	0.40	0.42	0.41	0.41	0.84	0.85	1.00											
KCYBuu2	0.29	0.32	0.32	0.29	0.34	0.32	0.30	0.29	0.34	0.39	0.34	0.32	0.31	0.32	0.32	0.33	0.35	0.34	0.31	0.32	0.31	0.33	0.22	0.23	0.23	0.24	0.35	0.36	0.35	0.37	0.39	0.39	0.39	0.37	0.84	0.82	0.89	1.00										
TELKoshin4	0.16	0.16	0.17	0.17	0.21	0.17	0.19	0.19	0.20	0.23	0.17	0.16	0.19	0.17	0.22	0.23	0.20	0.22	0.19	0.19	0.17	0.19	0.24	0.26	0.23	0.23	0.35	0.33	0.32	0.33	0.34	0.36	0.33	0.37	0.49	0.49	0.48	0.44	1.00									
JGYKoshin3	0.17	0.15	0.18	0.15	0.22	0.16	0.17	0.18	0.20	0.21	0.19	0.18	0.19	0.18	0.20	0.21	0.22	0.21	0.20	0.19	0.17	0.19	0.25	0.27	0.26	0.25	0.33	0.33	0.30	0.33	0.32	0.35	0.31	0.33	0.46	0.45	0.44	0.42	0.86	1.00								
MRYKoshin2	0.15	0.15	0.15	0.14	0.20	0.16	0.16	0.17	0.19	0.20	0.16	0.16	0.15	0.13	0.20	0.18	0.20	0.20	0.18	0.18	0.17	0.17	0.26	0.28	0.26	0.25	0.30	0.29	0.28	0.27	0.29	0.32	0.29	0.31	0.46	0.47	0.47	0.45	0.73	0.80	1.00							
KHKFang12	0.13	0.13	0.13	0.12	0.16	0.15	0.14	0.13	0.14	0.17	0.14	0.15	0.12	0.14	0.14	0.16	0.16	0.17	0.13	0.14	0.12	0.12	0.20	0.20	0.18	0.18	0.29	0.30	0.28	0.33	0.33	0.31	0.33	0.32	0.38	0.36	0.40	0.40	0.38	0.40	0.45	1.00						
DPNFang13	0.15	0.14	0.14	0.13	0.16	0.16	0.14	0.14	0.14	0.18	0.15	0.16	0.13	0.15	0.14	0.16	0.17	0.17	0.14	0.15	0.12	0.13	0.23	0.23	0.20	0.21	0.29	0.31	0.29	0.34	0.33	0.31	0.31	0.32	0.38	0.39	0.40	0.39	0.39	0.43	0.46	0.90	1.00					
KJSFang2	0.15	0.16	0.17	0.16	0.19	0.19	0.18	0.15	0.18	0.20	0.17	0.19	0.15	0.16	0.17	0.19	0.18	0.18	0.15	0.17	0.16	0.15	0.21	0.20	0.20	0.21	0.33	0.33	0.35	0.34	0.33	0.37	0.32	0.34	0.38	0.39	0.42	0.41	0.43	0.45	0.47	0.85	0.88	1.00				
KDVFang1	0.14	0.15	0.15	0.14	0.15	0.17	0.16	0.15	0.16	0.18	0.16	0.15	0.12	0.13	0.15	0.15	0.15	0.15	0.15	0.17	0.13	0.13	0.22	0.21	0.20	0.21	0.33	0.33	0.31	0.33	0.33	0.32	0.30	0.32	0.40	0.42	0.41	0.41	0.42	0.44	0.46	0.87	0.89	0.91	1.00			
NVIAjumbu1	0.24	0.24	0.23	0.22	0.26	0.29	0.22	0.22	0.24	0.26	0.21	0.23	0.21	0.23	0.25	0.27	0.24	0.25	0.18	0.19	0.20	0.18	0.23	0.24	0.28	0.27	0.20	0.20	0.19	0.23	0.20	0.20	0.22	0.23	0.34	0.34	0.34	0.37	0.25	0.22	0.24	0.30	0.34	0.36	0.32	1.00		
KMNAjumbu2	0.24	0.22	0.25	0.21	0.30	0.28	0.21	0.21	0.25	0.26	0.21	0.23	0.22	0.21	0.27	0.29	0.25	0.26	0.18	0.20	0.20	0.21	0.22	0.23	0.27	0.28	0.20	0.19	0.19	0.22	0.19	0.19	0.21	0.21	0.38	0.37	0.38	0.38	0.30	0.29	0.30	0.31	0.34	0.34	0.33	0.83	1.00						
NEMAjumbu9	0.21	0.19	0.21	0.18	0.23	0.23	0.18	0.17	0.20	0.22	0.18	0.17	0.19	0.18	0.21	0.22	0.19	0.20	0.15	0.15	0.19	0.17	0.23	0.24	0.26	0.27	0.24	0.22	0.22	0.23	0.22	0.23	0.20	0.26	0.39	0.40	0.39	0.39	0.33	0.32	0.32	0.32	0.36	0.37	0.35	0.76	0.83	1.00					
KDCAjumbu10	0.23	0.20	0.22	0.24	0.26	0.23	0.20	0.18	0.21	0.24	0.21	0.21	0.22	0.21	0.24	0.23	0.23	0.21	0.20	0.18	0.19	0.20	0.23	0.25	0.27	0.28	0.20	0.19	0.20	0.21	0.20	0.21	0.19	0.22	0.36	0.37	0.36	0.37	0.28	0.29	0.29	0.29	0.34	0.33	0.31	0.75	0.81	0.87	1.00				
NCMMashi5	0.19	0.19	0.18	0.18	0.18	0.20	0.18	0.18	0.17	0.19	0.17	0.17	0.18	0.17	0.18	0.21	0.20	0.19	0.20	0.21	0.19	0.20	0.21	0.21	0.21	0.22	0.22	0.23	0.22	0.23	0.21	0.21	0.23	0.22	0.30	0.30	0.33	0.32	0.32	0.33	0.34	0.21	0.23	0.23	0.26	0.20	0.22	0.21	0.21	1.00			
BAAMashi4	0.18	0.18	0.18	0.16	0.17	0.18	0.15	0.16	0.15	0.17	0.17	0.17	0.16	0.15	0.18	0.19	0.19	0.19	0.20	0.20	0.18	0.20	0.19	0.20	0.21	0.21	0.23	0.24	0.23	0.24	0.24	0.23	0.24	0.23	0.30	0.30	0.36	0.34	0.31	0.32	0.34	0.21	0.22	0.20	0.22	0.19	0.21	0.21	0.21	0.90	1.00		
KFKMashi1	0.17	0.18	0.19	0.16	0.19	0.19	0.18	0.17	0.17	0.20	0.16	0.16	0.16	0.16	0.18	0.21	0.18	0.19	0.19	0.21	0.19	0.21	0.18	0.19	0.18	0.19	0.24	0.23	0.24	0.24	0.24	0.23	0.23	0.23	0.31	0.31	0.34	0.34	0.32	0.31	0.31	0.21	0.22	0.22	0.23	0.19	0.22	0.22	0.21	0.88	0.88	1.00	
BKBMashi2	0.19	0.18	0.18	0.18	0.17	0.19	0.17	0.15	0.16	0.18	0.16	0.16	0.17	0.16	0.17	0.19	0.18	0.17	0.19	0.20	0.18	0.20	0.18	0.19	0.18	0.19	0.23	0.22	0.22	0.23	0.21	0.21	0.22	0.21	0.29	0.29	0.31	0.31	0.31	0.31	0.30	0.21	0.23	0.22	0.24	0.19	0.21	0.21	0.21	0.88	0.85	0.93	1.00

Table 4: Pairwise similarities among varieties based on shared similarity sets

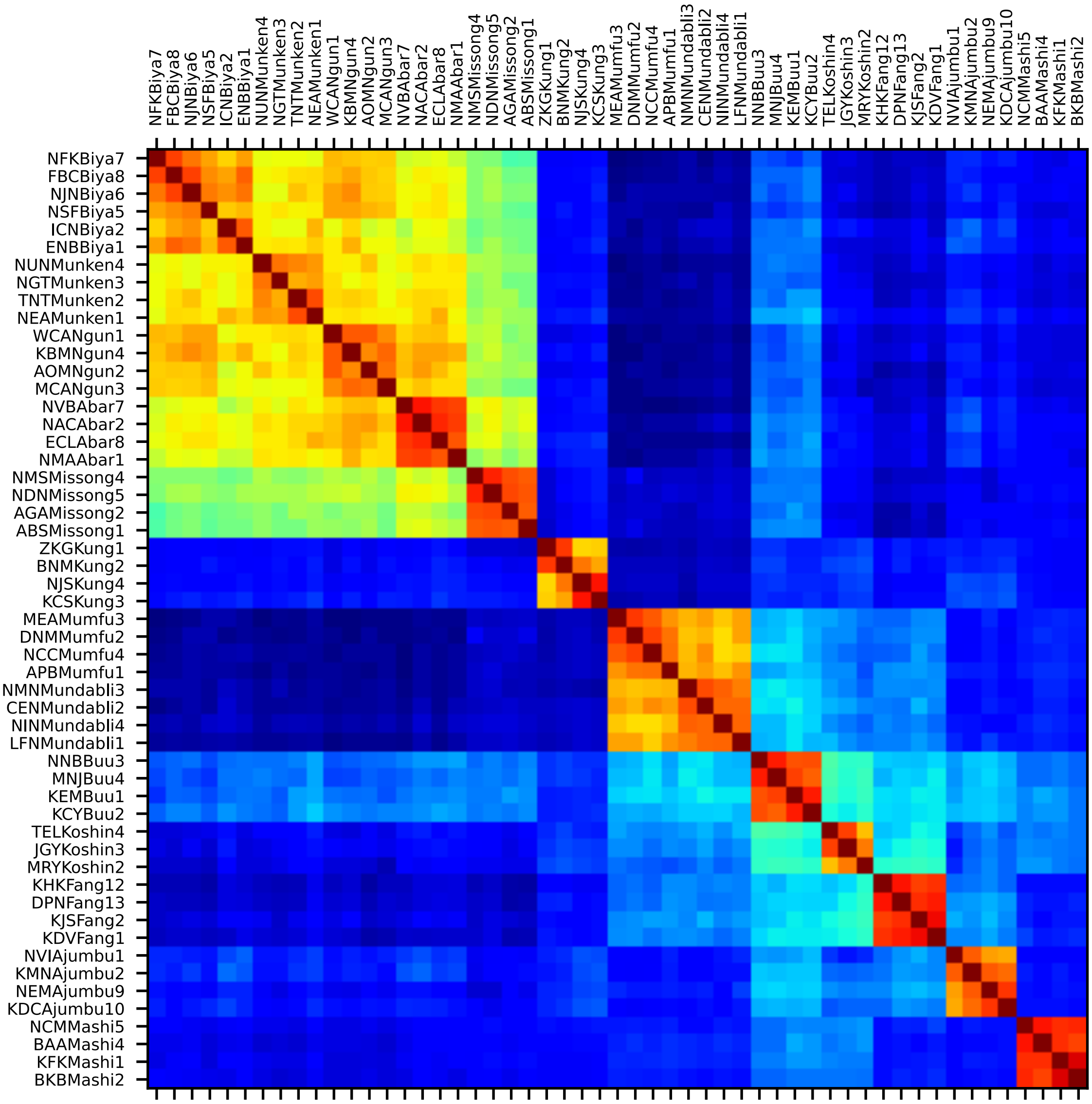

Figure 2: Heatmap representation of wordlist similarities using SCA comparison method

The fact that these clusters are present simply shows that wordlists that are supposed to be from the same variety are, in fact, quite similar to each other. Nevertheless, some interesting patterns emerge with respect to the variation seen within these sets of wordlists. The quantitative aspects of this variation will be considered in detail §3.5. However, a few qualitative remarks can be made at this point. (i) The Mungbam language cluster is clearly visible in the heatmap in the form of a large square in the upper-right quadrant of the heatmap, and, within this cluster, the divergence between the Missong variety, whose wordlists are in the bottom left corner of the Mungbam block, and the other varieties is also quite visible. This largely replicates what was known from available descriptions (see, e.g., Lovegren 2013) and is useful as a way of verifying that the approach used here is producing interpretable results. (ii) In the lower-right corner of the heatmap, there is an area that is relatively light in color comprising the Buu, Fang, and Koshin languages, though they are not especially close in a way that suggests a genealogical grouping, and, Buu, otherwise shows some evidence of being genealogically linked to Mufu and Mundabli (see Voll 2017: 5–7). This suggests the possibility of a southeastern contact area in Lower Fungom which had not been previously noted. Finally, (iii) Buu overall shows significant similarities to many Lower Fungom varieties (with the exception of Mashi and Kung, both recent entrants to the area), as indicated by the lighter blue bands associated with it, suggesting that it has been especially strongly impacted by contact in the region.

Not all of these results are specifically relevant to individual-based approaches to data collection. At the same time, even with respect to broad patterns, such as the apparent impacts of language contact on Buu vocabulary, this approach makes visible some cases of variation at the individual level of potential interest that suggest a need for further investigation, such as variation in the precise overlap of the vocabulary of each of the Buu wordlists with those of the other varieties which Buu has an apparent contact relationship with. Further investigation might reveal that these differences can be explained, at least in part, by differences in the multilingual repertoires of the individuals from whom the Buu wordlists were collected.

In §3.5 and §3.6, two aspects of the results of this work will be considered in more detail, patterns of similarity and difference within varieties and which concepts were most homogeneous and heterogeneous in terms of the number of similarity sets associated with them.

3.5 Similarities and differences within varieties

In Figure 3, the same information presented in §3.4 is visualized using a split graph representation produced using SplitsTree4 (Huson & Bryant 2006). Nodes are colored according to the variety that the individual-based wordlist is associated with (see Table 1) This kind of representation is often used in historical studies as a way of presenting both possible genealogical relationships and potential contact relationships, where reticulations in the network can sometimes be indications of lexical borrowing patterns (Heggarty, Maguire & McMahon 2010). Such relationships do appear to be visible in Figure 3. Wordlists associated with distinct varieties largely form distinct branches, except for those that have been previously described as part of dialect groups, such as the five Mungbam varieties seen at the right edge of the network or Mundabli and Mufu seen at its left edge. However, for present purposes, this diagram is presented for its value in interpreting patterns of synchronic variation. In particular, it is easier to see the different patterns of variation within varieties with this representation in comparison to the heatmap in Figure 2. A noteworthy pattern that is visible in the graph is the relative lack of clear differentiation between the Biya and Ngun varieties of Mungbam, discussed further below.

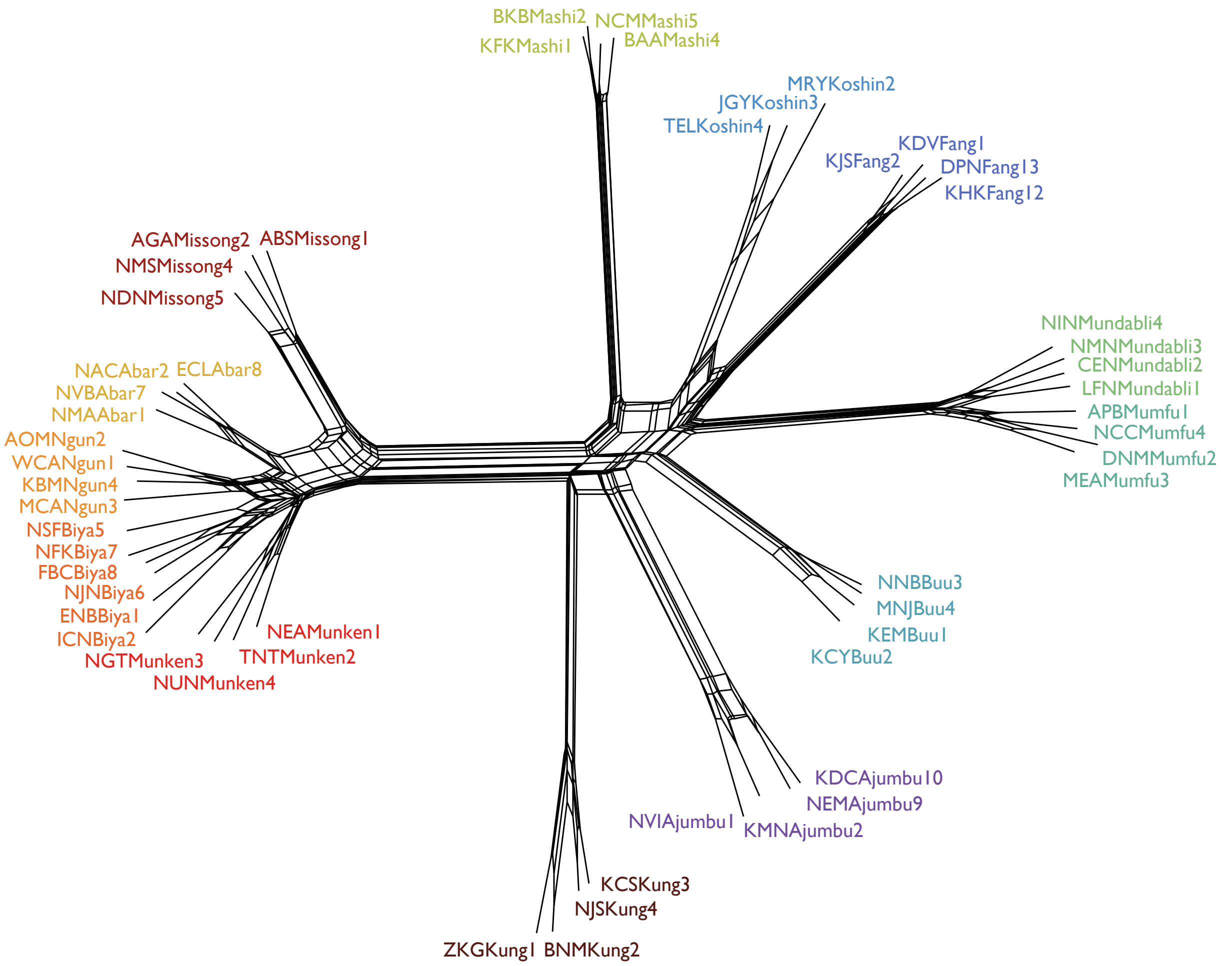

Figure 3: Network-based representation of wordlist similarities

The overall patterns of similarity within varieties are further detailed Table 5 which presents a subset of the data seen in Table 4, subdivided across the thirteen varieties. The data across varieties is not fully comparable both because there is variation in how many wordlists are available across them and because the sampling of individuals was opportunistic rather than based on some predetermined principle. Bearing this in mind, the overall patterns seen within varieties can be described as follows: The variety showing the greatest similarity across its wordlists is Mashi, where the two wordlists show a similarity of 0.89. The variety showing the greatest dissimilarity across its wordlists is Biya, where the mean similarity is 0.78. However, this low score may be at least partly an artifact of the fact that there are six Biya wordlists, the most of any variety. Of the varieties with four wordlists, Munken and Kung show the greatest dissimilarity, with a mean of 0.79. The greatest similarity between any two wordlists of the same variety is 0.93 for a pair of Mashi wordlists (KFKMashi1 and BKBMashi2), and the least similarity between any two wordlists of the same variety is 0.71, which is the case for a pair of Kung wordlists (ZKGKung1 and NJSKung4). This is actually lower than the similarity value found between two wordlists from different varieties, namely the Biya doculect NSFBiya5 and the Ngun doculect MCANgun3, which had a similarity score of 0.73. At the same time, it should be noted that, within the Lower Fungom context, Kung is very linguistically distinctive, as indicated by the clustering of the four Kung varieties on a long branch in Figure 3. Therefore, even with this level of variation, there is no indication that any of these varieties would not clearly be perceived as Kung locally, unlike Biya–Ngun where the boundary between these varieties is much less clear.

LANGUAGE	DOCULECT	PAIRWISE SIMILARITIES						MEAN	RANGE
Mungbam	NVBAbar7	1.00	0.90	0.86	0.85				
	NACAbar2	0.90	1.00	0.88	0.86				
	ECLAbar8	0.86	0.88	1.00	0.83				
	NMAAbar1	0.85	0.86	0.83	1.00			0.86	0.17
	NFKBiya7	1.00	0.85	0.80	0.76	0.71	0.76		
	FBCBiya8	0.85	1.00	0.85	0.78	0.76	0.82		
	NJNBiya6	0.80	0.85	1.00	0.80	0.78	0.81		
	NSFBiya5	0.76	0.78	0.80	1.00	0.72	0.74		
	ICNBiya2	0.71	0.76	0.78	0.72	1.00	0.83		
	ENBBiya1	0.76	0.82	0.81	0.74	0.83	1.00	0.78	0.29
	NUNMunken4	1.00	0.80	0.79	0.77				
	NGTMunken3	0.80	1.00	0.75	0.76				
	TNTMunken2	0.79	0.75	1.00	0.85				
	NEAMunken1	0.77	0.76	0.85	1.00			0.79	0.25
	WCANgun1	1.00	0.83	0.83	0.78				
	KBMNgun4	0.83	1.00	0.78	0.82				
	AOMNgun2	0.83	0.78	1.00	0.81				
	MCANgun3	0.78	0.82	0.81	1.00			0.81	0.22
	NMSMissong4	1.00	0.88	0.84	0.83				
	NDNMissong5	0.88	1.00	0.85	0.84				
	AGAMissong2	0.84	0.85	1.00	0.83				
	ABSMissong1	0.83	0.84	0.83	1.00			0.84	0.17
Ji group	MEAMumfu3	1.00	0.86	0.83	0.77				
	DNMMumfu2	0.86	1.00	0.85	0.81				
	NCCMumfu4	0.83	0.85	1.00	0.81				
	APBMumfu1	0.77	0.81	0.81	1.00			0.82	0.23
	NMNMundabli3	1.00	0.85	0.82	0.80				
	CENMundabli2	0.85	1.00	0.82	0.82				
	NINMundabli4	0.82	0.82	1.00	0.82				
	LFNMundabli1	0.80	0.82	0.82	1.00			0.82	0.20
	NNBBuu3	1.00	0.89	0.84	0.84				
	MNJBuu4	0.89	1.00	0.85	0.82				
	KEMBuu1	0.84	0.85	1.00	0.89				
	KCYBuu2	0.84	0.82	0.89	1.00			0.86	0.18
Fang	KHKFang12	1.00	0.91	0.85	0.86				
	DPNFang13	0.91	1.00	0.86	0.87				
	KJSFang2	0.85	0.86	1.00	0.92				
	KDVFang1	0.86	0.87	0.92	1.00			0.88	0.15
Koshin	TELKoshin4	1.00	0.86	0.73					
	JGYKoshin3	0.86	1.00	0.80					
	MRYKoshin2	0.73	0.80	1.00				0.80	0.27
Ajumbu	NVIAjumbu1	1.00	0.83	0.76	0.75				
	KMNAjumbu2	0.83	1.00	0.83	0.81				
	NEMAjumbu9	0.76	0.83	1.00	0.87				
	KDCAjumbu10	0.75	0.81	0.87	1.00			0.81	0.25
Naki	NCMMashi5	1.00	0.90	0.88	0.88				
	BAAMashi4	0.90	1.00	0.88	0.85				
	KFKMashi1	0.88	0.88	1.00	0.93				
	BKBMashi2	0.88	0.85	0.93	1.00			0.89	0.15
Kung	ZKGKung1	1.00	0.86	0.71	0.72				
	BNMKung2	0.86	1.00	0.80	0.76				
	NJSKung4	0.71	0.80	1.00	0.90				
	KCSKung3	0.72	0.76	0.90	1.00			0.79	0.29

Table 5: Patterns of similarity and difference within varieties across the individual-based wordlists

Further work in other parts of the world is needed to establish whether the kind of variation seen in the data above can be considered “normal” or not. To the best of our knowledge, no comparable dataset for any part of the world (even for European varieties) is available for consideration alongside the Lower Fungom data that we have collected.

3.6 Most and least homogenous concepts

An additional way that this data can be used to detect potentially interesting comparative patterns is through examination of the distribution of similarity sets across concepts. This can reveal which concepts show evidence of being more homogeneous in their expression in Lower Fungom and which show evidence of being more heterogeneous. From a synchronic perspective, this information can indicate which concepts are most likely to be associated with distinctive forms associated with relatively few varieties and, therefore, would be of more value for identifying the variety that an individual is using, potentially to the point of being an emblematic distinction.[17] From a diachronic perspective, this information can provide a potential indication of the stability of the expression of different concepts which could be valuable for the reconstruction of Bantoid prehistory by revealing roots which may be more indicative of older historical relationships.

In order to determine which concepts were associated with more homogenous sets of expression, a metric of expressional homogeneity was developed based on the notion of normalized entropy, as understood in the context of information theory. This is due to the known link between entropy and informativeness, which seemed appropriate for a study with an interest in examining the distinctiveness of a word in the local linguistic space. While information theory is being increasingly used in linguistic studies of various kinds (see, e.g., Mansfield 2021), its use here as a means to determine the potential emblematicity of a word in a multilingual context is, to the best of our knowledge, novel. The homogeneity scores that

[17] In looking at the data in this way, we see ourselves as building on Watson’s (2019) application of prototype theory to examine patterns of linguistic distinctiveness for varieties of the Casamance region of Senegal.

were calculated for the concepts used in this study are presented in Table 6.[18] A higher homogeneity score indicates that there is less variation in the allocation of words in similarity sets across the wordlists, and a lower score indicates that there is more variation. This analysis is conducted across all of the wordlists as a group, rather than within varieties, due to the relatively small number of wordlists available within each variety, but the basic approach could be extended to look at variation within a variety as well.

[18] The normalized entropy calculations that form the basis of the scores seen in Table 6 were first determined by calculating the distribution of similarity sets across a given concept and using this as the probability that a word from a given similarity set would be used to express that concept by an individual speaker. These probabilities were then used to determine the entropy associated with the words expressing that concept, using the formula $-\sum_{i=1}^{n} P(x_i) \cdot \ln\left(P(x_i)\right)$ where $P(x)$ represents the probability of a word from each similarity set being associated with a concept. The entropy was then normalized by dividing it by the natural logarithm of the number of different similarity sets associated with the concept. (The choice of the base for the logarithm in these calculations was arbitrary and, due to the fact that entropy was normalized, does not impact the scores seen in Table 6.) Roughly speaking, a high entropy score indicates that a given concept has forms distributed more evenly across a greater number of similarity sets and a low entropy score indicates that it has a less even distribution with more forms distributed into a smaller number of similarity sets. For purposes of presentation, the entropy scores were converted into what is being referred to here as a homogeneity score by subtracting them from 1.

grave	0.93
tongue	0.93
child	0.93
cow, cattle	0.90
axe	0.89
mother	0.88
ear	0.86
horse	0.83
bird	0.81
heart	0.81
song	0.80
father	0.79
fence	0.79
rope	0.78
tooth	0.78
ladder	0.77
war	0.77
chief	0.77
breast	0.76
soap	0.76
sieve	0.76
medicine	0.75
smoke	0.75
bag	0.75
stone	0.75
headpad	0.75
deity	0.75
tree	0.75
sheep	0.74
gong	0.74
hand	0.73
head	0.73
cat	0.72
nose	0.72
sand	0.71
hair	0.71
fowl	0.71
book	0.71
forest	0.70
basket	0.70
camwood	0.70
cricket	0.70
rain	0.69
road	0.69
goat	0.69
dance	0.69
umbrella	0.68
crab	0.68
devil	0.68
hailstone	0.68
mouth	0.67
raffia bamboo	0.67
horn (head)	0.67
friend	0.66
jaw	0.66
potato	0.66
fire	0.66
bitter leaf	0.66
caterpillar	0.66
xylophone	0.66
faeces	0.66
work (n)	0.66
eye	0.66
place	0.65
farm	0.65
dog	0.65
bridge	0.64
seed	0.64
house	0.64
toilet	0.64
dust	0.63
pineapple	0.63
corn	0.63
air	0.63
grass	0.63
monkey	0.62
person	0.62
stomach	0.62
animal	0.62
plantain	0.62
leaf	0.62
yam	0.61
fireside	0.61
belly	0.61
bat	0.61
case (court)	0.61
moon	0.60
snake	0.59
name	0.59
owl	0.58
broom	0.58
cutlass	0.58
root	0.58
zinc	0.58
dry season	0.58
gun	0.58
intestine	0.57
pig	0.57
chair	0.57
bed	0.56
gizzard	0.56
hill	0.56
egg	0.56
cloth	0.56
fish	0.56
hoe	0.56
sun	0.56
knife	0.56
water	0.56
cup	0.55
palm nut	0.55
oil	0.55
cap	0.55
grasshopper	0.55
palm tree	0.54
spider	0.53
feather	0.53
sky	0.52
blood	0.52
salt	0.52
banana	0.52
garden egg	0.52
day	0.52
soldier ant	0.52
drum	0.51
wingless termite	0.51
mushroom	0.50
storm (wind)	0.49
pap	0.49
story	0.49
pepper	0.46
elephant stalk	0.46
pot	0.45
trap	0.44
compound	0.44
rainbow	0.43
star	0.42
termite	0.34

Table 6: Concept homogeneity calculated as normalized entropy of similarity set distributions

To better understand the nature of the homogeneity scores, we can first compare the scores for ‘rain’ and ‘heart’ given that their similarity sets were presented above in Table 3. The concept ‘rain’ has a lower homogeneity score (0.69) than the concept ‘heart’ (0.81). The reason for this difference is due to the distribution of forms within the similarity sets for these concepts. While ‘heart’ is associated with six similarity sets, three of these have only one form in each, one has only three forms, and another has only four forms. The remaining forms collected for the concept are all found in a single, large similarity set. This means that there is a relatively low chance that the word for ‘heart’, as produced by a given speaker, will be strongly informative of the variety they are providing words from. By contrast, ‘rain’ is associated with five similarity sets, two of which are relatively large. This more even distribution is associated with a given word for the concept having greater informativity for the variety that it is associated with and, hence, a lower homogeneity score.[19]

Two more examples of similarity sets are provided in Table 7 for ‘tongue’, one of the three concepts with the highest homogeneity score in the dataset, and ‘termite’, the one with the lowest score. Fewer total forms were collected for ‘termite’ than ‘tongue’, which is why the list is shorter (see Table 2). The division of ‘tongue’ into two sound-based similarity sets is relatively straightforward. The forms in Kung, which all begin with a *kə* prefix followed by a root beginning with an *n*, are separated from all the other forms, where the root begins with an *l*, and some have a vocalic prefix. The forms for ‘termite’ are much more varied, and the similarity sets for ‘termite’ in Table 7 are ordered so that sets whose forms show some overlap are presented near each other.

[19] Because these scores are based on similarity sets, this means that words are grouped together based on their overall phonological similarity even though they may differ in salient ways. Even if two languages make use of a form found in the same similarity set, they may still differ in ways which makes them identifiable as belonging to a specific variety. We leave open the possibility of adjusting this metric to account for differences of this kind within similarity sets.

Forms collected for ‘tongue’

DOCULECT	FORM
BNMKung2	*kə^{1}nə^{5}m*
KCSKung3	*kə^{1}na^{5}m*
NJSKung4	*kə^{1}nə^{5}m*
ZKGKung1	*kə^{5}nə^{5}m*
ECLAbar8	*ɪ^{1}la^{5}m*
NACAbar2	*ɪ^{1}la^{5}m*
NMAAbar1	*ɪ^{5}la^{5}m*
NVBAbar7	*ɪ^{1}la^{5}m*
KDCAjumbu10	*la^{5}mə^{5}*
KMNAjumbu2	*la^{5}mə*
NEMAjumbu9	*la^{5}mə^{5}*
NVIAjumbu1	*la^{5}m*
ENBBiya1	*ɪ^{5}la^{5}m*
FBCBiya8	*ɪ^{1}la^{5}m*
ICNBiya2	*ɪ^{1}la^{5}m*
NFKBiya7	*ɪ^{1}la^{5}m*
NJNBiya6	*ɪ^{1}la^{5}m*
NSFBiya5	*ɪ^{1}la^{5}m*
KCYBuu2	*lɪ^{5}m*
KEMBuu1	*li^{5}m*
MNJBuu4	*lɪ^{5}m*
NNBBuu3	*lɪ^{5}m*
DPNFang13	*lɪː^{53}m*
KHKFang12	*lɪː^{53}m*
KDVFang1	*lɪ^{53}m*
KJSFang2	*liː^{13}m*
JGYKoshin3	*lə^{5}m*
MRYKoshin2	*lə^{5}m*
TELKoshin4	*lə^{5}m*
BAAMashi4	*li^{5}*
BKBMashi2	*lɪ^{5}*
KFKMashi1	*lɪ^{5}*
NCMMashi5	*li^{3}*
ABSMissong1	*ɪ^{1}la^{5}m*
AGAMissong2	*ɪ^{1}la^{5}*
NDNMissong5	*ɪ^{1}la^{5}m*
NMSMissong4	*ɪ^{1}la^{5}m*
APBMumfu1	*lje^{5}m*
DNMMumfu2	*ljə^{5}n*
MEAMumfu3	*ljɛ^{5}m*
NCCMumfu4	*lje^{5}m*
CENMundabli2	*ljə^{5}m*
LFNMundabli1	*ljə^{5}m*
NINMundabli4	*ljə^{5}m*
NMNMundabli3	*ljə^{5}m*
NEAMunken1	*ɪ^{5}la^{5}m*
NGTMunken3	*ɪ^{5}la^{5}m*
NUNMunken4	*la^{5}m*
TNTMunken2	*ɪ^{1}la^{5}m*
AOMNgun2	*i^{3}la^{5}m*
KBMNgun4	*ɪ^{1}la^{5}m*
MCANgun3	*ɪ^{1}la^{5}m*
WCANgun1	*ɪ^{5}laː^{5}m*

Forms collected for ‘termite’

DOCULECT	FORM
NVBAbar7	*kə^{1}ndʒi^{5}ndʒə^{5}ŋ*
KDCAjumbu10	*kə^{5}ndʒi^{51}n*
KMNAjumbu2	*kə^{1}ndʒi^{1}ɲ*
NEMAjumbu9	*kə^{1}ndʒi^{1}ɲ*
NVIAjumbu1	*kə^{1}ndʒi^{1}n*
ABSMissong1	*ki^{1}ndɛ^{15}ɛ^{3}*
AGAMissong2	*kɪ^{1}ndɛː^{53}*
NMSMissong4	*ki^{5}ndɛ^{15}ɛ^{3}*
FBCBiya8	*kə^{1}dzə^{5}dzoː^{15}*
NFKBiya7	*kə^{1}dzʊ^{5}dzoː^{15}*
NJNBiya6	*kə^{1}dzʊ^{15}dzoː^{15}*
NEAMunken1	*a^{1}zə^{1}zo^{5}lə^{5}*
TNTMunken2	*a^{1}dzə^{1}zɔː^{5}lə^{5}*
KBMNgun4	*kə^{1}zʊ^{1}zo^{5}*
MCANgun3	*kə^{1}kʊ^{1}k^{w}ɛ^{5}*
WCANgun1	*fje^{1}a^{5}zə^{5}zɔ^{5}*
DPNFang13	*dzəː^{15}ɣ*
KHKFang12	*dzɘ^{15}ɣ*
APBMumfu1	*dzɔː^{3}*
CENMundabli2	*ʃə^{5}ŋŋgə^{5}*
LFNMundabli1	*ø^{3}ʃə^{5}ŋgə^{5}lə^{5}*
NINMundabli4	*ʃə^{5}ŋə^{5}lə^{3}*
NMNMundabli3	*ʃə^{3}ŋgə^{3}lə^{3}*
TELKoshin4	*gɣə^{5}*
KCSKung3	*ɪ^{1}kə^{5}j*
NJSKung4	*ɪ^{1}kə^{5}ɪ^{5}*
KEMBuu1	*ʃjə^{5}kə^{5}*
NSFBiya5	*fi^{3}mkpɛ^{5}*
MNJBuu4	*fə^{1}ŋgɔː^{51}*
NNBBuu3	*fə^{1}ŋgɔː^{51}*
JGYKoshin3	*ŋgbʊ^{1}*
BAAMashi4	*ni^{3}*
KFKMashi1	*ni^{5}*
NCMMashi5	*ɲiː^{15}*
BKBMashi2	*ɲ^{5}ɲɪ^{5}*
DNMMumfu2	*zɔ^{3}l*
MEAMumfu3	*zɔ^{5}l*
NCCMumfu4	*zɔ^{5}l*
NGTMunken3	*ɪ^{1}zo^{5}*
NUNMunken4	*a^{5}zo^{5}*
NDNMissong5	*fi^{1}nʃɔ^{5}ha^{1}fɪ^{51}*
ECLAbar8	*mb^{w}o^{1}ɪ^{1}za^{3}m*

Table 7: Detected similarity sets for wordlist entries for ‘tongue’ and ‘termite’

The logic behind the precise grouping of the similarity sets for ‘termite’ is not as clear as it is for the other forms discussed here, and they illustrate some of the limitations of the approach that has been adopted. In particular, given the nature of the variation in the forms for ‘termite’, the grouping algorithm used by LingPy led to results that would probably be different from what would have been produced by human inspection. For example, the initial syllabic nasal in the form from BKBMashi2, *ɲ^{5}ɲɪ^{5}*, resulted in it being grouped separately from the other Mashi forms, even though it is clearly quite similar to them. The placement of the form from JGYKoshin3, *ŋgbʊ^{1}*, with the other Mashi forms also seems unusual, and its more natural grouping would probably be with the forms just above it in the table. (See §3.3 for additional discussion of this issue.)

As discussed above, we believe that homogeneity scores of this kind can be used to determine which lexical items may be more emblematic of specific varieties in the local linguistic space as well as which concepts are associated with words which, for whatever reason, may be more subject to processes of lexical replacement than others. As such, this presents a new technique for exploring the structure of individual-based lexical variation, though, at this stage, we have yet to consider the implications of the results presented in Table 6 beyond noting that there is a significant range of differences in the scores which we believe makes this a promising avenue for further investigation.

Moreover, we should be cautious about making detailed sociolinguistic inferences on the basis of the scores in Table 6 due to some of the limitations inherent to this study, for example the fact that noun class affix variation was not explored independently from variation in stems and the fact that it is based on elicited data rather than usage-based data. From a local sociolinguistic standpoint, an important next step will be to work with speakers to see whether they have any metalinguistic awareness of the homogeneity (or lack of homogeneity) for certain concepts and whether any expressions might be treated as a kind of shibboleth. It would also likely be valuable to develop ways to further examine these patterns of variation with respect to particular languages and varieties to see which concepts may be especially strong markers of a given variety and which are associated with the greatest individual-level (as opposed to variety-level) variation.

3.7 Advancing the approach

On the whole, we believe that the dataset and methods used here have yielded promising results. At the same time, future applications of this approach should consider potential adjustments. With respect to data collection, the main questions are: (i) What kind of concept list should be used for a study of this kind bearing in mind, in particular, that it would be valuable for the list to contain concepts with a range of diachronic stabilities since this will allow for investigation of which concepts may be especially prone to change in order to signal different sociolinguistic identities? And, (ii) What constitutes a representative sample of individuals for a study of this kind given that even two otherwise similar individuals from the same village may have quite distinct multilingual repertoires?

With respect to methods, while the basic techniques developed for finding cognates seem appropriate for finding similarity sets, it is likely the case that the parameters of the relevant algorithms should be adjusted to produce the best results for synchronic analysis. Work using automated cognate detection algorithms for historical purposes can be assessed by comparing their output to results obtained through traditional methods. However, for work of the kind described in this paper there is no accepted “gold standard” to serve as the basis for assessment, which means that new methods of assessment will need to be devised. Indeed, when it comes to the study of individual-level lexical variation, we do not even have a comparison set of comparable data to look at from outside of Lower Fungom. So, perhaps a simple step towards being able to assess the results of work of this kind would be to collect individual-based wordlists from better studied languages, such as English, Spanish, or French, to serve as at least an initial comparison for the Lower Fungom data.

4 Implications for comparative Bantoid linguistics

The bulk of the discussion of this chapter has focused on an analysis of synchronic lexical variation in the Lower Fungom region of Cameroon. However, some of the key motivations behind this work are diachronic in nature. In particular, we believe that Lower Fungom can serve as model for the prehistoric sociolinguistic situation of early Bantoid and, thus, provide an improved foundation for proposals regarding diversification and spread of Bantoid languages (and, by extension, Bantu languages as well). Two results of this study are particularly noteworthy from a methodological perspective, especially in light of the fact of the central role that data from wordlists has played in the historical analysis of the structure of the Bantoid family: (i) Individual-level lexical variation in Bantoid languages may be much higher than what has been implicitly assumed in previous work, and relying on a single wordlist to stand in for a "language" may be skewing the results of comparative analysis in significant ways. And, (ii) cognate detection tools developed for historical linguistics can be usefully adapted for synchronic investigation, yielding results that can both be used to provide a snapshot of variation within a given area and to detect which concepts may be associated with forms that serve as salient sociolinguistic markers of linguistic difference and which concepts may have forms that are more historically stable.

On a more conceptual level, other questions emerge. One of these is understanding what features are used to maintain linguistic difference within communities who categorize themselves as linguistically different, but whose varieties are quite close to each other in lexical terms. Another is whether communities whose varieties are quite distinct from those of other communities that they are in contact with can tolerate a higher degree of individual-level lexical variation due to the fact that they will still remain unambiguously distinctive in the local linguistic space (see §3.5). Finally, underlying all of the work presented here is understanding what conditions the observed individual-level linguistic variation and how it is tied to individuals' multilingual repertoires. More broadly, we hope that the work presented here will prompt more studies of individual-level variation in the Bantoid area since we believe that these will provide the necessary foundation for studies of the development of the family that incorporate accurate representations of the sociolinguistic contexts of its speaker communities into models of language change.

5 References

Akumbu, Pius W. & Roland Kießling. 2023. Variation in central ring: Convergence or divergence? *Linguistic Typology at the Crossroads* 3(1). 19–42. https://dx.doi.org/10.6092/issn.2785-0943/16312.

Anchimbe, Eric A. 2013. *Language policy and identity construction: The dynamics of Cameroon's multilingualism*. Amsterdam: Benjamins.

Blench, Roger M. 2015. The Bantoid languages. In *Oxford handbooks online: Linguistics*. Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780199935345.013.17.

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation* 7. 331–359. http://hdl.handle.net/10125/4606.

Di Carlo, Pierpaolo. 2011. Lower Fungom linguistic diversity and its historical development: Proposals from a multidisciplinary perspective. *Africana Linguistica* 17. 53–100.

Di Carlo, Pierpaolo & Jeff Good. 2014. What are we trying to preserve? Diversity, change, and ideology at the edge of the Cameroonian Grassfields. In Peter K. Austin & Julia Sallabank (eds.), *Endangered languages: Beliefs and ideologies in language documentation and revitalization*, 229–262. Oxford: OUP.

Di Carlo, Pierpaolo & Jeff Good. 2023. Language contact or linguistic micro-engineering? Feature pools, social semiosis, and intentional language change in the Cameroonian Grassfields. *Linguistic Typology at the Crossroads* 3(1). 72–125. https://doi.org/10.6092/issn.2785-0943/17231.

Di Carlo, Pierpaolo, Jeff Good & Rachel Ojong Diba. 2019. Multilingualism in rural Africa. In *Oxford research encyclopedia of linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.227.

Dimmendaal, Gerrit J. 2009. Esoterogeny and localist strategies in a Nuba mountain community. *Sprache und Geschichte in Afrika* 20. 75–95.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.

Esene Agwara, Angiachi Demetris. 2013. *Rural multilingualism in the North West Region of Cameroon: The case of Lower Fungom*. Buea, Cameroon: University of Buea MA thesis.

Esene Agwara, Angiachi Demetris. 2020. What an ethnographically informed questionnaire can contribute to the understanding of traditional multilingualism research: Lessons from Lower Fungom. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 181–203. Lanham, MD: Lexington Books.

Forkel, Robert & Johann-Mattis List. 2020. CLDFBench: Give your cross-linguistic data a lift. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Ishara, Bente Maegaard, Hélène Mazo

Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6995–7002. Marseille: European Language Resources Association (ELRA). https://doi.org/10.17613/8t0e-w639.

Good, Jeff. 2013. A (micro-)accretion zone in a remnant zone? Lower Fungom in areal-historical perspective. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols*, 265–282. Amsterdam: Benjamins.

Good, Jeff. under review. The micro-dynamics underlying large-scale areal patterns: Reviving early approaches to African linguistic prehistory. In Tom Güldemann & Jakob Lesage (eds.), *Between Niger-Congo and the Macro-Sudan Belt*. Berlin: Language Science Press. https://buffalo.edu/~jcgood/Good-WestermannVolume.pdf.

Good, Jeff, Jesse Lovegren, Jean Patrick Mve, Nganguep Carine Tchiemouo, Rebecca Voll & Pierpaolo Di Carlo. 2011. The languages of the Lower Fungom region of Cameroon: Grammatical overview. *Africana Linguistica* 17. 101–164.

Grollemund, Rebecca. 2012. *Nouvelles approches en classification: Application aux langues bantu du nord-ouest*. Lyon: Lumière University Lyon 2 doctoral thesis.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti & Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* 112(43). 13296–13301.

Hamm, Cameron, Jason Diller, Kari Jordan-Diller & Ferdinand Assako a Tiati. 2002. *A rapid appraisal survey of Western Beboid languages (Menchum Division, Northwest Province)* (SIL Electronic Survey Reports 2002-014). Dallas, TX: SIL International.

Hantgan, Abbie & Johann-Mattis List. 2022. Bangime: Secret language, language isolate, or language island? *Papers in Historical Phonology* 7. 1–43.

Heggarty, Paul, Warren Maguire & April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. 3829–3843.

Hombert, Jean-Marie. 1980. Noun classes of the Beboid languages. In Larry M. Hyman (ed.), *Noun classes in the Grassfields Bantu borderland*, 83–98. Los Angeles: University of Southern California Department of Linguistics.

Hombert, Jean-Marie & Rebecca Grollemund. 2018. Phylogenetic classification of Grassfields languages. In Eugene Buckley, Thera Crane & Jeff Good (eds.), *Revealing structure: Papers in honor of Larry M. Hyman*, 85–103. Stanford: CSLI.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23. 254–267.

Idiatov, Dmitry & Mark Van de Velde. 2021. The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa. *Language* 97. 72–107.

Kühl, Karoline & Kurt Braunmüller. 2014. Linguistic stability and divergence: An extended perspective on language contact. In Kurt Braunmüller, Steffen Höder & Karoline Kühl (eds.), *Stability and divergence in language contact: Factors and mechanisms*, 39–60. Amsterdam: Benjamins.

List, Johann-Mattis. 2012. SCA: Phonetic alignment based on sound classes. In Daniel Lassiter & Marija Slavkovik (eds.), *New directions in logic, language and computation: ESSLLI 2010 and ESSLLI 2011 student sessions: Selected papers*, 32–51. Berlin: Springer.

List, Johann-Mattis & Robert Forkel. 2021. *LingPy: A Python library for historical linguistics (version 2.6.8)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

List, Johann-Mattis, Simon J. Greenhill & Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). 1–18. https://doi.org/10.1371/journal.pone.0170046.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel.

2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3. 130–144. https://dx.doi.org/10.1093/jole/lzy006.

Lovegren, Jesse. 2013. *Mungbam grammar*. Buffalo, NY: University at Buffalo PhD dissertation.

Lüpke, Friederike. 2016. Uncovering small-scale multilingualism. *Critical Multilingualism Studies* 4. 35–74.

Lüpke, Friederike & Anne Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.

Mansfield, John. 2021. The word as a unit of internal predictability. *Linguistics* 59. 1427–1472.

Mba, Gabriel & Angela Nsen Tem. 2020. Ways to assess multilingual competence in small, unwritten languages: The case of Lower Fungom. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 205–224. Lanham, MD: Lexington Books.

Mve, Patrick, Nelson C. Tschonghongei, Pierpaolo Di Carlo & Jeff Good. 2019. Cultural distinctiveness and linguistic esoterogeny: The case of the Fang language of Lower Fungom, Cameroon. In Pius W. Akumbu & Esther P. Chie. (eds.), *Engagement with Africa: Linguistic essays in honor of Ngessimo M. Mutaka*, 163–178. Köln: Rüdiger Köppe.

Nurse, Derek & Gérard Philippson. 2003. Introduction. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 1–12. London: Routledge.

Pakendorf, Brigitte, Nina Dobrushina & Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25. 835–859.

Pommerolle, Marie-Emmanuelle & Hans De Marie Heungoup. 2017. The "Anglophone crisis": A tale of the Cameroonian postcolony. *African Affairs* 116. 526–538.

Roberts, James & Keith Snider. 2006. *SIL comparative African wordlist (SILCAWL)*. SIL Electronic Working Papers: SILEWP 2006-005.

Schadeberg, Thilo C. 2003. Historical linguistics. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 143–163. London: Routledge.

Slaska, Natalia. 2005. Lexicostatistics away from the armchair: Handling people, props and problems. *Transactions of the Philological Society* 103(2). 221–242.

Slaska, Natalia. 2006. *Meaning lists in lexicostatistical studies: Evaluation, application, ramifications*. Sheffield: University of Sheffield PhD thesis.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21. 121–137. http://www.jstor.org/stable/1263939.

Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55–75. Berlin: Mouton de Gruyter.

Thurston, Willam R. 1987. *Processes of change in the languages of north-western New Britain* (Pacific Linguistics Series B – No. 99). Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.

Thurston, Willam R. 1989. How exoteric languages build a lexicon: Esoterogeny in West Britain. In Ray Harlow & Robin Hooper (eds.), *VICAL 1, Oceanic languages: Papers from the fifth International Conference on Oceanic Linguistics*, 555–579. Auckland: Linguistic Society of New Zealand.

Trudgill, Peter. 2004. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305–230. https://doi.org/10.1515/lity.2004.8.3.305.

Voll, Rebecca. 2017. *A grammar of Mundabli: A Bantoid (Yemne-Kimbi) language of Cameroon*. Leiden: University of Leiden PhD dissertation.

Warnier, Jean-Pierre. 1980. Des précurseurs de l'école Berlitz: Le multilingualisme dans les Grassfields du Cameroun au 19ème siècle. In Luc Bouquiaux (ed.), *L'expansion bantoue: Actes du colloque international du CNRS, Viviers (France) 4–16 avril 1977. Volume III*, 827–844. Paris: SELAF.

Watson, Rachel. 2019. Language as category: Using prototype theory to create reference points for the study of multilingual data. *Language and Cognition* 11. 125–164. https://dx.doi.org/10.1017/langcog.2019.9.

Watters, John R. 2003. Grassfields Bantu. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 225–256. London: Routledge.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.

Wu, Mei-Shin & Johann-Mattis List. 2023. Annotating cognates in phylogenetic studies of Southeast Asian languages. *Language Dynamics and Change* 13(2). 161–197. https://dx.doi.org/10.1163/22105832-bja10023.