Chapter 1

Areal micro-dynamics as a window onto Niger-Congo prehistory

Jeff Good
University at Buffalo

In the latter half of the twentieth century, African historical linguistics was dominated by work that emphasized genealogical relatedness over the study of language contact despite clear evidence for the significance of contact in shaping the continent's linguistic diversity. Recent years have seen increased interest in the study of the role of language contact for understanding high-level patterns of language change in Africa, in particular in studies considering the significance of areality in the distribution of different grammatical features on the continent. This chapter seeks to expand and build on these developments by adding an element of historical sociolinguistic reconstruction to the study of African prehistory and proposing new methods to detect areal patterns in lexical data. Its methodological contribution focuses on a linguistic micro-area, the Lower Fungom region of the Cameroonian Grassfields, and it uses quantitative methods to simultaneously explore genealogical and areal patterns in individual-based lexical data. A key result is the extent to which the aggregated lexical data encodes information about the spatial relationships among Lower Fungom's villages, a clear indicator that areal factors have played a major role in the region's language dynamics.

Acknowledgements

The work underlying this paper is part of the Key Pluridisciplinary Advances in African Multilingualism–Cameroon Project, which has been supported by various grants, with US National Science Awards BCS-BCS-1761639 and BCS-2109620 offering key support for the work described here. This work has also benefited substantially from the input of Ling Bian, Pierpaolo Di Carlo, Clayton Hamre,

Yujia Pan, and Nelson C. Tschonghonghei, as well as the volume editors. None of the results described here would be possible without the additional support of members of the language communities of the Lower Fungom region of Cameroon and key local assistants, including Achuo Christopher Ikom and Nyoh Abang Charles.

1 The historical dynamics underlying areal patterns

In the latter half of the twentieth century, an idea took hold that the work of Greenberg (see, in particular, Greenberg 1963) represented a decisive shift towards a more rigorous approach to the classification of African languages than had previously been the case.[1] The work of Welmers (1973:2) offers a particularly critical appraisal of earlier work:[2]

> To an extent unparalleled in the study of languages anywhere else in the world, African language classification has been beset by persistent hypotheses of language mixture, intermediate or transitional languages, substrata, pervasive external influence far in excess of what is usually regarded as normal, and innovative exuberance unmatched in recorded language history.

In making his critique, Welmers (1973:1–19) shows an explicit bias towards modeling genealogical relationships over contact relationships, including what we would today refer to as areal patterns. For example, he questions the value of the characterization provided by Westermann (1952:253–254) of certain Benue-Congo languages of the Cross River region of Nigeria as being transitional in nature between Bantu languages and other Benue-Congo languages. This is apparently due to a belief that such characterizations ignore genealogical relations (Welmers 1973:6).

From a contemporary perspective, the position taken by Welmers (1973) must be considered too strong. While it may have been partly informed by the need to ensure that the racist dimensions of earlier approaches to African language classification were not carried forward into more modern scholarship (see Childs 2003:35 and Gerhardt 2020:132–133), if our concern is to understand the actual

[1] I refer to Greenberg (1963) in this paper as this seems to now be the standard citation for his proposals, though these were also discussed in earlier publications such as Greenberg (1949).

[2] See also Newman (1995) for a particularly strong defense of Greenberg's work on the classification of African languages. As an example of problematic scholarship, Welmers (1973) specifically cites Johnston (1919:27).

history of African languages, models of their development must necessarily incorporate both genealogical and contact relationships. Moreover, investigation of the latter class of relationships needs to encompass not only cases where contact effects among specific groups of languages can be identified and where the direction of influence is relatively clear, as is seen, for example, in the development of click consonants within languages of certain African language families (Güldemann & Stoneking 2008:99), but also those cases where the areal patterning of a feature can only be explained as the result of language contact even if it is not possible to determine the language or language family from which the feature originated. This is the case, in particular, for many of the features associated with a linguistic area that has been referred to as the Macro-Sudan Belt within Africa (Güldemann 2008).

One view of the critique offered by Welmers (1973) of work on African language classification that predated Greenberg (1963) might be to suggest that it is representative of the usual kinds of "pendulum shifts" found in any long-term research enterprise. His emphasis on traditional genealogical relations would have served as a kind of reaction to earlier approaches that did not make a rigorous distinction between genealogy and contact. More recent work focusing on areal relationships in African languages such as Güldemann (2008), Rolle et al. (2020), and Idiatov & Van de Velde (2021) could then, in turn, be understood as the "natural" counter-reaction to such a genealogical emphasis. However, in this paper, I will take a stronger position. Earlier claims regarding language "mixture", or similar processes, may have been lacking in the precision of the later genealogical analyses. Nevertheless, these claims were based on significant patterns found in the data, and dismissals of them were not merely too strong but also counterproductive. In particular, they hindered the development of historical approaches to African language classification reflecting the actual sociohistorical realities of the continent's communities, which, as argued below, were likely to have included much more complex patterns of language contact than standard genealogical approaches assume.

In the rest of this paper, I will defend this claim on both conceptual and methodological grounds. My focus will be on the relationship between Narrow Bantu languages and their closest Bantoid relatives. Notably, the analysis of the genealogical relationship holding between them has long been treated as one of the major results of Greenberg's (1963) comparative work (Newman 1995:4). In §2, I discuss my assumptions regarding the nature of historical linguistic analysis when used to reconstruct prehistoric linguistic relations in general terms, turning to specific issues of Bantu linguistic prehistory, with a focus on reconstructing the

family's early sociolinguistic patterns, in §3. The discussion moves to methodological concerns in §4, with a specific emphasis on the role of individual-based wordlists collected on the varieties of a linguistic micro-area as a tool for modeling genealogical and areal relationships simultaneously. Concluding remarks are presented in §7.

2 Reconstructing historical linguistic relationships

2.1 Language splits and "normal" transmission

Concerns about language relatedness have been among the most longstanding research areas within linguistics. Given this, it is surprising that key theoretical and methodological assumptions that underlie historical linguistic work are more often assumed than made explicit—and rarely presented systematically. Campbell & Poser (2008) is probably the most comprehensive relatively recent work on this topic, though its overall orientation is critical in nature rather than laying out the elements of a productive historical linguistic research program. Of particular relevance in the present context is the fact that much historical linguistic reasoning is based on an idealized model of "language" spread which involves either (i) a coherent language community being impacted by some kind of demographic process which causes it to split into multiple language communities, leading to tree-like change, or (ii) geographic spread paired with the maintenance of relative close social ties leading to a dialect continuum of the sort that is associated with wave-like change (see Heggarty et al. (2010) for further elaboration).

Greenberg (1972:196) clearly adopts the idealized "split-based" model in his interpretation of the prehistory of the Bantu family when stating, "I maintain that Bantu is merely a group within the vast Niger-Congo family and can in fact be placed within that family in the Benue-Congo branch and can be further pinpointed as a member of the Bantoid sub-branch of Benue-Congo." What is noteworthy about his claims in the present context is the degree of precision in which they are presented. This is despite the fact that, as Greenberg (1972:195) was clearly aware, the structure of Bantu was at the time unclear, in particular with respect to the relationship between the Bantu languages of the northwestern area of the family's distribution and the rest of the group.[3] In fact, this issue

[3] In providing this assessment, Greenberg (1972) may have been relying on the idea that the so-called Greenberg-Crabb hypothesis regarding nasal classes in Bantu cleanly delimited Bantu from non-Bantu Bantoid (see Hyman 2018:225 for updated discussion).

has yet to be resolved, even if substantial progress has been made in the following decades (Bostoen 2022:xvii–xix). Greenberg (1972) can only be sensibly interpreted if one assumes that the demographics of the Bantu spread (and, indeed, the development of all of Niger-Congo) were structured in a way such that the accompanying patterns of language diversification would show clear tree-like patterns, even though the actual evidence regarding internal variation within Bantu and the link of Bantu to its closest relatives does not show clear patterns of this kind (see, e.g., Schadeberg 2003:158–159 and Good 2023). Similar issues arise for the internal classification of Niger-Congo more broadly (Good 2020b:143). Thus, while many of the critiques that Greenberg (1972) makes regarding Guthrie's (1962) interpretation of Bantu prehistory would be widely accepted as valid today, the actual evidence in favor of key aspects of his alternative interpretation does not rest on as strong methodological footing as he appears to assume.[4] Put differently, the fact that Guthrie (1962) may have been "wrong" does not necessarily imply that Greenberg (1963) was "right" (*pace* Newman 1995).

Models of language change that are built around demographic splits rely on a related assumption, whether explicitly—or, more usually, implicitly. Namely, at any given historical period, each language is associated with a distinct community of users who view that language as their primary language of identity. This assumption is further connected to an idealized model of language transmission where acquisition takes place within households assumed to be primarily monolingual in nature. Welmers (1973:3) is helpfully explicit on this point (despite providing no evidence regarding the basis of his claims):

> Unquestionably the most normal, though by no means the only, variety of language history is the continuous, unbroken transmission of a system of communication from one generation to the next. Parents and children, grandparents and grandchildren, communicate with each other…Assume that a linguistically homogeneous community splits into two groups, through a process such as migration or invasion that creates a geographical separation between them. As long as neither group completely gives up its own language to adopt the language of some other people, there will now be two separate generation-to-generation continua.

As will be discussed in §3, there is no reason to believe that the situation described by Welmers (1973) is "normal" for sub-Saharan African settings, even if

[4] Moreover, even when the comparative data does show clear tree-like patterns, there are more potential sources to such patterns than simple demographic splits. See, e.g., Garrett (2006) on Indo-European.

it does characterize the situation for some communities. Recent work has made clear that, within many communities, individual-level multilingualism predominates and social structures are built around the fact that multilingualism is, in fact, the "norm" (see, e.g., Lüpke 2016, Evans 2017, Di Carlo et al. 2019, and Pakendorf et al. 2021). In such settings, monolingual transmission within a single household would, in fact, be decidedly "abnormal".[5]

Without denying its clear positive contributions to linguistic scholarship, it is difficult not to see in the work of Greenberg (1963) and his defenders how the problematic racist ideologies of earlier scholars were replaced with similarly problematic colonialist ideologies. In particular, it imposed models of language diversification developed in the context of work on Indo-European languages which not only had deep ties to European nationalism but where the nationalism itself had impacted the development of the European languages that were the focus of study (Mallory 1989:266–270; Garrett 2000). Moreover, this work was done in the larger context of the colonialist and imperialist enterprise that enabled the kind of large-scale comparative philology undertaken by scholars like Westermann (1927) and Greenberg (1963) in the first place (Errington 2001). In adopting an approach to change emphasizing demographic splits and monolingual language transmission, the multilingual social reality of African communities (see §3) was, in effect, erased and replaced with a model which treated small African language communities as taking the form of a "tribal homunculus" (Kopytoff 1987:4–7) that could be viewed as a kind of proto-nation. This model may have been conceptually comfortable from a Western perspective, but it was not justified by the observable sociolinguistic patterns on the ground.

Such abstract conceptual concerns may, at first, appear to have relatively little connection to concrete questions of language relatedness as determined through the comparison of lexical and structural linguistic features along the lines of the procedures outlined in Nichols (1996). However, at least in the case of African language classification, the ideological assumptions that inform the approaches taken by scholars have had concrete impacts on the research agendas of those

[5]While he may be unusually explicit in his assumptions, Welmers (1973) is not alone in holding a very narrow view of the typical patterns for language transmission. For example, in their important work on language contact, Thomason & Kaufman (1988:10) link genealogical relationships, to "normal" transmission that is defined in a comparable way to Welmers (1973). Similar claims have also been made more recently by Jacques & List (2019:129) in their defense of the tree model of language change, when they state that "[a]ll languages develop by descent with modification" and "that the primary process of language transmission is through childhood acquisition of a first language". This may be true for the majority of individuals living today, but it does not follow that it was necessarily the historical situation across the world. See §2.2.

working on the description and analysis of Africa's linguistic diversity. In this specific case, Greenberg's (1963) genealogical results were not merely seen as additive to early work which had recognized the significance of contact relations among African languages for understanding their prehistory but, rather, as replacive of such work. It would take some decades for work on language contact to again become prominent in Africanist linguistics (Lüpke 2010:2–3).[6] This has impacted not only high-level studies of linguistic relations on the continent but also locally oriented studies. If genealogical relations are seen as the key concern, then documentary and descriptive work can reasonably focus on one language at a time without considering the broader ecology of its speakers. If contact relations are seen as being of comparable interest to genealogical ones, that suggests the need for such work to adopt a more sociolinguistically grounded approach (see, e.g., Childs et al. 2014).

2.2 Determining the historical connections that hold between "related" languages

An additional relevant concern here relates to an underappreciated observation of Gensler (1994:12). This is that the fundamental distinction in comparative linguistic work is not whether resemblances between two languages are due to genealogy *or* contact but, rather, whether they represent a historical relationship of any kind between the users of these languages or if they are merely due to chance. First, chance resemblance must be ruled out. After that is done, it makes sense to consider whether a given set of resemblances are due to genealogy, contact, or a historical relationship of indeterminate type, all of which are important for the study of African prehistory.

The observations made by Gensler (1994:12) can, in my view, be extended further. Establishment of a historical relationship among a set of lexicogrammatical codes—i.e., the features of languages that emphasize their structural nature as a communicative system—should not be taken to automatically imply any specific kind of sociolinguistic reconstruction of the populations using those codes. Instead, the logic of any assumed connection between a historical lexicogrammatical code and its associated language community should be made explicit and

[6] Moreover, even though there has been a noticeable increase in work on language contact in Africa in recent years in specific settings, as will be made clear in §3, at higher levels, an emphasis on genealogical relations still dominates work on Bantu prehistory, with recent work, in particular, making use of phylogenetic methods adapted from evolutionary biology (Philippson & Grollemund 2019:343–347). The algorithms that these methods are based on are built on essentially the same assumptions as those outlined above about the nature of linguistic diversification.

supported with evidence that is as specific as possible with respect to the communities in question. Put differently, structural linguistic tools, such as the comparative method, only directly establish a historical relationship among lexicogrammatical codes. While one can infer from such kinds of comparison that there must have been historically related communities using those codes, further reconstruction and evidence is required to arrive at defensible proposals regarding the precise sociolinguistic character of those communities.

At present, this is not the standard approach, which instead adopts uniformitarian assumptions about the social structures in which lexicogrammatical codes are transmitted, as discussed in §2.1 (see Labov 1972:275, Christy 1983, and Bergs 2012 for discussion of the uniformitarian principle in linguistics). Consider, for example, the critique of Campbell & Poser (2008:352) regarding the claims of Ross (1996) (building on the work of Thurston 1989) that processes of esoterogeny can play a significant role in language change. These involve changes that add complexity to a language in the local linguistic context that would make it harder for those outside of a given language community to understand it. The nature of their criticism is summarized as follows (Campbell & Poser 2008:352):

> There are difficulties with this interpretation. It has been speculated that little-known, so-called exotic languages may change differently from others, attributed to various social and cultural characteristics of the speakers, such as small size of the speech community, the face-to-face interaction, isolation, lack of literacy, and so on…However, almost invariably, claims about general patterns of change correlated with these aspects of non-linguistic culture have proven indefensible. In the case of Thurston's and Ross' "esoteric" languages, it is attractive to imagine that speakers of these languages have initiated (or at least used) particular changes for the purpose of distinguishing themselves from and thus excluding outsiders. However, it is not clear how this hypothesized cultural motive for these changes—conscious exclusion of outsiders (Ross 1997:239)—could be tested or how the investigator might distinguish changes motivated for this purpose from changes that **just happen** with no such motive. [emphasis added]

The implicit assumption in Campbell & Poser's (2008) criticism of Ross (1996) is that the normal state of affairs is for changes to "just happen", in which case the fact that they happened does not need any explanation beyond, perhaps, invoking some notion of language "drift" (see Joseph 2013). However, this ignores the fact that a central concern of work within historical linguistics is precisely

the "actuation problem" (Weinreich et al. 1968:186)—i.e., understanding the cognitive and social factors that can account for why an attested change took place. Campbell & Poser (2008) appear to assume that *not* providing a sociolinguistic account for why a change has occurred is, somehow, preferable to hypothesizing a sociolinguistically grounded explanation for it. Taken to its logical conclusion, this reasoning, somewhat perversely (though, presumably, unintentionally), suggests that an account of a change which makes no attempt to address the actuation problem may be preferable to one that does.

This unusual argument can be made sensible if, like Greenberg (1972) and Welmers (1973), as discussed in §2.1, Campbell & Poser (2008) assume that language change primarily happens in contexts of so-called "normal transmission". That is, contexts with the following characteristics: (i) children learn their primary language in a household dominated by that language and within a community that also identifies that same language as its primary language and (ii) where many individuals are either monolingual or have one clearly dominant language which they use across their lifespan. In such settings, explaining a change in one language as the result of a specific sociolinguistic pressure to ensure that it is different in some salient way from another language in its ecology would indeed be unusual. However, as Trudgill (2020:8) points out, while the application of the uniformitarian principle in accounts of language change is justified for kinds of change that are connected to human characteristics that can reasonably be assumed to be universal across time (e.g., properties of the vocal tract), it cannot be extended to sociolinguistic aspects of change. The sociocultural contexts in which humans communicate are not uniform, and, in particular, they have, in many communities, changed drastically in the last few millennia as humans shifted from living in small-scale societies, where they mostly interacted only with people they already knew, to large-scale societies, where they routinely interact with strangers (see Lüpke 2016 and Pakendorf et al. 2021 for discussion in a linguistic context).

If we cannot assume the existence of a "default" sociolinguistic context in which language change takes place, then, when looking at general patterns of change within a family or linguistic area, we must also consider the sociolinguistic setting of the societies in which change is being modeled, to the extent that available data allows. In the next section, I will propose an initial reconstruction of the historical sociolinguistics of early Bantu communities en route to consideration of how we can develop sociolinguistically grounded models of Bantu linguistic diversification, at least in its earliest phases.

3 The historical sociolinguistics of early Bantu communities

As discussed in §1, recent scholarship such as Güldemann (2008, 2010, 2018), Rolle et al. (2020), and Idiatov & Van de Velde (2021) has firmly established the significance of areality in understanding Africa's linguistic diversity. Given this, it is now apparent that rigorous historical investigation of the continent's languages needs to be based on an understanding of the sociolinguistic relations holding among its language communities that is as explicit and accurate as possible. Otherwise, it will be difficult, and perhaps impossible, to answer crucial questions such as how the Macro-Sudan region can, at the same time, show extensive linguistic fragmentation—to use Dalby's (1970) term—while also showing evidence of shared grammatical features across thousands of kilometers. Moreover, this kind of investigation can clarify what types of sociolinguistic configurations could produce the synchronic linguistic patterns that would result in "persistent hypotheses of language mixture" (Welmers 1973:2). This would allow us to shift from dismissing them as being based on fundamentally flawed premises and facilitate research on the dynamics involved in creating the observable linguistic patterns which early scholars chose to characterize in that way.

The historical sociolinguistics of Bantu communities has not been an area that has seen significant dedicated research.[7] There have certainly been historical studies where sociolinguistic concerns are relevant, as seen in, for example, work on how click sounds entered Bantu languages from Khoisan languages (see, e.g., Herbert 1990 and Bostoen & Sands 2012). Closer to the likely Bantu homeland, comparative work on Chadic and Benue-Congo languages has established the ancient nature of contact among languages of these groups (see, e.g., Wolff & Gerhardt 1977 and Gerhardt 1983) as well as proposals for how contact relations have changed over time (Storch 2009:302–303). (See also Good 2017:482–485.) More recently, the study of Verkerk & Garbo (2022) considers the role of sociolinguistic factors in the development of Bantu noun class systems using quantitative methods. However, the goal of these studies is not to present a holistic view of the historical sociolinguistic dynamics of early Bantu communities.

In order to advance our understanding of the processes of change that impacted Bantu languages, in particular in the early stages of the Bantu expansion before the family had clearly spread beyond the southern periphery of the Macro-

[7] While interest in the study of historical sociolinguistics as a distinct subdiscipline has increased in recent decades, work in this area has been centered on the study of European languages and, in particular, English (Auer et al. 2015).

Sudan Belt, in (1) I propose an initial reconstruction of the historical sociolinguistics of early Bantu communities. This should be seen as an initial proposal rather than a definitive reconstruction. It is strongly informed by the linguistic situation of the Grassfields region, which is within the general area that has been proposed as the Bantu homeland (Nurse & Philippson 2003:5; Grollemund et al. 2015), and, within the Grassfields, the specific situation of the Lower Fungom region has been especially influential due to the extent to which its sociolinguistic dynamics have been studied from the perspective of their impact on language change (Good 2013, 2023, Di Carlo & Good 2014, 2023, Mve et al. 2019).[8] The Grassfields is chosen due to its plausibility as a model for the historical situation given its location and due to the fact that it is relatively well studied in both linguistic and anthropological terms, allowing for more precision in a reconstruction than would be possible for other potential candidate areas.[9]

(1) a. **Early Bantu communities were relatively small:** The area around the contemporary Cameroon–Nigeria border is characterized by the presence of many language communities with populations ranging from the hundreds to the (low) hundreds of thousands (Watters 2003, Blench 2015). The historical populations of the larger communities were presumably much lower than their contemporary populations, perhaps in the tens of thousands. Relatedly, the primary geographic area associated with these language communities is also typically quite small, resulting in high language density in the region (Stallcup 1980).

b. **Early Bantu communities were characterized by extensive multilingualism at the individual-level:** Work focusing on patterns of multilingualism in the Grassfields from before the period when the speech varieties that would become Cameroon Pidgin English were introduced to Cameroon (see Menang 2004:903–904) shows that the pre-contact sociolinguistic pattern was for individuals to be able to use a number of languages associated with the general region in which they lived, in a pattern that is likely to have extended to a large area of West Africa (Warnier 1980, Di Carlo et al. 2019). While presumably different in degree, widespread individual-level multilingualism also appears to be

[8] The fact this region has been a focus of my own research is also obviously relevant here. This bias is another reason why this should be viewed only as an initial proposal that will hopefully prompt debate leading to its improvement.

[9] The structure of the reconstruction in (1) takes its inspiration from the approach of Nichols (1990:496–499) to reconstructing the early social structure of the groups that led to the peopling of the Americas.

a significant historical characteristic of the larger Bantu area (Schadeberg 2003:158).

c. **Language choice was linked to the expression of relational identities rather than essentialist identities in early Bantu communities:** Given the fact that the nature of language ideologies in non-urban parts of Africa is relatively under-researched, any claims about early Bantu language ideologies must be made with caution. However, to the extent that current research allows historical inferences to be made, language ideologies in the early Bantu period were unlikely to be associated with essentialist or categorical identities, where use of a language was linked to a larger complex of identity features (e.g., some notion of "ethnicity"). Rather, they were more likely used to primarily express a more direct *relational* identity (see Brubaker & Cooper 2000) primarily signifying that the language user was part of the set of individuals who knew and used that language, e.g., as part of the expression of a kinship or a political relationship of some kind. This aspect of the reconstruction relies heavily on work on the languages of the Lower Fungom region (see, e.g., Di Carlo et al. 2020), but support can be found in detailed case studies of other regions such as the description of multilingualism among "Montagnard" communities in the Mandara mountains of Cameroon discussed by Moore (2004) or the rich case studies of multilingual individuals in Senegal presented in Lüpke & Storch (2013:22–33).

d. **Manipulated language was a regular part of the metalinguistic practices of early Bantu communities:** As documented in Storch (2011), African language communities attest a wide range of types of manipulated language associated with different social contexts. Relevant examples from the Grassfields and nearby areas include the royal register of the Bafut language (Di Carlo & Neba 2020), ritual language registers in Jukunoid languages (Storch 2011:38–41), and the construction of a special language within one generation of Limbum speakers (Ndamsah 2020). Practices associated with language manipulation demonstrate a high degree of metalinguistic awareness of language similarity and difference by members of these communities and suggest that language manipulation may have played an important role in language change when manipulated forms became part of the "regular" register of a language.

e. **Early Bantu language communities were characterized by the dynamics of the internal African frontier:** In addition to specifically sociolinguistic features, the reconstruction of broader cultural features also necessarily plays a role in the reconstruction of Bantu prehistory. Here, the work of Kopytoff (1987:16–17) on the so-called internal African frontier is relevant, in particular since it was strongly informed by the structure of Grassfields societies. In this model, the structure of certain African societies is in constant flux due to the movement of people, and their associated kin groups, into and out of different polities, which, in some cases, is accompanied by the formation of new polities. This formation and reformation of political units was further accompanied by processes where a given polity would adopt some of the cultural features from other nearby polities, including, potentially, linguistic features (see Di Carlo & Good 2023), as well as other practices such as different kinds of secret societies (see Röschenthaler 2011). This would have resulted in a mosaic-like pattern of similarities and differences among local communities in both linguistic and cultural terms.

Focusing on the Grassfields in general, and Lower Fungom in particular, may bias the reconstruction of the sociolinguistic situation with the result that (1) differs in crucial ways from the actual historical situation, which is an important reason why this should be treated as an initial proposal intended to advance work in this area rather than a reconstruction that has been firmly established. Just as it is now recognized that Bantu lexical and grammatical reconstructions need to consider the northwestern Bantu languages more carefully than was previously the case (Bostoen 2022:xviii), a sociolinguistic reconstruction of this kind will clearly be stronger if it can consider a wider range of regions and reconstruct how they might have developed from an initial sociolinguistic configuration. Relatedly, this reconstruction should be seen only as applying to the early Bantu situation, i.e., the period of the initial expansion of the languages into the area now associated with northwest Bantu. I do not assume that these features also characterize the historical situation of Bantu languages as they spread south and east within southern sub-Saharan Africa. Nevertheless, I believe that the proposal in (1) can serve as a useful starting point.

There are also sociolinguistic aspects of early Bantu communities that would clearly be useful to reconstruct for historical purposes but for which no proposal can easily be developed based on the available data. For instance, to understand how multilingualism might impact language contact phenomena, a reconstruc-

tion of the rules governing language choice within a given setting would be valuable. Available studies of this in non-urban African settings are limited, however, and do not show consistent patterns. In Lower Fungom, for example, code "regimentation" appears to be the norm where, in a given exchange, one local variety is chosen by the interactants and this remains stable unless the social conditions change (Ojong Diba 2020). In other nearby areas, code regimentation of this kind is not found (Tabe 2020). This is not to say that a reconstruction of the principles behind language choice cannot be undertaken but, rather, that this would require further detailed study.

If the sociolinguistic patterns of early Bantu communities were structured along the lines of what is described in (1), then we may have the beginnings of an account as to why, "African language classification has been beset by persistent hypotheses of language mixture, intermediate or transitional, substrata, pervasive external influence far in excess of what is usually regarded as normal (Welmers 1973:2)." (See §1.) In particular, societies characterized by frontier dynamics (1e) where individual-level multilingualism is the norm (1b), have foundational features in which language "mixture" could take place as individuals with knowledge of a number of local language used them to "establish distinct group identities…to foster good neighborship" (Schadeberg 2003:158) in a social context where the changing alliances of kin groups result in ongoing flux in the organization of local polities.

The additional reconstructed sociolinguistic features of these societies would have facilitated such processes. Small-sized language communities (1a) would both increase opportunities for contact among individuals with knowledge of different languages and facilitate the spread of new linguistic features within them, especially for the smaller linguistic communities of a few thousand speakers or fewer. The linkage of language choice to relational identities, rather than essentialist identities (1c), where linguistic change would be a tool to create new relational groupings is also relevant. In such contexts, the mixture of elements from different languages could serve as a means of creating a new language to express a new group identity (see, e.g., Di Carlo & Good 2014:243–246). Finally, the widespread use of language manipulation strategies (1d) also creates a situation where existing linguistic material can undergo semantic and formal shifts that would create irregular historical correspondence of a sort that could be easily confused with some kind of language mixture as well.

To this point, the arguments about historical change presented here have been largely conceptual. The rest of this paper shifts the discussion to more concrete matters of data collection and analysis. The goal is to introduce an analytical approach that can be used to model and investigate the dynamics of diversification

which would be predicted if we assume that early Bantu language communities had sociolinguistic characteristics along the lines of those presented in (1).

4 An individual-based dataset of Lower Fungom's varieties

4.1 Sociolinguistically grounded data collection and analysis

A leading idea of this paper is that, since all language change takes place in some kind of sociolinguistic context, our analysis of linguistic diversification for a given family or region should be based on an explicit analysis of the sociolinguistic dynamics in which that diversification took place. While there may be many situations where the sociolinguistic context favors tree-like diversification linked to migration and separation, this is not obviously the case for the reconstruction of the sociolinguistics of early Bantu communities offered in §3. This does not mean that it is clear what kinds of diversification should be found beyond the possibility that it might look like "mixture" of some kind. Mixture is, of course, a vague notion, and a convincing historical analysis requires more precision, especially when it is being offered as an alternative to the better-studied model of tree-like diversification under conditions of so-called "normal" transmission described in §2.1.

Arriving at a high level of precision requires attention both to the kinds of data that are collected on Bantu and Bantoid languages and to the methods of analysis applied to that data. Based on the study of the language dynamics of the Lower Fungom region of the Grassfields, a number of research strategies have been adopted to collect and interpret data from highly multilingual individuals occupying a geographic and social space that can plausibly serve as a model for the early Bantu situation, at least to a degree that data from that region can improve the present state of our understanding of the development of languages of the family. These strategies have emphasized the importance of documenting the multilingual competences of individuals, their linguistic knowledge, and patterns of language use. They are broadly summarized, from a documentary linguistic perspective in Good (2022). Here, I would like to focus on a particular dataset coming out of work on Lower Fungom. These are individual-based wordlists that represent the available dataset that is most closely associated with the kind of data traditionally used to establish historical relationships.

This dataset is introduced in §4.2. In §5, it is considered from a holistic perspective with an emphasis on evidence for areality in the distribution of lexical

items found within the data. Two more detailed analyses are presented on lexical patterns found in specific varieties in §6.

4.2 Individual wordlist data from Lower Fungom

The wordlist data analyzed here was collected from each of the thirteen linguistic varieties of the Lower Fungom region of the Grassfields (Good et al. 2011, Di Carlo 2011). Each of these varieties is associated with a specific village with the same name as the variety. Some of these varieties are classified as belonging to the same language, but all are clearly distinct based on both scholarly linguistic criteria and local perceptions. The data collection process and the initial analysis of that data, including a summary of its overall properties, are discussed in detail in Good et al. (to appear). The present paper is the result of secondary analysis based on the results of the earlier study.

A map of the region is provided in Figure 1, where Lower Fungom itself is encircled with a dotted line. A listing of its languages, their associated village varieties, and the identifiers used for the individual-level wordlists collected on these varieties is provided in Table 1.[10] All of the region's languages are classified within the Bantoid group, though, as indicated in Table 1, they are not part of a single genealogical unit within Bantoid.[11] 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 these do not yet have close relatives established outside of Lower Fungom, and they are classified within the referential Yemne-Kimbi group. Within this group, the language associated with the villages of Mufu and Mundabli and the one associated with the village of Buu appear to form a small subgroup (see §6.3). The Missong variety of Mungbam should, perhaps, be considered a separate language, though its close connection to the other Mungbam varieties is clear. Four of the region's languages, Ajumbu, Fang, Koshin, and Kung, are restricted to a single village. Kung has been classified with the Central Ring group of languages and has close relatives outside of Lower Fungom. 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 partly indicated in Figure 1. (Some Naki-speaking villages fall outside of the map's area.)

[10]The wordlist doculect identifiers are designed to be unique. They begin with a speaker identifier consisting of three capital letters, followed by the name of the variety, followed by a single digit. Within the identifiers, the Mufu variety is written as Mumfu, a local variant spelling of the village name in English.

[11]This table is also found in Good et al. (to appear).

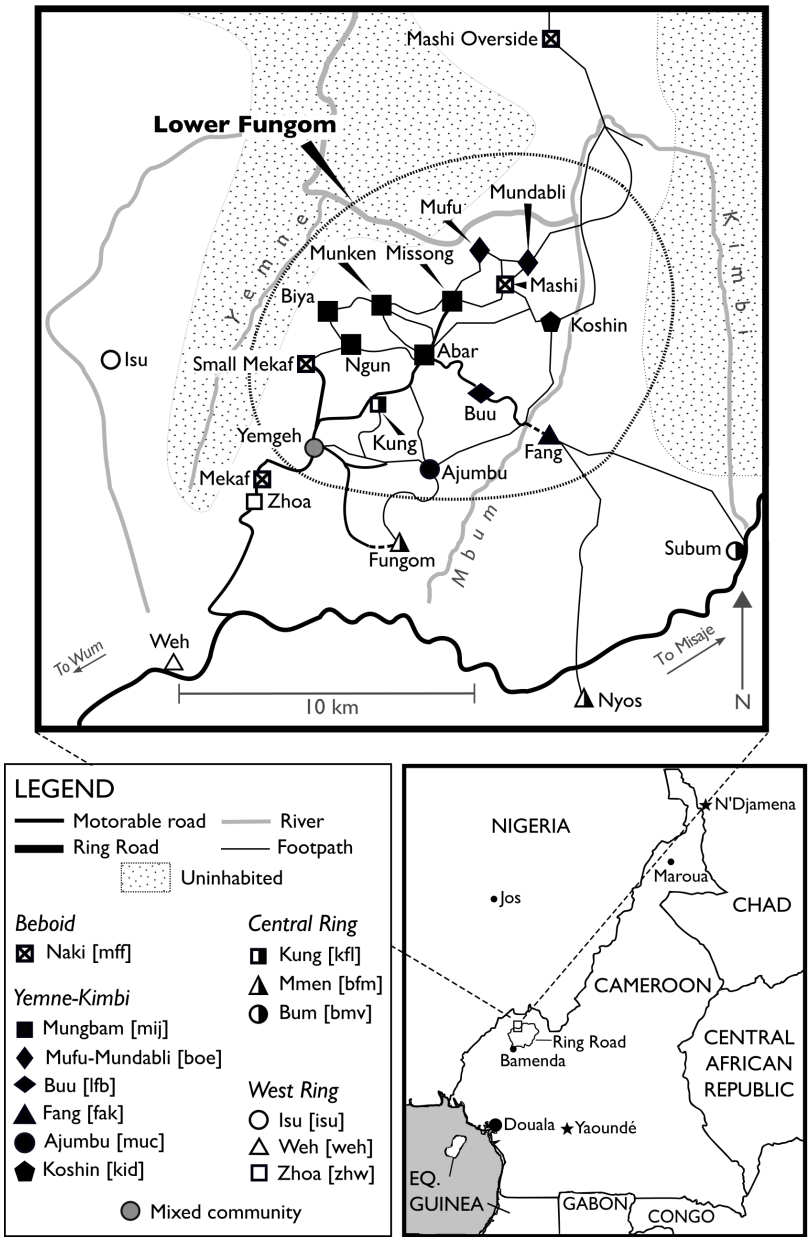

Figure 1: Lower Fungom and the surrounding region in context (map created by Pierpaolo Di Carlo)

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

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

The most distinctive characteristic of the dataset used in this paper is that multiple wordlists were collected from each of the Lower Fungom varieties from individual speakers who were not accompanied by other speakers during the elicitation process. Therefore, there was no attempt at convergence on a single form representing some kind of "consensus" on what word best corresponded to

a specific concept in a given variety. The same linguist, Nelson C. Tschonghongei, elicited and transcribed all the wordlists, limiting one potential source of variation in data collection. These wordlists will be treated as representing individual-level doculects in the analyses used here. The term variety will be reserved for the named linguistic varieties presented in Table 1, which, as mentioned above, are referred to using the names of villages that they are associated with (which is also how they are referred to in English by residents of Lower Fungom).

The present working dataset consists of more than 18,000 individual wordlist entries across fifty-three speakers. Verbs were collected for some speakers. However, only data on nouns was consistently collected for all speakers, and the data reported on here, therefore, only involves nouns. In addition, the data was filtered to only include concepts for which there was relatively high coverage across the wordlists. A cutoff was chosen where concepts were used if they had entries across at least forty of the fifty-three wordlists, resulting in a working dataset of around 7,000 total words. This step was taken to achieve a balance between ensuring there was decent coverage across all the concepts analyzed while also making use of enough data for robust patterns to emerge.[12]

Both singular and plural forms were collected for nouns (where semantically appropriate), in addition to variant forms provided by the speakers in some cases, but only one form was considered per speaker to simplify the analysis, namely the first transcribed form in an entry. 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 one variety, that of Koshin, due to the fact that the forms produced by one individual varied so extensively from those of all of the other speakers that it formed a kind of outlier.[13]

The collected forms were then entered into a database from handwritten notes by Charles Nyoh Abang and further processed using the CLDFBench framework

[12]Data and code used to produce the analyses used in this paper can be found in a Zenodo repository at https://doi.org/10.5281/zenodo.18364452. These analyses rely on additional data and code associated with Good et al. (to appear) available at https://doi.org/10.5281/zenodo.15814992

[13]Due to ongoing conflict in Cameroon referred to under the heading of the Anglophone Crisis (see Pommerolle & De Marie Heungoup 2017), 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, ideally involving working directly with speakers to obtain judgments about the different forms produced, but this is outside the scope of the present paper.

(Forkel & List 2020) to facilitate analysis using LingPy (List & Forkel 2021), a toolkit designed for historical linguistic analysis of wordlist data including the detection of cognate relationships. While some degree of semi-automated data cleaning was done using the tools provided by CLDFBench, individual forms have not been systematically checked by hand, and some errors almost certainly remain in the data. Nevertheless, the large size of the dataset should limit the extent to which these would significantly impact the results presented here. For data processing using LingPy, the work reported on in Hantgan & List (2022) was used 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. Because the analyses below emphasize the synchronic lexical similarity among contemporary Lower Fungom varieties, the vocabulary items treated as being "cognate" using the LingPy toolkit are referred to using the label *similarity sets* here rather than cognate sets, though, of course, many items in these similarity sets will clearly be cognates in the traditional sense.

In order to detect similarity sets, the Sound-Class-Based Phonetic Alignment (SCA) method of List (2012) was used, as implemented in LingPy. 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. Therefore, it was seen as the most appropriate method for understanding the synchronic areal linguistic situation of Lower Fungom. Further details regarding the analysis are provided in Good et al. (to appear). One important point to bear in mind when considering the results below is that, because the forms were not morphologically parsed—in particular to separate noun class prefixes from noun roots—in many cases, the similarity sets separated nouns into different similarity sets when they appeared with a noun class prefix on one set of varieties but without a prefix in other sets. This was treated as reasonable for this study to the extent that it represents a clearly notable synchronic difference in the realization of the nouns that can be used to differentiate varieties from each other. At the same time, it is clear that other approaches are possible, and the present work should be viewed as a preliminary means to analyze data of this kind from an areal perspective in the hopes that future work can refine the similarity measures in descriptively and theoretically motivated ways.[14] An ideal measure would be one that is based on an

[14]While it would have been possible to correct some of the results by hand (and this is done to a limited extent below), this would have led to potential biases relating to my own intuitions regarding comparative Bantu, which would not necessarily match the perceptions of members of the relevant language communities. For that reason, this was not attempted here, and the use of the LingPy methods was seen as a way to mitigate the potential for such bias.

understanding of how members of these language communities rate the perceptual similarity among different forms. However, I am not aware of any relevant research to make use of in this respect.

Obviously, the results presented here are limited to the extent that the data is limited (e.g., to nouns only) and that the automated processes that were used to compare the data have not been re-checked by hand or adapted specifically to this kind of dataset, which is different from datasets normally used in historical linguistic analysis. However, at the same time, no other comparable dataset is available, and it allows us to explore patterns of variation within and across varieties that are impossible using more traditional datasets. Good et al. (to appear) focuses on the individual-level variation within varieties found within the wordlists, finding, among other things, a relatively high range of variation within varieties. For example, the variety showing the least individual-level variation across the collected wordlists (Mashi) had around ninety-percent similarity, judged by the extent to which the collected forms were placed in the same similarity sets, while other varieties, such as Biya, Munken, and Kung) were just below eighty percent. Detailed results are provided in Good et al. (to appear).

For the purposes of this paper, however, individual-level variation within varieties is of less interest than how all of the collected varieties vary within the sociogeographical region of Lower Fungom, and the collected wordlists have two beneficial characteristics for the analyses to follow in §5 and §6. First, since each variety is represented by data from more than one speaker, we should be able to more reliably determine which lexical items are generally characteristic of that variety by, for example, focusing on those forms that were independently produced by all the sampled speakers.[15] Second, a dataset of this kind allows us to more reliably determine how clear the boundaries are between different linguistic varieties in lexical terms by providing a sample of lexical variation within a variety rather than just having a single set of lexical items representing each variety which would falsely imply a level of homogeneity that is not in line with actual language use.[16] While wordlist data, of course, does not represent naturalistic language use, an individual-based wordlist dataset of this kind is clearly

[15] An implicit assumption here is that the forms produced by a speaker working alone with a linguist are more representative of their typical patterns of usage than the forms that the forms that they might agree are "correct" in group elicitation, where social norms may cause them to defer to specific speakers. Testing this assumption will need to await further research.

[16] The selection of speakers for this process of data collection was opportunistic, i.e., working with individuals who were available for the research, rather than based on systematic principles. This could obviously impact the results. However, when dealing with small speaker communities of this kind, especially considering the fact that, in recent years, work can only continue with speakers displaced from their homes, due to the ongoing Anglophone crisis in

an improvement over having a wordlist from just one speaker or a "consensus" wordlist from a group of speakers in this regard (especially since the consensus is more likely to reflect social deference among members of the group as much as, if not more than, intuitions around actual patterns of usage).

5 Spatial and areal information in the dataset

5.1 An initial overview of the patterns in the data

Areal linguistic similarities can sometimes be partly accounted for in genealogical terms to the extent that genealogically related (and, therefore, lexicogrammatically similar) languages will often be spoken near to each other geographically. However, areal relationships are well known to transcend clear genealogical boundaries. In the present context, the existence of the large linguistic area of the Macro-Sudan Belt is of note since the Grassfields—and the broader area believed to contain the Bantu homeland—are contained within it. Moreover, there is evidence that the Grassfields and nearby areas form a smaller linguistic area within the Macro-Sudan belt, for example being a region characterized by the presence of a relatively high number of tone levels (Wedekind 1985:109) and the recent spread of numeral classifiers (Kiessling 2018). This is not surprising given the fact that the Grassfields have also formed a distinctive cultural area for a very long time, at least a millennium if not substantially longer (Warnier 1985:3).

While all of the languages of Lower Fungom can be classified as Bantoid, there is no evidence that they form a distinctive subgroup on their own, and there is good evidence that four varieties of Lower Fungom are relatively recent entrants to the region, namely Fang, Koshin, Kung, and Mashi (Di Carlo 2011). We can therefore analyze the similarities and differences among Lower Fungom's varieties not only in terms of shared inheritance but also in terms of possible areal patterning. The heatmap presented in Figure 2 provides an initial overview of the patterns of lexical similarity in the data, where both genealogical and apparent areal relationships are visible.[17] The overall grouping and ordering of doculects in the heatmap was calculated automatically using the default method in LingPy for creating a tree based on lexical distance data of this kind (see Hantgan & List 2022:16).[18] Individual-level doculects from the same variety are clustered to-

Cameroon (Pommerolle & De Marie Heungoup 2017), a more systematic sampling procedure was not practical.

[17] See Good et al. (to appear) for details on how the heatmap in Figure 2 was created.

[18] The tree is not included here since it cannot be considered a true genealogical analysis based on the nature of the input data. Its value is as a heuristic for grouping doculects by lexical similarity in a way that allows key relationships the be more readily visible.

gether using this method as are varieties known to be closely related (e.g., the varieties associated with the Mungbam language). The grouping of languages with respect to each other does not have as clear an interpretation, in particular since the languages being analyzed do not form a genealogical unit within Bantoid but are selected due to their location in a linguistic area.

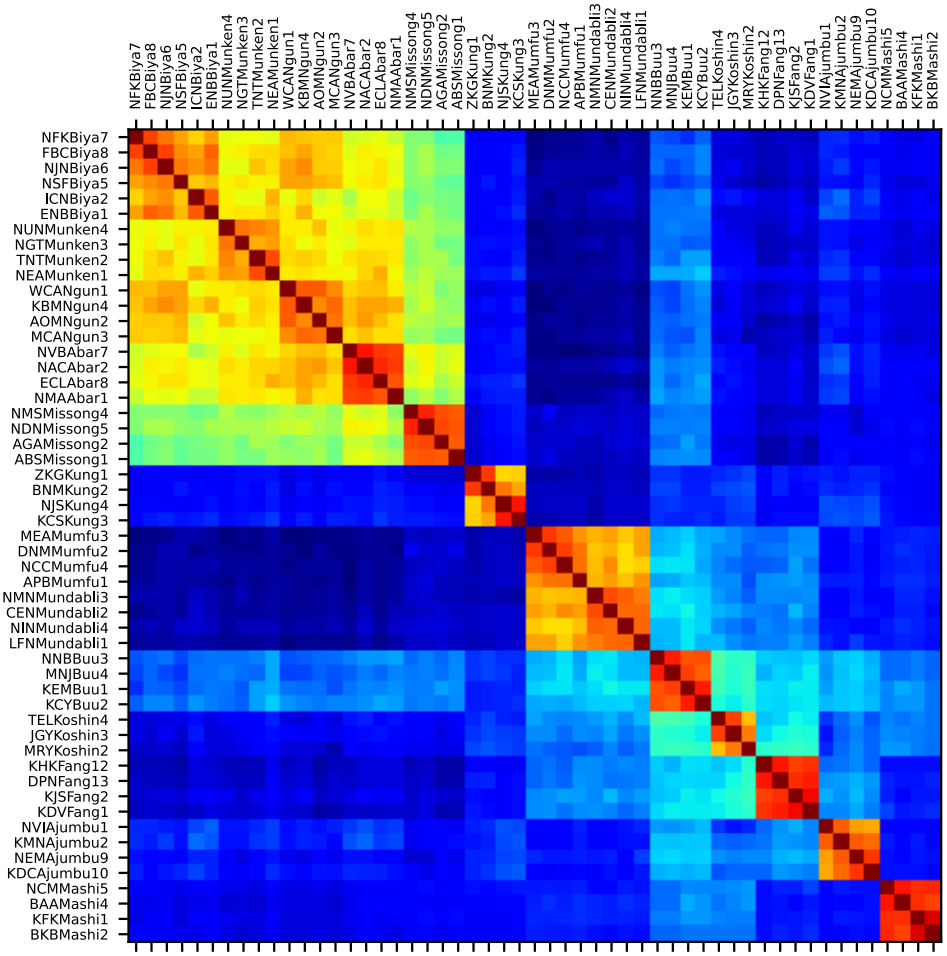

Figure 2: Lexical Similarity Heatmap: This heatmap represents the lexical overlap across wordlists, with warmer colors showing greater similarity and cooler colors indicating less similarity. Each wordlist is indicated with a unique identifier representing the speaker (consisting of the first three letters) followed by the variety name (see Table 1). The figure was created using the LingPy toolkit (List & Forkel 2021).

The heatmap in Figure 2 differs most strikingly from heatmaps generally used in comparative lexical studies with respect to the presence of the series of easily visible red-orange squares along the northwest–southeast diagonal line of the similarity matrix (see, e.g., Hantgan & List 2022:16 for a relevant comparison). These correspond to the collected doculects of the same variety, i.e., the collected wordlists across each of the villages of Lower Fungom, namely: the five Mungbam varieties (Abar, Biya, Missong, Munken, and Ngun), the two closely related varieties of Mufu and Mundabli, and the six other varieties of Ajumbu, Buu, Fang, Koshin, Kung, and Mashi. Four of the varieties of Mungbam—Abar, Biya, Munken, and Ngun—form a kind of "supercluster" in orange and yellow in the upper left portion of the heatmap that also shows a relatively close relationship to Missong variety of Mungbam (seen in greenish colors). Missong was previously known to be the most divergent of the Mungbam varieties (see, e.g., Lovegren 2013). The four varieties of Mungbam not including Missong will be referred to as Mungba here following the name for this group given in Glottolog (Hammarström et al. 2023).[19] Mufu and Mundabli also form a large cluster around the center of the heatmap. The rest of the varieties show more distinctive clusters.

Other relationships are visible in the heatmap that are most readily interpreted as being the result of contact. For example, there is a light blue area linking the doculects for Buu, Fang, and Koshin in the lower right section of the heatmp. There is no known evidence for a close genealogical connection among these varieties, but, as indicated in Figure 1, they are all close to each other geographically in the southeastern area of Lower Fungom. Further discussion of Buu's relationships is provided in §6.3.

5.2 Revealing the spatial information in the data

The fact that the heatmap in Figure 2 provides an overview of lexical similarity among the varieties of Lower Fungom makes it useful for informally distinguishing genealogical relationships from potential contact relationships. However, in order to examine this systematically, it is useful to consider other analytical methods. A pattern that emerges in the data using additional kinds of data visualization and clustering methods is that there is a clear spatial signal in the lexical data. As a first attempt to visualize this, Figure 3 presents variation in the lexical data

[19]The name Mungbam is used only for scholarly linguistic work since these five varieties are not recognized as linguistically unified locally. It is a quasi-acronym based on the names of the villages in the order: Munken, Ngun, Biya, Abar, and Missong. *Mungba* is formed by removing the final *m* associated with the village of Missong.

in a two-dimensional space using multidimensional scaling.[20] Multidimensional scaling is a data visualization technique that allows the similarities (or differences) among the measured values for a set of items to be represented using a small number of dimensions in a way that allows items that are more similar to each other overall to appear closer together than those that are more different from each other. There is necessarily some loss of information as a result of the transformation of the data into a low-dimensional representation, but this is offset by the ways that it allows major patterns in the data to become straightforwardly visible. In the case of the lexical similarity data of interest here, it can be used as a means to detect linguistically interest patterns of clustering among the doculects.[21] The data is reduced to two dimensions here since this allows for a straightforward spatial interpretation of the data in terms of north–south and east–west dimensions.

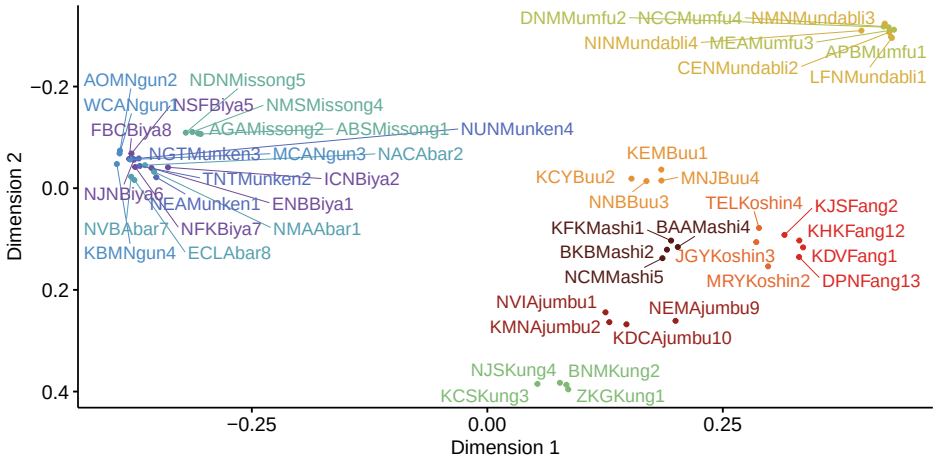

Figure 3: Multidimensional scaling of the lexical similarities among collected Lower Fungom doculects, presenting them in a two-dimensional space. Doculects collected from individuals from the same village variety appear in the same color.

[20] All figures in the remainder of this paper were produced using R (R Core Team 2023), making use of various additional libraries, including cluster (Maechler et al. 2022), factoextra (Kassambara & Mundt 2020), ggfortify (Tang et al. 2016), ggplot2 (Wickham 2016), ggpubr (Kassambara 2023), and khroma (Frerebeau 2023). The core method used to produce Figure 3 was cmdscale() as implemented in the R Stats Package.

[21] See Croft & Poole (2008) for a detailed discussion of multidmensional scaling from a linguistic perspective. Croft & Poole (2008:14) also contrast multidimensional scaling with principal component analysis, another technique for reducing the dimensionality of datasets which will be used in §5.3.

An immediate, and unsurprising, pattern that can be seen in the Figure 3 is that the points associated with each individual-level doculect cluster with the other doculects of the same variety. The situation is slightly more complicated for the Mungbam varieties (upper left) and Mufu–Mundabli (upper right). In those cases, there are not distinct clusters for each variety, given the high level of lexical similarity within those groups. At the same time, since Figure 3 is based entirely on lexical similarity data, the correspondence between the locations of the individual-level varieties and the geographic layout of the Lower Fungom villages is quite striking. The first dimension (corresponding to the x-axis in the figure) shows a strong match with the east–west positioning of the villages and the second dimension (corresponding to the y-axis in the figure) shows a strong match with the north–south positioning of the villages. (See Figure 1 for the actual geographic positioning of the villages.) There is some ambiguity in the possible information encoded in the x-axis, however, in that it likely represents a mix of spatial information as well as distance from Mungbam specifically.

Two of the individual doculect clusters do not correspond well to the spatial layout of Lower Fungom's villages. These are Kung, which appears much lower than the other clusters—especially below Ajumbu, even though Kung is north of Ajumbu—and Buu, which is positioned higher on the y-axis compared to its actual location. Kung's position in Figure 3 aligns with its recent history: its residents migrated into Lower Fungom from the south. Kung is genealogically linked to Central Ring languages spoken south of Lower Fungom, including in the village of Fungom, where Mmen is spoken, just outside the region's southern border (Di Carlo 2011:78; Yasho 2016:4–6). Put differently, the lexical distance between Kung and the other villages is in line with the migration trajectory of its speakers, if not its present-day location.

The case of Buu is more directly connected to the language dynamics of Lower Fungom itself and will be discussed in more detail in §6.3. In broad terms, its position in Figure 3 appears to reflect the fact that it has a genealogical connection to Mufu–Mundabli while having contact relationships with the Abar variety of Mungbam as well as Fang and Koshin. This causes its position in Figure 3 to deviate from what would be expected in purely spatial terms.

5.3 Looking for areality via a cluster analysis of the data

We can gain more insight into the areal and contact patterns that multidimensional scaling suggests are present in the lexical data by performing cluster analysis on it. This is shown in Figures 5–10, which provides the results of looking for between two, three, five, six, and thirteen clusters in the dataset using *k*-medoids

clustering.[22] This method is described in more detail in a linguistic context in Syrjänen et al. (2016) where it is shown to be successful in detecting dialect divisions in Finnish. The present study is comparable insofar as the goal is to look for clusters of varieties within a contemporary linguistic ecology rather than, for example, being focused on genealogical reconstruction.

When applying cluster methods, it is helpful to verify the overall cohesion of the clusters (i.e., the extent to which items in a cluster are more similar to each other than items in other clusters). This can involve finding the optimum number of clusters for the data, for example. However, in this case, the data will be considered in terms of the ways it is clustered when broken down into different numbers of clusters. In Figure 4, scores associated with grouping the data into different numbers of clusters are presented for clusters ranging from two to 25.[23] This provides a general means of assessing the validity of the clusters at different numbers of partitions.

A higher average silhouette width is interpreted to mean that the clustering is better than when there is a lower average. The range of clusters of interest here is from two to thirteen clusters given that there are thirteen distinct varieties under consideration. Within this range, the highest average silhouette width is found when the data is divided into seven clusters, but the extent of variation is relatively narrow, going from a maximum of 0.70 (for seven clusters) to a minimum of 0.54 (for three clusters), with a mean across two to thirteen clusters of 0.61. As seen in Figure 4, the average silhouette widths show a downward trend as the clusters increase past thirteen, and this trend continues up to the maximum possible number of clusters—which is 52, i.e., one less than the total number of datapoints—even though, for readability, the figure stops at 25 clusters. I take the relatively narrow range of the scores between two and thirteen clusters to mean that it is reasonable to compare the patterns of clustering within that range and the fact that all of the scores are above 0.50 within that range to mean that

[22] The figures were produced using the pam() function as implemented in the cluster package in R (Maechler et al. 2022). The clusters are presented on plots where the axes represent the values of the first and second principle components, on the x-axis and y-axis respectively, using the prcomp() function from the R Stats Package as rendered via the autoplot() function implemented in ggplot2 (Wickham 2016). This produces a layout that is comparable to the one in Figure 3, though somewhat different in the details since the two methods are not equivalent, with the principal component plots presenting less of the information about the overall structure of the dataset than the multidimensional scaling plots. However, since the first two principal components account for more than 75% of the variation in the data, plots based on them are still representing important patterns in the data.

[23] These scores are calculated using the silhouette width score as implemented in the fviz_nbclust() method in Kassambara & Mundt (2020).

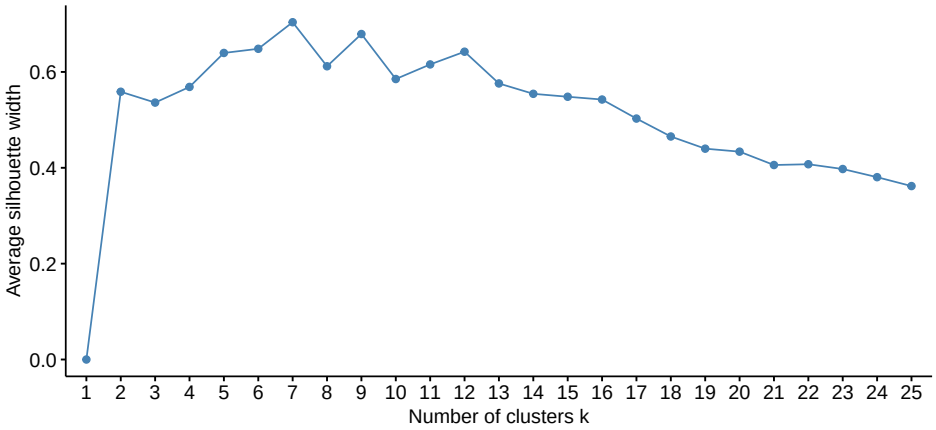

Figure 4: Average silhouette width scores across different numbers of clusters produced via *k*-medoids partitioning

genuine patterns are being detected in the data. I start by presenting the results for the detection of thirteen clusters in Figure 5, since it serves to validate the approach.

Given that there are thirteen varieties documented in the data, this provides an initial means to assess the validity of the detected clusters. Dividing the data into thirteen clusters using this method as seen in Figure 5, in fact, results in a perfect alignment between the thirteen clusters and the thirteen language varieties. This suggests that the clusters detected via this method are likely to be valid, or, at the very least, to represent reasonable hypotheses of significant clusters meriting further investigation.

It is important to bear in mind that Figure 4 is only showing the average scoring of the clusters for a given number of partitions. There will be variation among the clusters as well. In the data presented below, clusters representing known genealogical groupings (and, in particular, doculects from the same variety) show higher scores than those that represent apparent areal groupings. Assuming the general approach adopted here is valid, this represents a possible metric to be used for separting areal from genealogical clusters. This will be discussed below in this section when the division of the data into five clusters is discussed (see Figure 8 and Figure 9).

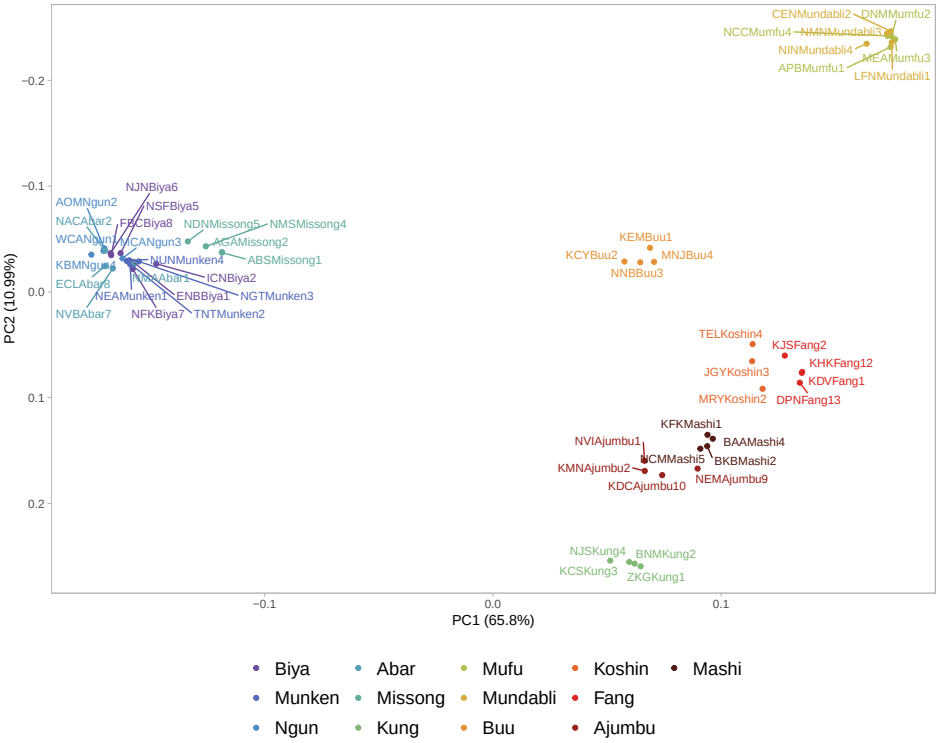

Figure 5: Division of the wordlist data into thirteen clusters based on similarity set distances. Each cluster corresponds to the varieties of each of Lower Fungom's thirteen villages.

A complete summary of the patterns detected when clustering the data from two to thirteen clusters is presented in Table 2. Illustrative figures depicting the detected clusters are presented below in Figures 6, 7, 8, and 10.[24]

The division of the data into two clusters in Figure 6 separates the Mungbam languages from all of the rest, which is not especially surprising given the closeness of those varieties to each other when set against the rest of Lower Fungom's varieties. The Mungbam group remains a coherent cluster until the data is divided into eight clusters, when its internal divisions first become apparent. The division of the data into three clusters in Figure 7 adds a cluster for the close genealogical grouping of Mufu-Mundabli. This cluster remains coherent until the data is divided into thirteen clusters in Figure 5, with the division between the

[24]The results of clustering the data into numbers of clusters between two and thirteen not presented in this section can be found in the supplementary materials.

Table 2: Patterns found when grouping data into 2–13 clusters

NUM	NOTABLE CLUSTERS	INTERPRETATION
2	Mungbam doculects grouped	Genealogical grouping of Mungbam detected
3	Mufu-Mundabli doculects grouped	Genealogical grouping of Mufu-Mundabli detected
4	Kung doculects grouped	Distinctive genealogical position of Kung detected
5	Mashi doculects grouped	Distinctive genealogical position of Mashi detected, potential southeast areal grouping identified
6	Koshin doculects split across clusters including Buu and Fang doculects respectively	Interpretation unclear, but lexical patterns of specific individuals worth investigating further
7	Ajumbu doculects grouped	Distinctive genealogical position of Ajumbu detected, possible core of southeast areal grouping identified (Buu, Fang, and Koshin)
8	Missong doculects grouped	Distinctive position of Missong within Mungbam detected
9	Buu, Fang, and Koshin doculects cleanly grouped	Earlier split pattern lost, but interpretation is not clear
10	Mungba divided into clusters of mixed varieties	Interpretation unclear, but revealing lexical patterns of specific individuals worth investigating further
11	Munken doculects grouped, other Mungbam doculects still in mixed clusters	Mungbam grouping becomes more consistent with expectations, but continued presence of mixed groups suggests need for further investigation of specific individuals
12	Mungba doculects all grouped by their variety	Known genealogical patterns in Mungbam cleanly emerge
13	Mufu and Mundabli split into two clusters and all 53 doculects are grouped perfectly by variety	Establishes validity of approach by aligning with ground truth

Mufu and Mundabli varieties the last to be detected. As summarized in Table 2, the division of the data into four clusters results in the identification of the Kung doculects as a distinct cluster, and its division into five clusters in Figure 8 results in the Mashi doculects identified as a distinct cluster.

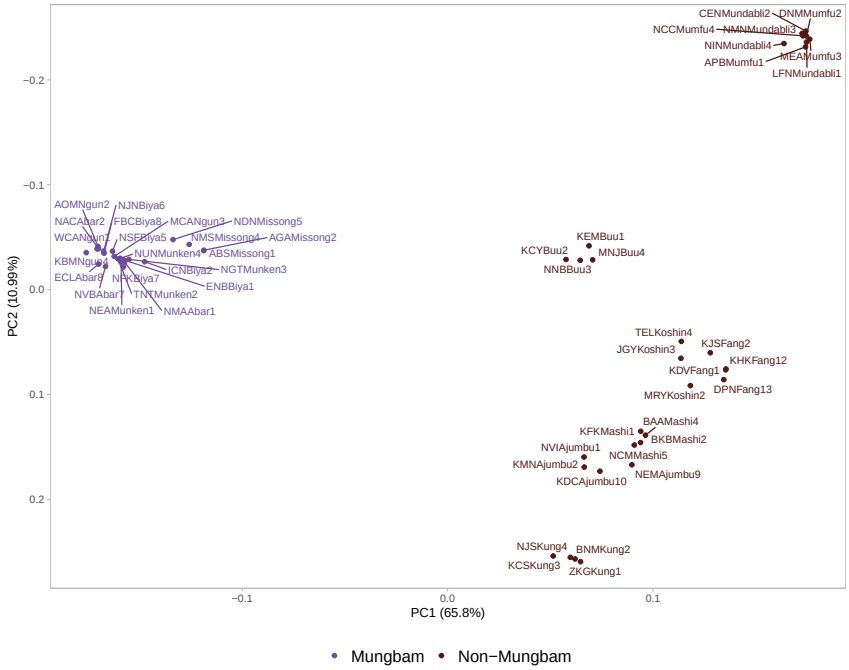

Figure 6: Division of the wordlist data into two clusters

These divisions of the data up to five clusters can be interpreted in terms of two clear genealogical groupings (Mungbam and Mufu–Mundabli) being identified followed by the identification of two varieties (Kung and Mashi) which are clearly lexically distinct from other varieties (as seen in Figure 2) and which are also known to represent relatively new entrants to Lower Fungom. Kung's presence in Lower Fungom is the result of a movement of speakers of a Central Ring variety, as discussed earlier in §5.2, and Mashi's presence results from a movement of speakers of the Naki language complex, most of whose communities are found outside of Lower Fungom (Di Carlo 2011:78–79). At the same time, when the data is divided into five clusters, the fifth cluster is a potential candidate for an areally based cluster, consisting of Ajumbu, Buu, Fang, and Koshin, all of which are found in the southeast of Lower Fungom. The silhouette widths for the doculects placed in this cluster are notably lower than those of the other

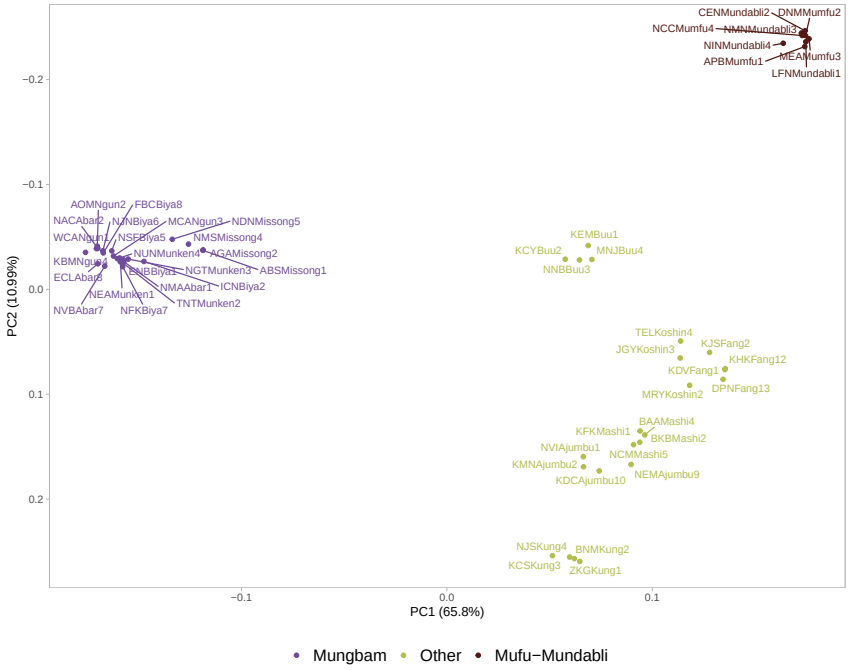

Figure 7: Division of the wordlist data into three clusters

clusters, which suggests that this is a less coherent cluster than the others, but this would not be unexpected for an areal grouping.[25] The average silhouette width across the genealogical clusters in Figure 8 is 0.65, while the average score within the Other cluster is 0.22. The fact that Mashi is included with these varieties when the data is collected into four clusters perhaps also suggests that it is part of this larger areal cluster as well, though, if so, it would seem to participate more weakly in this areal patterning.

The division of the data into six clusters in Figure 10 is largely similar to the division into five clusters with the addition of a cluster corresponding to the four Ajumbu doculects. The Buu, Koshin, and Fang doculects remain grouped together further suggesting an areal relationship among these three varieties which is seemingly stronger than the connection between the three of these varieties plus Ajumbu detected in Figure 8.

Division of the data into seven clusters represents the first case where the doculects of a single variety are divided across clusters. Two Koshin doculects

[25] Silhouette widths within clusters were examined using the fviz_silhouette() method implemented in Kassambara & Mundt (2020).

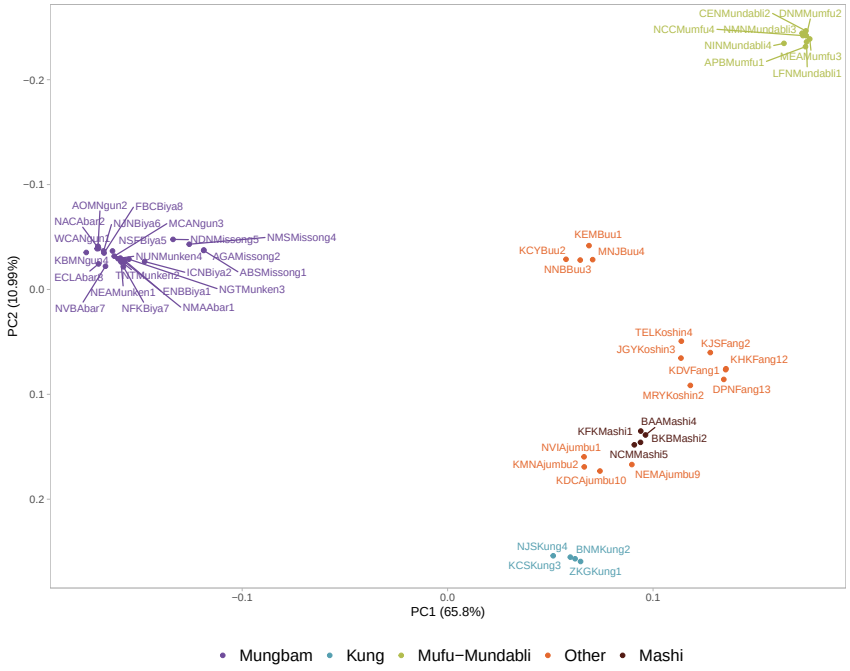

Figure 8: Division of the wordlist data into five clusters

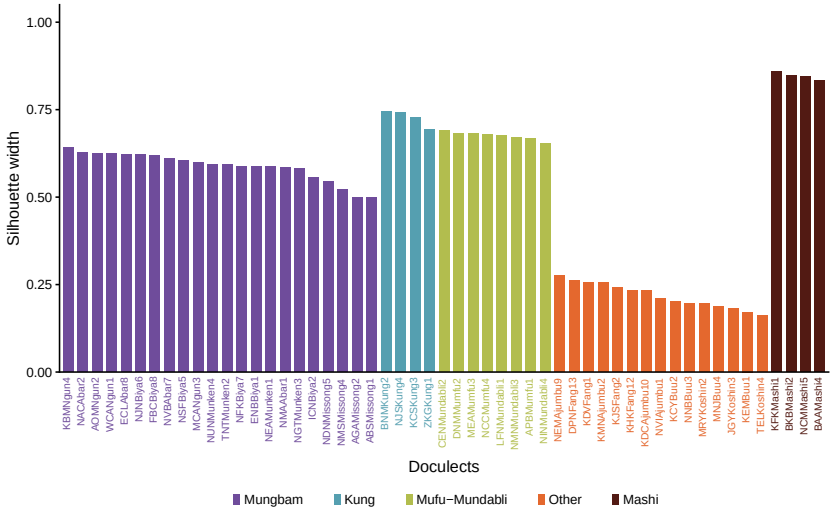

Figure 9: Silhouette widths for doculects grouped into five clusters

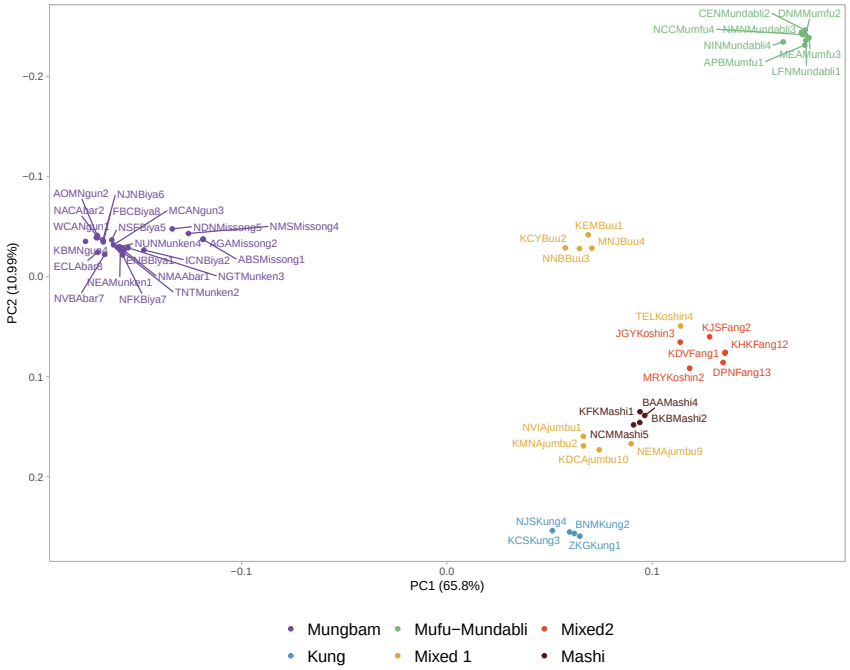

Figure 10: Division of the wordlist data into six clusters

are grouped with the Fang doculects, and one Koshin doculect is grouped with the Buu doculects. This raises interesting questions with respect to individual-based variation and potential contacts relationships, though these are outside of the scope of the present paper and further examination of these mixed clusters will need to await further study. Division of the data into eight clusters is similar to the division into seven clusters except that this is the first point at which internal diversification within Munbgam is detected, and, in particular, the Missong doculects are separated from the other Mungbam varieties, which are labeled Mungba here (see §5.1). Division of the data into nine clusters removes the first set of mixed groups, and results in a "clean" division between the Buu varieties, the Fang varieties, and the Koshin varieties.

Division of the data into ten and eleven clusters, results in further division of the Mungba varieties in ways that again mix doculects associated with different varieties. This suggests more detailed investigation of individual-level variation and areal patterning within the Mungba group is warranted, but, for reasons of scope and space, they will not be further considered here. Finally, division of the data into twelve clusters results in the five Mungbam varieties each falling into

their own cluster. The only remaining larger cluster is Mufu-Mundabli which, as discussed above, is separated into its two constituent varieties when the data is divided into thirteen clusters, as seen in Figure 5.

5.4 Interpreting the results of the cluster analysis

What emerges from this examination of the clustering of the lexical data is that, not unexpectedly, clearly distinct genealogical groupings (e.g., Mungbam, Mufu–Mundabli, and Kung) appear fairly quickly as the number of clusters increases, but, at the same time, this method can help detect additional clusters that are indicative of other historical relationships, such as a contact relationship among varieties such as Buu, Fang, and Koshin in the southeast of Lower Fungom. More broadly, some kind of contact relationship based on areal proximity seems likely to be relevant to understanding the distribution of lexical items across the entire region given the close overall relationship between the lexical similarity data and the spatial layout of the thirteen Lower Fungom villages, as seen in Figure 3. While, strictly speaking, the clustering analysis did not require the use of individual-based wordlist data, as opposed to adopting a more standard approach where each locally salient variety was represented by a single wordlist, the presence of multiple wordlists from each variety served as a useful way to validate the methods insofar as the clustering, on the whole, detected groupings of doculects from the same variety, providing some assurance that the other clusters represented real linguistic patterns rather than being mere artifacts of the application of a statistical method.

Returning to the higher-level concerns that are the focus of this paper, these results show us that the analysis of genealogical relationships and areal relationships can be conducted simultaneously on the same lexical dataset.[26] Moreover, prioritizing genealogical relationships as somehow "primary", comes at the risk of missing important historical generalizations including the possibility that there is a spatial dimension to the distribution of lexical diversity and the importance of looking for contact clusters within a dataset. If these patterns in the Lower Fungom data are indicative of the early Bantu situation, then this also means that lexical studies of the Bantu expansion should assume that the lexical

[26] It is possible to reach this conclusion here precisely because the genealogical relationships were already known in advance, allowing the results of the analysis to be validated against ground truth. In principle, the same methods could be extended to cases where these relationships are not known, though, of course, the results seen here should ideally be further verified to ensure that they are broadly reliable since it is possible that they will not generalize well beyond Lower Fungom.

data is strongly influenced by low-level spatial factors and that there were small areas of significant language contact impacting lexical patterns as the family diversified and spread. Of course, this proposal must be taken with caution given the largely exploratory nature of this study and the fact that it relies on analysis of individual-based doculects whose representativeness of patterns of variation of larger areas has not yet been established.

These observations are directly connected to two aspects of the sociolinguistic reconstruction of early Bantu communities presented in §3, namely that they were relatively small (1a) and characterized by high degrees of individual-level multilingualism (1b). Both of these would facilitate spatially-conditioned patterns of lexical distribution where multilingual individuals would serve as the conduit for lexical variation to enter a community, and the small size of communities would facilitate the spread of new norms.

While outside of the scope of the present paper, a more detailed study of lexical patterning of specific individuals in these communities is also likely to yield more fine-grained insights about how specific lexical items are contributing to the detection of genealogical vs. areal clusters. Analyses moving partly in that direction will be presented in §6 which looks at the distinctive lexical items in specific varieties in the context of understanding dynamics of language change in Lower Fungom.

On some level, the general conclusions reached here are not especially surprising. It is well known that language relationships are shaped by genealogy and contact. However, the key contribution of the studies above is to demonstrate how the use of certain quantitative methods to analyze a dataset like the one of interest here can provide a high degree of precision with respect to claims of areality and contact so that they can be more directly tested and analyzed rather than being subordinated to the analysis of genealogical relations.[27]

6 Considering lexical patterns in more detail

6.1 Lexical divergence in Missong and lexical overlap in Buu

The results presented in §5 looked at overall patterns of lexical similarity without considering the specific lexical items that contributed to similarity or dissimilarity among varieties. In this section, the dataset will be explored in more detail

[27] Historical analysis making use of phylogenetic networks can involve simultaneous exploration of patterns of diversification involving significant contact against those that do not involve contact (see, e.g., Heggarty et al. 2010:3834). However, such approaches do not also directly bring in a spatial element to the analysis, as has been the case here.

with a focus on patterns of lexical divergence and convergence for two varieties, Missong and Buu, each of which is chosen because of previous work suggesting that they have distinctive contact histories within Lower Fungom. Missong represents a relatively newly constructed variety, and Buu appears to have an old, presumably genealogical, lexical connection to Mufu-Mundabli and more recent connections to Mungbam, resulting in a lexicon with contributions from both groups.

While these studies will not examine the data at an individual level, the information in the individual wordlists will be used as a means to achieve greater reliability of the results insofar as lexical overlap among varieties will only be considered in cases where all of the collected individual doculects of a variety were determined to have forms belonging to the same similarity set for a given concept. This should help ensure that the lexical items considered are truly representative of an entire variety rather than being idiosyncrasies of the lexical choices of only one individual.

The case of Missong will be discussed in §6.2, and the case of Buu will be discussed in §6.3.

6.2 The distinctive lexical features of Missong

As discussed in Di Carlo & Good (2014:243–246), ethnographic and linguistic evidence suggests that the distinctiveness of the Missong variety of Mungbam, as is clearly visible in Figure 2, is not due to simple "drift". Rather, it results from a historical process where the formation of the Missong village was accompanied by the deliberate formation of a distinct linguistic variety. This variety was constructed with a primarily Mungbam base. Other linguistic material was introduced to ensure that Missong was clearly different from the other Mungbam varieties. More details about Missong's divergence can be found in (Good 2020a:51–53). It manifests not only in lexical distinctiveness but also idiosyncratic differences in verbal morphology and a second-person pronoun form that is not cognate with the equivalent forms of the other varieties. Missong also uses a reflex of Proto-Bantu Class 7 *ki- for a set of nouns that can be linked to Proto-Bantu Class 12 *ka- in the other Mungbam varieties. Further details are provided in Di Carlo & Good (2023:93–110), and Maho (1999:247) provides a relevant general overview of reconstructions of the Proto-Bantu noun classes.

The current dataset allows us to build on this earlier work to extract a list of specific nouns where Missong diverges from the other Mungbam varieties, and

the results of this are presented in Table 3.[28] In the table, informal "consensus" forms are presented for the relevant varieties.[29]

Table 3: Lexical divergence between Missong and the other varieties of Mungbam in the dataset, including comparison forms from other Lower Fungom varieties whose roots are similar to those found in Missong, where applicable (indicated in the COMP column).

GLOSS	MUNGBA	MISSONG	COMP	SOURCE
'bird'	*-ɲÁ(NÁ)*	*kímwá*	—	—
'cup'	*-b(w)Âm*	*ŋwɛ̀*	—	—
'deity, god, God'	*kə̀yì*	*kìgbə̄m*	*gbə̄m*	Mufu-Mundabli
'fish'	*ìSu*	*ìʃàbə̀yə̌*	*bîyə̌ŋ*	Buu
'hoe'	*-dɔ̂ŋ*	*ŋkwō̤m*	*ŋkɔ̄m*	Mufu-Mundabli
'horse'	*-ŋkéɲām*	*kìkwʊ́m*	*-kwÓm*	widespread
'pig'	*ìWɔ̀ŋ*	*kìfí*	*kə̀fí*	widespread
'tree'	*-tV*	*ukwa*	*fəkâʔ*	Kung

The data in Table 3 presents cases where: (i) all four Missong varieties showed forms which were grouped together into a single similarity set, abstracting away from some low-level variation, (ii) the forms of all of the Mungba varieties for which data was collected for the relevant concept were all grouped into the same similarity set as each other but a different one from Missong, and (iii) forms from other Lower Fungom varieties that were placed into the same similarity set as the Missong forms, where applicable. In those cases where the Missong form was grouped into the same similarity set as another group of non-Mungbam Lower Fungom varieties, this is indicated in the table. In two cases, the forms found in Missong but not Mungba were associated with a number of other Lower Fungom varieties, namely for 'horse', where the Missong forms overlapped with forms found in Buu, Fang, and Koshin, and for 'pig', where the Missong forms overlapped with forms found in Koshin, Mashi, Mufu, and Mundabli.

[28] The forms for 'cup' were not all grouped into the same similarity set due to the fact that, in one form, tone was not transcribed for the initial nasal, while it was for the other three. This was noted via manual inspection, which is why the form is included here.

[29] In these consensus forms, an *N* represents a nasal of variable place in the relevant varieties, an *A* represents a non-high vowel, *S* represents an *s* or an *ʃ*, *O* represents a non-low back vowel, *W* represents variation between a *w* and a *g*, and *V* represent a vowel of varying qualities. Stems beginning with a hyphen represent cases where the noun class prefix varied significantly across varieties. Vowels on which tone is not marked represent cases where a consensus tone could not be easily determined.

The same basic procedure was applied to the other Mungbam varieties, where the doculects from one variety were compared with the remaining four varieties (including Missong) to see if similar patterns of divergence were found. The pattern seen in Missong was not replicated. No other Mungbam variety showed more than one form that diverged from the other varieties, against the eight divergent forms found in Missong. This underscores Missong's distinctiveness.

Returning to the specific Missong forms in Table 3, two general points stand out. The first is that the divergent words are in relatively basic vocabulary and do not, for example, represent new concepts which might be expected to be associated with recent borrowings. The second is their non-systematic overlap with other forms found in Lower Fungom, indicating that these divergences are not due to a single borrowing event from one nearby variety but, rather, must have entered the language in some other way. Building on ideas of Mufwene (2001), Di Carlo & Good (2023) provide an account for patterns like this in terms of the presence of a pool of variant forms in a multilingual setting like Lower Fungom that are generally available for the construction of social meaning and, in particular, for the deliberate construction of similarity and difference between one language community and another.

That study did not consider the Missong vocabulary in constructing its arguments. However, the data in Table 3 lends further support to them. It suggests that the formation of Missong's distinctiveness involved changes to its vocabulary that drew on lexical variants for common concepts present in the multilingual setting of Lower Fungom. This provided an additional means, beyond grammatical differences, to distance Missong other Mungbam varieties. With respect to the Missong words for 'bird' and 'cup', while the Missong forms are not presently attested within other Lower Fungom varieties, without doing an extensive lexical survey of languages in the general region outside of Lower Fungom, it is not possible to say at this point whether they represent truly distinctive Missong formations or are shared with other languages of the area not examined during the course of this study.

Given the nature of the meanings in Table 3, the data in the table is consistent with the idea that the Missong lexicon has been impacted by a degree of lexical "mixing" which would be expected to be impossible under conditions associated with so-called normal language transition (see §2.1). While the list of words in Table 3 is relatively small in its absolute number, it should be kept in mind that: (i) the procedure for identifying these words was a strict one where all of the non-Missong forms had to belong to the same similarity set and all of the Missong doculects had to use roots drawn from a single other similarity set and (ii) a limited set of vocabulary items was considered, with around 130 words on average

per variety. If the only data available suggesting that the formation of Missong involved some form of deliberate change were the presence of the forms in Table 3, that would clearly be relatively weak evidence. However, when data from the individual-based wordlists is added to the previously collected evidence, both linguistic and ethnographic, the case for this becomes increasingly compelling and provides concrete data related to the idea that the "persistent hypotheses of language mixture" that were unequivocably dismissed by Welmers (1973:2) do indeed merit careful consideration in the kinds of multilingual sociolinguistic settings that characterize any parts of West Africa as described in §3.[30]

6.3 Lexical overlap: Buu, Abar, and Mufu-Mundabli

The lexicon of the Buu language offers another useful case study for the application of the dataset to questions of the historical development of languages in the kinds of contexts assumed to represent the Proto-Bantu situation as summarized in §3. Buu has been genealogically classified together with Mufu and Mundabli due to a combination of lexical and grammatical similarities. On the grammatical side, for example, there are clear similarities between the Mundabli and Buu noun class systems that are strongly indicative of a close genealogical connection, including some idiosyncratic singular/plural alternations whose presence would not be expected to result from common patterns of borrowing (see, e.g., Good 2023:394–397 and Ngako Yonga 2013:108–114). However, the lexical data is somewhat equivocal since Buu shows strong lexical connections to both Mufu–Mundabli and Mungbam, as first explicitly observed (Hamm et al. 2002:12).[31]

Arriving at a definitive genealogical classification of Buu (if this is even a sensible goal) is outside the scope of the present paper. However, using the current wordlist database, we can now readily examine the lexical overlap among nouns in Mungbam, Buu, and Mufu–Mundabli. To do this, it is useful to think of the similarity sets as constructing a network of related words encoding specific concepts in the Lower Fungom region. This is depicted visually in Figure 11, where each of the doculects is linked to the other doculects by a line whose thickness and

[30] While Welmers (1973:3–4) acknowledges the role of contact in language change, in particular with respect to its impact on a language's vocabulary, he explicitly adopts a very narrow view of "normal" language change which would not encompass the historical scenario proposed for Missong here.

[31] The study of Botne (2021:29) of tense-mood-aspect encoding in Yemne-Kimbi languages of Lower Fungom, a referential grouping covering all of the varieties discussed here other than Kung and Mashi, did not find evidence of a close connection between Buu and Mundabli. Mufu is not as well described as Mundabli, which is why it is not considered in as much detail with respect to potential grammatical parallels to Buu.

darkness is directly related to the number of their lexical items that are grouped into the same similarity set.[32]

As can be seen in the Figure 11, doculects associated with the same variety or closely related varieties form relatively dense networks, while the connections between some pairs of doculects with relatively few words in the same similarity sets are much lighter, to the point of being barely visible on a casual examination.[33] Overall, this network representation provides a visual snapshot of the ways in which the doculects are lexically connected to each other, and is especially helpful for uncovering details in the relationships of doculects that are not from the same variety.

In the case of the Buu doculects, as expected, they show relatively strong lexical connections to both Mungbam varieties and Mufu-Mundabli, but another pattern is also evident that might have been difficult to otherwise detect, namely the relatively weak lexical relationships between Mungbam varieties and Mufu-Mundabli, suggesting that the lexical connections between Buu and Mungbam, on the one hand, and Buu and Mufu–Mundabli, on the other, are relatively distinct in nature, rather than primarily representing forms that are broadly similar across Lower Fungom. In Figure 12, this pattern is highlighted by extracting and emphasizing the part of the network connecting the varieties of Abar, Buu, and Mufu-Mundabli.[34]

[32]The networks in Figure 11 and Figure 12 were created using the R packages igraph (Csárdi et al. 2023) and ggnetwork (Briatte 2023). As is the case with Figure 3, the layout of the nodes is based on a multidimensional scaling analysis of the similarities among the varieties. In the case of Figure 11 and Figure 12 the specific similarity metric used was the inverse of the square root of the number of lexical items belonging to the same similarity set across each pair of doculects. This resulted in a similar overall layout to what was seen in Figure 3, with minor differences, the most notable of which is the fact that Buu and Mashi doculects in Figure 11 and Figure 12 cluster more closely together. The specific metric chosen was determined via trial and error, but appears to partly replicate results from dialectometry where a different formula with similar properties has been proposed to describe the relationship between lexical and spatial distances among dialects (Séguy 1971, Nerbonne 2010).

[33]This is the case, for example, for the link between KDVFang1 and BNMKung2 in the lower right of Figure 11, which share only nineteen words across their similarity sets, against the maximum attested overlap of 127 words found for the two Mashi doculects, BKBMashi and KFKMashi1. The minimum overlap is ten words, for the pair DNMMumfu2 and ICNBiya2 and the pair DNMMumfu2 and NMAAbar1.

[34]As discussed in §5.3, there is evidence for a linguistic micro-area in the southeast of Lower Fungom, including the villages of Buu, Fang, Koshin, and, possibly also, Ajumbu and Mashi. Similar analysis could be done among those varieties as well, The comparison between Buu and Abar is done here due to previous work suggesting a contact relationship between Buu and Mungbam and because Buu's closest geographic link to Mungbam is via its Abar variety.

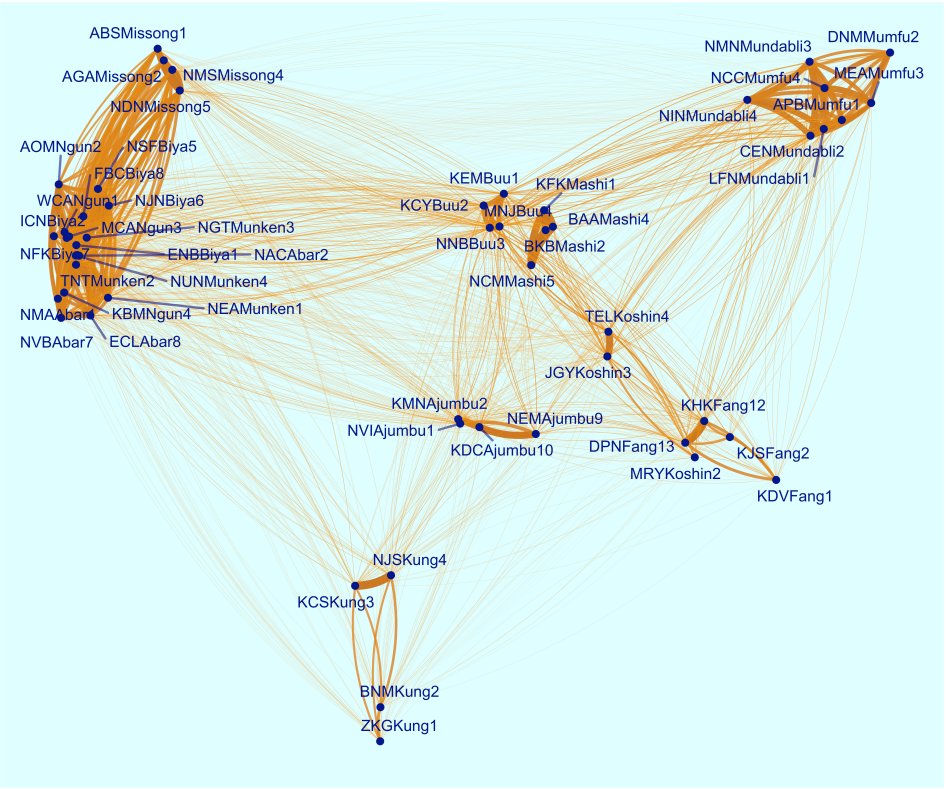

Figure 11: Lower Fungom lexical similarity network: Line darkness and thickness reflect the number of words in shared similarity sets across varieties, enhanced by an exponential thickness formula. The node layout, where each node represents a doculect, is derived from multidimensional scaling using a metric based on shared word count, which partly corresponds to the spatial arrangement of the villages.

The network visualization in Figure 12 is comparable to the one in Figure 11 but highlights the distinctive overlap between the Abar doculects and the Buu doculects via connections in blue, and the Buu doculects and the Mufu-Mundabli doculects in red. Forms common to all of the doculects are indicated with links directly between Abar doculects and Mufu-Mundabli doculects in purple. The relative weakness of those connections is clear from the figure, as is the fact that there is, broadly speaking, more overlap in the similarity sets among the Buu and Mufu-Mundabli varieties, which generally appear with thicker lines.[35]

[35]The range for the number of shared items in the similarity sets for the Buu doculects with the Abar and Mufu-Mundabli doculects is a high of 57 for the pair MNJBuu4 and NMNMundabli3

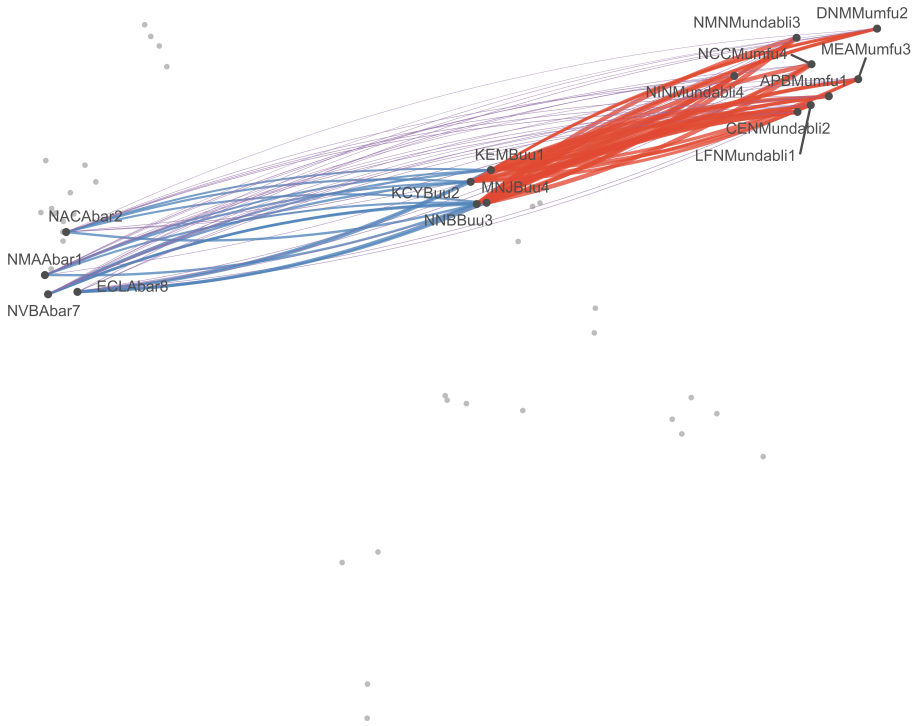

Figure 12: Abar, Buu, and Mufu-Mundabli lexical similarity subnetwork: Purple edges show words common to Abar and Mufu-Mundabli that are also found in Buu. Blue and red edges indicate words found only across Abar and Buu doculects and across Buu and Mufu-Mundabli doculects, respectively. Thicker lines mean the doculect pairs have more words in common, enhanced by an exponential thickness formula. Node positions match Figure 11 for comparison

Adopting a similar procedure to what was done for the analysis of Missong in §6.2, we can consider the specific lexical items involved in these relationships, and relevant data is presented in Table 4. The concepts where the words across all of the Buu, Abar, and Mufu-Mundabli doculects were grouped into the same similarity set are found in Table 4a.[36] The concepts where the words across only the Buu and Abar doculects were grouped into the same similarity set are found in Table 4b. The concepts where the words across only the Buu and Mufu-Mundabli doculects were grouped into the same similarity set are found in Table 4c. As with the case of the analysis in §6.2, the threshold for determining if a set of forms could be considered representative of Buu, Abar, and Mufu-Mundabli for the purposes of collecting the data in Table 4 was that they were all grouped together into the same similarity set for a given concept across all of the relevant individual-based doculects. Put differently, if the individual doculects were not all in "agreement" with respect to the lexical item associated with a given concept, then the concept is not included in the table.

In determining what data to collect in Table 4, only the Abar varieties of Mungbam were used used. This is due to the close, asymmetric contact relationship between Buu and Abar. Abar is the more socioeconomically influential village in Lower Fungom. If all Mungbam varieties had been considered, their internal variation would have obscured lexical overlap between Buu and Mungbam and Buu and Mufu-Mundabli. In particular, that would greatly reduce the number of similarity sets found between Buu and all of the Mungbam doculects. All of the doculects of Mufu-Mundabli were used due to grammatical evidence for a closer connection between Buu and Mufu-Mundabli and because the lower total number of Mufu-Mundabli doculects, in comparison to Mungbam doculects, did not result in as great a reduction in shared similarity sets. The concepts are listed with a homogeneity score. This score provides a measure of how uniform the expression of a concept is among the Lower Fungom varieties in the dataset, as determined by how many different similarity sets it is associated with.[37] A higher score means that there is less variation in the expression of the concept

and a low of 36 for KEMBuu1 with both NMAAbar1 and NACAbar2. The pair NNBBuu3 and ECLAbar8 has the highest number of shared similarity sets for a Buu and Abar doculect pairing, at 47. KCYBuu2 and DNMMumfu2 have the lowest number of shared similarity sets for a Buu and Mufu-Mundabli doculect pairing, at 41.

[36]The forms for 'breast' were not all grouped into the same similarity set due to the fact that, in one Mundabli doculect form, a high front vowel was transcribed instead of a high front glide in forms that were otherwise clearly very similar. This was noted via manual inspection, which is why the form is included here.

[37]The homogeneity metric used to arrive at this score is an entropy-based measure that is described in detail in Good et al. (to appear).

Table 4: Concepts in the same similarity sets for Abar, Buu, and Mufu–Mundabli, with a measure of concept homogeneity (HOMOGEN) within Lower Fungom included.

(a) Concepts in the same similarity set across the three sets of doculects

CONCEPT	HOMOGEN
grave	0.93
tongue	0.93
child	0.93
axe	0.89
breast	0.76
sieve	0.76

(b) Concepts in the same similarity set across only **Abar** and **Buu** doculects

CONCEPT	HOMOGEN
mother	0.88
ear	0.86
horse	0.83
bird	0.81
song	0.80
soap	0.76
cat	0.72
hair	0.71
book	0.71
forest	0.70
umbrella	0.68
devil	0.68
jaw	0.66
bitter leaf	0.66
caterpillar	0.66
faeces	0.64
dust	0.63
air	0.63
yam	0.61
intestine	0.57

(c) Concepts in the same similarity set across only **Buu** and **Mufu-Mundabli** doculects

CONCEPT	HOMOGEN
father	0.79
tooth	0.78
war	0.77
smoke	0.75
bag	0.75
stone	0.75
head	0.73
nose	0.72
sand	0.71
fowl	0.71
basket	0.70
goat	0.69
friend	0.66
farm	0.65
house	0.64
plantain	0.62
belly	0.61
court case	0.61
snake	0.59
egg	0.56
oil	0.55

(in terms of similarity sets) in Lower Fungom, and a lower score means that there is more variation.

As was the case with Missong discussed in §6.2, the sets of shared words do not show any semantic patterns that suggest that the partitioning can be explained in terms of something as straightforward as borrowings having entered Buu from Abar for words denoting recently introduced concepts. The words involve basic vocabulary with no obvious grouping into semantic domains across the sets. Again, we see a pattern consistent with some kind of language mixture and which would be difficult to explain as a result of the kind of "normal" transmission often assumed in work on language change.

In considering results like those seen in Table 4, it is important to bear in mind that the similarity sets were detected using a method that was specifically focused on surface similarity rather than deeper historical relationships (see §4.2). In many cases the classification of the words for the Buu equivalents with either Abar or Mufu–Mundabli reflect whether or not the Buu words appeared with or without a prefix since a noteworthy difference between Mungbam varieties like Abar from Mufu-Mundabli in a Lower Fungom context is the fact that Mungbam nouns generally show prefixes coding class on the noun in all classes, while Mufu-Mundabli nouns lack prefixes in most noun classes. While Buu's noun class system is structurally closer to that of Mufu-Mundabli, prefixes appear more frequently on nouns since, in some classes, where they are generally missing in Mufu–Mundabli on nouns, they do appear in Buu. For further discussion, see (Good 2023:394–397).

The fact that Buu nouns are "divided" in their resemblances to Abar and Mundabli both lexically and with respect to the encoding of noun class strengthens the characterization of the language as being mixed in character. The precise mechanism for the development of this pattern would require further study and could presumably involve a mix of three factors: (i) contact-induced convergence where Buu became more like Mungbam, (ii) contact-induced stability, where Buu retained features lost in Mufu-Mundabli that were independently found in Mungbam, due to contact with Mungbam varieties like Abar, (iii) contact-induced convergence and stability in Mufu-Mundabli with languages (to be determined) that those varieties are in closer contact with than Buu. (See Connell (2023) for relevant discussion of these possibilities in a West African context that shares key resemblances to the Lower Fungom situation.) Of course, this is not to say that all the ways in which Buu diverges from Mufu-Mundabli must be connected to contact, but, rather, that the particular divergences seen in Table 4 appear to have a straightforward explanation in terms of some kind of contact even if the precise historical details are not yet clear.

A more detailed examination of the data presented in this section is outside of the scope of this paper. Nevertheless, in light of the general points made here, it is clear that, based on the Lower Fungom data, looking at the development of a variety in terms of a general hypothesis of "mixture" is not only plausible but also likely to yield significant insights into its actual history. Moreover, using methods like those seen here can allow such hypotheses to be examined in a principled way rather than being simply dismissed as untestable (see §2.2).

7 Grounding models of change in local social structures

This paper has made two interrelated sets of arguments, the first conceptual and the second methodological. On the conceptual side, the key claims were: (i) Previous historical work privileging an approach to linguistic diversification in Africa based on the assumption that monolingual language acquisition within household is the norm cannot be supported by what is known about the sociolinguistics of societies in many parts of Africa. And, as a result, (ii) analyses of language "mixture" should be taken seriously in sociolinguistic settings characterized by small communities and high degrees of individual-level multilingualism. Associated with these arguments was a proposed reconstruction of the early sociolinguistics of the Bantu communities in §3.

On the methodological side, the paper considered what kinds of analytical approaches might help us identify potential cases of language mixture, and it employed a database of individual-level wordlists collected from speakers associated with a geographically compact region characterized by extensive multilingualism. Using this dataset, on the one hand, it examined the extent to which data of this kind shows patterns best explained in spatial terms, which is seen as a proxy for patterns of contact, and in line with recent work on the areal typology of African languages. On the other hand, it also looked at lexical patterns in potentially "mixed" varieties as a means of coming to a better understanding of the historical forces that shaped them. The paper also demonstrates that it is possible to consider hypotheses of mixture in a way that pays attention to concrete data rather than vague hypotheses. Importantly, while the quantitative methods used in this paper were not available to earlier scholars, the lexical patterns that were encountered could have been found using more traditional methods if this had not been dismissed as a worthy area of investigation.

With these points in mind, it is important not to overstate the results of this work, especially given that the treatment of the data remains relatively shallow in some cases. I certainly do not mean to claim that, for example, deliberate language mixture has been proven here to be the primary mechanism of change in

African languages. Rather, I think there is sufficient evidence for it to be taken seriously alongside more traditional models of change and that it is very likely an important mechanism of change in some cases, even if the extent of its significance has not been established.

In addition, it should be emphasized that the conceptual and methodological aspects of this study are strongly biased by the linguistic situation of the Cameroonian Grassfields. It may be reasonable to assume that the situation of the Grassfields is similar to other areas characterized by similar degrees of linguistic diversity, such as nearby areas of Nigeria (see, e.g., Connell 2023 for a recent relevant discussion of Cross River languages). However, the Grassfields clearly cannot serve as a complete model for all of Africa, or even only the Bantu-speaking area given the very different patterns of language distribution found throughout the continent.[38] I would therefore, hope, that the arguments presented here may prompt more detailed consideration of the sociolinguistic dynamics of more African regions so that this information could be used to arrive at a fuller picture of the forces that have shaped language diversification and areal patterns on the continent. That is, the arguments here are intended to expand the conceptual and methodological toolkit for understanding Africa's linguistic prehistory rather than to narrowly circumscribe the set of questions that are considered "worthwhile" to ask in the first place.

References

Auer, Anita, Catharina Peersman, Simon Pickl, Gijsbert Rutten & Rik Vosters. 2015. Historical sociolinguistics: The field and its future. *Journal of Historical Sociolinguistics* 1(1). 1–12. https://doi.org/10.1515/jhsl-2015-0001.

Bergs, Alexander. 2012. The Uniformitarian Principle and the risk of anachronisms in language and social history. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The handbook of historical sociolinguistics*, 80–98. Oxford: Blackwell.

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

[38] In this regard, however, it should be noted that there is at least one instance of a language in the southern part of the Bantu area that has been described as linguistically "hybrid" in a way that is in line with its sociolinguistic history. This is the case of siPhuthi, which has been treated as a Sotho-Nguni hybrid, in particular (Donnelly 2009:8–13). This suggests that dynamics claimed here to be active in the Grassfields may have a wider distribution than might be otherwise expected.

Bostoen, Koen. 2022. An introduction to *Reconstructing Proto-Bantu grammar*. In Koen Bostoen, Gilles-Maurice de Schryver, Rozenn Guérois & Sara Pacchiarotti (eds.), *On reconstructing Proto-Bantu grammar*, v–xlviii. Berlin: Language Science Press.

Bostoen, Koen & Bonny Sands. 2012. Clicks in south-western Bantu languages: Contact-induced vs. language-internal lexical change. In Matthias Brenzinger & Anne-Maria Fehn (eds.), *Proceedings of the 6th World Congress of African Linguistics (WOCAL 6)*, 121–132. Köln: Rüdiger Köppe.

Botne, Robert. 2021. The organization and evolution of tense systems in the Yemne-Kimbi languages. *Africana Linguistica* 27. 3–36.

Briatte, François. 2023. *ggnetwork: Geometries to plot networks with 'ggplot2'*. https://CRAN.R-project.org/package=ggnetwork.

Brubaker, Rogers & Frederick Cooper. 2000. Beyond "identity". *Theory and Society* 29(1). 1–47.

Campbell, Lyle & William J. Poser. 2008. *Language classification: History and method*. Cambridge: CUP.

Childs, G. Tucker. 2003. *An introduction to African languages*. Amsterdam: Benjamins.

Childs, G. Tucker, Jeff Good & Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation & Conservation* 8. 168–191. http://hdl.handle.net/10125/24601.

Christy, T. Craig. 1983. *Uniformitarianism in linguistics*. Amsterdam: Benjamins.

Connell, Bruce. 2023. Language contact and evidence of divergence and convergence in the morphology of Usaghade. *Linguistic Typology at the Crossroads* 3(1). 43–71. https://doi.org/10.6092/issn.2785-0943/16309.

Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics* 34(1). 1–37.

Csárdi, Gábor, Tamás Nepusz, Vincent Traag, Szabolcs Horvát, Fabio Zanini, Daniel Noom & Kirill Müller. 2023. *igraph: Network analysis and visualization in R*. https://doi.org/10.5281/zenodo.7682609.

Dalby, David. 1970. Reflections on the classification of African languages: With special reference to the work of Sigismund Wilhelm Koelle and Malcolm Guthrie. *African Language Studies* 11. 147–171.

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, Angiachi Demetris Esene Agwara & Rachel Ojong Diba. 2020. Multilingualism and the heteroglossia of ideologies in Lower Fungom (Cameroon): Use and attitudes. *Sociolinguistic Studies* 14(3). 321–345. https://doi.org/10.1558/sols.38799.

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.

Di Carlo, Pierpaolo & Ayu'nwi N. Neba. 2020. The so-called royal register of Bafut within the Bafut language ecology: Language ideologies and multilingualism in the Cameroonian Grassfields. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 29–51. Lanham, MD: Lexington Books.

Donnelly, Simon Scurr. 2009. *Aspects of tone and voice in Phuthi*. Urbana-Champaign, IL: University of Illinois PhD dissertation (corrected version). https://doi.org/doi:10.7282/T3NZ86JB.

Errington, Joseph. 2001. Colonial linguistics. *Annual Review of Anthropology* 30. 19–39.

Evans, Nicholas. 2017. Did language evolve in multilingual settings? *Biology & Philosophy* 32. 905–933.

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.

Frerebeau, Nicolas. 2023. *khroma: Colour schemes for scientific data visualization*. Pessac, France: Université Bordeaux Montaigne. https://doi.org/10.5281/zenodo.1472077.

Garrett, Andrew. 2000. A new model of Indo-European subgrouping and dispersal. In Steve S. Chang, Lily Liaw & Josef Ruppenhofer (eds.), *Proceedings of the twenty-fifth annual meeting of the Berkeley Linguistics Society: General session and parasession on loan word phenomena*, 146–156. Berkeley: Berkeley Linguistics Society.

Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 139–151. Cambridge: McDonald Institute for Archaeological Research.

Gensler, Orin D. 1994. On reconstructing the syntagm S-Aux-O-V-Other to Proto–Niger-Congo. In Kevin E. Moore, David A. Peterson & Comfort Wentum (eds.), *Proceedings of the twentieth annual meeting of the Berkeley Linguistics Society: Special session on historical issues in African linguistics*, 1–20. Berkeley: Berkeley Linguistics Society.

Gerhardt, Ludwig. 1983. Lexical interferences in the Chadic/Benue-Congo border-area. In H. Ekkehard Wolff & Hilke Meyer-Bahlberg (eds.), *Studies in Chadic and Afroasiatic linguistics*, 301–310. Hamburg: Helmut Buske.

Gerhardt, Ludwig. 2020. Reflections on the history of African language classification. In Rainer Vossen & Gerrit J. Dimmendaal (eds.), *The Oxford handbook of African languages*, 125–136. Oxford: OUP.

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. 2017. Niger-Congo languages. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 471–499. Cambridge: CUP.

Good, Jeff. 2020a. Interdisciplinarity in areal documentation: Experiences from Lower Fungom, Cameroon. *Language Documentation & Conservation* Special Publication No. 21. 43–71. http://hdl.handle.net/10125/24941.

Good, Jeff. 2020b. Niger-Congo, with a special focus on Benue-Congo. In Rainer Vossen & Gerrit J. Dimmendaal (eds.), *The Oxford handbook of African languages*, 139–160. Oxford: OUP.

Good, Jeff. 2022. Adapting methods of language documentation to multilingual settings. *Journal of Language Contact* 15(2). 341–375. https://doi.org/10.1163/19552629-15020006.

Good, Jeff. 2023. Language change in small-scale multilingual societies: Trees, waves, or magnets? In Darya Kavitskaya & Alan C. L. Yu (eds.), *Life cycle of language: Past, present, and future*, 386–398. Oxford: OUP.

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.

Good, Jeff, Nelson C. Tschonghongei, Pierpaolo Di Carlo & Clayton Hamre. to appear. Bantoid lexical diversity from an individual-based perspective. In Rebecca Grollemund, Derek Nurse & John Watters (eds.), *Bantoid and Bantu in Cameroon: An historical re-assessment*. Berlin: De Gruyter Mouton. https://www.acsu.buffalo.edu/~jcgood/GoodEtAl-IndividualWordlists.pdf.

Greenberg, Joseph H. 1949. Studies in African linguistic classification: III. The position of Bantu. *Southwestern Journal of Anthropology* 5. 309–317.

Greenberg, Joseph H. 1963. *The languages of Africa*. Bloomington, IN: Indiana University.

Greenberg, Joseph H. 1972. Linguistic evidence regarding Bantu origins. *The Journal of African History* 13. 189–216.

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.

Güldemann, Tom. 2008. The Macro-Sudan belt: Towards identifying a linguistic area in northern Sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A linguistic geography of Africa*, 151–185. Cambridge: CUP.

Güldemann, Tom. 2010. Sprachraum and geography: Linguistic macro-areas in Africa. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation: Volume 2: Language mapping*, 561–585. Berlin: De Gruyter Mouton.

Güldemann, Tom. 2018. Areal linguistics beyond contact, and linguistic areas of afrabia. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 448–545. Berlin: De Gruyter Mouton.

Güldemann, Tom & Mark Stoneking. 2008. A historical appraisal of clicks: A linguistic and genetic population perspective. *Annual Review of Anthropology* 37. 93–109.

Guthrie, Malcolm. 1962. Bantu origins: A tentative new hypothesis. *Journal of African Languages* 1. 9–21.

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.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2023. *Glottolog 4.8*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.813108.

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.

Herbert, Robert K. 1990. The sociohistory of clicks in southern Bantu. *Anthropological Linguistics* 3/4. 295–315.

Hyman, Larry M. 2018. More reflections on the nasal classes in Bantu. In John R. Watters (ed.), *Proto–East-Benue-Congo*, 223–236. Berlin: Language Science Press.

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.

Jacques, Guillaume & Johann-Mattis List. 2019. Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics* 9(1). 128–166.

Johnston, Harry H. 1919. *A comparative study of the Bantu and Semi-Bantu languages: Volume 1*. Oxford: Clarendon.

Joseph, Brian D. 2013. Demystifying drift. In Martine Robbeets & Hubert Cuyckens (eds.), *Shared grammaticalization, with special focus on the Transeurasian languages*, 43–66. Amsterdam: Benjamins.

Kassambara, Alboukadel. 2023. *ggpubr: 'ggplot2' based publication ready plots*. https://rpkgs.datanovia.com/ggpubr/.

Kassambara, Alboukadel & Fabian Mundt. 2020. *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Kiessling, Roland. 2018. Niger-Congo numeral classifiers in a diachronic perspective. In William B. McGregor & Søren Wichmann (eds.), *The diachrony of classification systems*, 34–75. Amsterdam:Benjamins.

Kopytoff, Igor. 1987. The internal African frontier: The making of African political culture. In Igor Kopytoff (ed.), *The African frontier: The reproduction of traditional African societies*, 3–84. Bloomington: Indiana University Press.

Labov, William. 1972. Some principles of linguistic methodology. *Language in Society* 1. 97–120.

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.

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

Lüpke, Friederike. 2010. Multilingualism and language contact in West Africa: Towards a holistic perspective. *Journal of Language Contact* 3(1). 1–12. https://doi.org/10.1163/19552629-90000002.

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.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2022. *cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster.

Maho, Jouni. 1999. *A comparative study of Bantu noun classes*. Göteburg: Acta Universitatis Gothoburgensis.

Mallory, J. P. 1989. *In search of the Indo-Europeans: Language, archaeology, and myth*. London: Thames & Hudson.

Menang, Thaddeus. 2004. Cameroon Pidgin English (Kamtok): Phonology. In Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds.), *Handbook of varieties of English: Volume I: Phonology*, 902–917. Berlin & New York: Mouton.

Moore, Leslie C. 2004. Multilingualism and second language acquisition in the northern Mandara Mountains. In George Echu & Samuel Gyasi Obeng (eds.), *Africa meets Europe: Language contact in West Africa*, 131–148. New York: Nova Science.

Mufwene, Salikoko S. 2001. *The ecology of language evolution*. Cambridge: CUP.

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.

Ndamsah, Gratiana. 2020. Multilingualism and the paradox of language creation: The case of Lítâ (research note). In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, chap. 105–112. Lanham, MD: Lexington Books.

Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3821–3828. https://doi.org/10.1098/rstb.2010.0048.

Newman, Paul. 1995. *On being right: Greenberg's African linguistic classification and the methodological principles which underlie it.* Bloomington, Indiana: Institute for the Study of Nigerian Languages & Cultures, African Studies Program, Indiana University.

Ngako Yonga, Monique Doriane. 2013. *Ébauche phonologique et morphologique de la langue bu.* Yaoundé: University of Yaoundé MA thesis.

Nichols, Johanna. 1990. Linguistic diversity and the first settlement of the New World. *Language* 66. 475–521.

Nichols, Johanna. 1996. The comparative method as heuristic. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed*, 39–71. Oxford: OUP.

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

Ojong Diba, Rachel. 2020. Nuances in language use in multilingual settings: Code-switching or code regimentation in Lower Fungom? In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 15–28. Lanham, MD: Lexington Books.

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

Philippson, Gérard & Rebecca Grollemund. 2019. Classifying Bantu languages. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu Languages* (second edition), 335–354. Abingdon, UK: Routledge.

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

R Core Team. 2023. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Rolle, Nicholas, Florian Lionnet & Matthew Faytak. 2020. Areal patterns in the vowel systems of the macro-sudan belt. *Linguistic Typology* 24(1). 113–179. https://doi.org/10.1515/lingty-2019-0028.

Röschenthaler, Ute. 2011. *Purchasing culture: The dissemination of associations in the Cross River Region of Cameroon and Nigeria.* Trenton, NJ: Africa World Press.

Ross, Malcolm. 1996. Contact-induced change and the comparative method: Cases from Papua New Guinea. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed*, 180–217. Oxford: OUP.

Ross, Malcolm. 1997. Social networks and kinds of speech-community event. In Roger Blench & Matthew Spriggs (eds.), *Archaeology and language I: Theoretical and methodological orientations*, 209–261. London: Routledge.

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

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35. 335–357.

Stallcup, Kenneth. 1980. La géographie linguistique des Grassfields. In Larry M. Hyman & Jan Voorhoeve (eds.), *L'expansion bantoue: Actes du colloque international du CNRS, Viviers (France) 4–16 avril 1977. Volume I: Les classes nominales dans le bantou des Grassfields*, 43–57. Paris: SELAF.

Storch, Anne. 2009. Cultured contact—Ritualisation and semantics in Jukun. In Wilhelm J.G. Möhlig, Frank Seidel & Marc Seifert (eds.), *Language contact, language change and history based on language sources in Africa* (SUGIA 20), 297–319. Köln: Rüdiger Köppe.

Storch, Anne. 2011. *Secret manipulations: Language and context in Africa*. Oxford: OUP.

Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino & Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change* 6(2). 235–283. https://doi.org/10.1163/22105832-00602002.

Tabe, Florence A. E. 2020. Multilingualism in rural Africa: A case study of Ossing village in Cameroon. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 115–136. Lanham, MD: Lexington Books.

Tang, Yuan, Masaaki Horikoshi & Wenxuan Li. 2016. ggfortify: Unified interface to visualize statistical results of popular R packages. *The R Journal* 8(2). 474–485. https://doi.org/10.32614/RJ-2016-060.

Thomason, Sarah G. & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: UC Press.

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. 2020. *Millennia of language change: Sociolinguistic studies in deep historical linguistics*. Cambridge: CUP.

Verkerk, Annemarie & Francesca Di Garbo. 2022. Sociogeographic correlates of typological variation in northwestern Bantu gender systems. *Language Dynamics and Change* 12(2). 155–223. https://doi.org/10.1163/22105832-bja10017.

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.

Warnier, Jean-Pierre. 1985. *Échanges, développement et hiérarchies dans le Bamenda pré-colonial (Cameroun)*. Wiesbaden: F. Steiner.

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

Wedekind, Klaus. 1985. Thoughts when drawing a map of tone languages. *Afrikanistische Arbeitspapiere* 1. 105–124.

Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium*, 97–195. Austin: University of Texas Press.

Welmers, William E. 1973. *African language structures*. Berkeley: University of California Press.

Westermann, Diedrich. 1927. *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu*. Berlin: Walter de Gruyter.

Westermann, Diedrich. 1952. African linguistic classification. *Africa: Journal of the International African Institute* 22. 250–256.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wolff, Ekkehard & Ludwig Gerhardt. 1977. Interferenzen zwischen Benue-Kongo- und Tschad-Sprachen. In Wolfgang Voigt (ed.), *XIX. deutscher Orientalistentag*, 1518–1543. Wiesbaden: F. Steiner.

Yasho, Tatang Joyce. 2016. *Aspects of Kung grammar*. Buea, Cameroon: University of Buea MA thesis.