# The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa

Dmitry Idiatov, Mark van de Velde

# THE LEXICAL DISTRIBUTION OF LABIAL-VELAR STOPS IS A WINDOW INTO THE LINGUISTIC PREHISTORY OF NORTHERN SUB-SAHARAN AFRICA

Dmitry Idiatov
*LLACAN (CNRS – USPC/INALCO)*

Mark L. O. Van de Velde
*LLACAN (CNRS – USPC/INALCO)*

Using a very large lexical database and generalized additive modeling, this article reveals that labial-velar (LV) stops are marginal phonemes in many of the languages of Northern Sub-Saharan Africa that have them, and that the languages in which they are not marginal are grouped into three compact zones of high lexical LV frequency. The resulting picture allows us to formulate precise hypotheses about the spread of the Niger-Congo and Central Sudanic languages and about the origins of the linguistic area known as the Sudanic zone or Macro-Sudan belt. It shows that LV stops are a substrate feature that should not be reconstructed into the early stages of the languages that currently have them. We illustrate the implications of our findings for linguistic prehistory with a short discussion of the Bantu expansion. Our data also indirectly confirm the hypothesis that LV stops are more recurrent in expressive parts of the vocabulary, and we argue that this has a common explanation with the well-known fact that they tend to be restricted to stem-initial position in what we call C-emphasis prosody.*

*Keywords*: areal linguistics, Bantu expansion, generalized additive modeling, historical linguistics, labial-velar stops, Northern Sub-Saharan Africa, substrate interference

**1.** Introduction. Labial-velar stops (LV stops), such as /k͡p/, /g͡b/, and /ŋ͡m/, are speech sounds that are produced with almost simultaneous gestures of velar and labial closure (Ladefoged & Maddieson 1996:332–43). LV stops are found in many languages in the west and center of Northern Sub-Saharan Africa (NSSA), while they are rare elsewhere (Cahill 2008, 2018, Maddieson 2011, 2018). Pointing out that the set of languages with LV stops is geographically coherent but genealogically diverse, Güldemann (2008: 156–58) and Clements and Rialland (2008) use their presence as one of the defining features of a linguistic area, which they respectively call the 'Macro-Sudan belt' and the 'Sudanic zone'. The preponderance of such a typologically unusual feature in a large and genealogically diverse area raises the questions of where and how it originated and by what mechanism it spread. The hypotheses proposed in the literature rely on the usual explanatory tools of areal linguistics, such as inheritance, innovation through sound change, borrowing of phonemes through loanwords, substrate interference, and a more abstract and less well-defined concept of diffusion. Thus, Westermann (1911, 1927) suggested that the feature is an innovation through sound change from labialized velars to LV stops. Applying the majority wins rule, Greenberg (1983:8–9) suggested that the feature should be reconstructed to Proto-Niger-Congo and that it was retained in some of the daughter languages and diffused through borrowing of loanwords or 'convergent sound change', whose vector presumably was bilingual speakers bringing LV stops into their primary language communities. The same majority-wins logic is applied by Cahill (2017, 2018), who generally argues for inheritance from the proto-language not only in Niger-Congo and its major branches, but also in other groups, such as Central Sudanic.

* The research reported in this article is part of the projects LC2 'Areal phenomena in Northern Sub-Saharan Africa' and GL7 'Reconstruction, genealogy, typology and grammatical description in the world's two biggest phyla: Niger-Congo and Austronesian' of the Labex EFL ('Investissements d'Avenir' program, overseen by the French National Research Agency, reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001. Special thanks with respect to the present paper are due to Guillaume Segerer and Benoît Legouy for their assistance in using the RefLex database. Last but not least, we are grateful to the three anonymous referees and the editors for their constructive criticism.

An unpublished paper by Vogler (2014) argues against the plausibility of the majority argument for the inheritance hypothesis and suggests an unknown substrate origin of the feature in the western part of the West African rainforest (the area centered on the modern domain of the Kru languages), from which it diffused elsewhere.

In order to try to answer the 'where' and 'how' questions for both the origin and the spread of LV stops, we propose to go beyond the usual practice in areal linguistics of looking at the geographical distribution of a feature in terms of its presence versus absence in given languages. Following the new methodology that we adopt in this article, we investigate how deeply LV stops are entrenched in the languages where they are attested by estimating their lexical frequencies in a large lexical database, and we analyze the spatial distribution of these frequencies. We also compare these findings to data from a number of other relevant fields, such as geography; paleo-, molecular, and cultural anthropology; archaeology; paleoclimatology; and paleobotany. By exploiting the variation in the data instead of trying to reduce it, we arrive at a much richer and more detailed historical account than would have been possible in the traditional reductionist approach.

We start from the observation that individual languages that have LV stops in their phoneme inventories differ greatly in terms of the prominence of these consonants. LV stops are rather marginal phonemes in most of the NSSA languages we are familiar with, in the sense that they are less frequent than the other stops and/or are restricted to specific positions in the word and/or to parts of the vocabulary. In contrast, we also know languages such as Yoruba [yoru1245],[1] where LV stops are normal consonant phonemes. Primarily thanks to the existence of the RefLex database (Segerer & Flavier 2011–2021), we were able to estimate the lexical frequency of LV stops in 315 African languages that have them in their phonological inventories and compare the estimate for each language to an estimate of the canonical situation in which each consonant phoneme is equally frequent in the lexicon. We subsequently studied the spatial distribution of the interval between both estimates. The geographical patterns that emerge lend themselves to a clear and interesting historical interpretation. We show that there are three hotbeds of high lexical frequency of LV stops, which have all of the characteristics of refuge zones. They are separated from each other by zones with low lexical frequencies of LV stops that correspond to areas with a climate and vegetation to which migrants from a savanna habitat are better adapted. We therefore conclude that the current areal distribution of LV stops is due to areal retention of these consonants themselves and/or of other phonetic features that facilitate their emergence.

The article is organized as follows. In §2, we discuss the quantity, quality, and origin of our data (§2.1), and we explain how we estimated the lexical frequency of LV stops in the languages of Africa and show the results of these estimates (§2.2). We also test the hypothesis that LV stops are more common in expressive words than in the general vocabulary, a hypothesis that we derived from looking at individual languages and that is relevant for our explanation of the origin and spread of LV stops (§2.3). In §3, we present the spatial distribution of lexical frequencies of LV stops using two different visualization methods, spatial interpolation (§3.1) and generalized additive modeling (§3.2–§3.5). The two methods converge on the same spatial structure, consisting of three hotbeds, two of which are separated only by a narrow discontinuity. We further cross-validated these findings by plotting African toponyms spelled with an LV stop, which produces three clusters that closely correspond to our three hotbeds (§3.6). Our explanation of the origin

[1] The code composed of a combination of four letters and four numbers between square brackets after the name of a language is its Glottolog identifier (Hammarström et al. 2019).

and spread of LV stops crucially involves the phenomenon of C-emphasis prosody, which facilitates both the emergence of LV stops and their transfer through language contact, and which accounts for the higher frequency of LV stops in the more expressive parts of the lexicon. We briefly discuss the phenomenon of C-emphasis prosody in §4. In §5, we discuss the historical implications of our findings, namely, that LV stops, or at least the features that contribute to their emergence, were present in and around the current hotbeds before the arrival of the language families that are currently spoken there (§5.1), but that they are an innovation in all or most of those language families (§5.3). This geographical retention is mainly due to language shift in the hotbeds of high lexical frequency and to borrowing in the other areas. By comparing our findings against data from a number of other relevant fields, we propose a detailed scenario for the initial emergence and spread of the LV stops and/or the phonetic features that facilitate their emergence on the macrolevel of NSSA (§5.1). Our findings also allow us to adjust and refine the scenarios proposed in the literature for the Bantu expansion, one of the biggest language expansion events in recent human history (§5.2). Finally, the geographical distribution of high lexical frequencies of LV stops suggests that LV stops should not be reconstructed into the proto-languages of the major subbranches of Niger-Congo and Central Sudanic, unless they were spoken in one of the hotbeds, nor into Proto-Niger-Congo and Proto-Central Sudanic (§5.3).

## 2. Estimating the frequency of LV stops in the lexicon.

**2.1.** The database. Our main source of data is RefLex (Segerer & Flavier 2011–2021), an online database of over a thousand lexicons of African languages, which comes with a number of useful tools for reconstruction and statistical analysis. The accuracy of its contents is easily verifiable, thanks to integrated links to its published sources. We have left out sources published before 1900, because the notation of LV stops is unreliable in many of them. Moreover, we have disregarded sources with fewer than one hundred entries. This partly arbitrary cutoff point was chosen in order to include sources that are sufficiently large to be minimally representative for a language, while not excluding the lists of basic vocabulary items that are the only existing sources of information on many African languages.[2] Whenever the same language is represented by more than one source in RefLex, we chose the best source in terms of size and reliability. Furthermore, we have added some lexical sources of Bantu and Mande languages that are not (yet) integrated into RefLex, and we have used the information on the presence versus absence of LV stops in the phonological inventories of African languages that is available in PHOIBLE (Moran et al. 2014).

Our sample contains 1,110 languages, of which 545 have LV stops in their inventory and 565 do not. We have data on the lexical frequency of LV stops for 315 of the 545 languages that have them. Figure 1 represents the languages of our sample and highlights the areas of their concentration in space: that is, it shows their spatial intensity.[3]

[2] Based on an experimental study using data from Australian languages, Dockum and Bowern (2019:50) argue that 'we are likely to be incorrectly representing basic facts about the phonology of the language' if we use 'wordlists below [a] 400-item threshold'. From this perspective, even our threshold of 100 items is really a poor man's solution. Nevertheless, for the purposes of our research the inclusion of sources below the threshold of 400 items proved not to affect the results significantly, and this despite the fact that such sources constitute around 40% of our sample of languages with data on the lexical frequency of LV stops (cf. §3.3). Thus, using only the sources with 400 items and more produced very similar results to the sample as a whole.

[3] All charts and calculations in this article were made with the program R (R Core Team 2015) using RStudio IDE (RStudio Team 2016). We used the 'spatstat' package to produce the spatial intensity and spatial interpolation graphs (Baddeley & Turner 2005). Note that this figure and several of the others are presented in

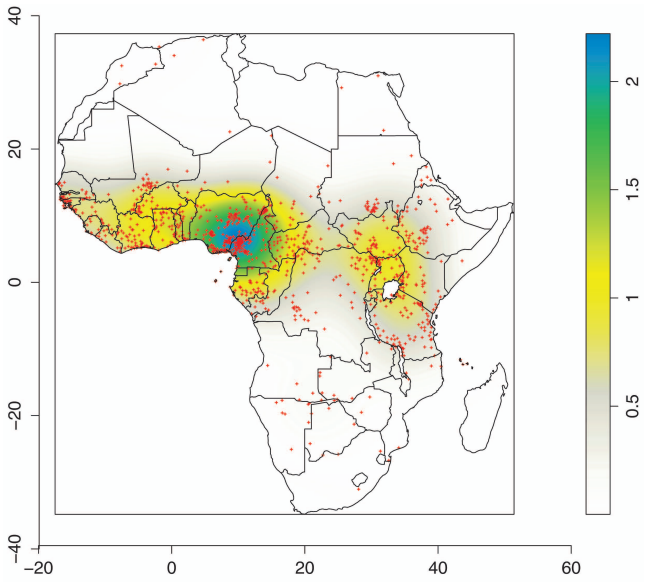


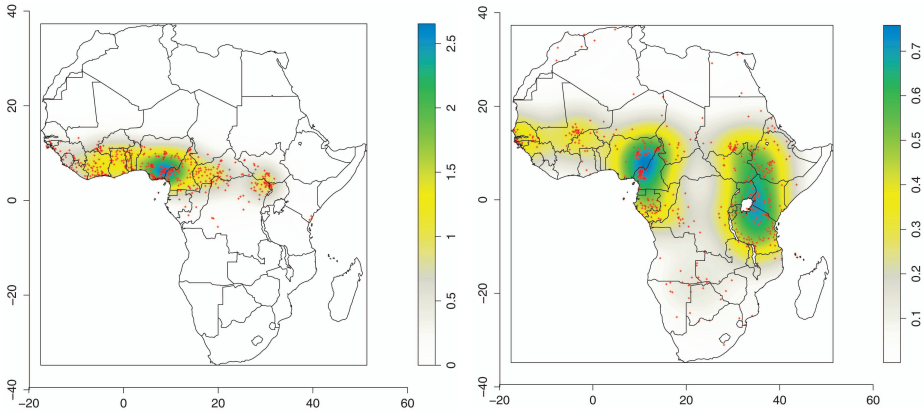

FIGURE 1. The geographical distribution of the 1,110 languages of our sample and their spatial intensity (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using the mean square error).

There are two areas with a high concentration of languages, both situated in the so-called linguistic fragmentation belt. The most important one is situated at the border between Cameroon and Nigeria. The second covers a large part of East Africa and centers around Lake Victoria.

Figure 2 shows the geographical distribution of languages with (Fig. 2a) and without (Fig. 2b) LV stops, together with their spatial intensity. As expected, the distribution of languages with LV stops shown in Fig. 2a corresponds to Güldemann's (2008) Macro-Sudan belt and to Clements and Rialland's (2008) Sudanic zone. Moreover, Fig. 2a shows that languages with LV stops are especially strongly concentrated in an east-west-oriented area in the south of Nigeria. Figure 2b shows areas of concentration of languages without LV stops at the extreme west and east of NSSA, with a band of somewhat lower spatial intensity of such languages stretching west to east across the western part of NSSA to the north of the area, with many languages with LV stops closer to the coast. There is also a strong concentration of languages without LV stops in an area stretching from the northeast of Nigeria to the west of Cameroon, which partly overlaps with the area of strongest concentration of LV languages. This partial overlap is due to the high linguistic fragmentation in that area. Note, however, that the spatial orientation of the two overlapping zones is different. The LV area is situated further to the west and has a horizontal orientation, whereas the LV-less area has a north-south orientation. These configurations will turn up again in §3, where we look at the spatial distribution of the lexical frequency of LV stops in the languages that have these consonants in their phoneme inventories.

**2.2.** ESTIMATING THE FREQUENCY OF LV STOPS IN THE LEXICON. We now turn to the 315 languages in our database that have LV stops and for which we have lexical infor-

full color in the electronic versions of this article, but in grayscale in the print version; color versions of the figures are also available open access along with the other supplementary materials at http://muse.jhu.edu/resolve/115.



a. Languages with LV stops. b. Languages without LV stops.

FIGURE 2. The geographical distribution of (a) the 545 languages with LV stops and (b) the 565 languages without LV stops and their spatial intensity (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using the mean square error).

mation. In order to estimate the frequency of LV stops in their lexicons, we started by cleaning up the data harvested in RefLex in order to normalize them and remove any errors. We removed some languages in which the symbol combinations *kp* and/or *gb* are used for successions of a velar and a bilabial stop. Then we recoded digraphs not recognized as such by RefLex. We also corrected for some orthographic conventions related to the representation of LV stops that differ from the IPA, such as Yoruba [yoru1245] *p* corresponding to /k͡p/. Finally, we separated clusters that RefLex automatically treats as complex phonemes, such as so-called prenasalized stops (successions of a homorganic nasal and oral stop, e.g. *nd*, *mb*), consonants followed by marks of labialization (e.g. *bw*) or palatalization (e.g. *by*), and successions of a stop and a labiodental fricative (e.g. *bv*).

The formula we used to estimate the lexical frequency of LV stops in a language is given in 1. It expresses the frequency of LV stops as a percentage, so that 0% corresponds to the absence of LV stops and 100% corresponds to the number of LV stops that would be expected in the canonical situation where all consonant phonemes of the language have exactly the same probability of occurrence in the lexicon. We call this frequency (i.e. $F_{LV} = 100\%$) the REFERENCE FREQUENCY.[4]

$$(1)\quad F_{LV} = \frac{LV_O}{LV_E} * 100\% = \frac{\sum T_{LV}}{\frac{\sum T_C}{\sum P_C} * \sum P_{LV}} * 100\%$$

As shown in 1, the estimated frequency of LV stops in a language ($F_{LV}$) is the quotient of the observed number of LV stops in our lexical source for that language ($LV_O$) and the expected number of LV stops in the canonical situation where every consonant phoneme is equally frequent in the lexicon ($LV_E$). Needless to say, $LV_E$ is a purely the-

[4] In the interests of mapping and statistical analysis, we consider the relative lexical frequency of the full set of LV stops in any given source, rather than considering the lexical frequencies of every individual LV separately. In theory, this choice could lead to misleading results if there were areas where many languages have a set of two or more LV stops, of which one has a high lexical frequency, whereas the others are marginal. There is no such area, however, and there are only a handful of languages in our sample where one LV has a lexical frequency of more than 66% of the reference frequency, while the frequency of the others is less than 33%.

oretical calibration point and is in no way expected to exist. It is calculated by dividing the total number of consonants observed in the data source ($\sum T_C$) by the total number of consonant phonemes of the language ($\sum P_C$) and then multiplying this quotient by the number of LV stop consonants in the phoneme inventory of the language ($\sum P_{LV}$).

The results of the estimate of $F_{LV}$ (as percentages) in the 315 languages of our sample that have LV stops and for which we have lexical data are summarized in Figure 3 by means of a probability density plot truncated at zero. The probability density plot is overlaid on the same data presented by means of a histogram. Figure 3 also shows the median of the distribution of $F_{LV}$ and the reference frequency of 100%. The median is greatly inferior to the reference frequency, showing that LV stops are relatively rare phonemes in the great majority of the languages that have them.

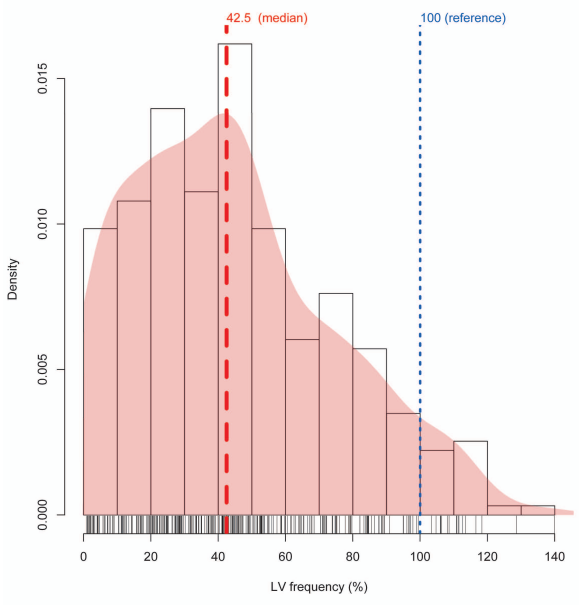



FIGURE 3. The probability density of all $F_{LV}$ frequencies in our sample in percentages (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using two-stage plug-in bandwidth selector; the data closest to zero were up-weighted to account for the truncation of the distribution at zero). Median $F_{LV}$ = 42.50%; reference $F_{LV}$ = 100%. The probability density of $F_{LV}$ is overlaid on the same data presented by means of a histogram. The rug plot at the bottom shows the distribution of the data points.

**2.3.** LV STOPS AND EXPRESSIVITY: ESTIMATING THE FREQUENCY OF LV STOPS IN BASIC VOCABULARY. Descriptions of individual languages or language groups sometimes mention that LV stops are more common in expressive parts of the vocabulary than in the general vocabulary. For instance, Bostoen and Donzo (2013) point out that LV stops in Ngombe (Bantu, DRC; [ngom1268]) are much more frequent in ideophones, words derived from ideophones, and sound-symbolic adverbs than in the general lexicon. They conclude that LV stops are associated with expressivity in this language. Likewise, Martin (2015) points out that LV stops are much more frequent in ideophones than elsewhere in Wawa [wawa1246], a Mambiloid language spoken in Cameroon, close to the border with Nigeria. Ngombe and Wawa differ from each other in that LV stops are common phonemes in Ngombe, but rare in Wawa. There may therefore be a

preference for LV stops to occur in expressive parts of the vocabulary irrespective of their overall lexical frequency in a language.

A particularly telling example of such a preferential association between LV stops and expressivity is provided by the Nkundo variety of the Mongo dialect cluster (Bantu, DRC; [nkun1238], [mong1338]), as described by Hulstaert (1957, 1961, 1965, 1966). Across the Mongo dialect cluster, we find between-dialect and sometimes within-speaker variation between LV stops /k͡p/ and /g͡b/ and the corresponding labialized velars /kw/ and /gw/. Most Mongo lects strongly prefer the LV stop realization, but Nkundo, the reference lect of Hulstaert's grammatical and lexical description, prefers the labialized velar realization (1957:xiii, 959). Yet there are a number of words in Nkundo where only an LV stop realization is possible (Hulstaert 1957:xiii). Remarkably, the overwhelming majority of these approximately forty words with only LV realization in Hulstaert's (1957) dictionary are ideophones, such as *kpótókpòtò* 'ideophone expressing the sound made by slippers', complemented by a few nouns derived from ideophones, such as *lì-kpòtò* 'slippers', a few emotionally charged nouns for which no corresponding ideophone is recorded, such as *lì-kpéké* 'swindle, swiz, rip off" and *ngbàngà* 'quarrel, brawl', and a few subordinate-level terms, such as *ngbàà* 'type of chisel with a long handle for woodcarving' and *è-ngbélé* 'special type of cassava cake'.

We are interested in finding out whether the higher frequency of LV stops in expressive parts of the vocabulary is a general tendency in NSSA, as this may shed light on the way these consonants are transferred from one language to another and thus contribute to an explanation for the areal pattern. Since there is no way to automatically extract reliable lists of expressive vocabulary from our lexical sources, we used a proxy to the hypothesis by testing whether LV stops are less common in the basic vocabulary of the languages of our sample than in their general vocabulary, under the generally held assumption that expressive items do not make up part of basic vocabulary. In order to do this, we restricted our sample to the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries.[5] From these sources, we automatically extracted lists of 200 basic vocabulary items based on the 'Swadesh 200' list (Swadesh 1952). Since all sources lacked various numbers of entries whose translation corresponds to an item in the Swadesh 200 list (from 21–139 missing entries, mean = 67), we ended up with a sample of 178 quasi-Swadesh 200 lists of uneven length, altogether using 196 of the 200 concepts on the Swadesh 200 list. We then calculated the lexical frequencies of LV stops in the quasi-Swadesh 200 lists as a percentage of their reference frequency, and we compared these to the lexical frequencies of LV stops in the full sources from which they were extracted.

The results are shown in Figure 4, which represents the probability densities of the $F_{LV}$ frequencies in percentages in the subsample of 178 RefLex sources with minimally 400 entries (in pink) and in the quasi-Swadesh 200 lists derived from this subsample (in blue). The median $F_{LV}$ of the latter (dot-dashed blue line to the left) is, at 24.95%, much lower than the median $F_{LV}$ (42.49%) of the former (larger-dashed red line toward the center), which is itself largely inferior to the reference frequency of 100% (smaller-dashed black line to the right). The two distributions are not normal, but their variances are similar, which allows for comparing them using the Wilcoxon signed-rank test (paired U-test). This test confirms that it is very unlikely that the two distributions represent the same population ($p < 0.001$) and that the difference between the median $F_{LV}$

[5] Our threshold of 400 items is in agreement with the findings of Dockum and Bowern (2019:50), discussed in n. 2.

values of these two data sets is significant. We also did bootstrap validation (repetitions = 999), which equally confirmed this result (100% of the values $p < 0.05$, 87% of the values $p < 0.001$). The results summarized in Fig. 4 confirm our hypothesis: LV stops are less common in the basic vocabulary than in the general vocabulary.

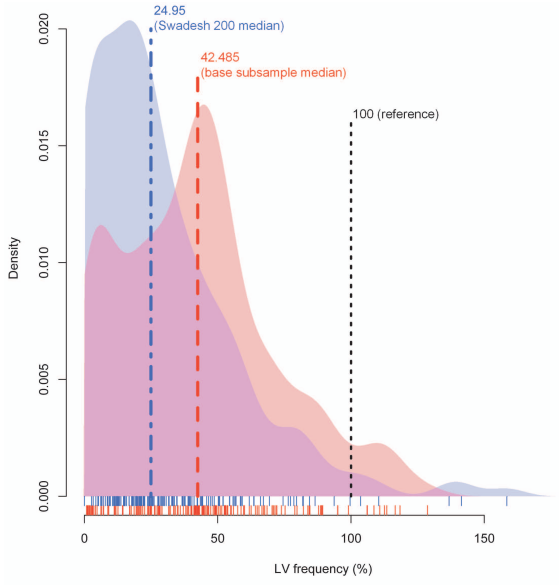



FIGURE 4. The probability densities of the $F_{LV}$ frequencies in percentages in the subsample of 178 RefLex sources with minimally 400 entries and in the quasi-Swadesh 200 lists derived from this subsample (estimated in the same way as in Fig. 3). The median $F_{LV}$ is 42.49% in the base subsample and 24.95% in the derived quasi-Swadesh 200 lists; reference $F_{LV}$ = 100%. The rug plots at the bottom show the distribution of the data points in the two data sets.

Importantly, the data of the quasi-Swadesh 200 lists can also be viewed as providing support for our hypothesis that LV stops are more common in expressive parts of the vocabulary than in the general vocabulary. This support comes from the distribution of words with and without LV stops in 178 quasi-Swadesh 200 lists by the part-of-speech category of the concepts, as summarized in Table 1.[6] The total proportion of words with LV stops in 178 quasi-Swadesh 200 lists is 4.81%, and we find a similar ratio of 4.75% for nouns and verbs taken together. At the same time, the frequency of words with LV stops is almost double, 8.52%, for qualifiers and quantifiers taken together, the two categories that are most likely to contain expressive and evaluative concepts. Remarkably, we find the reverse situation with the part-of-speech categories that are usually described as functional categories, viz. conjunctions, adpositions, demonstratives, personal pronominals, negation markers, and indefinites. The frequency of words with LV stops in these functional categories taken together is as low as 0.48%. Numerals, which are more of an open category similar to nouns and verbs, and interrogatives, which are similar to functional categories but canonically are associated with a

[6] Since, strictly speaking, part-of-speech categories are necessarily language-specific distributional classes of lexemes, the part-of-speech classification used in Table 1 is basically a semantic classification of concepts enriched with some rather broad morphosyntactic features. The part-of-speech tag of a given concept need not correspond to the part-of-speech tag used in the description of a given language for the word translating this concept.

prominent information-structural status, have somewhat higher frequencies of words with LV stops, 1.05% and 2.92%, respectively, than the more canonical functional categories taken together.

| PART OF SPEECH | CONCEPT | WORDS WITH LV | WORDS W/O LV | WORDS WITH LV (%) | |
|---|---|---|---|---|---|
| conjunction | and, because, if | 0 | 185 | 0.00% | |
| adposition | at, in, with | 1 | 147 | 0.68% | |
| demonstrative | here, that, there, this | 3 | 360 | 0.83% | |
| pronoun | 1PL, 1SG, 2PL, 2SG, 3PL, 3SG | 1 | 437 | 0.23% | **0.48%** |
| negator | not (NEG) | 0 | 52 | 0.00% | |
| indefinite | other, some | 1 | 63 | 1.56% | |
| interrogative | how?, what?, when?, where?, who? | 5 | 472 | 1.05% | |
| numeral | five, four, one, three, two | 24 | 798 | 2.92% | |
| noun | animal, ashes, back, bark, belly, bird, blood, bone, child, cloud, day, dog, dust, ear, earth, egg, eye, fat, father, feather, fire, fish, flower, fog, foot, fruit, grass, guts, hair, hand, head, heart, husband, lake, leaf, leg, liver, louse, man, meat, mother, mountain, mouth, name, neck, night, nose, person, river, road, root, rope, salt, sand, sea, seed, skin, sky, smoke, snake, star, stick, stone, sun, tail, tongue, tooth, tree, water, wife, wind, wing, woman, woods, worm, year | 573 | 10,839 | 5.02% | |
| verb | to bite, to blow, to breathe, to burn, to come, to cut, to die, to dig, to drink, to eat, to fall, to fear, to fight, to float, to flow, to fly, to give, to hear, to hit, to hold, to hunt, to kill, to know, to laugh, to lie, to live, to play, to pull, to push, to rain, to rub, to say, to scratch, to see, to sew, to sing, to sit, to sleep, to smell, to spit, to split, to squeeze, to stab, to stand, to suck, to swell, to swim, to think, to throw, to tie, to turn, to vomit, to walk, to wash, to wipe | 300 | 6,684 | 4.30% | **4.75%** |
| qualifier | bad, big, black, cold, dirty, dry, dull, far, good, green, heavy, left, long, narrow, near, new, old, red, right, right (correct), rotten, sharp, short, small, smooth, straight, thick, thin, warm, wet, white, wide, yellow | 188 | 2,190 | 7.91% | |
| quantifier | all, few, many | 43 | 290 | 12.91% | **8.52%** |
| TOTAL | | *1,139* | *22,517* | ***4.81%*** | |

TABLE 1. The distribution of words with and without LV stops in 178 quasi-Swadesh 200 lists by the part-of-speech category of the concepts.

**3.** THE SPATIAL DISTRIBUTION OF LEXICAL FREQUENCIES OF LV STOPS. The map in Fig. 2a above is a point-pattern representation of the geographical distribution of African languages with LV stops. It shows the areas where many versus few versus no languages have LV stops, adding detail to a general picture that was already largely known. We are here mostly interested in finding any patterns in the geographical distribution of differences in the lexical frequency of LV stops among the languages that have them. We therefore coupled the results of our frequency estimates presented in §2 with the geographical coordinates of the languages of our sample. We first visualize the results using two types of spatial interpolation plots in §3.1 and subsequently try to quantify our results in a more rigorous way by modeling and visualizing them using generalized additive modeling in §3.2–§3.5. Finally, in §3.6, we replicate our results by looking at the geographical distri-

bution of African toponyms that contain LV stops. The plots and the models in the remainder of the article focus on the area for which we have data on the lexical frequency of LV stops (longitude interval [−18°, 36°], latitude interval [−9°, 16°]).

**3.1.** SPATIAL INTERPOLATION. Spatial interpolation is a tool for visualizing the distribution of a variable in space by estimating the value of a variable at any specific location based on a weighted average of the known values at sampled locations. There exist a wide range of spatial interpolation methods. Here, we apply two commonly used deterministic spatial interpolation methods, inverse distance weighting (IDW) and kernel smoothing. Both methods give closer points higher weights, but IDW considers all known data points, while kernel smoothing considers only the neighboring observed data points within a certain window. Another important difference between the two methods is that for the sampled locations IDW produces values identical to the observed values (it is an exact interpolator), while kernel smoothing is comparable to regression in that it may produce values that are different from the observed values (it is an inexact interpolator). As a result, kernel smoothing is better at highlighting the general trends in the spatial structure of the data by smoothing sharp peaks and troughs (see below in this section), while IDW is better at highlighting the finer detail of the spatial structure of the data (cf. §3.5). IDW produces somewhat grainier plots that may occasionally become disturbed by some spurious visualization artifacts because it considers all known data points (cf. Idiatov 2018:139–41).

Figure 5a shows a spatial interpolation plot of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops) produced by means of Gaussian kernel smoothing, while Figure 5b shows a spatial interpolation plot of the $F_{LV}$ frequencies in percentages by means of IDW. Since kernel smoothing is an inexact interpolator, the smoothed $F_{LV}$ frequencies in Fig. 5a reach only 100%, as opposed to the actual maximum of 140% faithfully reflected by the exact interpolator IDW in Fig. 5b.

Despite these minor differences, both plots show the same general structure in the spatial distribution of the $F_{LV}$ frequencies, with kernel smoothing in Fig. 5a being naturally more adapted for visualizing general trends. Thus, the plots clearly distinguish three major regions of high lexical frequency of LV stops: (i) the Gulf of Guinea coast west of Ghana, which we refer to as the UPPER GUINEA HOTBED, (ii) the Gulf of Guinea coast east of Ghana, which we refer to as the LOWER GUINEA HOTBED, and (iii) an area centered on the Ubangi River basin and covering the Central African Republic (CAR) and the north of the Democratic Republic of the Congo (DRC), which we refer to as the

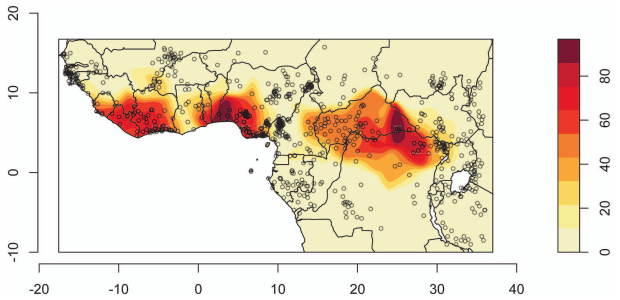



FIGURE 5a. A spatial interpolation plot of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops) produced by means of Gaussian kernel smoothing with the Nadaraya-Watson smoother. The bandwidth of the kernel was optimized to maximize the point process likelihood cross-validation criterion. The languages of the sample are indicated by black circles. The ribbon to the right of the plot shows the color scheme used to represent the smoothed $F_{LV}$ frequencies in percentages.

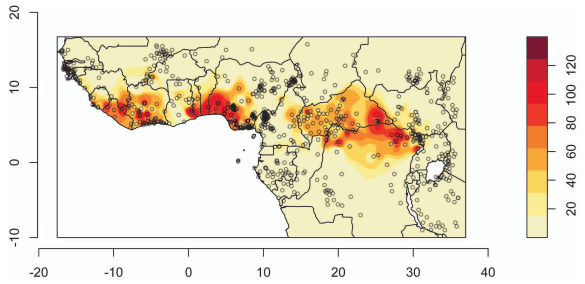



FIGURE 5b. A spatial interpolation plot of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops) produced by means of inverse distance weighting (power = 5). The languages of the sample are indicated by black circles. The ribbon to the right of the plot shows the color scheme used to represent the $F_{LV}$ frequencies in percentages.

UBANGI BASIN HOTBED. The Lower Guinea hotbed and the Ubangi Basin hotbed are separated by a large discontinuity located in Cameroon and the northeast of Nigeria, which we refer to as the CAMEROON GAP. The two hotbeds along the Gulf of Guinea coast are separated from each other by a less important discontinuity situated in and around Ghana, which we refer to as the GHANA GAP. Since in its southern part this discontinuity corresponds mainly to the area of diffusion of Akan [akan1250], a big language without LV stops, one could have assumed that this discontinuity is only apparent due to the accidental presence of a big language without LV stops. However, the two plots clearly show the existence of a good number of languages with low lexical LV frequency north of the Akan-speaking area, which contribute to the emergence of this discontinuity. Furthermore, the kernel-smoothing interpolation in Fig. 5a suggests the existence of two hotbed extensions with LV frequencies that are clearly lower than in the hotbeds from which they protrude. The first extension, which we refer to as the BANFORA EXTENSION, protrudes from the Upper Guinea hotbed into the southeast of Mali and southwest of Burkina Faso, roughly following the Banfora Escarpment. The second, less prominent extension, which we refer to as the DJA-NTEM EXTENSION, protrudes from the Ubangi Basin hotbed into the lowlands of southern Cameroon toward Equatorial Guinea.

What is interesting about these spatial interpolation plots is that they show that the Macro-Sudan belt/Sudanic zone is not a homogeneous area when it comes to the distribution of lexical frequencies of LV stops. In §5, we provide a historical interpretation of its complex internal structure, but first we try to quantify our results in a more rigorous way (§3.2–§3.5) and then cross-validate them with another data set (§3.6).

**3.2.** GENERALIZED ADDITIVE MODELING VS. SPATIAL INTERPOLATION. Generalized additive models (GAMs) are a statistical tool that is particularly well adapted to our data. Originally, GAMs are an extension of multiple regression that provide flexible tools for modeling complex interactions describing wiggly surfaces. Baayen 2013, Tamminga et al. 2016, Winter & Wieling 2016, and Wieling 2018 are useful introductions to using GAMs in linguistics. Examples of the use of GAMs in linguistics in relation to spatial analysis can be found in Wieling et al. 2011, Wieling et al. 2014. In addition to their ability to deal with highly nonlinear data, a great advantage of GAMs is that they are a tool that allows complex data to speak for themselves without having to recode or bin them first. However, the freedom that GAMs offer and their ability to handle very complex data also have a certain side effect. Thus, GAMs do not come with coefficients that can be easily interpreted in a direct way, and their visualization is very important for their evaluation.

The major conceptual difference between generalized additive modeling and the spatial interpolation methods used in §3.1 is that the latter methods produce deterministic models whose results are fully determined by the input values and the mathematical functions used, while the former produces statistical models describing a distribution of possible outcomes. In practical terms with respect to our research, generalized additive modeling generally produces clearer visualizations that importantly are much more stable to variations in coverage of the input data. Thus, as we show in §3.3, even if we remove all languages without LV stops from our data set, or if we subset our frequency data from the full set of 315 languages to only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (see §2.3), generalized additive modeling still produces very similar visualizations (Figs. 7, 8a, 8b) to the visualization based on the full data set (Fig. 6). This provides additional confirmation of the robustness of our findings.

Another practical difference between GAMs and the two spatial interpolation methods is that generalized additive modeling produces quantified results that are useful for at least two reasons. First, GAMs allow us to evaluate in a more objective way the quality of the visualization and the accuracy of the statistical model that produces it. There are two main issues here, which we discuss in §3.4: the level of precision of the models' coefficient estimates that can be realistically expected with our type of data, and whether the produced models are qualitatively robust. Second, as we show in §3.5, the quantified results of GAMs may help us find interesting patterns in the data that would be much more difficult to identify otherwise.

**3.3.** GAM visualizations. The GAMs and their visualizations were produced with the 'mgcv' package for R (Wood 2006, 2019). The GAM visualizations in this article are contour plots representing the regression surface of the lexical LV frequencies ($F_{LV}$) as a function of the combination of longitude and latitude using thin-plate regression splines with the heat-map color scheme. Lighter shades correspond to higher $F_{LV}$ values. Contour lines are isopleths that mark deviations from the mean in terms of standard deviation. The parameter too.far, which controls the size of the area to be plotted around each data point, was set to 0.05 (half of its default value of 0.1) in order to strike a balance between the accurateness in the representation of the spatial continuity between the data points and the ease of perception of the visualization as a whole, avoiding too much patchiness in the contour plot. The remaining parameters of the models, their full textual results, and the residuals plots are provided in the supplementary materials.[7]

Figure 6 is a visualization of the GAM regression surface of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops). The visualization in Fig. 6 is largely comparable to the spatial interpolation plots in Fig. 5a and Fig. 5b and lends itself to observations on the spatial distribution of the $F_{LV}$ frequencies in the languages of our sample, similar to those that were formulated in §3.1. As compared to the spatial interpolation plots, the GAM visualization further confirms the presence of the Ghana gap and particularly highlights the internal structure of the three hotbeds. It also highlights the particular prominence of high $F_{LV}$ values in the Lower Guinea hotbed as compared to the other two hotbeds. Both the Banfora extension and the Dja-Ntem extension are virtually absent from Fig. 6.

The qualitative robustness of the GAM visualization in Fig. 6 is further strengthened by the fact that GAMs based on significantly smaller subsets of the full data set still produce very similar visualizations. Thus, compare the visualization based on the full

[7] The supplementary materials can be accessed online at http://muse.jhu.edu/resolve/115.

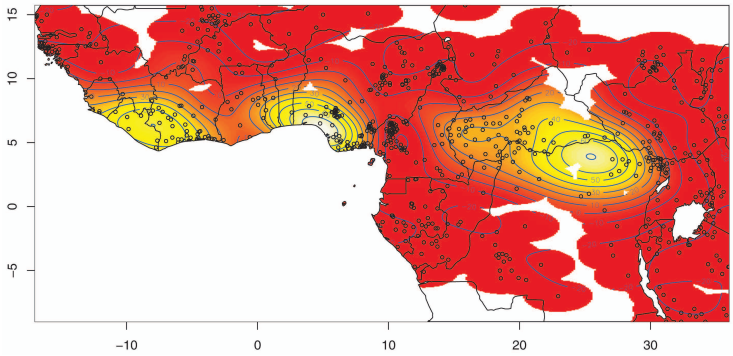



FIGURE 6. The heat-map color scheme contour plot of the GAM regression surface of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: k = 13 (k-index = 0.99, $p = 0.39$, k′ = 195), family = Gaussian, edf = 70.76, deviance explained = 77.60%, AIC = 6048, intercept $F_{LV} = 19.95\%$, $p < 0.001$.

data set in Fig. 6 with Figure 7, which visualizes the GAM regression surface of the $F_{LV}$ frequencies in percentages for only the 315 languages for which we have lexical frequency data. Although Fig. 7 is inevitably less accurate than Fig. 6, it does show the same three hotbeds separated by the same two discontinuities.

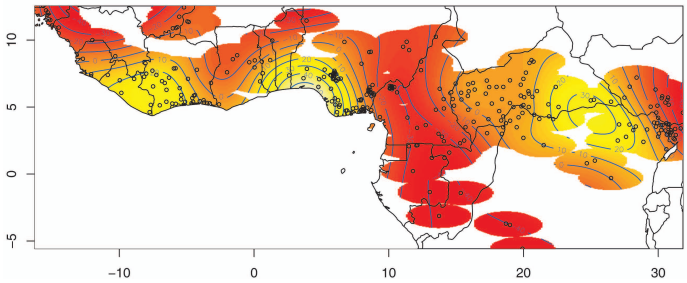



FIGURE 7. The heat-map color scheme contour plot of the GAM regression surface of the $F_{LV}$ frequencies in percentages (excluding languages without LV stops with $F_{LV}$ of 0%) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: k = 15 (k-index = 0.99, $p = 0.41$, k′ = 224), family = Gaussian, edf = 29.83, deviance explained = 56.40%, AIC = 2831, intercept $F_{LV} = 45.92\%$, $p < 0.001$.

Figure 8a plots the GAM regression surface of the $F_{LV}$ frequencies in percentages for the subset of the full data set that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (see §2.3) plus languages without LV stops with $F_{LV}$ of 0%. Figure 8a show the same three hotbeds separated by the same two discontinuities, although the Ubangi Basin hotbed has a simpler shape and structure.

Figure 8b plots the GAM regression surface of the $F_{LV}$ frequencies in percentages for the same subset as in Fig. 8a, but the $F_{LV}$ frequency values are those of the quasi-Swadesh 200 lists (see §2.3). The spatial structure in Fig. 8b is very similar to that in Fig. 8a, but in line with the effects of focusing on the basic vocabulary discussed in §2.3, all three hotbeds become less prominent in terms of their $F_{LV}$ frequency values. Furthermore, the Ubangi Basin hotbed also slightly shrinks in its expanse.

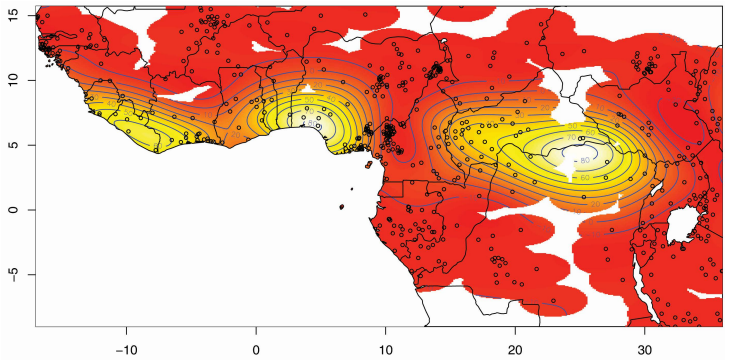



FIGURE 8a. The heat-map color scheme contour plot of the GAM regression surface of the $F_{LV}$ frequencies (in percentages, as a function of the combination of longitude and latitude using thin-plate regression splines) for the subset of the full data set that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries plus languages without LV stops with $F_{LV}$ of 0%. Model summary: k = 11 (k-index = 1.03, $p = 0.76$, k′ = 120), family = Gaussian, edf = 64.54, deviance explained = 76.80%, AIC = 4760, intercept $F_{LV}$ = 13.40%, $p < 0.001$.

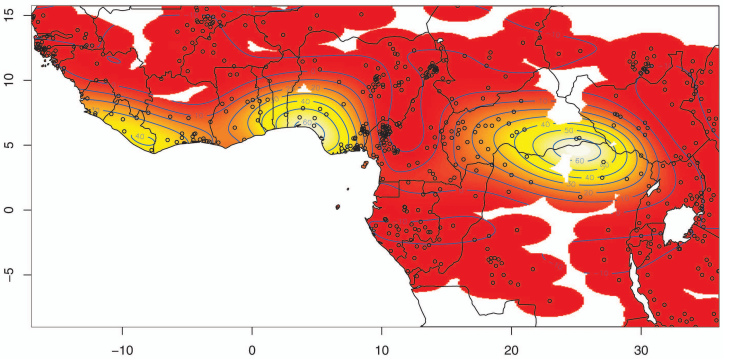



FIGURE 8b. The heat-map color scheme contour plot of the GAM regression surface of the $F_{LV}$ frequencies in the quasi-Swadesh 200 lists (in percentages, as a function of the combination of longitude and latitude using thin-plate regression splines) for the subset of the full data set that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries plus languages without LV stops with $F_{LV}$ of 0%. Model summary: k = 11 (k-index = 1.05, $p = 0.78$, k′ = 120), family = Gaussian, edf = 54.49, deviance explained = 61.30%, AIC = 4823, intercept $F_{LV}$ = 9.74%, $p < 0.001$.

Finally, let us consider Figure 9, which is a visualization of the GAM regression surface of the log-transformed $F_{LV}$ frequencies, including the languages without LV stops.[8] The most important effect of log-transformation for our $F_{LV}$ frequency data is that it transforms the original right-tailed distribution (see Fig. 3) into a left-tailed distribution, spreading out the lower values of $F_{LV}$ frequencies while at the same time condensing the higher values. Due to this effect of zooming in on the lower values of $F_{LV}$ frequencies, the GAM based on the log-transformed $F_{LV}$ frequencies is much better at visualizing the transitions between the hotbeds and the areas without LV stops. At the same time, it levels out the internal structure of the hotbeds and the differences in prominence between the hotbeds, which are better visualized with the nontransformed $F_{LV}$ frequencies in percentages, as in Fig. 6. Figure 9 highlights the transitional nature of the Ghana

[8] Because the logarithm of zero is undefined, we scaled up the $F_{LV}$ values prior to log-transformation by adding 0.83, the minimal $F_{LV}$ value different from zero.

gap. It also confirms the presence of the two hotbed extensions, the Banfora extension of the Upper Guinea hotbed and the Dja-Ntem extension of the Ubangi Basin hotbed. Finally, Fig. 9 clearly suggests the presence of a link between the Ubangi Basin hotbed and the Lower Guinea hotbed along the Benue River Valley, which in Fig. 6 is only vaguely hinted at by the shape of the isopleths. We refer to this link between the two hotbeds as the BENUE RIVER LINK. Another interesting observation with respect to the log-transformed GAM visualized in Fig. 9 is that it explains by far the highest percentage of deviance (even after the somewhat higher number of basis dimensions that it uses is taken into account), viz. 85.8%. In all probability, this higher proportion of the deviance explained is due to the fact that there is more complexity in the spatial distribution of the lower $F_{LV}$ frequencies as compared to the higher $F_{LV}$ frequencies, which are concentrated in the three hotbeds, and by zooming in on the lower $F_{LV}$ frequencies, log-transformation makes it easier for the model to account for the complexity in the spatial distribution of these lower values. All of these properties make the visualization of Fig. 9 the most informative overall, and for this reason we use it in the remainder of the article for reference purposes.

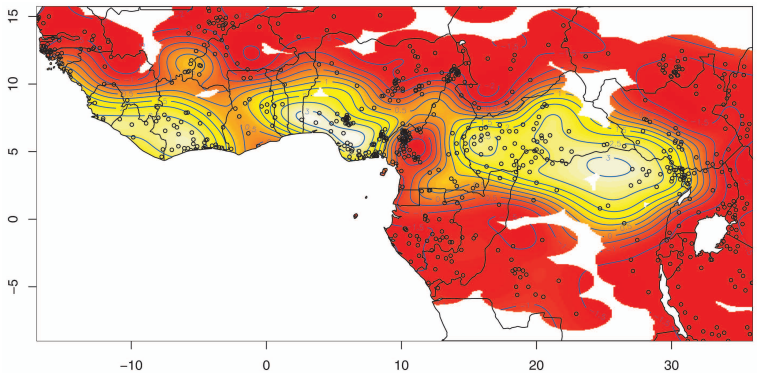



FIGURE 9. The heat-map color scheme contour plot of the GAM regression surface of the log-transformed (after scaling up by 0.83) $F_{LV}$ frequencies (including the languages without LV stops) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: k = 18 (k-index = 1, $p = 0.53$, k′ = 323), family = Gaussian, edf = 108.1, deviance explained = 85.80%, AIC = 1764, intercept log-transformed (after scaling up by 0.83) $F_{LV} = 1.54837$, $p < 0.001$.

**3.4.** MODEL CRITICISM: THE LEVEL OF PRECISION AND QUALITATIVE ROBUSTNESS. The level of precision of the coefficient estimates of a GAM based on the Gaussian distribution depends on how well the model satisfies the assumptions that the residuals are normally distributed and that their variance is constant (homoscedastic) across the values of the linear predictor. For some research questions, such as studies on the significance of minute differences in some particularly fine phonetic phenomena, the precision of the coefficient estimates is very important since even a minor change in the value of the coefficient estimates may affect the essence of our inferences (for instance, see Wieling 2018). A high quantitative precision is much less relevant for the type of data we are looking into here, where a lot of imprecision is inherently built into the data, as both our dependent variable, viz. the estimations of the lexical frequencies of LV stops, and the independent variable, viz. the combination of longitude and latitude values taken to represent the location of the languages of our sample, are necessarily rough approximations. What matters most is the qualitative robustness of our findings, which is confirmed

through cross-validation using different methods (spatial interpolation and GAM), different types of subsamples (see §3.3), and also different types of data (see §3.6).

The coefficient estimates of most GAMs presented in §3.3 are not precise, because two properties of our data cause a violation of the assumptions of normality and homoscedasticity of the residuals. The first relevant property is the presence of a large number of data points with zero values (languages without LV stops). Thus, as illustrated in Figure 10, which shows the four residuals plots for the GAM of the full data set in percentages visualized in Fig. 6 in §3.3, these zero data points form the heavy oblique line in the residuals vs. linear predictor plot (bottom left) and the heavy horizontal line at 0 on the response axis in the response vs. fitted values plot (bottom right), and they strongly disturb the normality of the distribution in the histogram of residuals (top right).[9]

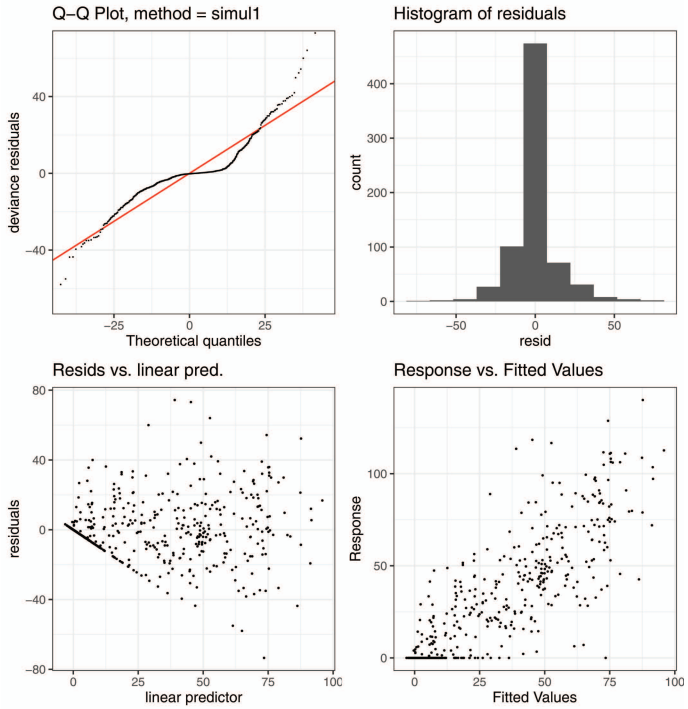



FIGURE 10. The four residuals plots for the GAM of the full data set in percentages visualized in Fig. 6 (§3.3).

It is not surprising then that removing all data points with zero values (languages without LV stops) from the data set, as in the GAM visualized in Fig. 7 in §3.3, significantly improves the distribution of the residuals, as illustrated in Figure 11.

In fact, it would now suffice to remove just six outlier residuals, viz. the four most extreme positive residuals and the two most extreme negative residuals on the residuals vs. linear predictor plot, to achieve normality in the distribution of the residuals (Shapiro-Wilk normality test: W = 0.99168, $p$ = 0.07979).

The outliers in the residuals highlight the second property of our data that causes a violation of the assumptions of normality and homoscedasticity of the residuals, espe-

[9] The combined sets of four GAM residuals plots were produced with the package 'mgcViz' for R (Fasiolo et al. 2018).

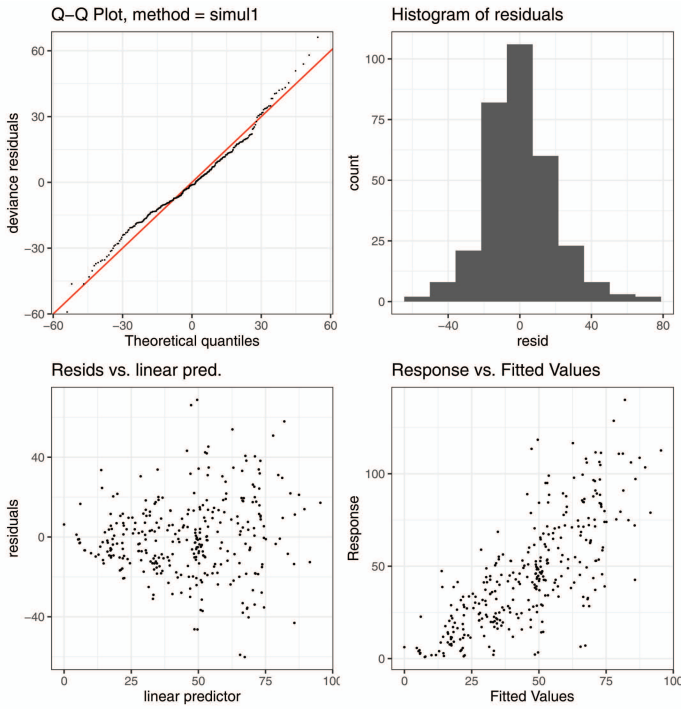



FIGURE 11. The four residuals plots for the GAM of the data set in percentages with all languages without LV stops removed, as visualized in Fig. 7 (§3.3).

cially in the models including languages without LV stops. This property is the presence of cases of abrupt local fluctuations of $F_{LV}$ frequencies, which we discuss next in §3.5.

**3.5.** RESIDUALS OUTLIERS: ABRUPT LOCAL FLUCTUATIONS OF $F_{LV}$ FREQUENCIES AND THEIR INTERPRETATION. Although generalized additive modeling is a good tool for describing wiggly surfaces, it still may have trouble with big, abrupt changes in the value of the dependent variable, viz. $F_{LV}$ frequency, against only very minor changes in the value of the independent variable, viz. the combination of longitude and latitude. As a result, abrupt local jumps or dips in the values of the dependent variable may result in outliers in the residuals of the regression surface produced by a GAM. Figure 12 is a version of the residuals vs. linear predictor plot for the GAM of the full data set in percentages (visualized in Fig. 6 in §3.3) presented in Fig. 10 (§3.4) that highlights some such more extreme values of the residuals (marked with triangles instead of circles). Table 2 links the indexes used in Figure 12 with the names of the respective languages.

In fact, the presence of the data points that are likely to cause difficulties for generalized additive modeling because they bring in abrupt local fluctuations of $F_{LV}$ frequencies can to a certain extent already be gleaned from the simple spatial interpolation plot of the $F_{LV}$ frequencies in percentages by means of inverse distance weighting in Fig. 5b (§3.1), as this method of spatial interpolation is good at highlighting the finer detail of the spatial structure of the data. To illustrate this point and to show the location of the languages in question within NSSA, we reproduce Fig. 5b here as Figure 13 and mark the more extreme values of the residuals in the residuals vs. linear predictors plot in Fig. 12 with triangles (see Table 2 for the meaning of the indexes).

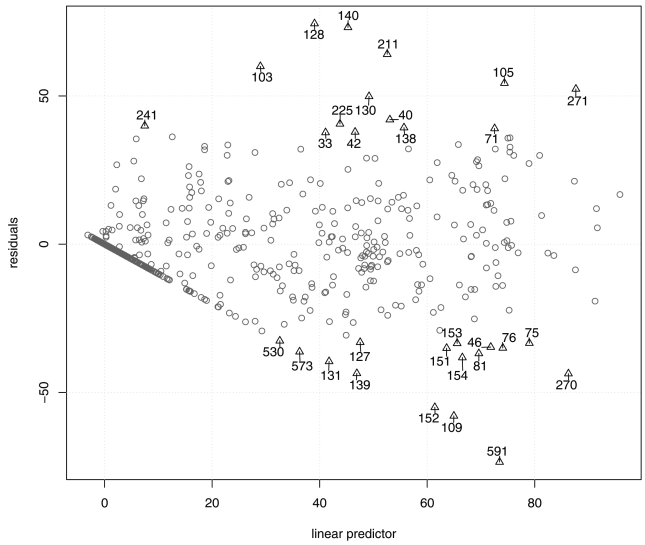



FIGURE 12. The residuals vs. linear predictor plot for the GAM of the full data set in percentages visualized in Fig. 6 (§3.3) with some of the more extreme values of the residuals highlighted as triangles as opposed to the circles for the less extreme values of the residuals. The indexes are explained in Table 2.

| INDEX | LANGUAGE | GLOTTO-CODE | GENUS | $F_{LV}$ (%) | INDEX | LANGUAGE | GLOTTO-CODE | GENUS | $F_{LV}$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| 33 | Monzombo | monz1249 | Mundu-Baka | 78.75 | 138 | ItuMbuso | itum1245 | Cross River | 94.99 |
| 40 | Sherbro | sher1258 | Mel | 95.06 | 139 | Okobo | okob1241 | Cross River | 3.34 |
| 42 | Mbandja | mban1263 | Bandaic | 84.44 | 140 | Oro | oroo1241 | Cross River | 118.39 |
| 46 | Nzakara | nzak1247 | Zandic | 37.13 | 151 | Ikaan | ukaa1243 | Ukaan | 28.57 |
| 71 | Daloa Bete | dalo1238 | Kru | 111.56 | 152 | Iyinnọ | ukaa1243 | Ukaan | 6.40 |
| 75 | Guere | weno1238 | Kru | 45.63 | 153 | Iigau | ukaa1243 | Ukaan | 32.15 |
| 76 | Jrwe | yrew1238 | Kru | 39.08 | 154 | Ishẹu | ishe1239 | Ukaan | 28.32 |
| 81 | Wobe | weno1238 | Kru | 32.76 | 211 | Logba | logb1245 | Kwa | 116.62 |
| 103 | Lendu | lend1245 | Lendu | 88.94 | 225 | Hai | haii1241 | Bandaic | 84.34 |
| 105 | Birri | birr1240 | Birri | 128.68 | 241 | Baka | baka1272 | Mundu-Baka | 47.39 |
| 109 | Gouro | guro1248 | Southeastern Mande | 7.05 | 270 | Eruwa | eruw1238 | Edoid | 42.61 |
| 127 | Ebughu | ebug1241 | Cross River | 14.46 | 271 | Isoko | isok1239 | Edoid | 140.00 |
| 128 | Efai | efai1241 | Cross River | 113.44 | 530 | Akan | akan1250 | Kwa | 0.00 |
| 130 | Ekit | ekit1246 | Cross River | 99.03 | 573 | Ebira | ebir1243 | Nupoid | 0.00 |
| 131 | Enwang | enwa1245 | Cross River | 2.22 | 591 | Ikwere | ikwe1242 | Igboid | 0.00 |

TABLE 2. The explanation of the indexes of the more extreme values of the residuals highlighted as triangles in Fig. 12.

These abrupt local fluctuations of $F_{LV}$ frequencies that we can pinpoint with such ease by using the residuals vs. linear predictor plot of a GAM are particularly interesting for the analysis of the data for two reasons. First, they may highlight the sources in our data that may be giving us less accurate estimates of the $F_{LV}$ frequencies. The estimates from these sources need to be cross-validated with better sources, if available. The major possible origin of such inadequate estimates is the particularly small size of some of the sources. When a source is small, even minor changes may have a significant effect on its $F_{LV}$ value. Thus, an accidental increase or decrease in the number of roots with an LV stop by one or two roots is likely to have little effect on the $F_{LV}$ value of a source with many entries, while the consequences may be much more significant in a source with a small

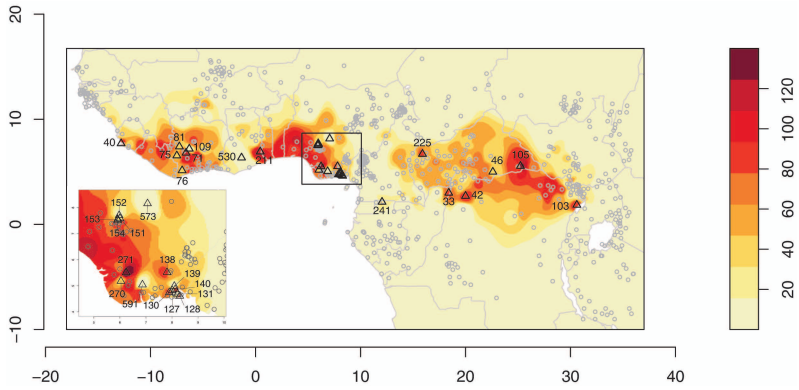



FIGURE 13. A spatial interpolation plot of the $F_{LV}$ frequencies in percentages (including 0% for languages without LV stops) produced by means of inverse distance weighting (power = 5), similar to Fig. 5b. The triangles here mark the data points that correspond to the more extreme values of the residuals in the residuals vs. linear predictors plot in Fig. 12. See Table 2 for the meaning of the indexes.

number of entries. Among the languages listed in Table 2, this possibility needs to be considered for Nzakara, whose source has just 225 entries, Monzombo, whose source has just 240 entries, the two Edoid languages, as they have around 172 and 177 entries, the four Ukaan languages, as our Ukaan sources have around 200 words per source, and three of the four Kru languages, viz. Guere, Jrwe, and Wobe, as all four Kru sources have just around 300 words per source, but the three languages in question have a significantly lower $F_{LV}$ then the surrounding languages in our data set.

Second, besides such potential imperfections in the data used, the abrupt local fluctuations of the $F_{LV}$ values highlighted by the GAM may also provide a window on the actual spatial and temporal dynamics as they most likely reflect relatively shallow historical events. If such abrupt fluctuations had happened at more significant time depths, we would expect them to have been more smoothed out by now. There are two major types of spatiotemporal processes that are apparent in our data, viz. recent events of local loss or emergence of LV stops and recent events of language spread or relocation that may have brought together languages with significantly different $F_{LV}$ profiles (including the complete absence of LV stops) in one small area. As we briefly discuss in the remainder of this section, these processes may produce two types of patterns in the spatial distribution of the $F_{LV}$ values, which can be described as positive or negative spikes and cliffs, respectively.

If the process of a partial or complete loss of LV stops happens to affect only one language in the sample within an area with relatively high $F_{LV}$ values, it will result in a negative spike in the $F_{LV}$ values at this data point. For example, this is the reason behind the abrupt dip in the $F_{LV}$ value down to zero with the Igboid language Ikwere, surrounded by data points with high $F_{LV}$ values, which results in the most extreme (negative) outlier in our GAM.[10] A less radical outlier in the GAM results is represented by the Nupoid language Ebira. In the dialect of Ebira in our sample, LV stops similarly have recently changed to labial stops (cf. Scholz 1976:8),[11] while the languages around Ebira in our sample have relatively high $F_{LV}$ values. A local negative spike in the $F_{LV}$

[10] In fact, Ikwere is not the only Igboid language that has undergone the process of partial or complete loss of LV stops (cf. Blench & Williamson 2016:13), but it is the only such language in our sample in this area.

[11] In this respect, note an alternative spelling of the name of this language, *Egbira*, that reflects the presence of an LV stop in other dialects.

values may also reflect the situation where LV stops have developed within a group of related languages spoken in the same general area, but one language has been trailing behind in this evolution. In our data, an example of such evolution may be represented by the Iyinno variety of Ukaan. Possibly the same explanation applies to the negative spike in the $F_{LV}$ values in the Edoid language Eruwa, whose relatively low $F_{LV}$ value furthermore contrasts sharply with a local positive spike in the $F_{LV}$ values in the neighboring Edoid language Isoko. As mentioned above, however, in both the Ukaan and the Edoid case these abrupt local fluctuations of the $F_{LV}$ values may eventually turn out to be due to the small size of the respective sources. A local spike in the $F_{LV}$ values may also reflect a recent migration of a language into an area with a different $F_{LV}$ profile. This is the explanation behind the local positive spike in the $F_{LV}$ values produced by the Bandaic language Hai in the northwest of the CAR. The $F_{LV}$ value of Hai is comparable to that of other Bandaic languages in our sample but is significantly higher than the $F_{LV}$ values of the languages in our sample that surround Hai, and we know that the speakers of Hai migrated to their current location from an area at least 500 km to the east in the first half of the nineteenth century (cf. Moñino 2004:28). The positive spike in the $F_{LV}$ values in the southeastern corner of the CAR produced by Birri may be particularly relevant from the historical perspective because Birri is a potential isolate. Unfortunately, very little is currently known about this language and its history (cf. Güldemann 2018b: 269, 359).

In southeastern Nigeria close to the border with Cameroon (see the inset in Fig. 13), we observe an unusual crisscross pattern of positive and negative spikes in the $F_{LV}$ values within a very limited area produced by several Cross River languages of the Lower Cross branch. This crisscross pattern is a result of a process of loss of LV stops that has affected several of these languages, such as Ebughu, Enwang, and Okobo to various degrees, while high $F_{LV}$ values have been preserved in the remaining languages, such as Efai, Ekit, ItuMbuso, and Oro.[12] The process of loss of LV stops among Lower Cross languages must have occurred relatively recently, as it has not yet affected all of these languages to a similar extent.

Transitions between the hotbeds of high $F_{LV}$ values to the areas without LV stops are generally gradual, resembling a slope of a hill. However, in a number of cases we observe transitions that are abrupt, more like the face of a cliff. Like positive and negative spikes in the $F_{LV}$ values, such cliffs cause difficulties for GAMs, as apparent in the residuals vs. linear predictors plot. Like spikes, cliffs may also reflect a variety of relatively recent historical events. For example, a number of Kwa languages in our data, such as Akan, that have partially or completely lost LV stops, thus contributing to the emergence of the Ghana gap, happen to border on the Kwa languages from the western periphery of the Lower Guinea hotbed with high $F_{LV}$ values, such as Logba, with a cliff-like fluctuation in the $F_{LV}$ values as a result. See §5.3 for a discussion of the population movements that may have contributed to the emergence of this $F_{LV}$ cliff. A comparable example is provided by the Southeastern Mande language Guro, which appears to have lost a good deal of its LV stops and has experienced an event of rather significant spread within the northeastern periphery of the Upper Guinea hotbed, both relatively recently. Another $F_{LV}$ cliff is found at the southeastern extremity of the Ubangi Basin hotbed, where the Lenduic language Lendu, with a high $F_{LV}$ value, located to the west of the

[12] Consider, for example, the word 'leopard', reconstructed by Connell (1991) for Proto-Lower Cross as **é-kpè* because its reflexes have *kp* in almost all Lower Cross languages, while Ebughu has *é-píè*, Enwang *é-pè*, and Okobo *é-pì*.

mountain range on the border between the DRC and Uganda, neighbors Bantu languages with no LV stops to the east and south, and Moru-Madi and Nilotic languages with low $F_{LV}$ values or no LV stops to the east and north. All of these neighbors are known to have moved into the area to the east of the mountain range from either the south (Bantu) or the north (Nilotic and Moru-Madi).

**3.6.** Cross-validating the results: LV stops in african toponyms. The fine-grained spatial analysis of lexical frequencies of LV stops in African languages that we presented in the previous sections was made possible by the existence of the RefLex database. RefLex provided us with lexical data of sufficient quantity and quality for sixty percent of the 545 languages in our general sample that have LV stops. Unfortunately, that leaves out forty percent of the relevant languages, either because no appropriate lexical source exists for them, or because no appropriate source has yet been included in RefLex. Moreover, the size of the sources we did have at our disposal varies greatly, and some areas and families are relatively better represented than others. We therefore found it useful to cross-validate our findings with an alternative data set, viz. the African toponyms included in the GeoNames database (GeoNames.org n.d.). More precisely, we studied the spatial distribution of unique toponyms spelled with an LV stop (e.g. *kp*, *gb*, or Yoruba *p*),[13] assuming that the frequency of LV stops in this part of the lexicon should roughly correlate with their frequency in the general lexicon. The results are shown in Figure 14, which we can compare with the GAM visualization in Fig. 9, reproduced here in the inset.

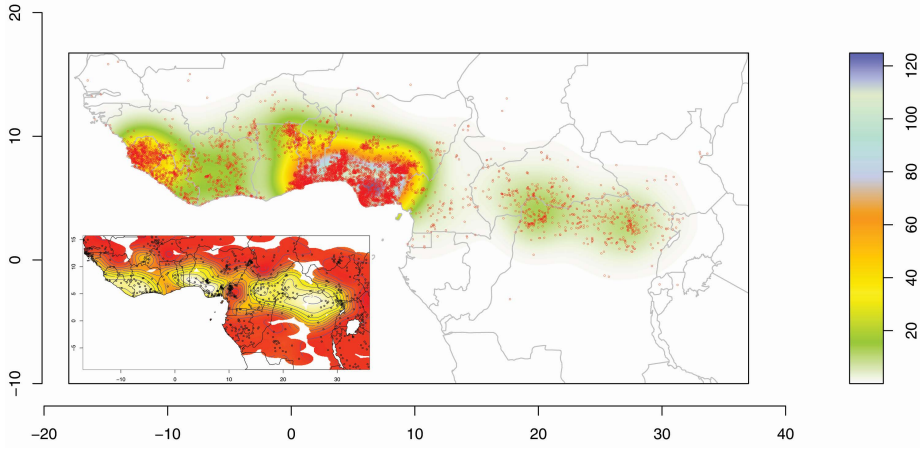



Figure 14. The geographical distribution of the unique toponyms spelled with an LV stop in the GeoNames database (GeoNames.org) (red circles) and their spatial intensity. For comparison, the inset reproduces Fig. 9, visualizing the GAM of the log-transformed $F_{LV}$ values.

As we can see, the spatial distribution of toponyms containing an LV stop in Fig. 14 is characterized by the presence of three regions of high spatial intensity of toponyms with

[13] There are quite a few false toponym doublets in the GeoNames database, many more so than true toponym doublets. To eliminate the effect of false toponym doublets in our data set, we arbitrarily kept only the first data point for any set of toponym doublets. We also removed all toponyms spelled with *gb* or *kp* in South Africa and Namibia, such as *Springbok*, since these toponyms of Germanic origin do not contain an LV stop. At the same time, we did not take the trouble to manually remove the few toponyms spelled with *gb* and *kp* on the northern periphery of NSSA which cannot contain an LV stop either, since there are no languages with LV stops in these areas (and there have not been any since the start of the European colonization either, when the first maps of the region started to be produced).

an LV stop that clearly correspond to the three hotbeds of high lexical frequency of LV stops in the GAM visualization. As in our lexical-frequency data set, we can again observe a stronger connection between the Upper Guinea hotbed and the Lower Guinea hotbed and a much weaker link between the Lower Guinea hotbed and the Ubangi Basin hotbed. One major difference is that neither the two hotbed extensions nor the Benue River link are visible in the toponyms data set. However, at least for the Benue River link this discrepancy is merely apparent and is due to two factors. First, the data coverage of the GeoNames database is deficient for toponyms of the Benue River Valley area in Nigeria.[14] Second, and more importantly, large swathes of land along the Benue River link have been settled in the last few centuries by speakers of Fula [fula1264], a language without LV stops that was originally spoken in Senegal and neighboring regions. The Fula incursion into the Benue River link was both a migration of nomadic herders and a military and religious expansion, the result of which was the creation of the Adamawa Fula emirate, which controls most of the relevant region. Given the dominant sociopolitical status of the Fula in much of the region, it is highly likely that many of the earlier toponyms in the areas currently inhabited by the Fula were replaced by new Fula ones, or at least adapted to Fula pronunciation. Furthermore, an important number of toponyms in the areas still inhabited by speakers of minority languages that often have LV stops appear on official maps in simplified forms without LV stops.[15]

Like in the lexical-frequency data set (cf. Fig. 6 in §3.3), in the toponyms data set the Lower Guinea hotbed is particularly prominent as compared to the other two hotbeds. At the same time, the Ubangi Basin hotbed looks much weaker in the toponyms data set in Fig. 14 than in the lexical-frequency data set. In order to appreciate the true prominence of the Ubangi Basin hotbed in the toponyms data set, we need to plot the geographical distribution of the unique toponyms spelled with an LV stop against the distribution of the toponyms spelled without an LV stop, as in Figure 15. Figure 15 clearly suggests that the seeming weakness of this hotbed in the toponyms data set in Fig. 14 is an artifact of the general low population density in Central Africa, as reflected in the low density of populated places (compare Idiatov 2018:147 on a roughly similar area). Figure 15 similarly confirms the significance of both discontinuities between the three hotbeds.

**4.** C-EMPHASIS PROSODY CAN EXPLAIN THE EMERGENCE, SPREAD, AND INTRALINGUISTIC DISTRIBUTION OF LV STOPS. In §2.3 we showed that LV stops are less common in the basic vocabulary represented in Swadesh 200 lists than in the general vocabulary of individual languages, and that within Swadesh 200 lists, LV stops are particularly rare in function words, and significantly less common in nominal and verbal concepts as compared to more expressive qualifying and quantifying concepts. We argued that this is an indirect quantitative confirmation of the observation that LV stops are relatively more frequent in expressive parts of the vocabulary. Another type of intralinguistic skewing in the distribution of LV stops that has often been noted in the literature is their strong tendency to be restricted to stem-initial position (see Connell 1994:468, Cahill 2018: 151). This tendency is a specific case of a more general tendency of phonotactic skewing in the languages of NSSA, which is manifested in a decreasing number of phonological oppositions and an increasing application of lenition rules as one moves away

[14] Our comparison of a few areas in the Benue River Valley in the GeoNames database versus on 1:100,000 maps by Nigeria Federal Surveys (1958–1973), such as sheet 175 *Shellen*, suggests that the coverage of the GeoNames database may sometimes go as low as 50% of the toponyms for this part of Nigeria.

[15] See for example Shimizu 1979:64 on Mumuye place names with /k͡p/ misrepresented as *p* on the maps.

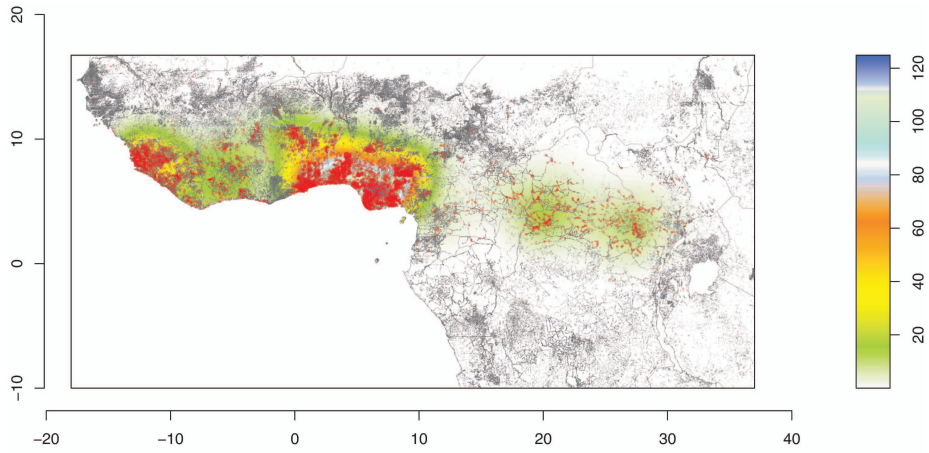



FIGURE 15. The geographical distribution of the unique toponyms spelled with an LV stop (darker red circles), with their spatial intensity, and without an LV stop (gray circles) in the GeoNames database (GeoNames.org).

from the stem-initial consonant position ($C_1$), whether to the right or—in languages with prefixes—to the left (e.g. Hyman 2004:80–81, Lionnet & Hyman 2018:652–55).

Both types of intralinguistic skewing can be explained by what we call C-EMPHASIS PROSODY, a type of word- or utterance-level prosodic prominence whose primary phonetic correlate is consonant length. In a number of Northwestern Bantu languages, this phenomenon has been described in terms of word accent on the initial consonant of stems ($C_1$), by Paulian (1975) for Kukuya [teke1280] and by Van de Velde (2008) for Eton [eton1253]. Figure 16 illustrates such a word accent realized by the length of the stem-initial consonant in the nonsense word *mà-màmà* as pronounced by an Eton speaker, where *mà-* is a noun class prefix and *-màmà* a nonsense stem.

The initial findings of our ongoing research in a number of West and Central African languages strongly suggest that C-emphasis prosody is in origin an utterance-level prosodic/intonational phenomenon marking a particular emphasis on a given element within the utterance (cf. Idiatov & Van de Velde 2016). Consonant lengthening facilitates the emergence of LV stops out of labialized velars in two ways. First, the extra articulatory effort needed for its production increases the chances of overshoot in the realization of the labial approximation of a labialized velar as a full closure (compare Bybee & Easterday 2019:294–95 on the articulatory basis of the strengthening of glides through overshoot). Second, the extra duration increases the chances that the initial velar gesture does not end up being masked by the subsequent labial gesture. The intonational use of C-emphasis prosody as a means of marking emphasis on an element within the utterance accounts naturally for the preponderance of LV stops in expressive parts of the vocabulary. Moreover, the intonational use of C-emphasis prosody and the link between LV stops and expressivity enhances the borrowability of both (see Matras 2009:231 on the position of prosodic features in the borrowability hierarchy within phonology).

**5.** HISTORICAL INTERPRETATION OF THE FACTS.

**5.1.** THE HOTBEDS ARE RETENTION ZONES. We now know that, within the general area where LV stops are found, there are three separate hotbeds where they are normal phonemes, surrounded by areas with languages in which their lexical frequency is very low. Arguably, LV stops and/or the phonetic features that facilitate their emergence

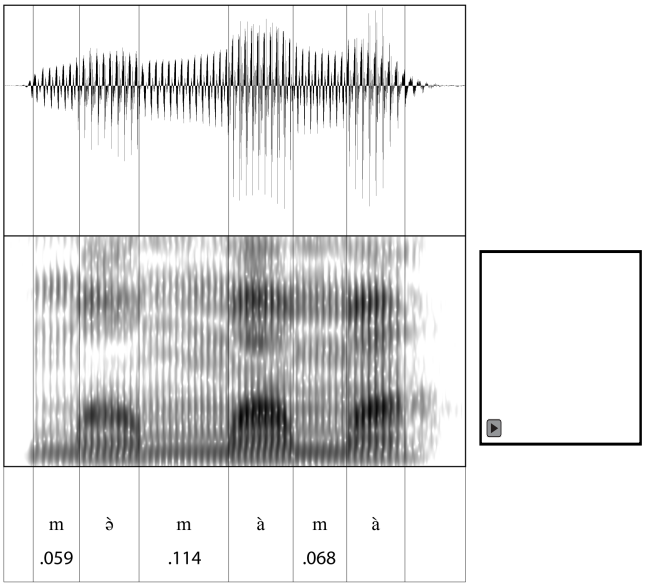



Figure 16. A random repetition by an Eton speaker of the nonsense word *mə-màmà*, where *mə-* is a noun class prefix and *-màmà* a nonsense stem. The duration of the three *m* consonants is measured in seconds.[16]

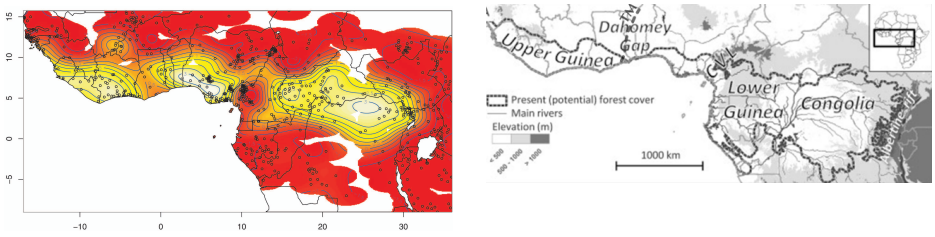
must have greater time depth in the populations that currently occupy the hotbeds than in the surrounding populations. Given the rarity of LV stops in the languages of the world outside of NSSA, it is highly unlikely that the three contemporary hotbeds correspond to areas where LV stops have emerged independently. Instead, they must be areas currently occupied by populations who have retained the feature. In contrast, the areas surrounding the hotbeds must be occupied by populations that have acquired LV stops more recently, or that have less perfectly retained the feature.

The geographical position of the hotbeds is consistent with their identification as areas of retention. All three of them are typical refuge zones, where refuge should be understood negatively in the sense of a place of last resort.[17] The two western hotbeds are tropical forest zones delimited in the south by the Atlantic Ocean, in the west by the mountain ranges of the Guinea Highlands, and in the east by the mountain ranges of the Cameroon Volcanic Line at the border between Nigeria and Cameroon (cf. Figures 17a and 17b). They are separated from each other by the Ghana gap. The Ghana gap roughly corresponds to the so-called Dahomey Gap, a zone of wooded savanna that interrupts the coastal rainforest and that became established around 4,500 BP at the onset of the late Holocene (see Salzmann & Hoelzmann 2005). The match with the Dahomey Gap is closer on the eastern border of the Upper Guinea hotbed where there are no other major topographic boundaries. The western border of the Lower Guinea hotbed is delimited by the Togo Mountains, which cut through the Dahomey Gap, separating the lower-lying River Volta Basin to their west from a plateau sloping gradually toward the

[16] A sound file accompanying this figure is included with the electronic versions of this article, and can be accessed along with the other supplementary materials at http://muse.jhu.edu/resolve/115.

[17] A refuge zone is characterized by environmental conditions that present significant subsistence challenges, whose negative effect is offset by the fact that these challenges make it also less attractive for outsiders. From the perspective of language dynamics, refuge zones can be both 'residual (or accretion) zones' and 'spread zones', in the terms of Nichols (1992).

coast to their east. The Dahomey Gap has the same type of vegetation and climate as the areas with low lexical frequency of LV stops that surround the two western hotbeds. Like the two other hotbeds, the southern part of the Ubangi Basin hotbed in Central Africa is a tropical forest zone, which goes from a forest-savanna mosaic into the lowland and swamp forests of the Congo Basin. Its northern part is a kind of geographical cul-de-sac with lots of marshy and seasonally flooded areas. Both ecological features contribute to the low population density that we already highlighted in Fig. 15 in §3.6.[18] In the west, the Ubangi Basin hotbed is delimited by the highlands on the border between Cameroon and the CAR, which form an eastern extension of the Adamawa Plateau and separate the Ubangi River catchment area from that of the Sangha River, two tributaries of the Congo River. In the east, this hotbed is delimited by the mountain ranges of the Congo-Nile Divide on the borders between the CAR and the DRC with South Sudan and Uganda.




a. b.

FIGURE 17. (a) The same as Fig. 9, visualizing the GAM of the log-transformed $F_{LV}$ values. (b) Guineo-Congolian forest delimitation, subdivision, and topography (adapted with permission from Hardy et al. 2013). CVL: Cameroonian Volcanic Line. TM: Togo Mountains.

The historical scenario that emerges from these facts is one in which populations that were adapted to living in a savanna habitat and that spoke languages without LV stops migrated southward, where they encountered populations of speakers of languages that had LV stops and/or the phonetic features that facilitate their appearance, which for brevity's sake we refer to as PRIMARY LV POPULATIONS. Migration was easiest and fastest in the savanna areas. The marginal presence of LV stops in these areas is most likely due to borrowing in the first place. However, it may also be due to incorporation of smaller subgroups of primary LV populations (e.g. through sex-biased intermarriage)[19] and occasional later migrations in the opposite direction out of the refuge zones (cf. §5.3). The newly arrived populations spread much more slowly in the tropical forest refuge zones, where the high lexical frequency of LV stops is much more likely the result of a language shift of primary LV populations to the languages of the newcomers. It is an instance of 'shift-induced substrate interference', in the terms of Thomason & Kaufman 1988.

[18] As discussed by Idiatov (2018:146–51), besides being sparsely populated, this region of Central Africa is also rather homogenous linguistically. It is occupied by a small number of linguistic groups that are all rather shallow. Furthermore, most, if not all, of these groups are very likely to have moved into this region of Central Africa relatively recently.

[19] This possibility is suggested, for instance, by some recent interdisciplinary work on the transfer of click stops from the so-called Khoisan languages, where clicks are regular phonemes, to a small group of Bantu languages in southwestern Zambia, such as Fwe [fwee1238], where clicks are very marginal. See Pakendorf 2014:634–35 for a concise summary and Bostoen & Sands 2012 and Barbieri et al. 2013 for more details.

Typically, substrate effects on a target language are directly proportional with the ratio of the shifting population among its speakers (Thomason 2017). Furthermore, although so-called borrowed phonemes may be transferred along with lexical borrowings containing them, such phoneme transfer is usually believed to be subject to strong constraints, which can be overruled only through intense contact and a high degree of bilingualism (cf. Winford 2003:54–56, 2005, Dimmendaal 2011:182). Based on a similar line of argument, Bostoen and Donzo (2013:458) argue, for example, with respect to the emergence of LV stops in Ngombe, a Bantu language of the DRC, that 'the integration of labial-velar stops [in Ngombe] could only happen through advanced Bantu-Ubangi bilingualism/multilingualism, probably accompanied by language shift of entire Ubangi language groups to neighbouring Bantu languages involving phonological substrate influence of their first language on the target language'.

Chances are very high that LV stops and/or the phonetic features that facilitate their emergence originated in a language family or families that have disappeared today. Thus, according to a recent appraisal of the status of language classification in Africa by Güldemann (2018b), the overwhelming majority of the language groups that are confined to the hotbeds are either 'robust' or 'promising' members of the Niger-Congo family. The only 'weak' candidate for Niger-Congo affiliation in the hotbeds is Ijoid, a small language group spoken in the Niger delta. A few language groups found on the eastern fringes of the Ubangi Basin hotbed are members of the Central Sudanic family. Importantly, some groups from the same Central Sudanic family and numerous groups from the same Niger-Congo family are also found outside of the hotbeds, and these often lack LV stops altogether. As mentioned in §3.5, the language with a very high lexical frequency of LV stops that may be particularly interesting in this area is Birri, as it is a potential isolate.

The fact that most language groups currently found both within and outside of the hotbeds in NSSA are Niger-Congo clearly indicates that the populations that spoke languages without LV stops, which were adapted to living in a savanna habitat and migrated southward, were by and large speakers of Niger-Congo languages. Interpreting the spatial distribution of high lexical frequencies of LV stops allows us to formulate detailed hypotheses regarding prehistoric migration patterns of Niger-Congo-speaking populations. By way of an example, we concentrate on the oft-discussed low-level subgroup of Bantoid in the following section.

**5.2.** Scenarios for the Bantu expansion. Our data on the lexical frequency of LV stops has particularly interesting historical implications for Bantoid, a major subgroup of the Benue-Congo branch of Niger-Congo, and especially for its biggest subgroup, (Narrow) Bantu, currently spread over a large part of the continent, from northeastern Nigeria and Kenya in the north all the way down to South Africa. Together with the unrelated Chadic languages and a number of small related Niger-Congo language groups currently subsumed under the label of Adamawa, Bantoid languages are the languages responsible for the existence of the Cameroon gap between the Lower Guinea and Ubangi Basin hotbeds. Furthermore, an overwhelming majority of the Bantu languages are spoken outside of the hotbeds of high lexical frequency of LV stops, with the noticeable exception of the Bantu languages in the north of the DRC and Congo, which participate in the Ubangi Basin hotbed. This general distribution suggests that Bantoid populations must have passed the area of the Benue River link on their way south, without much interaction with the primary LV populations. This passage must have occurred somewhere in the timeframe between around 4,500 BP, the presumed period of the initial diversification of

Bantu, and around 6,900 BP, the presumed period of the initial diversification of Bantoid as a whole (cf. Bostoen et al. 2015, Grollemund et al. 2015).

It is currently assumed that the initial diversification of both Bantoid and later Bantu happened in the general region of the Grassfields Plateau in western Cameroon (see Greenberg 1972, Bostoen et al. 2015, Grollemund et al. 2015). However, we believe that it is more likely to have started in a more northerly location, possibly somewhere to the north of the western end of the Adamawa Plateau closer to the Alantika Mountains. The reason is that in the Early Holocene, from about 11,000 to 6,000 BP, the African rainforest extended as far north as the Adamawa Plateau and possibly the middle Benue Valley, well beyond the Cameroonian Grassfields, and it is only by the end of this climatic optimum that the forest started fragmenting on the Adamawa Plateau (see Vincens et al. 2010). That is, given the preference of the Benue-Congo groups in general and Bantoid groups in particular for savanna environments,[20] any significant migration of these populations further south before the onset of the forest fragmentation is less likely. This interpretation also matches better the paleoanthropological data from the only two Grassfields sites for the relevant periods of the Early and Middle Holocene, namely the Mbi Crater and Shum Laka rock shelters (Asombang 1988, Orban et al. 1996, Lavachery 2001, Lipson et al. 2019).[21]

After their initial diversification, most Bantoid groups remained clustered in the same general area. The major exception is represented by the Bantu expansion, one of the biggest known expansion events in the history of the continent. Two principal scenarios for the Bantu expansion have been proposed in the literature, 'East out of West' and 'East separate from West', which differ in whether they consider Eastern Bantu a later offshoot from a Western Bantu node or a primary branch of the Proto-Bantu node, respectively (for an overview see Bostoen & Grégoire 2007, Pakendorf et al. 2011, Bostoen et al. 2015, Grollemund et al. 2015). The general migration routes associated with the two scenarios are illustrated in Figures 18a and 18b, while Figure 18c shows the approximate location of the major Bantu subgroups referred to in the text.

Our data on the lexical frequency of LV stops clearly support the 'East out of West' model of the Bantu expansion, with the Eastern Bantu break-off point somewhere south of the rainforest. According to the alternative 'East separate from West' model, the Proto-Eastern Bantu populations would have migrated to the north of the Congo Basin rainforest, which would be right through the core of the Ubangi Basin hotbed of high lexical frequency of LV stops, as illustrated by Figure 19. If this had been the case, we should have found many Eastern Bantu languages with LV stops and with a relatively high lexical frequency of those. However, LV stops are particularly marginal in Eastern Bantu, as can be easily observed by comparing the Eastern Bantu domain in Fig. 18c (where it is labeled as F) with the GAM visualization of the spatial distribution of the LV frequencies in Fig. 9, reproduced here in Figure 19.

[20] Thus, as concluded by Grollemund et al. (2015:13299) with respect to the Bantu expansion that followed the initial diversification of Bantoid, 'the Bantu expansion was characterized by a measureable preference for following familiar savannah habitats … [and] avoided rainforest habitats', and presently 'rainforest-dwelling Bantu cultures … retained some cultural knowledge of how to exploit the savannah environment'.

[21] The adult skeleton from the Mbi Crater site, dated to around 9,000–8,400 BP, has a height comparable to that of the modern-day Pygmy populations, while the adult skeletons from the two funeral phases at the Shum Laka site, dated to around 8,000–7,500 BP and 3,900–3,000 BP, respectively, vary in their height between that of the modern-day Pygmy populations and that of the modern-day Bantu populations of the region. However, Lipson et al. (2019) report that their analysis of the genome-wide DNA data from the individuals buried at Shum Laka from 8,000–3,000 BP shows that their ancestry profiles are more similar to West-Central African hunter-gatherers. In this respect, the height variation in the Shum Laka individuals may provide additional evidence for the possibility, suggested by Skoglund et al. (2017:67, e14), that the shorter stature of present-day rainforest hunter-gatherer populations is a 'relatively recent evolution'.


a. East out of West. b. East separate from West. c. Major Bantu subgroups.

FIGURE 18. (a)–(b) The two prevailing models of the Bantu expansion. 1: Proto-Bantu nucleus, 2: Western Bantu nucleus, 3: Eastern Bantu nucleus. (c) Approximate location of the major Bantu subgroups: A: Northwestern, B: Lebonya-Boan, C: Inner Congo Basin, D: West-Coastal, E: Southwestern, F: Eastern. The gray area in (a) and (b) shows the Bantu-speaking area, while in (c) it marks the part of Southern Africa outside of the Bantu domain. The tree symbols in all three figures show the current extent of the rainforest. (Adapted with permission from Pakendorf et al. 2011:55, 57.)

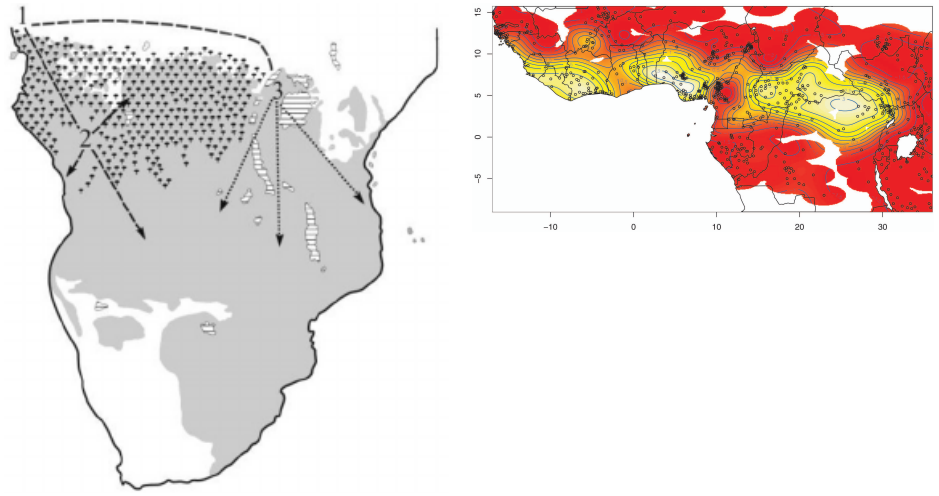
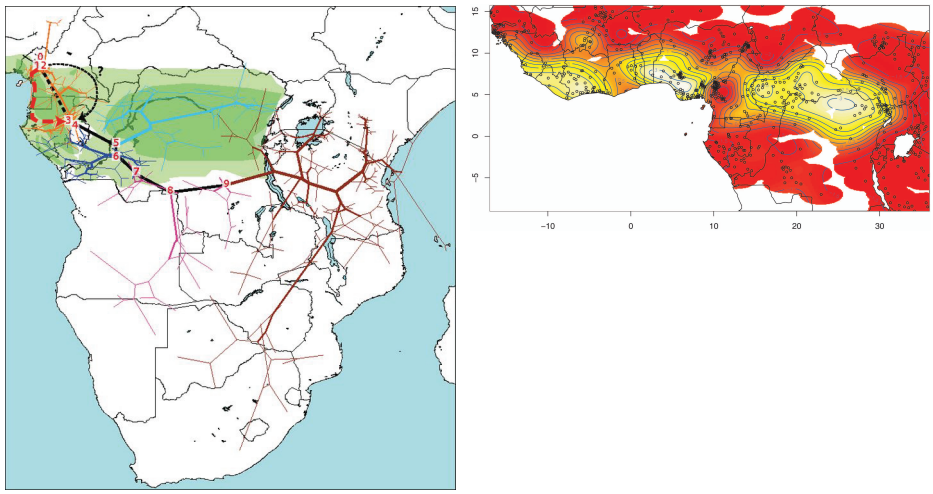



FIGURE 19. The 'East separate from West' hypothesis of the Bantu expansion (adapted with permission from Pakendorf et al. 2011:57), compared to the visualization of the spatial distribution of the lexical frequencies of LV stops in Fig. 9 (the GAM using the log-transformed $F_{LV}$ values).

The 'East out of West' scenario for the Bantu expansion has been recently further supported by the results of a number of interdisciplinary studies combining phylogenetic linguistic, population genetic, archaeological, and paleoclimatic data (de Filippo et al. 2012, Bostoen et al. 2015, Grollemund et al. 2015). Bostoen et al. (2015) and Grollemund et al. (2015) also propose a detailed reconstruction of the Bantu migration route, illustrated in Figure 20a. Bostoen et al. (2015) focus on the route that the Bantu populations took to first cross the equatorial rainforest. They argue that this early migration proceeded through savanna corridors of the Sangha River Interval. To the north of the rainforest around the Sanaga-Mbam confluence area, these savanna corridors started opening around 4,000–3,500 BP, but the passage through the core of the rainforest in the Sangha River Interval itself was freed only by ~2,500 BP. This suggested migration route through savanna corridors is indicated in Fig. 20a with a black curved dashed arrow joining nodes 2 and 3 with a question mark next to it. However, the mi-

gration route through the Sangha River Interval does not square well with our data on the lexical frequency of LV stops (and incidentally, with our data on C-emphasis prosody), because it would have the early Bantu populations pass through the western periphery of the Ubangi Basin hotbed of high lexical frequency of LV stops (cf. Figure 20b). Accordingly, we would have expected to find a significant number of Bantu languages with LV stops to the south of the rainforest among West-Coastal, Southwestern, and Eastern Bantu. Yet this expectation is not borne out, as can be observed by comparing Fig. 20a and Fig. 20b. Our own data on the lexical frequency of LV stops suggest that the migration between nodes 2 and 3 is more likely to have happened through the savannas on the coastal plains that opened from 4,000 BP onward. This possible passage is indicated in Fig. 20a with a red curved larger-dashed arrow along the coast. In this respect, it is telling that the Bantu languages presently spoken along the coast of southwestern Cameroon and Gabon, viz. most Bantu A20 and A30 languages and Bantu B10 languages, look much more like Eastern Bantu languages in their phonology and morphosyntax. All either lack or have a low frequency of LV stops, and have no or limited C-emphasis prosody.




a. b.

FIGURE 20. (a) Bantu migration route reconstructed by Grollemund et al. (2015) on the consensus tree by using geographical locations of contemporary languages and connecting ancestral locations by straight lines (adapted with permission from Grollemund et al. 2015:13298). Numbered positions correspond to major diversification nodes on the consensus tree. Lighter green shading corresponds to the delimitation of the rainforest at 5,000 BP; the darker green corresponds to the delimitation of the rainforest at 2,500 BP. The black curved smaller-dashed arrow indicates the migration route through the Sangha River Interval proposed by Bostoen et al. (2015). The red curved larger-dashed arrow indicates the migration route through the savannas on the coastal plains that better matches our data on the lexical frequency of LV stops. (b) The same as Fig. 9, visualizing the GAM of the log-transformed $F_{LV}$ values.

From the perspective of our data, the Sangha River Interval route was rather a secondary route taken later (from ~2,500 BP) by the ancestral populations of the Bantu groups A80 and A90, the other clad that split off at node 2. These Bantu languages must have been affected by shift-induced substrate interference by LV populations. Some of these Bantu A groups that took the Sangha River Interval route are likely to have later shifted to the West-Coastal Bantu languages of the Bantu groups B50–B80 spoken in the area centered on the Bateke Plateau, the highland savannas in the east of Gabon and

the center of Congo. This is suggested by the presence of C-emphasis prosody (see §4) in some of these languages, such as the B70 language Kukuya [teke1280] (Paulian 1975). Our data on the frequency of LV stops and C-emphasis prosody further suggest that a number of other Northwestern Bantu groups are likely to have initially taken a similar eastbound migration route toward Central Africa, which probably lay somewhat more northerly along the Adamawa Plateau, but then either, as Bantu A70, turned back in a southwestern direction along the northern fringes of the tropical forest or, as Jarawan Bantu, turned back toward the Northwest along the Benue River into northeastern Nigeria.[22] Bostoen et al. (2015:365) also briefly entertain the possibility that the Bantu populations took a coastal route to first cross the equatorial rainforest, but ultimately reject it in favor of the Sangha River Interval scenario because of lack of sufficient archaeological evidence on the existence of village communities on the coastal plains between 4,000 and 3,000 BP. At the same time, as Bostoen et al. (2015:366) also acknowledge, no archaeological data on the existence of village communities is available for the Sangha River Interval for the period ~2,500 BP either.

**5.3.** LV STOPS SHOULD NOT BE RECONSTRUCTED INTO PROTO-NIGER-CONGO AND PROTO-CENTRAL SUDANIC. LV stops have been claimed to be reconstructable into the proto-languages of the major subbranches of Niger-Congo and in Proto-Niger-Congo itself in a number of publications summarized and endorsed by Cahill (2017, 2018). These claims are generally based on the correct but historically irrelevant observation that LV stops can be found in many of the daughter languages of these families. However, the geographical distribution of high lexical frequencies of LV stops strongly suggests that they should not be reconstructed into the proto-language of Niger-Congo or its major subbranches, except in those that may have been spoken in one of the hotbeds. Niger-Congo languages expanded into the hotbeds from the north. The further contemporary Niger-Congo languages are removed from the hotbeds, the less likely they are to have LV stops, showing that they must have acquired the feature as they spread into the hotbeds. The same considerations apply to Central Sudanic.

Comparative evidence confirms this. In the majority of cases of cognacy between an LV stop and another consonant, this other consonant is a velar stop that is either labialized or followed by a rounded vowel. Reconstructing such cognate sets with an LV stop would involve lenition and loss of the labial release, but for phonotactic and perceptual reasons such an evolution is highly unlikely. The labial release of LV stops is perceptually more salient (cf. Ladefoged & Maddieson 1996:336–39, Connell 1994, Cahill 2018), which makes its loss in favor of the velar articulation unlikely. A generalized loss of the labial release in the languages outside of the hotbeds is even more unlikely because LV stops are typically restricted to stem-initial position, which is often also word-initial. In word-initial position, it is precisely the labial gesture that is more likely to mask the velar gesture because in LV stops the labial release follows the velar release. In contrast, reconstructing such cognate sets with a velar stop naturally explains why LV stops tend to be restricted to stem-initial ($C_1$) position, as the prosodic prominence of this position facilitates their emergence (see §4). The inverse scenario calls for the lenition of inherited LV stops in exactly the prosodic environment where lenition is least likely to occur.

[22] With respect to the proposed Jarawan migration route, note that two Jarawan populations, Nagumi [nagu1244] and Mbonga [mbon1252], are reported to have been found in the beginning of the twentieth century along this route in northern and eastern Cameroon (close to the border with the CAR), respectively (see Maddieson & Williamson 1975). The two languages appear to already be extinct. Oral traditions claiming a downstream migration along the Benue River have also been reported for Mbula, another Jarawan population in northeastern Nigeria (Meek 1931:57–68).

For all of these reasons, we strongly side with the reconstructions positing the emergence of LV stops from labialized velars, such as Creissels 2004 for Manding languages [mand1435] and Hyman 2011:13–14 for Bantu languages.

Most convincing examples of the loss of LV stops involve cognate sets in which LVs correspond to labial consonants, due to the higher perceptual prominence of the labial release in LV stops. They tend to be rarer, are mostly found outside of the hotbeds, and tend to be due to recent evolutions. In our scenario, the loss of LV stops is most likely to come about when speaker communities move out of the hotbeds and incorporate significant numbers of speakers of languages without LV stops. Akan [akan1250] and Supyire [supy1237] are interesting potential examples. Both are spoken outside of the West African hotbeds and are unusual in their lower-level linguistic groups for their lack of LV stops (according to Cahill 2018:155, this is a result of a merger of LV stops with labials in these languages).[23]

An even more spectacular example of the correlation between location in the hotbeds and presence of LV stops is represented by the Salaga variety of Dendi [dend1243], as presented by Zima (1985). In this variety, LV stops were acquired upon moving into a hotbed and then lost again upon a subsequent migration outside of the hotbed during the mid-nineteenth century. Dendi is the southernmost member of the Songhay family whose other members are spoken in the Sahel, mostly along the Niger River bend. The core of the Dendi domain is situated to the north of the Lower Guinea hotbed at the border region between Niger and Benin. When a community of Dendi speakers moved south to the town of Djougou in western Benin, inside of the Lower Guinea hotbed, LV stops developed in this variety of Dendi by a regular sound change out of labialized velars. After a subsequent migration in the mid-nineteenth century, a part of the Dendi community from Djougou settled in the town of Salaga in central Ghana, inside the Ghana gap of low lexical frequency of LV stops. By the 1980s, LV stops in the Salaga variety were preserved only in the speech of older speakers in a limited number of fixed expressions, while outside of these fixed expressions speakers of all ages use the corresponding labialized velars in the same words. This seeming reversal of a regular sound change suggests that at the moment of the migration from Djougou to Salaga, the sound change from labialized velars to LV stops was not yet complete and there was still variation in the realization between LV stops and labialized velars.

**6.** Conclusions. The skewed areal distribution of LV stops has been known for decades and has been included in sets of features used to characterize large linguistic areas in Northern Sub-Saharan Africa, such as the Sudanic zone (Clements & Rialland

[23] Supyire, a Senufo language, and Akan, a Potou-Tano language, are both situated outside of the hotbeds, and their respective families are likely to have expanded northward away from the Upper Guinea hotbed. Thus, the Senufo domain as a whole is generally south-north oriented, largely following the Banfora extension of the Upper Guinea hotbed (see §3.1), while Supyire is also one of the most northern Senufo languages. The Potou-Tano domain has a general southwest-northeast orientation, extending from the southeast of Côte d'Ivoire (in the Upper Guinea hotbed) to the north of Togo and Benin. The area of the greatest linguistic diversity within Potou-Tano (its center of gravity) and presumably its homeland lies in the southwest of its domain. The northward expansion of the Tano branch of Potou-Tano (the branch including Akan) into the present-day Dahomey Gap is likely to be related to a temporary return to more humid climatic conditions and a renewed spread of forests into the savanna in this area from c. 3,300–1,100 BP (see Salzmann & Hoelzmann 2005). Besides negatively affecting the frequency of LV stops, the northward expansion of Tano also appears to have negatively affected the frequency of implosives in the phonological inventories of these languages, as suggested by the data in Clements & Rialland 2008:58. The Tano expansion itself may be largely responsible for a major gap in the distribution of another prominent areal feature in NSSA, viz. clause-final negation marking (cf. Idiatov 2018:145).

2008) or the Macro-Sudan belt (Güldemann 2008). The recent development of the large lexical database RefLex allowed us to go beyond listing the languages that have LV stops in their phoneme inventories and to estimate the degree of lexical entrenchment of these phonemes in a very large sample of 315 languages that have them. As we had expected, based on our knowledge of individual languages, LV stops turned out to be marginal phonemes in a sizeable proportion of our sample. When we subsequently studied the geographical distribution of lexical frequencies of LV stops, an unexpected and highly interesting pattern emerged of three hotbeds of high lexical LV frequency surrounded by areas of low lexical LV frequency. These hotbeds cut across genealogical groupings and are most straightforwardly characterized in geographical terms: they have low elevation and forest or swampy habitats, and they are separated from each other by areas with higher elevation and/or a savanna habitat. This detailed picture clearly points to the origin and current distribution of LV stops as a substrate phenomenon: a feature of language communities who found refuge in the hotbeds when Niger-Congo and Central Sudanic languages were spreading southward and who eventually shifted toward the languages of the incoming groups. This scenario implies that LV stops were not part of the phoneme inventories of the proto-languages of the families currently attested in NSSA, which is in line with purely linguistic arguments in terms of the perceptual and phonotactic naturalness of the evolution of labialized velars toward labial-velars in word- or stem-initial position, versus the opposite evolution.

Our data on the spatial distribution of high lexical frequencies of LV stops also allowed us to formulate detailed hypotheses regarding prehistoric migration patterns of Niger-Congo-speaking populations. In particular, we were able to adjust and refine the scenarios proposed in the literature for the Bantu expansion, one of the biggest language expansion events in recent human history.

Finally, our quantitative data allowed us to strengthen the empirical basis of the claim that LV stops are more frequent in expressive parts of the vocabulary than in the general vocabulary. We did this indirectly by showing that they are less frequent in the basic, non-expressive vocabulary of Swadesh 200 lists than in the general vocabulary. This distributional fact has a common explanation with the seemingly unrelated fact that LV stops are preponderant in stem-initial position. We argue that both are due to C-emphasis prosody, the prosodic prominence of stem-initial consonants, whose typical phonetic correlate is consonant length. Our ongoing research shows that C-emphasis prosody is used at the utterance level in many languages of NSSA to mark particular emphasis on a given element.

## REFERENCES


ASOMBANG, RAYMOND NEB'ANE. 1988. *Bamenda in prehistory: The evidence from Fiye Nkwi, Mbi Crater and Shum Laka rockshelters*. London: University of London dissertation.

BAAYEN, R. HARALD. 2013. Multivariate statistics. *Research methods in linguistics*, ed. by Robert J. Podesva and Devyani Sharma, 337–72. Cambridge: Cambridge University Press.

BADDELEY, ADRIAN, and ROLF TURNER. 2005. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(6).1–42. DOI: 10.18637/jss.v012.i06.

BARBIERI, CHIARA; ANNE BUTTHOF; KOEN BOSTOEN; and BRIGITTE PAKENDORF. 2013. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *European Journal of Human Genetics* 21.430–36. DOI: 10.1038/ejhg.2012.192.

BLENCH, ROGER, and KAY WILLIAMSON. 2016. A reconstruction of the phonology of Proto-Igboid. Cambridge: McDonald Institute for Archaeological Research, MS. Online: http://llacan.vjf.cnrs.fr/nigercongo2/discussions/Proto-Igboid_phonology.pdf.

Bostoen, Koen; Bernard Clist; Charles Doumenge; Rebecca Grollemund; Jean-Marie Hombert; Joseph Koni Muluwa; and Jean Maley. 2015. Middle to Late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of western Central Africa. *Current Anthropology* 56(3).354–84. DOI: 10.1086/681436.

Bostoen, Koen, and Jean-Pierre Donzo. 2013. Bantu-Ubangi language contact and the origin of labial-velar stops in Lingombe (Bantu, C41, DRC). *Diachronica* 30(4).435–68. DOI: 10.1075/dia.30.4.01bos.

Bostoen, Koen, and Claire Grégoire. 2007. La question bantoue : Bilan et perspectives. *Mémoires de la Société de Linguistique de Paris* 15.73–91.

Bostoen, Koen, and Bonny Sands. 2012. Clicks in south-western Bantu languages: Contact-induced vs. language-internal lexical change. *Proceedings of the 6th World Congress of African Linguistics, Cologne 2009*, ed. by Matthias Brenzinger and Anne-Maria Fehn, 129–40. Cologne: Rüdiger Köppe.

Bybee, Joan, and Shelece Easterday. 2019. Consonant strengthening: A crosslinguistic survey and articulatory proposal. *Linguistic Typology* 23(2).263–302. DOI: 10.1515/lingty-2019-0015.

Cahill, Michael. 2008. Why labial-velar stops merge to /gb/. *Phonology* 25(3).379–98. DOI: 10.1017/S0952675708001541.

Cahill, Michael. 2017. Labial-velars: A questionable diagnostic for a linguistic area. *Proceedings of the 8th World Congress of African Linguistics, Kyoto 2015*, ed. by Shigeki Kaji, 13–24. Tokyo: Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Cahill, Michael. 2018. Labial-velars of Africa: Phonetics, phonology, and historical development. *The Routledge handbook of African linguistics*, ed. by Augustine Agwuele and Adams Bodomo, 150–67. Abingdon: Routledge.

Clements, G. N., and Annie Rialland. 2008. Africa as a phonological area. *A linguistic geography of Africa*, ed. by Bernd Heine and Derek Nurse, 36–85. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486272.004.

Connell, Bruce. 1991. *Phonetic aspects of the Lower Cross languages and their implications for sound change*. Edinburgh: University of Edinburgh dissertation. Online: http://hdl.handle.net/1842/19642.

Connell, Bruce. 1994. The structure of labial-velar stops. *Journal of Phonetics* 22(4).441–76. DOI: 10.1016/S0095-4470(19)30295-5.

Creissels, Denis. 2004. L'occlusive vélaire sonore *g* et les labio-vélaires (*w*, *gw*, *kw*, *gb*, *kp*) en mandingue. *Mandenkan* 39.1–22.

de Filippo, Cesare; Koen Bostoen; Mark Stoneking; and Brigitte Pakendorf. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences* 279(1741).3256–63. DOI: 10.1098/rspb.2012.0318.

Dimmendaal, Gerrit J. 2011. *Historical linguistics and the comparative study of African languages*. Amsterdam: John Benjamins.

Dockum, Rikker, and Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16.35–54. Online: http://www.elpublishing.org/PID/168.

Fasiolo, Matteo; Raphaël Nedellec; Yannig Goude; and Simon N. Wood. 2018. Scalable visualisation methods for modern generalized additive models. arXiv:1809.10632 [stat.ME]. Online: https://arxiv.org/abs/1809.10632.

GeoNames.org. n.d. GeoNames database dump. Online: http://download.geonames.org/export/dump/, accessed March 14, 2016.

Greenberg, Joseph H. 1972. Linguistic evidence regarding Bantu origins. *The Journal of African History* 13.189–216. DOI: 10.1017/S0021853700011427.

Greenberg, Joseph H. 1983. Some areal characteristics of African languages. *Current approaches to African linguistics*, vol. 1, ed. by Ivan R. Dihoff, 3–21. Dordrecht: Foris.

Grollemund, Rebecca; Simon Branford; Koen Bostoen; Andrew Meade; Chris Venditti; and 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).13296–301. DOI: 10.1073/pnas.1503793112.

Güldemann, Tom. 2008. The Macro-Sudan belt: Towards identifying a linguistic area in Northern Sub-Saharan Africa. *A linguistic geography of Africa*, ed. by Bernd Heine and Derek Nurse, 151–85. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486272.006.

Güldemann, Tom (ed.) 2018a. *The languages and linguistics of Africa*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110421668.

Güldemann, Tom. 2018b. Historical linguistics and genealogical language classification in Africa. In Güldemann 2018a, 58–444. DOI: 10.1515/9783110421668-002.

Hammarström, Harald; Robert Forkel; and Martin Haspelmath (eds.) 2019. *Glottolog 4.1*. Jena: Max Planck Institute for the Science of Human History. Online: http://glottolog.org.

Hardy, Olivier J.; Céline Born; Katarina Budde; Kasso Daïnou; Gilles Dauby; Jérôme Duminil; Eben-Ezer B. K. Ewédjé; et al. 2013. Comparative phylogeography of African rain forest trees: A review of genetic signatures of vegetation history in the Guineo-Congolian region. *Comptes Rendus Geoscience* 345(7).284–96. DOI: 10.1016/j.crte.2013.05.001.

Hulstaert, Gustaaf. 1957. *Dictionnaire lɔmɔ́ngɔ-français*. 2 vols. Tervuren: Musée Royal du Congo Belge.

Hulstaert, Gustaaf. 1961. *Grammaire du lɔmɔ́ngɔ, vol. 1: La phonologie*. Tervuren: Musée Royal de l'Afrique Centrale.

Hulstaert, Gustaaf. 1965. *Grammaire du lɔmɔ́ngɔ, vol. 2: La morphologie*. Tervuren: Musée Royal de l'Afrique Centrale.

Hulstaert, Gustaaf. 1966. *Grammaire du lɔmɔ́ngɔ, vol. 3: La syntaxe*. Tervuren: Musée Royal de l'Afrique Centrale.

Hyman, Larry M. 2004. How to become a 'Kwa' verb. *Journal of West African Languages* 30(2).69–88. Online: https://main.journalofwestafricanlanguages.org/index.php/downloads/download/101-volume-30-number-2-new/424-how-to-become-a-kwa-verb.

Hyman, Larry M. 2011. The Macro-Sudan belt and Niger-Congo reconstruction. *Language Dynamics and Change* 1(1).3–49. DOI: 10.1163/221058211X570330.

Idiatov, Dmitry. 2018. An areal typology of clause-final negation in Africa: Language dynamics in space and time. *Aspects of linguistic variation*, ed. by Daniël Van Olmen, Tanja Mortelmans, and Frank Brisard, 115–63. Berlin: De Gruyter Mouton. Online: 10.1515/9783110607963-005.

Idiatov, Dmitry, and Mark L. O. Van de Velde. 2016. Stem-initial accent and C-emphasis prosody in north-western Bantu. Paper presented at the 6th International Conference on Bantu Languages, Helsinki. Online: http://idiatov.mardi.myds.me/talks/2016_BANTU6_C-emphasis_Idiatov_Van_de_Velde_SLIDES.pdf.

Ladefoged, Peter, and Ian Maddieson. 1996. *The sounds of the world's languages*. Oxford: Blackwell.

Lavachery, Philippe. 2001. The Holocene archaeological sequence of Shum Laka rock shelter (Grassfields, Western Cameroon). *African Archaeological Review* 18(4).213–47. DOI: 10.1023/A:1013114008855.

Lionnet, Florian, and Larry M. Hyman. 2018. Current issues in African phonology. In Güldemann 2018a, 602–708. DOI: 10.1515/9783110421668-006.

Lipson, Mark; Mary Prendergast; Isabelle Ribot; Carles Lalueza Fox; and David Reich. 2019. Ancient human DNA from Shum Laka (Cameroon) in the context of African population history. Paper presented at the 84th annual meeting of the Society for American Archaeology, Albuquerque, NM.

Maddieson, Ian. 2011. Presence of uncommon consonants. *The world atlas of language structures online*, ed. by Matthew Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://wals.info/feature/19A.

Maddieson, Ian. 2018. Phonetics and African languages. In Güldemann 2018a, 546–601. DOI: 10.1515/9783110421668-005.

Maddieson, Ian, and Kay Williamson. 1975. Jarawan Bantu. *African Languages/Langues Africaines* 1.124–63.

Martin, Marieke. 2015. Wawa ideophone phonetics vs. Wawa phonology. Paper presented at the World Conference of African Linguistics 8, Kyoto.

Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press.

Meek, Charles K. 1931. *Tribal studies in northern Nigeria*. 2 vols. London: Kegan Paul, Trench, Trubner & Co.

Moñino, Yves. 2004. Prête-moi ta langue, que je dise un mot : Emprunts banda au gbaya. *Langues et cultures : Terrains d'Afrique (hommage à France Cloarec-Heiss)*, ed. by Pierre Nougayrol and Pascal Boyeldieu, 25–31. Louvain: Peeters.

Moran, Steven; Daniel McCloy; and Richard Wright (eds.) 2014. *PHOIBLE online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://phoible.org, accessed July 15, 2015.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nigeria Federal Surveys. 1958–1973. Nigeria 1:100.000. Lagos: Nigeria Federal Surveys.

Orban, Rosine; Isabelle Ribot; Sylvie Fenaux; and Pierre de Maret. 1996. Les restes humains de Shum Laka (Cameroun, LSA-Age du fer). *Anthropologie et Préhistoire* 107.213–25.

Pakendorf, Brigitte. 2014. Historical linguistics and molecular anthropology. *The Routledge handbook of historical linguistics*, ed. by Claire Bowern and Bethwyn Evans, 627–41. London: Routledge. DOI: 10.4324/9781315794013.ch29.

Pakendorf, Brigitte; Cesare de Filippo; and Koen Bostoen. 2011. Molecular perspectives on the Bantu expansion: A synthesis. *Language Dynamics and Change* 1(1).50–88. DOI: 10.1163/221058211X570349.

Paulian, Christiane. 1975. *Le kukuya : Langue teke du Congo*. Paris: SELAF.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org/.

RStudio Team. 2016. RStudio: Integrated development for R. Boston: RStudio, Inc. Online: http://www.rstudio.com/.

Salzmann, Ulrich, and Philipp Hoelzmann. 2005. The Dahomey Gap: An abrupt climatically induced rain forest fragmentation in West Africa during the late Holocene. *Holocene* 15(2).190–99. DOI: 10.1191/0959683605hl799rp.

Scholz, Hans Jürgen. 1976. *Igbira phonology*. Dallas: SIL International.

Segerer, Guillaume, and Sébastien Flavier. 2011–2021. *RefLex: Reference lexicon of Africa*. Version 1.1. Online: http://reflex.cnrs.fr/.

Shimizu, Kiyoshi. 1979. *A comparative study of the Mumuye dialects (Nigeria)*. Berlin: Reimer.

Skoglund, Pontus; Jessica C. Thompson; Mary E. Prendergast; Alissa Mittnik; Kendra Sirak; Mateja Hajdinjak; Tasneem Salie; Nadin Rohland; et al. 2017. Reconstructing prehistoric African population structure. *Cell* 171(1).59–71.e21. DOI: 10.1016/j.cell.2017.08.049.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4).452–63. Online: https://www.jstor.org/stable/3143802.

Tamminga, Meredith; Christopher Ahern; and Aaron Ecay. 2016. Generalized additive mixed models for intraspeaker variation. *Linguistics Vanguard* 2(s1).1–9. DOI: 10.1515/lingvan-2016-0030.

Thomason, Sarah G. 2017. On establishing ancient shift-induced interference: Problems and prospects. Paper presented at the workshop Language Shift and Substratum Interference in (Pre)history, Max Planck Institute for the Science of Human History, Jena.

Thomason, Sarah G., and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Van de Velde, Mark L. O. 2008. *A grammar of Eton*. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110207859.

Vincens, A.; G. Buchet; M. Servant; and ECOFIT Mbalang Collaborators. 2010. Vegetation response to the 'African Humid Period' termination in Central Cameroon (7° N)—New pollen insight from Lake Mbalang. *Climate of the Past* 6(3).281–94. DOI: 10.5194/cp-6-281-2010.

Vogler, Pierre. 2014. La formation des labiales-vélaires à double occlusion en Niger-Congo. Illkirch-Graffenstaden, ms. Online: https://hal.archives-ouvertes.fr/hal-01183115/document, accessed October 19, 2018.

Westermann, Diedrich. 1911. *Die Sudansprachen: Eine sprachvergleichende Studie*. Hamburg: L. Friederichsen.

Westermann, Diedrich. 1927. *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu*. Berlin: De Gruyter.
Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70.86–116. DOI: 10.1016/j.wocn.2018.03.002.
Wieling, Martijn; Simonetta Montemagni; John Nerbonne; and R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and Standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90(3).669–92. DOI: 10.1353/lan.2014.0064.
Wieling, Martijn; John Nerbonne; and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9):e23613. DOI: 10.1371/journal.pone.0023613.
Winford, Donald. 2003. *An introduction to contact linguistics*. Malden, MA: Blackwell.
Winford, Donald. 2005. Contact-induced changes: Classification and processes. *Diachronica* 22(2).373–427. DOI: 10.1075/dia.22.2.05win.
Winter, Bodo, and Martijn Wieling. 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution* 1(1).7–18. DOI: 10.1093/jole/lzv003.
Wood, Simon N. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall–CRC.
Wood, Simon N. 2019. mgcv: Mixed GAM computation vehicle with automatic smoothness estimation. Online: http://CRAN.R-project.org/package=mgcv.
Zima, Petr. 1985. Labiovelar stops in the Djougou Dendi dialect of Songhay. *Acta Universitatis Carolinae – Philologica 3: Phonetica pragensia* 7.97–104.

LLACAN – UMR 8135 du CNRS
7, rue Guy Môquet – BP 8
94801 Villejuif Cedex, France
[dmitry.idiatov@cnrs.fr]
[mark.vandevelde@cnrs.fr]

[Received 6 August 2019;
revision invited 21 December 2019;
revision received 22 June 2020;
accepted pending revisions 18 August 2020;
revision received 9 September 2020;
accepted 9 September 2020]