Title: **Linguistic Diversity and Human Migrations in Gabon** [1]

Authors: Franz Manni (1)* and John Nerbonne (2)

(1) CNRS UMR 7206, Département « Homme et Environnement », Musée de l'Homme, National Museum of Natural History, University Paris-Diderot, Paris, France.
(2) Germanistische Linguistik, Albert Ludwigs University of Freiburg, Freiburg im Breisgau, Germany.

(*) *Corresponding author:*
Franz Manni,
CNRS UMR 7206, National Museum of Natural History of Paris
Musée de l'Homme
17, place du Trocadéro
75116 Paris - France
0033 1 44 05 73 01 ; franz.manni@mnhn.fr

Short title: Linguistic Diversity in Gabon

Keywords: Africa, Gabon, Bantu languages, Bantu dispersal, linguistics, Levenshtein distance, shared vocabulary, human migrations, phylogeny.

[1] A longer version of this book chapter appears in the doctoral dissertation in linguistics by F. Manni (2017).

ABSTRACT

Gabon is an African country located very close to the homeland of Bantu languages (Cameroun). Starting about 5,000 years ago, Bantu-speaking populations diffused into almost all sub-Saharan Africa. By processing with computational linguistic methods (Levenshtein distance) two independently-collected lexical datasets recording the pronunciation of 88 and 158 words in more than 50 linguistic varieties spoken in Gabon, we obtain a numerical classification of the major linguistic groups. We compare this classification to available ones based on historical linguistics methods (cognate-sharing defined by experts), and find them overlapping, which indicates that the two methods capture the same signal of linguistic difference (and relatedness). To focus on the historical relatedness between major linguistic clusters, we control for the linguistic similarity related to contact, proportional to geographic vicinity, and suggest that the first Bantu-speaking groups to people Gabon where those speaking KOTA-KELE (B20) languages. The other varieties concern five different immigration waves (B10; B30; B40; B50-B60-B70 – Guthrie nomenclature) that penetrated Gabon later in history. To conclude, we suggest a peopling scenario that incorporates available paleoclimatic, archaeological and population genetic evidence.

1. INTRODUCTION

1.1 Bantu languages and classifications

Bantu languages belong to the Niger-Congo phylum and include about 600 varieties spoken in almost all of sub-Saharan Africa. Their geographic continuity is nearly perfect, interrupted only by the Khoisan languages spoken in South Africa. Wilhelm Heinrich Immanuel Bleek (1862) was the first to hypothesize the genetic unity of Bantu languages. Concerning them, the most important classification, still used as a practical taxonomic reference, is Guthrie (1967), but more recent ones are available (Mann and Dalby 1987; Grimes 2000).

Guthrie defined the geographical boundary of the Bantu linguistic domain and divided it in a eastern and a western zone. The western region includes Cameroon, Gabon, Congo, the west of the Democratic Republic of Congo (DRC), Angola and a part of Zambia. The eastern region includes the eastern part of DRC and all the eastern countries of sub-Saharan Africa. This split, although debated, has been consensual for a long time. Using a lexicostatistics approach, Heine (1973) corroborated the division of Bantu languages into two clusters, a western and an eastern one: the eastern group derives from the western one, while the languages of Gabon and Cameroun are independent lineages. Later, Ehret (1999) suggested that eastern, central and southern Bantu languages should be merged into a single group called *Savannah Bantu*. Other studies (Bastin *et al.* 1999; Nurse and Philippson 2003; Rexova *et al.* 2006) confirm a western/eastern division of the entire Bantu linguistic domain suggesting that the western part is older because the languages spoken in the zones A, B and C emerged first from proto-Bantu varieties initially spoken in the middle Benue river valley, located between Cameroon and Nigeria about 5000 ybp (Greenberg 1955). The identification of a homeland is essential to describe how Bantu languages later disseminated, and the middle Benue river valley is a good candidate because it includes speakers of the only linguistic varieties close enough to another branch (Benue-Congo, spoken in Nigeria) of the Niger-Congo linguistic family to which Bantu languages belong as well.

1.2 Migrations waves

Concerning the diffusion of Bantu languages, Bastin *et al.* (1979) suggested two migration waves: *1)* a western wave from the south of Cameroon, then to the equatorial forest along the rivers and, finally, progressing further southwards along the Atlantic coast; and *2)* an eastward wave avoiding the equatorial forest. Of course the general picture is blurred by secondary contact between the languages, by a different migration speed according to the route (to the south or to the east), by secondary migrations and by continuous population displacements until today.

The attention of the reader should be attracted to the extremely fast spread of Bantu languages, which have disseminated throughout half of Africa and have replaced almost all pre-existing languages, similarly to what Latin did in Europe. A possible explanation is related to the lifestyle of early Bantu-speaking populations: proto-Bantu lexical roots[2] show that they generally were agriculturalists and farmers. In contrast to hunting and gathering, agriculture requires a considerable work-force before it can be sustained viably, which then leads to large societies. Once populations begin growing, migration processes and population diffusion follow, which are further promoted when the soil rapidly (but only temporarily) depletes due to its exploitation. This is how the Bantu expansion could have progressed.

1.3 Archaeology and linguistics

A first attempt to combine archaeological and linguistic evidence to understand the dispersal process can be traced back to Oliver (1966). Later, Phillipson (1976, 1977a, 1977b, 2002) provided an ambitious reconstruction by arguing for the development of early Bantu language varieties in Cameroon from about 3000 ybp on. According to his scenario, when Bantu-speakers dispersed eastwards along the northern fringe of the equatorial forest, they met other farmers, probably speaking Central Sudanic languages. After a long phase of contact, they started herding domestic cattle and sheep, learned about the cultivation of certain cereal crops and acquired metal-working techniques. Vansina (1984, 1990

[2] For a reference database including about 10,000 entries proposed for Proto-Bantu reconstructions see www.africamuseum.be/collections/browsecollections/humansciences/blr

pp. 49-57, 1995) criticized Phillipson about the need for a reliable dispersal route, and stressed that the major driving force for the Bantu linguistic divergence was a phenomenon of linguistic fission between varieties that had diverged in an earlier phase, with outermost dialects developing into languages after each fission (Heggert 2004, p. 315). According to Vansina, if indeed a first migration happened eastwards to the Great Lakes Region, a *simultaneous* movement took place to the south of Cameroon and Gabon, and more southwards. Later on, the continuous pattern of habitat resulting from the first major migration was disrupted by the presence of both a dense forest and large stretches of marshlands. New computational analyses (Holden 2002, Holden and Gray 2006) on 75 Bantu languages extracted from the 542 Bantu languages published by Bastin *et al.* (1999) were in agreement with the archaeological hypothesis about the large migration eastwards of Bantu-speaking agriculturalists, meaning that Bantu languages diffused together with their speakers.

1.4 Population genetics and linguistics

An interesting study bringing together genetic and linguistic evidence over the whole Bantu linguistic domain (de Filippo *et al.* 2013) was aimed at testing two models of Bantu population-language dispersal : *i)* Early Split, north of the rainforest, of the eastern and western groups, about 4000 ybp; *ii)* Late Split, south of the rainforest, of the eastern group from the western group about 2000 ybp. The authors measured DNA genetic diversity as function of the geographic distance from the Bantu homeland, along possible inferred itineraries of migration, finding a progressive reduction of the genetic diversity from the homeland. This pattern supports the demic diffusion[3] of the Bantu expansion and better correlates with the Late Split model. Li *et al.* (2014) estimated the first expansion of the Bantu-speaking groups to have occurred at around 5600 years ago but found that a migration to the east and then to the south is statistically *as likely* as other models.

While the DNA genetic diversity of Gabon populations is very low and detectable only by genome sequencing (Patin *et al.* 2017), the linguistic differences are not negligible when compared to those in other Bantu-speaking areas, suggesting than the peopling — by populations having a similar genetic background — happened early in the dispersal from Cameroon, leaving time for later differentiation *in situ*.

1.5 Current linguistic diversity in Gabon

All the ethnic groups living in Gabon speak Bantu languages (with the exception of some Pygmy groups), but the use of French (the official language) is widespread in the increasingly multiethnic towns and many indigenous language varieties are now threatened.

[3] *Demic diffusion* is a demographic term referring to a migratory model of population diffusion into and across an area that had been previously uninhabited by that group, possibly, but not necessarily, displacing, replacing, or intermixing with pre-existing populations.

The Bantu varieties of Gabon include languages from the Guthrie zones A (*Benga* A34; *Fang* A75; *Shiva* A83; *Bekwil* A85b) and B (B10/20/30/40/50/60/70) (see Fig. 1). According to grammatical and lexical traits, the languages of the group B10 (MYENE) and B30 (TSOGO) are related and distinct from other languages in the region, but it is not clear if this is the consequence of a common genealogical origin or the result of linguistic convergence due to contact (Nurse and Philippson, 2003; Mouguiama-Daouda and Van der Veen, 2005). The languages B20 (KOTA-KELE) have an ambiguous status too, because their genetic unity is unclear. While the languages belonging to the three groups B50 (NJABI), B60 (MBETE) and B70 (TEKE) are close to those of the zone C and might be classified into a single cluster B50-B60-B70, languages of the group B40 (SIRA) are related to those spoken in the zone H (south of Gabon),.

According to Clist (2005, p. 490), Gabon has been progressively peopled by waves of Bantu-speaking populations coming from the north-east, but also from the south and the east starting about 2600 ybp. This peopling scenario is more complex than the simple southwards movement from Cameroon proposed by Vansina (1995) in which Gabon would have been crossed only by the western Bantu expansion wave moving to the south. Clist (2005) suggests that, in reality, the main migration wave from the Benue river valley to the south might not have passed through Gabon, but rather further to the east, in a savannah corridor created by dry climatic conditions, in what previously was a dense equatorial forest. This corridor is believed to have lasted from 2800 ybp to 2100 ybp, that is about seven centuries (Maley 2001): a time-span long enough to enable continued human migrations. Whatever the general migration scenario, it is known that the Fang languages (A75) correspond to a rather recent migration wave from Cameroon started 500 ybp and continued until the 1930s (Hombert *et al.* 1989).

1.6 The aims of this study

By processing word lists accounting for the linguistic diversity of Gabon with a computational linguistics method measuring the phonetic difference between two words, the Levenshtein distance (Heeringa 2004), we compute distance matrices accounting for the aggregate lexical difference. Then, we analyze the distance matrices using both bootstrap phylogenetic trees and Multidimensional Scaling (MDS) to identify major linguistic groups that we relate to specific migration waves.

We process two independently-collected and largely overlapping datasets: the ALGAB (*Atlas Linguistique du Gabon,* Hombert 1990) and the MRAC (*Musée Royal de l'Afrique Centrale*, Tervuren, Belgium).[4] The ALGAB dataset lists 158 words for 53 linguistic varieties, while the MRAC is based on 88 words and accounts for 64 varieties (Fig. 1).

The Levenshtein method is different from the cognate-sharing approach used so far in studying the diversity of Bantu languages. The earlier work (Bastin *et al*. 1999, Hol-

[4] The MRAC is a subset of the database processed by Bastin *et al.* (1999).

den 2002, Holden and Gray 2006, Grollemund *et al*. 2015) proceeded by establishing 0/1 matrices, where a '0' pairwise difference is attributed to a pair of words having the same meaning and sharing a common ancestor term (cognates), while a difference of '1' is attributed to a pairwise comparison where the words originated independently (not cognates). This method relies on having expert judgments of cognacy and is less sensitive than Levenshtein distances, which measure the difference in pronunciation according to string alignment.

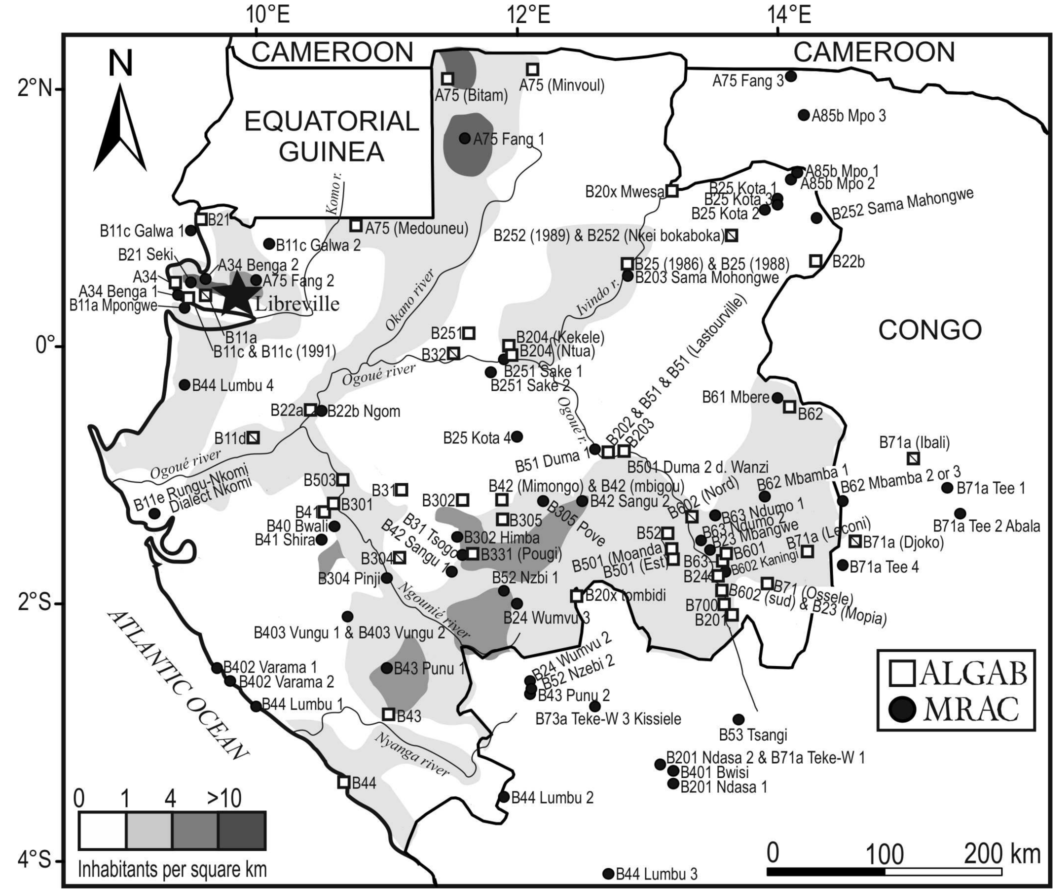

Figure 1 ▸ The Republic of Gabon (see map) is an equatorial African country largely covered by rainforest with a total population of about a million and half inhabitants. More than half of the population lives in the bigger cities, population density outside urban areas is low (see shades of gray). In the map we show the location *a)* of the 53 varieties reported in the ALGAB database (empty squares; when a diagonal appears inside them the position is approximated) and *b)* of the 64 varieties of the MRAC database (see section 2.1 for details).

2. Results of the Levenshtein classification

2.1 ALGAB and MRAC datasets

Concerning the ALGAB (wordlists of 158 items), the consensus tree (Fig. 2A) shows that there are five main clusters: {A75}; {B10, B30}; {B20}; {B40}; {B50, B60, B70}. When we represent these groups on a geographic map (Fig. 2B), we note a striking degree of geographic coherence suggesting a significant correlation between geographic and linguistic distances (r = 0.461**; Mantel test). When MRAC data (wordlists of 88 items) are processed in the same way (figure not shown) the five major clusters are not exactly the same ones: {A75}; {B10, B30}; {B20*a*}; {B20*b*}; {B20*c*, B40, B50, B60, B70}. While the varieties labelled as A75 and B10/B30 are classified in a same way, the latter dataset shows that the linguistic group B20 is split in three clusters and that heterogeneous varieties belonging to the groups B20, B40, B50, B60 and B70 are clustered together as in Bastin *et al.* (1999): they called it *North-central Bantu*. It is reasonable to find similar results because we are processing the *same* dataset used by Bastin..

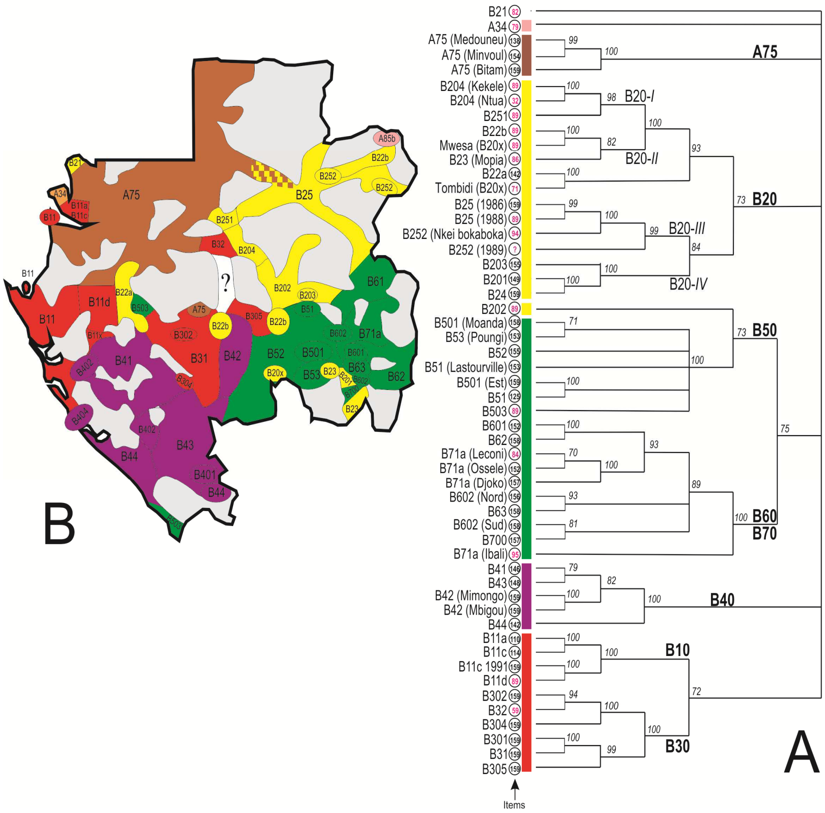

Figure 2 ▸ *A:* UPGMA bootstrap consensus tree concerning the classification of the 53 varieties listed in the ALGAB. Nodes supported by fewer than 70% of the bootstrap sub-replicates have been collapsed. The number of lexical items available for each language is reported after the labels. *B:* Mapping of the major clusters on the consensus map we obtained according to several references (Grimes 2000, Simons 2016, Maho 2009).

Are the two clusterings different because the two databases, the ALGAB and the MRAC, account for word lists of a different length? A closer look shows that the wordlists of the MRAC are a subset of those used in the ALGAB, the latter including 65 additional items (50 nouns + 15 verbs), meaning that ALGAB data is likely to provide a better classification, as more information is available.

2.2 Merging datasets

To verify whether identical varieties, documented independently in the two databases, would cluster together, we have *merged the two datasets* to compute a unique consensus bootstrap tree (not shown) only based on shared lexical items. It points to:

1. The unity of the cluster of languages B10-B30 (bootstrap score = 99);
2. The close association of the languages B60-B70 (bootstrap score = 82);
3. The coherence of the languages B50 that form a cluster (bootstrap score = 79) distinct from B60/B70;
4. The looseness of the cluster B40 (bootstrap score= 60);
5. The "explosion" of the group B20, split in five different and independent clusters.

Two other aspects of the classification of the merged dataset are interesting:

6. Identical varieties, documented by two independent databases, are generally clustered next to each other in the tree, thus suggesting that the discrepancies between the classifications are related to the different lengths of the wordlists. This leads us to trust more ALGAB data (158 words) than the MRAC (88 words).
7. The bootstrap support for the *North-central western Bantu* cluster advocated by Bastin *et al.* 1999 becomes weak (bootstrap score = 54).

To compare the outcomes of our analysis based on a new method, the Levenshtein distance, with the classical cognate-sharing approach, we align our classification of the ALGAB dataset to the classification of Grollemund *et al.* (2015) that process a subset of the same varieties. We find that clusters and subclusters do correspond (not shown), meaning that the Levenshtein algorithm captures the *same* signal of linguistic relatedness (or difference) that a method based on cognate-sharing does, but with the immense advantage of not requiring the aid of experts to provide judgments about cognacy. The only discrepancy between the two approaches concerns, again, the clustering of the group B20.

3. Discussion and Conclusions

3.1 The peopling of Gabon: Savannah corridors versus the rainforest

The question of the Bantu dispersal has vigorously resurfaced thanks to the work of the team of Koen Bostoen (University of Ghent). The routes of Bantu expansion they suggest rely on the geographical plot of consensus Bayesian phylogenetic trees (see Bostoen *et al.* 2015 for a review and Grollemund *et al.* 2015). These authors code linguistic diversity according to cognacy and provide a temporal frame for the different splits calibrated according to archaeological dates. By interpreting the topology of trees as reliable migrations routes,[5] they highlight that savannah corridors were migration routes preferable to rainforest crossings, the rationale being that Bantu speaking societies were adapted to this kind of environment, since their earlier homeland was savannah. According to palynological evidence, a progressive formation of savannah corridors took place through the rainforest during the Middle and Late Holocene, which is starting from 4000 ybp to 2500 ybp, when the surface of savannah was at its maximum extension (Lézine *et al.* 2013; Bostoen *et al.* 2015). Concerning Gabon, the formation of savannah corridors can be inferred as in Fig. 3; if the Bantu migration proceeded through them, the peopling would have been possible from the east and the south (the west being the Atlantic seashore), but not from the north.

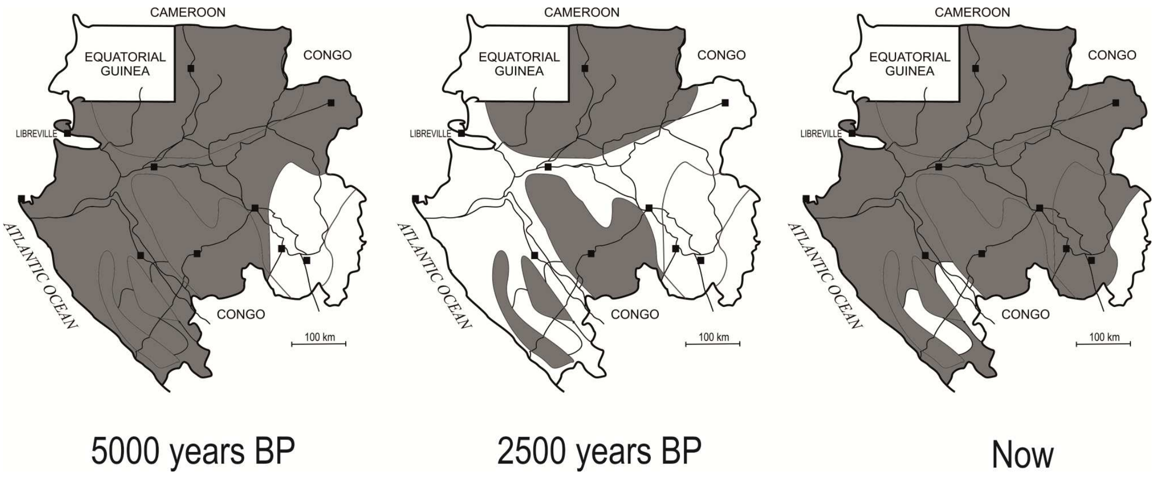

Figure 3 ▸ Gabon: the progressive appearance of savannah corridors (white) in the rainforest (gray) according to palaeoenvironmental data (adapted from Grollemund *et al.* (2015). Equatorial Guinea in not shaded. BP means "before present" = years ago.

But, did the Bantu speaking populations enter Gabon following the progressive formation of savannah corridors starting 4000 ybp or were their migrations independent from them? We

[5] By hypothesizing that the present-day location of languages corresponds to the location they had in the past.

cannot answer directly, because the linguistic analyses we conducted do not explicitly address temporal issues related to peopling phases. To set a timeframe according to archaeological sites excavated in Gabon, we refer to the calibrated radiocarbon ^{14}C dates compiled by Oslisly *et al.* (2013) and to their classification in four main stages of technological knowledge (Fig. 4). While some dating uncertainty cannot be excluded, the temporal sequence of occupation addressed by Oslisly *et al.* (2013) overlaps with the timeframe of the Bantu expansion and, interestingly, points to a population decline starting about 2400 ybp and lasting until recent centuries (Oslisly 2001; Wozka 2006), to finally reach a new maximum five centuries ago after which it declines again until the colonial period (Fig. 4). The Neolithic Stage corresponds to the transition between the Late Stone Age and the Early Iron Age, that is, when people started to become sedentary, made polished stone tools and pottery, and used stone hoes and axes to practice slash-and-burn agriculture. This phase is related to the arrival of Bantu migrations, with a demographic explosion in the subsequent period, the Iron Age. Because the Neolithic stage started 3500 ybp, we date the first arrival of Bantu speakers to Gabon at this point. This timeframe fits well with the scenario and theory of savannah corridors (Grollemund *et al.* 2015) but is also compatible with earlier Bantu migrations southwards, directly though the rainforest. Interestingly, the maximum extension of the savannah corridors (Fig. 3) corresponds to a possible population maximum (Fig. 4).

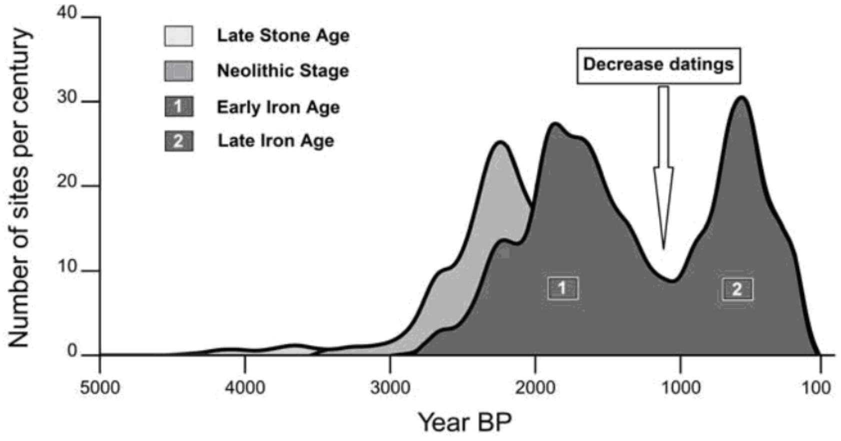

Figure 4 ▸ Survey of radiocarbon dates over the past 4000 years in central Africa including Gabon (from Oslisly *et al.* 2013). Late Stone Age (5500-3500 ybp)—14 sites, shown between 4400 and 3500 ybp; Neolithic Stage (3500-1900 ybp)—33 sites; Early Iron Age (2800-1000 ybp) —79 sites; Late Iron Age (1000-100 ybp) —40 sites. The periods sometimes overlap because two technological phases can coexist at the same time, like typewriters and computers in the late 1990s.

Was the rainforest a real impediment to migrations from Cameroon? While a migration route from Cameroon southwards, along the generally sandy seashore advocated by Bastin *et al.* (1979), is possible, we suggest that the rivers crossing the rainforest could have been potential paths of displacement and that the practice of slash and burn agriculture (characterizing today many Bantu-speaking groups) would also have been practiced in the past, in a forestall environment. For all of the above reasons, we will not assume that the Bantu peopling of Gabon was necessarily dependent on the climatic change that led the rainforest to shrink and the savannah habitat to increase its surface.

3.2 The current challenge, going beyond geography

In linguistics, it has long been admitted that the geographic distance between varieties has an effect on their evolution, namely that closer varieties are generally more similar than distant ones. Lyle Campbell (1995) summarizes this succinctly "[. . .] neighboring languages often turn out to be related." The first model about the spread of linguistic innovations, a form of contact, was the WAVE THEORY of Johannes Schmidt (1872). While Seguy (1971) presented linguistic distances as function of the square root of geographic distance, Trudgill (1974) suggested that the spread of innovations declines quadratically. Nerbonne and Heeringa (2007) and Nerbonne (2010) found a logarithmic model to better function, similarly to the models of population genetics concerning the biological differences of neighbouring populations that are function of migration processes (Wright 1943, Malécot 1948).

When a matrix of Levenshtein aggregate linguistic distances is found to be significantly correlated with the corresponding matrix of geographic distances, as it case for the ALGAB dataset, it is possible to compute a linear regression (*linguistic distance* vs. log [*geographic distance*]) in order to compute, from it and for each pairwise comparison, the linguistic distance that is *expected* between two linguistic varieties according to the geographic distance separating them. This procedure leads to a matrix of *expected* pairwise linguistic distances that can be *subtracted* from the linguistic distances obtained from the original data. The matrix that results *after* the subtraction consists in *residual* distances that can be positive, negative or null. They will be positive when the linguistic distance computed on original data is higher than the one expected from the regression; they will be negative when two varieties at a given distance exhibit a linguistic distance that is lower than what is expected according to the regression. The idea is that residual distances represent the fraction of the linguistic variability that is *not* explained by "normal" linguistic contact between neighbours (geography).

We can then analyse these residual distances via multimensional scaling (MDS). Our goal in this is to detect latent influences on linguistic similarity beyond simple proximity. If varieties are assigned nearby positions in the multidimensional space, then those varieties are similar for reasons other than the contact promoted by nearness. One obvious candidate influence is, of course, the history of the varieties, that is, their genealogy. The MDS plot of the residuals of the geographic analysis is likely to reflect the relations among the varieties before they drifted apart due to migration or were subject to convergence/divergence influences related to contact. Residual distances (Fig. 5) convey clues about the possible historical scenario of linguistic diversification of Gabon. They point to linguistic diversity between different languages that long-lasting linguistic contact and convergence has progressively defaced (see Table 1).

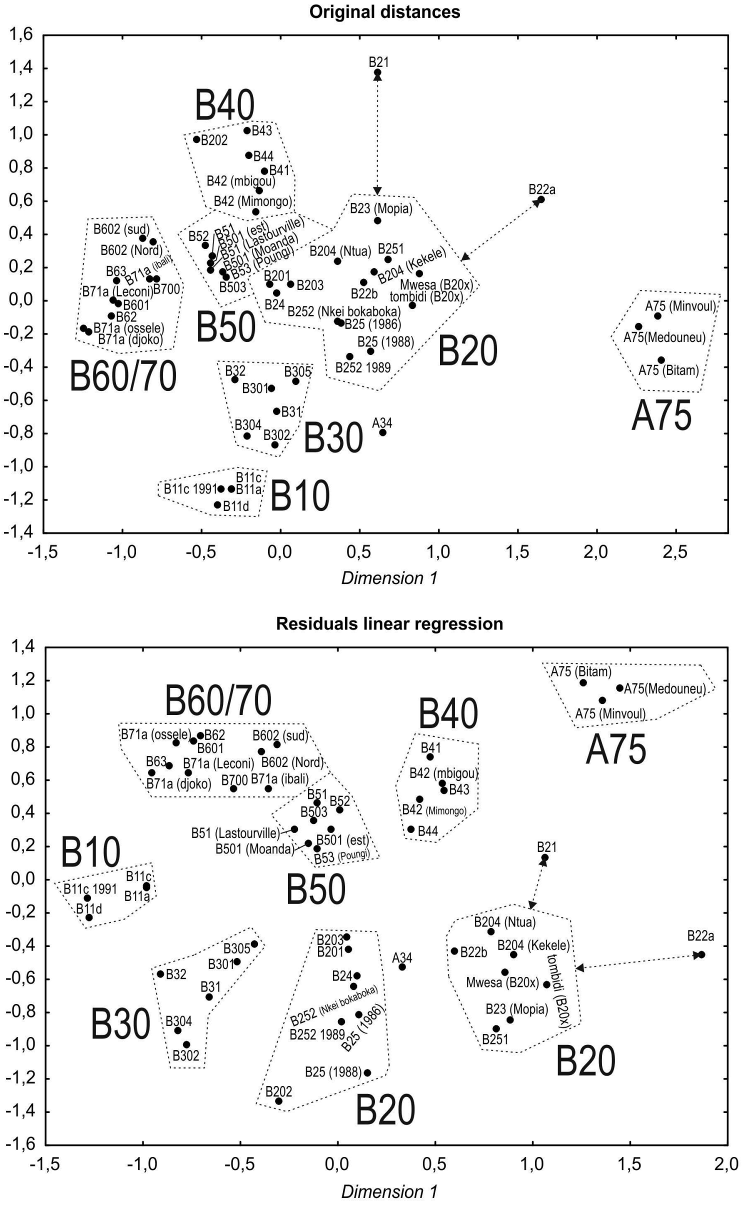

Figure 5 ▸ Multidimensional scaling projections concerning the 53 linguistic varieties listed in the ALGAB. *Top:* Original Levenshtein distances. Stress values: in 1 dimension = 0.3247, in 2 dim. = 0.1641 (plot reported), in 3 dim. = 0.1215. Correlation between geographic and linguistic distances = 0.478**. *Bottom:* Residual distances after computing the regression (R^2=0.216; the logarithmic transformation of geographical distances makes almost no difference: R^2=0.222) between the kilometric distances and the corresponding Levenshtein distances. Stress values: in 1 dimension = 0.399, in 2 dim. = 0.249 (plot reported), in 3 dim. = 0.171. Residuals are normally distributed. See Tab. 1.

Two features of the analysis deserve particular note. First, the MDS plot of the residuals confirms the historical division proposed by Guthrie in that his groups are in general, close to one another. Given the status of Guthrie's work, we hasten to add that it is perhaps better to regard the correspondence between the two as a confirmation of the step of applying MDS to the residuals of the geographic analysis. Second, despite the overall excellent correspondence between Guthrie's classification and the MDS-residuals analysis, the B20 group stands out in reflecting the Guthrie classification less faithfully. This leads us to hypothesize that the B20 group has a more complex history, and that it includes two groups that were distinct in earlier history.

Table 1 ▸ Summary of the possible effects on the Bantu varieties in Gabon that linguistic contact has determined (ALGAB data)**.** This scenario is inferred by comparing the two plots of Fig. 5.

Varieties	***Scenario inferred from residual distances***
B10	Initially separate, later converged with B30
B30	Initially separate, later converged with B10
B20	Two initially separate groups converged together
B40	Initially separate, later converged with B50
B50	Initially part of same group including B60 and B70, later slightly diverged from those, becoming closer to B40 and to some varieties B20
B60/B70	Initially part of a single group, together with B50, later diverged from B50
A75	A75 arrived in Gabon recently (~5 centuries ago), when B40 was already spoken. Their closeness might correspond to a similar geographic origin in Cameroon: today they are very different.

3.3 Gabon: Immigration scenarios

Palaeoenvironmental and archaeological studies show that the opening of savannah plains on the coastal region of Gabon started about 4000 ybp (Bostoen *et al.* 2015), with a Neolithization process dated at around 3500 ybp (Oslisly 2001) and a detectable sedentarization starting at 2700 ybp in northern Gabon (Bostoen *et al.* 2015). According to linguistic cartography we suggest that B20 varieties correspond to a early migration southwards of Cameroon, through the rainforest or along savannah corridors, to the north of Gabon in at least two independent and early waves: the DNA genetic diversity of B20-speakers is the highest among all the Bantu speakers of Gabon (Patin *et al.* 2017), meaning that these populations and the languages they spoke had more time to evolve than varieties brought by more recent migration waves. Based on the genetic diversity reported by Patin *et al.* (2017), we propose the hypothesis that other early migrations took place by following the Atlantic coast from Cameroon to Gabon. These migrations might correspond to two different and separate waves corresponding to B10 and B30. The other two groups emerged or arrived later, i.e., B40 and the single group B50/B60/B70.

3.4 Conclusions

We do not know whether the present-day location of each group of languages corresponds to the position they had in the past, and we consider extensive migrations over millennia more than likely, meaning that ancestral languages might have been spoken elsewhere, not necessarily where they are today. We also judge it possible that, after an initial stage of peopling, some Bantu languages diffused from one Bantu group to another, in the absence of population movements. Finally, we recognize that the vast majority of African populations today are multilingual, and there is no reason to think that the past situation was different. Multilingualism, in itself, is a source of language diversification and this is not a recent phenomenon. We also find it reasonable to admit that borrowing and secondary contact between differentiated languages might have been a major force in the process of linguistic differentiation. These phenomena might well explain the high degree of correlation existing between linguistic and geographic distances and that we partially took into account by dealing with residual distances, which we consider to be closer to the historical scenario of the peopling of Gabon.

The Levenshtein distance measure the signal of historical relatedness *and* the contact between the languages, its ability to match classifications based on shared cognates identified by experts is much higher than past criticisms would suggest (see Manni 2017, pp. 278-286 for a review). The very good match between the clusters identified by Grollemund *et al.* 2015 and the corresponding Levenshtein classifications has been reported. This result might be explained by the fact that Bantu languages are linguistically quite close, often forming dialect-chains: this is a scenario closer to the initial application of the Levenshtein method to dialectology, nevertheless, the Levenshtein classification of more distantly related languages not forming dialect chains (Mennecier *et al.* 2016) turned equally convincing. The Levenshtein distance captures the *same* historical signal that a cognate-based approach does, without the need to seek for the assistance of an expert assessing the historical (genetic) relatedness of lexical items (shared vocabulary); this is a remarkable advantage because it allows the phylogenetic classification of linguistic corpora that might otherwise be neglected, whilst they can lead to straightforward hypotheses about past migrations and peopling stages that are poorly documented.

References:

Bastin Y., Coupez A., de Halleux B. 1979. Statistique lexicale et grammaticale pour la classification historique des langues bantoues. *Bulletin des séances de l'Académie royale de Sciences d'Outre Mer*, **3**: 375-387.

Bastin Y., Coupez A., Mann M. 1999. *Continuity and Divergence in the Bantu Languages: Perspectives from a Lexicostatistic Study*. Tervuren: MRAC.

Batini C., Ferri G., Destro-Bisol G., Brisighelli F., Luiselli D., Sánchez-Diz P., Rocha J., Simonson T., Brehm A., Montano V., Elwali N.E., Spedini G., D'Amato M.E., Myres N., Ebbesen P., Comas D., Capelli C. 2011. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Molecular Biology and Evolution*, **28**: 2603-2613.

Bleek, W. H. I. 1862. *A comparative grammar of South African languages* (Vol. 1). Gregg International.

Bostoen K., Clist B., Doumenge C., Grollemund R., Hombert J. M., Muluwa J. K., Maley Jean. 2015. Middle to late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa. *Current Anthropology*, **56**: 354-384.

Campbell, L. 1995. The Quechumaran Hypothesis and Lessons for Distant Genetic Comparison. *Diachronica* XII(2):157–200.

Clist B. 2005. Des premiers villages aux premiers européens autour de l'estuaire du Gabon : quatre millénaires d'interaction entre l'homme et son milieux. PhD dissertation. Bruxelles, Université Libre de Bruxelles.

de Filippo C., Barbieri C., Whitten M., Mpoloka S.W., Gunnarsdóttir E.D., Bostoen K., Nyambe T., Beyer K., Schreiber H., de Knijff P., Luiselli D., Stoneking M., Pakendorf B. 2011. Y-chromosomal variation in Sub-Saharan Africa: Insights into the history of Niger-Congo groups. *Molecular Biology and Evolution*, **28**: 1255–1269.

de Filippo C., Bostoen K., Stoneking M., Pakendorf B. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B*. **279**: DOI: 10.1098/rspb.2012.0318

Ehret C. 2002. Language family expansion: broadening our understandings of cause from an African perspective. In: P. Bellwood and C. Renfrew (eds.). *Examining the farming / language dispersal hypothesis*. Cambridge (UK): McDonald Institute Monographs, pp. 163-176.

Greenberg J.H. 1955. *Studies in African linguistics classification*. New Haven: The Compass Publishing Company.

Grimes B. F. (ed.). 2000. Ethnologue**.** Dallas: SIL International. 2 vols, (14th edition).

Grollemund R., Branford S., Bostoen K., Meade A., Venditti C., Pagel M. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences USA*, **112**: 13296–13301.

Guthrie M. 1967. *Comparative Bantu***.** Farnborough: Gregg International Publishers Ltd. Vols. 1-4.

Heeringa W. 2004. Measuring dialect pronunciation differences using Levenshtein distance. PhD Doctoral disserationthesis. Groningen: Rijksuniversiteit Groningen.

Heeringa W., Joseph B. 2007. The Relative Divergence of Dutch Dialect Pronunciations from their Common Source: An Exploratory Study. In: J. Nerbonne, T. Mark Ellison, G. Kondrak (eds.), *SigMorPhon 07 ACL 2007, Computing and Historical Phonology, Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology.* Prague, (Czech Republic), Stroudsburg (PA): The Association for Computational Linguistics (ACL), June 28, pp. 31-39.

Heggert M. 2004. The Bantu problem and African archaeology. In: A.B. Stahl (ed.) *African archaeology a critical introduction*. Blackwell Publishing, pp.301-326.

Heine B. 1973. Zur genetischen Gliederung der Bantu-sprachen. *Afrika und Űrbersee*, **56**: 164-195.

Holden C.J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society B. Biological Sciences*, **22**: 793-9.

Holden C.J., Gray R.D. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In: P. Forster and C. Renfrew (eds.) *Phylogenetic methods and the prehistory of languages*. Cambridge (UK): McDonald Institute Monographs, pp. 19-32.

Hombert J.M., 1990a. Atlas linguistique du Gabon. *Revue gabonaise des Sciences de l'homme*, **2**: 37-42.

Hombert J.M., Medjo Mvé P. Nguéma, R. 1989. Les Fangs sont-ils Bantu? *Pholia*, **4**: 133-147.

Lézine A-C., Assi-Khaudjis C., Roche E., Vincens A., Achoundong G. 2013. Towards an understanding of West African montane forest response to climate change. *Journal of Biogeography*, **40**: 183–196.

Li S., Schlebusch C., Jakobsson M. 2014. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society B*. **281**: DOI: 10.1098/rspb.2014.1448

Maho J.F. 2009. NUGL online: The online version of the New Updated Guthrie List, a referential classification of the Bantu languages. goto.glocalnet.net/mahopapers/nuglonline.pdf

Malécot G. 1948. Les mathématiques de l'hérédité. Paris: Masson.

Maley J. 2001. La destruction catastrophique des forêts d'Afrique centrale survenue il y a environ 2500 ans exerce encore une influence majeure sur la répartition actuelle des formations végétales. *Systematic and Geography of Plants*, **71***:* 777-796.

Mann M., Dalby D. 1987. A Thesaurus of African Languages**.** London: Hans Zell Publishers.

Manni F. 2017. Linguistic Probes into Human History. Groningen: University of Groningen. 320 pp. Groningen dissertations in linguistics n° 162. ISBN: 978-90-367-9871-6 (print version); ISBN: 978-90-367-9872-3 (electronic version).

Mennecier P., Nerbonne J., Heyer E., Manni F. 2016. A Central-Asian survey. *Language Dynamics and Change*, **6**: 57-98.

Mouguiama-Daouda P., Van der Veen L.J. 2005. B10-B30 : conglomérat phylogénétique ou produit d'une hybridation. In: K. Bostoen K., J. Maniacky J. (eds.), *Studies in African Comparative Linguistics, with special focus on Bantu and Mande*. Tervuren: Royal Museum for Central Africa (RMCA/MRAC), Sciences Humaines, pp. 1781-9857.

Nerbonne J. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B*, **365**: 3821-3828.

Nurse D., Philippson G. 2003. Towards a historical classification of the Bantu. In: D. Nurse and G. Philippson (eds.), *The Bantu languages*. London: Routledge, pp. 164-179.

Oliver R. 1966. The problem of the Bantu expansion. *Journal of African History*, **7**: 361-376.

Oslisly R. 2001. The history of human settlement in the middle Ogooué valley (Gabon): implications for the environment. In: Weber W., White L.J.T., Vedder A. Naughton-Treves L. (eds) Afriacan rain forest ecology and conservation. New Haven: Yale University Press, pp. 101-18.

Oslisly R., Bentaleb I., Favier C., Fontugne M., Gillet JF. 2013. West Central African peoples: survey of radiocarbon dates over the past 4000 years, Proceedings of the 21st International Radiocarbon Conference, A.J.Jull & C. Hatté Eds., *Radiocarbon*, **55**: 1377–1382.

Patin E., Laval G., Barreiro L.B., Salas A., Semino O., Santachiara-Benerecetti S., Kidd K.K., Kidd J.R., Van der Veen L., Hombert J.M., Gessain A., Froment A., Bahuchet S., Heyer E., Quintana-Murci L. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics,* **5** :e1000448

Patin et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North-America. Science **356**: 543-546.

Phillipson D.W. 1976. Archaeology and Bantu linguistics. *World Archaeology*, **8**: 65-82.

Phillipson D.W. 1977a. Later prehistory of eastern and southern Africa. London: Heinemann.

Phillipson D.W. 1977b. The spread of the Bantu languages. *Scientific American* **236**: 106-14.

Phillipson D.W. 2002. Language and farming dispersals in Sub-Saharan Africa. In: P. Bellwood and C. Renfrew (eds.). *Examining the farming / language dispersal hypothesis*. Cambridge (UK): McDonald Institute Monographs.

Rexova K., Bastin Y., Frynta D. 2006. Cladistics analysis of bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* **93**: 189-194.

Schmidt J. 1872. Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar: H. Böhlau.

Séguy J. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, **35**: 335-357.

Simons G. (ed). 2016. *Ethnologue. Languages of the world.* Dallas (TX): SIL International. Internet publication accessible at www.ethnologue.com

Trudgill P. 1974. Linguistic Change and Diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, **2**: 215-246. pp.

Vansina J. 1984. Western Bantu Expansion. *Journal of African history*, **25**: 129-145.

Vansina J. 1990. *Paths in the rainforest*. London: Currey.

Vansina J. 1995. New linguistic evidence of the Bantu expansion. *Journal of African history*, **36**: 173-195.

Verdu P., Austerlitz F., Estoup A., Vitalis R., Georges M., Thery S., Froment A., Le Bomin S., Gessain A., Hombert J.M., Van der Veen L., Quintana-Murci L., Bahuchet S., Heyer E. 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19**: 312–318.

Whiteley W.H. 1971. Introduction. In: Wilfred H. Whiteley (ed.) *Language use and social change: Problems of multilingualism with special reference to Eastern Africa*. Oxford: Oxford University Press, pp. 1–23.

Wotzka H-P. 2006. Records of activity: radiocarbon and the structure of iron age settlement in central Africa. In: H-P. Wotzka (ed.). *Grundlegungen. Beiträge zur europäischen und afrikanischen Archäologie für Manfred K.H. Eggert*. Tübingen: Francke Attempto Verlag and Co., pp. 271-289.

Wright S. 1943. Isolation by distance. *Genetics*, **28**: 114-138.