PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

**Cite this article:** Currie TE, Meade A, Guillon M, Mace R. 2013 Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. Proc R Soc B 280: 20130695.
http://dx.doi.org/10.1098/rspb.2013.0695

Received: 19 March 2013
Accepted: 15 April 2013

**Subject Areas:**
evolution, taxonomy and systematics, ecology

**Keywords:**
language evolution, cultural diversity, language diversity, phyloplasty, phylogenetic inference, phylogenetic comparative methods

**Author for correspondence:**
Thomas E. Currie
e-mail: t.currie@ucl.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2013.0695 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

# Cultural phylogeography of the Bantu Languages of sub-Saharan Africa

Thomas E. Currie[1], Andrew Meade[2], Myrtille Guillon[1] and Ruth Mace[1]

[1]Human Evolutionary Ecology Group, Department of Anthropology, University College London, 14 Taviton St, London WC1H 0BW, UK
[2]Evolution research group, School of Biological Sciences, University of Reading, Lyle Building, Whiteknights, Reading, Berkshire RG6 6BX, UK

There is disagreement about the routes taken by populations speaking Bantu languages as they expanded to cover much of sub-Saharan Africa. Here, we build phylogenetic trees of Bantu languages and map them onto geographical space in order to assess the likely pathway of expansion and test between dispersal scenarios. The results clearly support a scenario in which groups first moved south through the rainforest from a homeland somewhere near the Nigeria–Cameroon border. Emerging on the south side of the rainforest, one branch moved south and west. Another branch moved towards the Great Lakes, eventually giving rise to the monophyletic clade of East Bantu languages that inhabit East and Southeastern Africa. These phylogenies also reveal information about more general processes involved in the diversification of human populations into distinct ethnolinguistic groups. Our study reveals that Bantu languages show a latitudinal gradient in covering greater areas with increasing distance from the equator. Analyses suggest that this pattern reflects a true ecological relationship rather than merely being an artefact of shared history. The study shows how a phylogeographic approach can address questions relating to the specific histories of certain groups, as well as general cultural evolutionary processes.

## 1. Introduction

It is estimated that there are more than 500 Bantu languages spoken in sub-Saharan Africa [1], making this one of the largest and most widespread language groupings in the world. Genetics, archaeology and linguistics all point to the extensive distribution of this language family being the result of a population dispersal that began around 3000–5000 years ago [2–4]. While there is now consensus that the homeland of Bantu speakers was in the region of the border between Nigeria and Cameroon, where the Bantoid languages most closely related to narrow Bantu are spoken [5,6], there is less agreement about the route taken by Bantu groups as they spread out over the rest of the continent. These debates have implications regarding the origin and spread of important cultural innovations, such as metallurgy and cattle-keeping [3]. Two main dispersal scenarios have been debated in the literature [4,7,8]. The first sees an early split into East and West branches, in which the East branch travelled north of the rainforest as far as the Great Lakes before heading south to cover the eastern half of Africa, while the West branch expanded through the rainforest, emerging on the south side to cover the western half of the continent. The second scenario envisages no such early split but an initial movement through the rainforest and subsequent divergence once groups had emerged on the other side.

Large-scale migrations of human populations are thought to be a major feature of human history during the Holocene. An influential theory sees the presence of widespread language families as indicative of large-scale population movements fuelled by the invention of agriculture [9]. Hypotheses relating to these migrations are often presented as plausible narratives. Computational phylogenetic methods can be used to go beyond verbal arguments and make more rigorous assessments of competing ideas [10]. The structure of

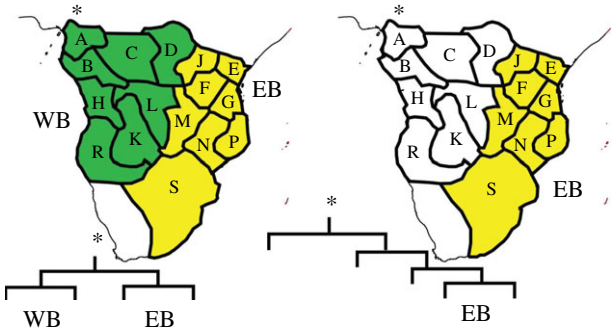


**Figure 1.** Alternative Bantu dispersal hypotheses and the different phylogenetic tree structures they imply. Alphabetically labelled regions refer to the Guthrie zones commonly used in Bantu studies. Asterisks (*) represent the approximate location of the ancestral Bantu society. Under the 'deep-split' scenario, there was an early split between East (EB, yellow) and West (WB, green) Bantu clades as one group went along the northern edge of the rainforest (approx. regions A–D). In contrast, the 'pathway through the rainforest' scenario implies a more chained topology as groups spread and diversified through the rainforest, with the later formation of a monophyletic East Bantu clade. Guthrie zone map is adapted from map created by I. Edricson (http://en.wikipedia.org/wiki/File:Bantu_zones.png).

phylogenetic trees built using language data can be used to test between dispersal scenarios and have previously been used to study the history of a number of language groupings, e.g. Austronesian [11], Semitic [12], Indo-European [13,14] and the languages of island Melanesia [15]. The assumption these approaches make is that the hierarchical relationships represented in language trees reflect the splitting of groups as populations diverge and move into new areas. This same approach can be applied to test between competing hypotheses about Bantu migrations. Under the early split scenario, East Bantu and West Bantu would represent distinct, monophyletic clades (i.e. each would descend from their own unique common ancestor). Under the second scenario, the phylogenetic trees would exhibit a more chained topology. East Bantu would represent a valid, monophyletic group while West Bantu would be considered paraphyletic (i.e. having no unique ancestor that was not also an ancestor to East Bantu) [4] (figure 1).

Previous work has employed quantitative methods to address questions relating to the historical relationships between Bantu languages, but has not produced conclusive results. Bastin *et al*. [16] used distance methods to assess the overall similarity between Bantu languages based on a sample of 542 comparative wordlists, but resisted drawing any firm conclusions about historical relationships between languages. Such distance-based techniques, or 'lexicostatistics', have been criticized because they do not distinguish between retentions and innovations, and implicitly assume a constant rate of evolutionary change, making phylogenetic assessments problematic [17,18]. Holden and co-workers [19–21] used a sample of these data and applied character-based methods of phylogenetic inference (Maximum-Parsimony and Bayesian Markov chain Monte Carlo (MCMC)), while Rexova *et al*. [22] also added grammatical data to a sample of the lexical data and analysed them using similar techniques. These analyses were inconclusive with respect to testing between the two main dispersal scenarios. The maximum-parsimony analyses in particular exhibit a large degree of uncertainty present in the deeper relationships between language groups, which are essential for distinguishing between these scenarios. While the topology of the trees constructed using Bayesian methods lack a deep-split between East and West Bantu and have evidence for a more chained topology, the results from these studies remain tentative owing to the relatively small sample of languages used, and the uneven sampling of languages. Holden's sample of languages was selected in order to carry out cross-cultural comparative analyses, and Rexova *et al*.'s sample was based on the availability of lexical and grammatical data. Both these samples contain relatively few languages from the rainforest and northwest region even though these make up a sizable proportion of the full Bantu grouping and are important for understanding the early diversification patterns owing to their proximity to the proposed Bantu homeland [23]. de Filippo *et al*. [8] recently examined a larger sample of 412 languages and correlated linguistic distances with geographical distances calculated under simplified versions of both dispersal scenarios. The results showed significant correlations with both scenarios, although stronger correlations with the rainforest-first model. Correlations between genetic distances and linguistic distances supported the rainforest-first model over the deep-split model. However, these latter analyses are based on only a small number of populations ($n = 21–36$), which had both genetic and linguistic data available.

To address these issues here, we employ character-based Bayesian methods to build phylogenetic trees using the complete comparative linguistic dataset of Bastin *et al*. [16], comprising 542 wordlists. Using more taxa can improve the performance of phylogenetic methods [24], and the more complete sampling of languages from the whole of the Bantu region helps to overcome the problems associated with the more biased samples of previous studies. Here, we make use of the structure of these trees and explicitly map them in geographical space. This approach can reveal the likely pathway of movement of Bantu groups and allows us to assess these alternative dispersal hypotheses.

Mapping phylogenetic trees in space can also enable us to shine a light on the possible relationship between migrations and general processes involved in the diversification of human ethnolinguistic groups. Previous work has demonstrated a latitudinal gradient in the diversity of human languages, with more languages near the equator than towards the poles [25,26], a pattern that parallels well-known latitudinal gradients in biological species diversity and suggests an ecological basis to language diversification [25,27]. However, if groups tend to spread out from an initial origin in the tropics (as appears to be the case for Bantu), it is possible that this association may simply reflect the fact that these groups have had less opportunity to diverge [28]. Knowing the phylogenetic relationships between languages and the route taken during expansion and diversification allows us to test whether this process can explain the observed latitudinal gradient.

## 2. Material and methods

Investigating the phylogeography of Bantu languages involved several steps. First, we inferred the historical relationships between languages using Bayesian MCMC phylogenetic techniques. Then, using information about the geographical location of these languages, we inferred the geographical location of ancestral nodes in these trees using a phylogenetic comparative method. This allowed us to map the pathways of expansion as Bantu languages spread out and diversified from their original location. Finally, we tested whether there was a

latitudinal gradient in language diversity in this region. By combining data on the area covered by Bantu languages with the phylogenies, we can test the hypothesis that any latitudinal pattern is due solely to a historical process whereby these languages originated in a tropical region and have simply had less time to diversify at higher latitudes, or the failure to adequately control for the non-independence of the languages. Details of each of these steps are given below.

## (a) Phylogenetic inference

### (i) Lexical data

Here, we use linguistic data from 542 spoken varieties of Bantu [16]. The names and alpha-numeric codes of all these languages are listed in the electronic supplementary material, table S1. Linguistic data in the form of different lexical items ('words') were taken from Bastin *et al*. [16]. These data code whether these basic vocabulary words from different languages can be considered cognate (i.e. they share a common origin). To facilitate phylogenetic analyses, these data were recoded into binary cognate sets reflecting the presence or the absence of each cognate in each language. In the original dataset, the wordlists contained 92 words. However, the published data contain an error in that the cognate judgements for 'nose' and 'one' are printed twice at the expense of the words for 'name' and 'neck'. The dataset used here therefore contains coded data for 90 words, represented by 2908 cognate sets.

### (ii) Outgroup selection

Our phylogenetic method requires the selection of certain taxa to act as an outgroup. The original dataset designates 12 languages as Bantoid but non-Bantu: *Ejagham* (language code 800), *Tiv* (802), *Amasi* (805), *Ambele* (806), *Asumbo* (894), *Ngyemboong* (951), *Yemba* (952), two varieties of *Ghomala* (906/1, 960/2) and three varieties of *Fe'fe* (970/1, 970/2, 970/3). The more recent classification scheme of Lewis [1] also classifies five other languages as being non-Bantu Bantoid languages: *Nen* (A44), *Tuki* (A61), two varieties of *Yambasa* (A62/1, A62/2) and *Nu Gunu* (A66) These 17 languages were therefore used as an outgroup during phylogenetic inference. To assess the effects of outgroup specification on our results, we also ran confirmatory analyses using only the 12 Bantoid languages specified in the original dataset as an outgroup.

### (iii) Model of linguistic evolution and MCMC analysis

The cognate data were used to reconstruct phylogenetic trees using Bayesian MCMC techniques in the program BayesPhylogenies (http://www.evolution.reading.ac.uk/BayesPhy.html). In order to infer the phylogenetic relationships between languages using these methods, we need a model of lexical evolution that specifies how cognates are gained and lost [21,29]. Here, we use a model of evolution with equal rates of cognate gains and losses. Historical linguists have long argued that languages do not change at a constant rate [30], and recent empirical work has demonstrated how rates of change vary across time, space and different elements of language [31–33]. We therefore tested whether incorporating rate heterogeneity significantly improved the fit to the lexical data by running covarion models, which model within-site rate variation, allowing the rate of change to switch between different rate categories in different regions of the tree [34]. The covarion is a good model for language evolution because it can reflect historical changes in the rate of evolution of different words in different languages at different points in time. This may potentially model possible causes of rate variation such as changes in population size and structure, movements into new environments or when different languages come into contact [4,11,35]. Subsequent analyses were based on a posterior sample of 500 phylogenetic trees built using the best model of lexical evolution (see the electronic supplementary material, section 1.2).

## (b) Reconstructing ancestral locations

To test between dispersal scenarios, we inferred the locations of ancestral Bantu societies and examined the likely pathways of movement during the Bantu expansion. In order to map the expansion through space, we combined the language phylogenies with information about the present-day location of these languages. We inferred the geographical location of ancestral Bantu languages using a phylogenetic comparative method (a modified version of independent contrasts), which works backwards from the information at the tips of the tree to reconstruct the likely location of the nodes in these trees. This process therefore makes explicit what is currently often only done intuitively when relating the pattern of branching in linguistic phylogenetic trees to geographical spread.

In independent contrasts, continuous traits (here, geographical location) are mapped on to a phylogenetic tree and the values at the nodes of the tree are inferred by assuming that the traits are evolving at a constant rate under a model of Brownian motion [36]. These inferred values are then used to calculate values at deeper nodes. Modelling geographical spread using Brownian motion may be inappropriate if present-day geographical location is the result of a population expansion. In such a situation, some groups remain close to the homeland region while others can end up vast distances away, meaning overall rates of movement will be different for different groups (e.g. a lineage that moved only 200 km from an ancestral homeland 5000 years ago would have a lower overall rate of movement than a lineage that travelled 3000 km). To enable us to incorporate regions where rates of movement differ, we employed the recently developed phyloplasty (PP) implementation of the independent contrasts method [37], using Bayesian reversible-jump MCMC estimation in a modified version of the program BayesTraits (http://www.evolution.rdg.ac.uk/BayesTraits.html). In short, this method searches for departures from Brownian motion by transforming individual branches or whole clades to reflect faster or slower rates of change in the particular trait being modelled. For our analyses, we would expect those groups that have moved further from the homeland to be associated with branches that have been stretched relative to those that have remained closer. These branch length transformations improve the fit of the model to the data but each separate transformation adds an extra parameter to the model, which is taken into account when deciding on the optimum number of transformations (e.g. if Brownian motion is an adequate description of the rate of movement then no branch length transformations are required).

Bastin *et al*. [16] provide longitude and latitude data of the location where each language list was collected. For some languages, visual inspection revealed that the geographical location provided was somewhat outside the range indicated by the Global Mapping International (GMI) language maps (probably owing to opportunistic sampling). These languages were pruned from the phylogenetic trees and were not included in the comparative analyses, leaving 507 languages. As the relationship between distance and points of longitude and latitude is not constant, but varies with latitude, locations were mapped in the Geographic Information System ArcGIS v. 9.2 under a projection that preserves distance between points (African conic equidistant projection). To infer values of longitude and latitude at the nodes in the tree, we ran two separate PP analyses using BayesTraits, one each for the projected values of longitude and latitude.

## (c) Latitude and language area

In order to assess the potential effect of migration history on patterns of language diversity, we can take advantage of the fact

that as languages tend to divide up a region with little or no overlap [25], the diversity of languages in a region is related to the area covered by those languages (i.e. more diverse regions have languages covering smaller areas). We therefore tested whether there is a latitudinal gradient in Bantu language areas, and if so, whether this has arisen from groups that migrated south simply having had less opportunity to diverge? Comparative methods also enable us to control for the fact that individual languages (similar to biological species) cannot be assumed to be independent for the purposes of statistical analyses owing to their historical relationships (e.g. socio-cultural traits that led to larger language areas may have been inherited by numerous languages from a common ancestor that happened to also be present in higher latitudes without there being a functional connection between area and latitude) [38]. Inferring the ancestral locations of nodes in our phylogenetic trees allows us to calculate the geographical distance between nodes and therefore the total distance travelled by Bantu groups from the original Bantu homeland.

Language area data were calculated in ArcGIS from digital language maps produced by Global Mapping International http://www.gmi.org [25]. There were 186 languages in the sample for which language area data were available.

Phylogenetic Generalized Least-Squares (PGLS) analyses of the relationship between language area, latitude and distance travelled were conducted using maximum-likelihood estimation. To assess whether a phylogenetic correction needed to be made in these analyses, the phylogenetic signal in the residuals was assessed using the lambda parameter.

# 3. Results

## (a) Phylogenetic inference

### (i) Model testing

Covarion models were initially run with up to five covarion categories and their fit to the data was compared via their average log likelihood. Table S1 in the electronic supplementary material shows that allowing for rates of linguistic change to vary in this way substantially increases the likelihood of the data. The model with five covarion categories showed the best fit to the data, and as the improvement in likelihood from each additional covarion category was beginning to tail-off, we selected this model for further analysis. We also ran a confirmatory analysis with six covarion categories. Although the extra parameters in this model mean that the likelihood is improved in comparison to the five covarion model, the degree of improvement is smaller again and could be capitalizing on chance. Importantly, the six covarion model produced trees that were varied little from the five covarion trees. We therefore focus on the results of trees built under the five covarion model.

### (ii) Phylogenetic tree topologies

Bayesian analyses do not produce a single phylogenetic tree but instead return a posterior sample of trees that reflect uncertainty in the tree topology and branch lengths given the data and the particular model of evolution. One sample of 500 trees built under the five covarion category model is summarized in figure 2 as a consensus tree showing nodes only present in 70 per cent or more of the tree sample (a more complete consensus tree is shown in the electronic supplementary material, figure S1). No tree in the sample has a deep-split separating East and West Bantu into monophyletic clades. While East Bantu languages do form a well-supported monophyletic clade, West Bantu languages are paraphyletic. The posterior sample shows that the historical relationships between some languages cannot be reconstructed with a large degree of confidence. However, in contrast to the trees presented by Holden *et al*. [21], this uncertainty does not affect the major sub-groupings, meaning we can be more confident in rejecting the deep-split between East and West Bantu scenario. Apart from one or two exceptions, the alphabetically labelled Guthrie zones that are commonly used in Bantu studies (see the electronic supplementary material, figure S1) do not relate well to specific monophyletic clusters. This confirms the idea that these zones should be thought of as a short-hand for describing geographical locations but not historical linguistic sub-groupings [39].

### (iii) Outgroup specification

The overall tree topologies differ very little between trees built under the more-inclusive and less-inclusive outgroup specifications (see the electronic supplementary material, figure S3). Unsurprisingly, the languages that are not included in the outgroup under the more restrictive specification (*Nen*, *Tuki*, *Yambasa* and *Nu Gunu*) take up a basal position in these trees. These results suggest that the broad pattern of relationships between Bantu languages identified in our main analyses, which supports the rainforest-first expansion hypothesis, is not solely dependent on the particular specification of the outgroup.

## (b) Geographical mapping

The PP analyses of the geographical data were a substantial improvement over the default implementation of independent contrasts, leading to between 31 and 67 transformations (mean = 48) for the latitude data and between 37 and 72 transformations (mean = 54) for the longitude data. This indicates that movement has occurred at faster rates in some regions than others. Figure S2 in the electronic supplementary material compares the original tree with summary trees from the PP analyses showing how branch lengths have been modified by the rate transformations. The relative root-to-tip distances of these modified branch lengths are indicative of how far groups have moved from the ancestral homeland latitudinally and longitudinally. Overall, these transformations are consistent with the migration scenario in which groups have sequentially spread out from a homeland in the border region of Nigeria/Cameroon. The branch lengths of the PP trees indicate large latitudinal movements in the K and R groups, and the M, N, P and S groups of East Bantu consistent with southward expansions. The longitude trees show a more even distribution of rate transformations. Of particular note are long branches indicative of eastern movement into the rainforest, and eastern movement after leaving the rainforest eventually leading to the establishment of the East Bantu clade.

The likely pathway of the Bantu expansion can be more clearly revealed by plotting the inferred location of well-supported nodes from the backbone of the consensus phylogeny (figure 2). The results suggest that, beginning from a homeland somewhere in the Nigeria–Cameroon border, Bantu groups moved through the rainforest. Emerging on the south side, one branch moved south and west, while another moved east, south of the rainforest, towards the Great Lakes region. Some groups diversified in this region while another branch moved south once again to inhabit Southeastern Africa.

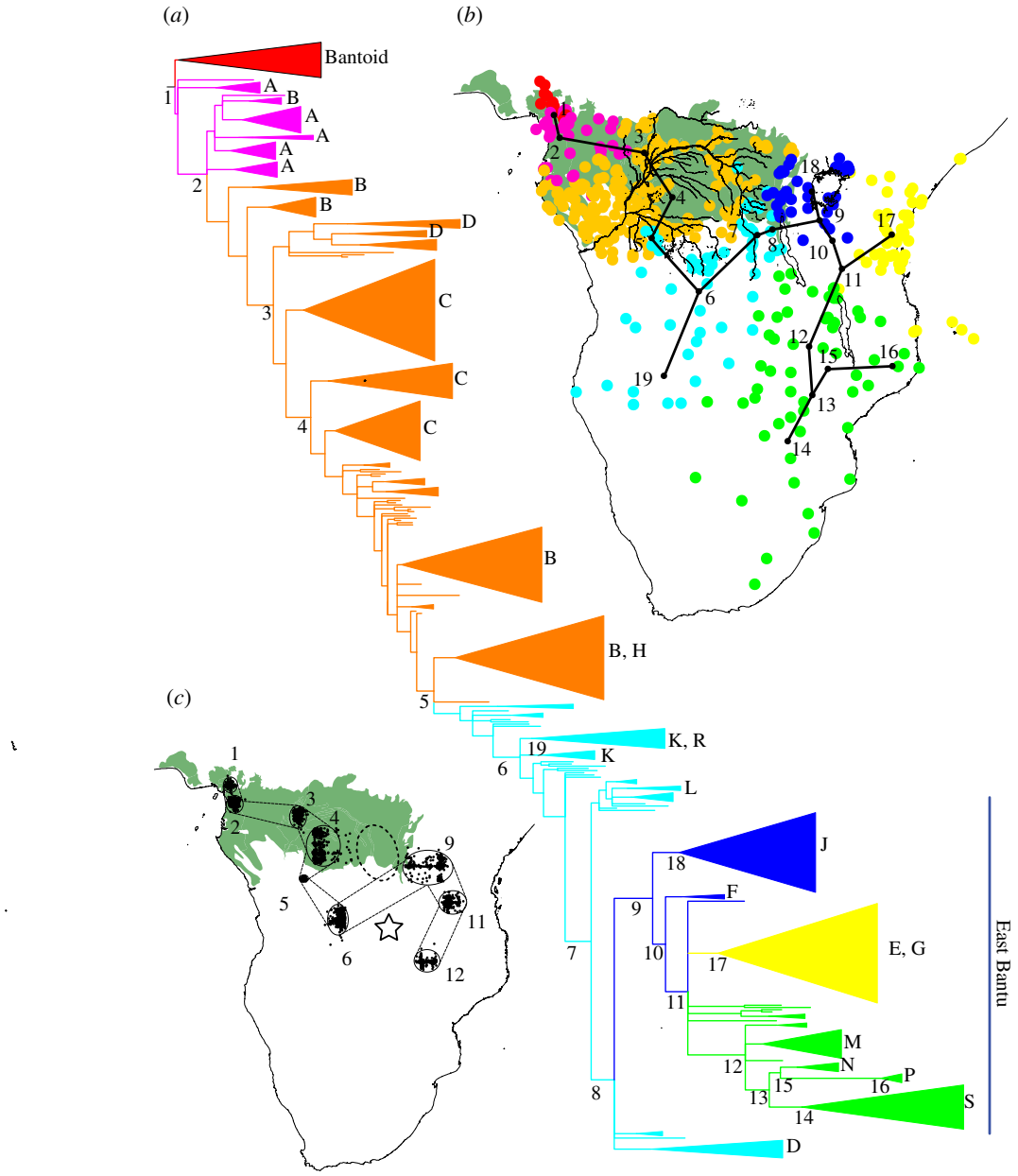


**Figure 2.** Phylogenetic tree structure and explicit phylogeographic mapping support the 'pathway through the rainforest' dispersal scenario. (*a*) Simplified consensus phylogenetic tree of 542 Bantu languages constructed from posterior sample of 500 trees. Triangle size is proportional to number of languages. Letters refer to Guthrie zones (figure 1). (*b*) Ancestral locations of numbered nodes in the tree and pathway of expansion are shown in relation to extant languages. (*c*) Distribution of inferred ancestral location of some nodes across PP trees shows uncertainty about precise locations but the overall pathway implied by these analyses remains the same. The dashed-line ellipse refers to the location of dispersal for languages leaving the rainforest as hypothesized by Rexova *et al*. [22] and the star represents the location of dispersal for East Bantu languages under the simplified rainforest-first dispersal scenario of de Filippo *et al*. [8]. These locations are not supported by the dispersal pathway inferred from our analyses.

There is some variation in the estimates of exact geographical location of these ancestral groups due to differences in the branch lengths of the PP trees. Figure 2 shows that the broad outlines of the inferred pathway of movement remain the same. Therefore, the main finding of a 'rainforest-first' route is robust to this uncertainty.

## (c) Language area and latitude

Consistent with findings from global studies, the area covered by individual Bantu languages tends to increase with increasing latitude (figure 3). Ordinary least-squares (OLS) regression suggests that latitude explains around 21 per cent of the variance in the area covered by languages (see figure 3 and electronic supplementary material, table S3). As discussed earlier, this association between latitude and area may not reflect a functional ecological relationship between these variables but could result from other historical processes. PGLS analyses reveal that the lambda parameter is significantly greater than 0 ($p < 0.001$). This indicates that there is indeed a phylogenetic signal in the residuals of the regression of latitude and language area, which violates the assumptions of OLS. When using PGLS to control for phylogeny, latitude is still a significant predictor of language area; however, the proportion of variation explained is reduced ($R^2 = 0.07$; figure 3). OLS regressions also show a significant positive relationship between language area and distance travelled from the homeland (see the electronic supplementary material, table S3). However, PGLS analyses indicate that this relationship is no longer significant after controlling

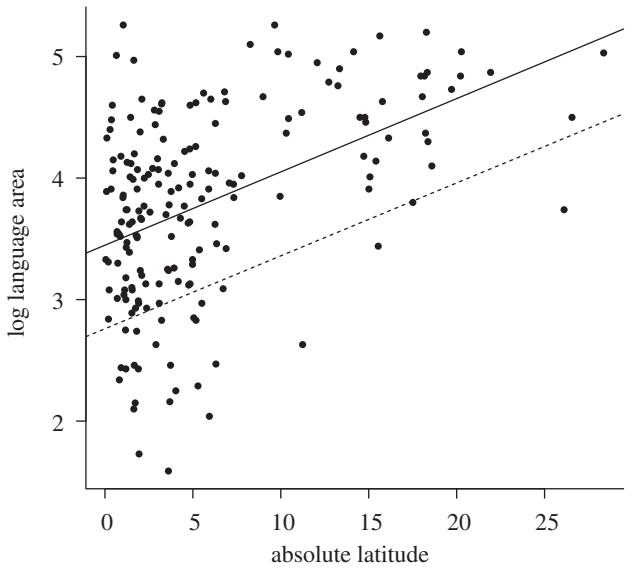


**Figure 3.** Language area increases with increasing latitude. The solid line is the ordinary least-squares (OLS) regression line ($R^2 = 0.21$), while the dotted line shows the regression line based on parameter values of the PGLS analysis, which controls for phylogenetic relatedness ($R^2 = 0.07$).

for phylogeny. This is true if distance is entered into the model alone, or in combination with latitude (see the electronic supplementary material, table S3). These results suggest that the association between the area covered by a language and latitude reflects a true ecological relationship rather than merely being the artefact of shared history. The $R^2$-value from the PGLS analyses indicating the effect of latitude (or some ecological variable related to latitude) on language area is consistent with previous effect sizes from broader-scale analyses of ecological predictors of language area in agricultural societies, which used hierarchical linear models to control for non-independence between languages [25,26].

## 4. Discussion

In this study, we have used language phylogenies in conjunction with the explicit mapping of ancestral locations to make inferences about the specific route taken during the dispersal of Bantu languages. The results clearly support the 'pathway through the rainforest' scenario for the expansion of Bantu through much of sub-Saharan Africa. There is no support in these analyses for an early, deep split between East and West Bantu languages and a movement by one branch north of the rainforest. This finding is robust to a number of sources of uncertainty, including the phylogenetic relationships between languages, the designation of which languages act as an outgroup, and the rates of movement through geographical space.

Our proposed expansion scenario also differs in other ways from those previously proposed. Rexova *et al*. [22] use an informal assessment based on the topology of their language tree to place the location of divergence of all non-forest Bantu groups near the eastern part of the Congo river (figure 2*c*). This is further east than is indicated in our analyses. This difference may be explained by the relative lack of northwest and forest languages (and from the B zone, in particular) in the Rexova *et al*. [22] sample, which may have led to a biased assessment of the location of dispersal of these languages. de Filippo *et al*. [8] place the locus of expansion of East Bantu groups west of lake Tanganyika (figure 2). This location was assigned *a priori* and was not inferred from the linguistic or genetic data. Our analyses, however, indicate that the initial divergence of East Bantu languages occurred further north and east of this point (figure 2, node 9), and that there were subsequent sequential divergence events as groups spread south from this region (e.g. the trees indicate that the divergence of the more southerly M, N, P and S groups occurred after earlier divergence from the E and G groups). This inferred pathway is in agreement with the archaeological evidence, which indicates an eastern archaeological stream linked to the development of Urewe pottery around Lake Victoria, and the spread south of the Chifumbaze complex [3].

Recently, an innovative study by Bouckaert *et al*. [14] used linguistic data to model the Indo-European language expansion through time, inferring the geographical locations of ancestral Indo-European languages at the same time as inferring phylogenetic relationships. Such an approach requires a way of translating the branches of phylogenetic trees, which are proportional to the degree of lexical change, into units of time. For Indo-European, this is facilitated by the presence of attested historical languages or other historical information that can act as reliable calibration points. Unfortunately, Bantu has been studied with much less intensity than Indo-European, and such prior information is currently absent for Bantu. An important area for future research will be to identify suitable calibration points. Potentially, archaeological data can play a role in this endeavour as it has done for studies of Austronesian-speaking societies [11]. However, it is important to ascertain that the archaeological events can be reliably matched to linguistic events, and that constraining tree topologies in this way does not introduce biases that automatically privilege one expansion scenario over another. This work is likely to require drawing further on the specialist knowledge of linguists, archaeologists and historians.

Here, we have interpreted the phylogenetic trees produced by our analyses as indicating the historical spread and diversification of populations speaking Bantu languages. An alternative hypothesis is that Bantu languages spread by a process of linguistic diffusion in the absence of population movements. There are a few well-attested cases in which populations speaking Bantu languages are the result of language-shift, e.g. Pygmy populations such as the Aka [7]. However, such an explanation is unlikely for Bantu as a whole owing to the low levels of genetic diversity across this region, and correlations between genetic and linguistic distances [8]. Another possibility is that subsequent contact and borrowing between languages are the dominant processes in shaping the present-day diversity of these languages. In support of this idea, de Filippo *et al*. [8] tested a model of isolation-by-distance (IBD) and found that geographical distance between languages was a better predictor of linguistic distance than either migration scenario. However, under real migration scenarios, we would also expect closely related languages to be close together in space. The migration scenario models used by de Filippo *et al*. [8] may not be tapping into this finer-scale correlation between linguistic and geographical distance due to the way they were simplified and specified, with distances being calculated according to a small number of way-points. Therefore, it is unclear to what extent the superior fit of the IBD model really does represent general processes of contact and borrowing.

If borrowing has been a dominant process then it may be expected to erase the phylogenetic signal in the linguistic data, making a tree-model a poor fit [40]. There are some rake-like regions in our consensus phylogeny, which possibly indicate that some degree of borrowing or non-tree-like processes such a dialect continua are likely to have been important during the history of these languages [16,20]. However, the strong support for the pattern of branching predicted by the rainforest-first model in our phylogenetic trees suggests that a tree-model is a good fit for these data. It is important to point out that simulation studies have demonstrated that phylogenetic methods can still make accurate inferences about the historical patterns of branching even if some degree of borrowing has occurred [41]. Borrowing can be conceptualized as leading to higher or lower effective rates of change in certain branches depending on how it occurs [42]. Therefore, explicitly allowing rates of lexical change to vary, as we have done here, may further limit the impact of borrowing on the ability of these methods to recover the phylogenetic history of these languages.

Our findings in this study support results from previous analyses that have shown links between proxies of ethnolinguistic group diversification and ecological variables such as latitude [25,27,43]. This study demonstrates that this relationship holds within a single language grouping. Our analyses using spatial and phylogenetic information suggest that this relationship is not the result of more southerly groups, who migrated from the north, having had less opportunity to split. Instead, these analyses strengthen the case for true ecological rules governing the distribution of ethnolinguistic groups. It is important to note, however, that the amount of variation explained by latitude is quite small. This suggests that there are other important factors affecting diversification. Previous analyses have shown how subsistence strategies and socio-political organization can affect language area and its relationship with ecological predictors [25,26]. Phylogenetic methods and spatially explicit models [44], which can deal with issues relating to non-independence due to spatial autocorrelations that may not be completely addressed by phylogenetic comparative methods, can be used to further explore how these and different aspects of ecology and social organization contribute to the diversification of human groups.

Phylogeographic studies are of key importance in understanding the origin and evolution of biological diversity [28]. Similarly, phylogenies derived from culturally transmitted information such as language in combination with explicit geographical information enable a more formal approach to testing theories in prehistory and cultural evolution [14,26]. The phylogenies generated in this study can be used to further investigate the manner and speed of spread of groups and the development of important cultural traits such as metallurgy and cattle-keeping, which are thought to have amplified the competitive advantage provided by agriculture [45,46,3]. The present study shows how cultural phylogeographic studies can provide insights into when, where and why different aspects of cultural diversity arise and are an important complement to other fields, such as genetics, archaeology and history, offering independent lines of evidence [11,47].

T.E.C. and R.M. were supported by a European Research Council Advanced Grant ADG 249347, *The Evolution of Cultural Norms in Real-World Settings*. A.M. was supported by the European Research Council Grant no. 268744, *Mother Tongue*. We thank Mark Pagel and Rebecca Grollemund for useful discussions and feedback, and two anonymous reviewers for helpful suggestions on the manuscript.

## References


1. Lewis MP. 2009 *Ethnologue: languages of the world*, 16th edn. Dallas, TX: SIL International.
2. Huffman TN. 2007 *Handbook to the Iron Age: the archaeology of pre-colonial farming societies in Southern Africa*. Scottsville, South Africa: University of KwaZulu-Natal Press.
3. Phillipson DW. 2005 *African archaeology*. Cambridge, UK: Cambridge University Press.
4. Ehret C. 2001 Bantu expansions: re-envisioning a central problem of early African History. *Int. J. Afr. Hist. Stud.* **34**, 5–41. (doi:10.2307/3097285)
5. Nurse D. 2003 *Bantu languages*. London, UK: Taylor & Francis.
6. Hombert JM, Hyman LM. 1999 *Bantu historical linguistics: theoretical and empirical perspectives*. Stanford, CA: Center for the Study of Language and Information.
7. Vansina JM. 1990 *Paths in the rainforests: toward a history of political tradition in equatorial Africa*. London, UK: James Currey Ltd.
8. de Filippo C, Bostoen K, Stoneking M, Pakendorf B. 2012 Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc. R. Soc. B* **279**, 3256–3263. (doi:10.1098/rspb.2012.0318)
9. Diamond J, Bellwood P. 2003 Farmers and their languages: the first expansions. *Science* **300**, 597–603. (doi:10.1126/science.1078208)
10. Greenhill SJ, Gray RD. 2005 Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and Austronesian languages. In *The evolution of cultural diversity: a phylogenetic approach* (eds R Mace, C Holden, S Shennan), pp. 31–52. London, UK: UCL Press.
11. Gray RD, Drummond AJ, Greenhill SJ. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858).
12. Kitchen A, Ehret C, Assefa S, Mulligan CJ. 2009 Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. B* **276**, 2703–2710. (doi:10.1098/rspb.2009.0408)
13. Gray RD, Atkinson QD. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)
14. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD. 2012 Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960. (doi:10.1126/science.1219669)
15. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC. 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075. (doi:10.1126/science.1114615)
16. Bastin Y, Coupez A, Mann M. 1999 *Continuity and divergence in the Bantu languages: perspectives from a lexicostatic study*. Tervuren, Belgium: Musée royal de l'Afrique centrale.
17. Steel MA, Hendy MD, Penny D. 1988 Loss of information in genetic distances. *Nature* **336**, 118. (doi:10.1038/336118a0)
18. Blust R. 2001 Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics* (eds C Renfrew, A McMahon, RL Trask), pp. 311–331. Cambridge, UK: McDonald Institute for Achaeological Research.
19. Holden CJ. 2002 Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* **269**, 793–799. (doi:10.1098/rspb.2002.1955)

20. Holden CJ, Gray RD. 2006 Rapid radiation borrowing and dialect continua in the Bantu languages. In *Phylogenetic methods and the prehistory of languages* (eds P Forster, C Renfrew), pp. 19–31. Cambridge, UK: McDonald Institute for Archaeological Research.
21. Holden CJ, Meade A, Pagel M. 2005 Comparison of maximum parsimony and Bayesian Bantu language trees. In *The evolution of cultural diversity: a phylogenetic approach* (eds R Mace, CJ Holden, S Shennan), pp. 53–65. London, UK: Left Coast Press.
22. Rexova K, Bastin Y, Frynta D. 2006 Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* **93**, 189–194. (doi:10.1007/s00114-006-0088-z)
23. Marten L. 2006 Bantu classification, Bantu trees and phylogenetic methods. In *Phylogenetic methods and the prehistory of languages* (eds P Foster, C Renfrew), pp. 43–55. Cambridge, UK: McDonald Institute for Archaeological Research.
24. Zwickl DJ, Hillis DM. 2002 Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **51**, 588–598. (doi:10.1080/10635150290102339)
25. Currie TE, Mace R. 2009 Political complexity predicts the spread of ethnolinguistic groups. *Proc. Natl Acad. Sci. USA* **106**, 7339–7344. (doi:10.1073/pnas.0804698106)
26. Currie TE, Mace R. 2012 The evolution of ethnolinguistic diversity. *Adv. Complex Syst.* **15**, 1150006. (doi:10.1142/S0219525911003372)
27. Nettle D. 1999 *Linguistic diversity*. Oxford, UK: Oxford University Press.
28. Wiens JJ, Donoghue MJ. 2004 Historical biogeography, ecology and species richness. *Trends Ecol. Evol.* **19**, 639–644. (doi:10.1016/j.tree.2004.09.011)
29. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314. (doi:10.1126/science.1065889)
30. McMahon A, McMahon R. 2005 *Language classification by numbers*. Oxford, UK: Oxford University Press.
31. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008 Languages evolve in punctuational bursts. *Science* **319**, 588. (doi:10.1126/science.1149683)
32. Pagel M, Atkinson QD, Meade A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)
33. Greenhill SJ, Atkinson QD, Meade A, Gray RD. 2010 The shape and tempo of language evolution. *Proc. R. Soc. B* **273**, 2443–2450. (doi:10.1098/rspb.2010.0051)
34. Tuffley C, Steel M. 1998 Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63–91. (doi:10.1016/S0025-5564(97)00081-3)
35. Lupyan G, Dale R. 2010 Language structure is partly determined by social structure. *PLoS ONE* **5**, e8559. (doi:10.1371/journal.pone.0008559)
36. Felsenstein J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15. (doi:10.1086/284325)
37. Venditti C, Meade A, Pagel M. 2011 Multiple routes to mammalian diversity. *Nature* **479**, 393–396. (doi:10.1038/nature10516)
38. Mace R, Pagel M. 1994 The comparative method in anthropology. *Curr. Anthropol.* **35**, 549–564. (doi:10.1086/204317)
39. Maho J. 2003 A classification of the Bantu languages: an update of the Guthrie referential system. In *Bantu languages* (ed. D Nurse, G Philippson), pp. 639–651. London, UK: Taylor & Francis.
40. Holden CJ, Shennan S. 2005 Introduction to part I. How tree-like is cultural evolution? In *The evolution of cultural diversity: a phylogenetic approach* (eds R Mace, CJ Holden, S Shennan), pp. 13–29. London, UK: Left Coast Press.
41. Greenhill SJ, Currie TE, Gray RD. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)
42. Currie TE, Greenhill SJ, Mace R. 2010 Is horizontal transmission really a problem for phylogenetic comparative methods? A simulation study using continuous cultural traits. *Phil. Trans. R. Soc. B* **365**, 3903–3912. (doi:10.1098/rstb.2010.0014)
43. Mace R, Pagel M. 1995 A latitudinal gradient in the density of human languages in North-America. *Proc. R. Soc. Lond. B* **261**, 117–121. (doi:10.1098/rspb.1995.0125)
44. Rangel TF, Diniz JAF, Bini LM. 2010 SAM: a comprehensive application for spatial analysis in macroecology. *Ecography* **33**, 46–50. (doi:10.1111/j.1600–0587.2009.06299.x)
45. Holden CJ, Mace R. 2003 Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proc. R. Soc. Lond. B* **270**, 2425–2433. (doi:10.1098/rspb.2003.2535)
46. Silva F, Steele J. 2012 Modeling boundaries between converging fronts in prehistory. *Adv. Complex Syst.* **15**, 1150005. (doi:10.1142/S0219525911003293)
47. Renfrew C, Forster P. 2006 Introduction to phylogenetic methods and the prehistory of languages. In *Phylogenetic methods and the prehistory of languages* (eds P Forster, C Renfrew), pp. 1–8. Cambridge, UK: McDonald Institute for Archaeological Research.