THE ROYAL
SOCIETY

Received 20 July 2001
Accepted 21 December 2001
Published online 25 March 2002

Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis

Clare Janaki Holden

Department of Anthropology, University College London, Gower Street, London WC1E 6BT, UK (*c.holden@ucl.ac.uk*)

Linguistic divergence occurs after speech communities divide, in a process similar to speciation among isolated biological populations. The resulting languages are hierarchically related, like genes or species. Phylogenetic methods developed in evolutionary biology can thus be used to infer language trees, with the caveat that 'borrowing' of linguistic elements between languages also occurs, to some degree. Maximum-parsimony trees for 75 Bantu and Bantoid African languages were constructed using 92 items of basic vocabulary. The level of character fit on the trees was high (consistency index was 0.65), indicating that a tree model fits Bantu language evolution well, at least for the basic vocabulary. The Bantu language tree reflects the spread of farming across this part of sub-Saharan Africa between *ca.* 3000 BC and AD 500. Modern Bantu subgroups, defined by clades on parsimony trees, mirror the earliest farming traditions both geographically and temporally. This suggests that the major subgroups of modern Bantu stem from the Neolithic and Early Iron Age, with little subsequent movement by speech communities.

Keywords: Bantu; linguistic evolution; maximum parsimony

1. INTRODUCTION

Linguistic divergence occurs after speech communities divide, in a process similar to speciation among isolated biological populations. In consequence, many languages are related in a hierarchical, tree-like pattern like genes or species. As among biological taxa, descent groups among languages (i.e. languages sharing a unique common ancestor) are defined by the presence of shared linguistic innovations, equivalent to derived characters in biology. It is therefore possible to apply biological phylogenetic methods to linguistic data to infer language trees.

Despite the similarities between linguistic and biological evolution, formal cladistic methods for estimating phylogeny, developed in evolutionary biology, have rarely been applied to linguistic data. An exception is the parsimony tree of Austronesian languages of Gray & Jordan (2000). To construct language trees, linguists use either the 'comparative method' or lexicostatistical methods. In the comparative method, linguistic innovations are used to define descent groups. This method is thus cladistic, only counting derived not primitive characters as evidence for descent. However an explicit, computer-implemented optimality criterion is not used. In lexicostatistical methods, linguists build trees by comparing overall similarity among language pairs across a standard vocabulary list (e.g. Henrici 1973). In biological terms, this method is phenetic, meaning that no distinction is made between derived and ancestral similarity. Therefore, if the rate of evolution varies across the tree, overall similarity trees produce a misleading estimate of genealogical relationships (Sober 1988).

In this analysis, maximum-parsimony trees of 75 Bantu and Bantoid languages were constructed. Bantu is a large group of about 450 related languages, spoken across Africa south of 5° N (figure 1). Bantu is a subset of the Bantoid linguistic group (Williamson & Blench 2000). Maximum parsimony is an optimality criterion that operates directly on discrete data (such as a list of cognate words), minimizing the number of character changes on the tree, or tree length. Bootstrap analysis was used to investigate the level of support in the data for individual clades on the tree.

Guthrie (1967–71) classified Bantu languages into 15 zones based on geographical and linguistic criteria, labelled A to S. His coding system is still used today. Previously published lexicostatistical trees indicated that the northwestern Bantu languages, comprising Guthrie's zones A and B, were the most divergent (Heine 1973; Henrici 1973; Bastin *et al.* 1999). Non-northwestern Bantu languages are usually divided into West and East Bantu. West Bantu includes zones H, J, K, L, R, parts of D and usually M. These languages are spoken in the equatorial forest, modern Zambia and southwest Africa. East Bantu comprises languages of zones E, F, J, N, P and S. These languages are spoken in East and southeast Africa (Williamson & Blench 2000).

Parsimony trees of Bantu languages were compared with archaeological evidence for the spread of farming in this part of Africa. The aim was to investigate how far the Bantu language tree may reflect broader cultural history in this region.

Some authors have questioned whether a tree model can describe language evolution, because there is some borrowing or diffusion of linguistic elements between neighbouring speech communities (Bastin *et al.* 1999; Hinnebusch 1999). If borrowing is widespread then relationships among languages will be reticulate or net-like, not tree-like. How far relationships among languages are tree-like is part of a wider debate on the level of interconnection between human cultures (Moore 1994; Bellwood 1996).

DOI 10.1098/rspb.2002.1955

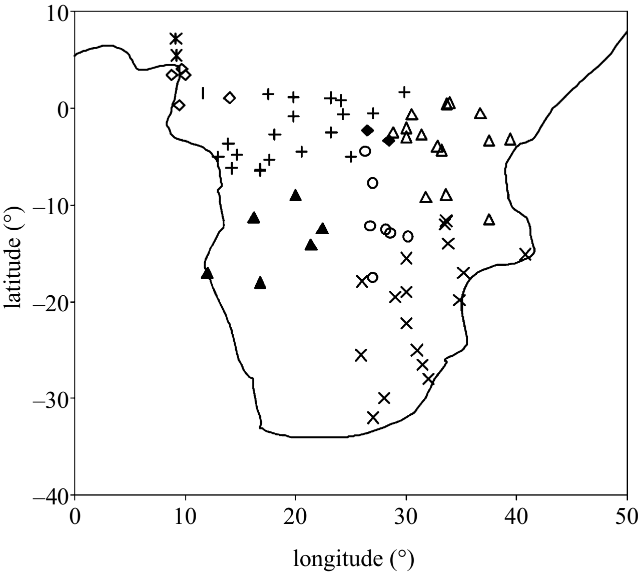

Figure 1. Map of modern Bantu languages. Subgroups were defined following the maximum-parsimony tree (cf. figure 2).

✳ Bantoid
◇ northwest
+ West Bantu, equatorial
▲ West Savannah
◆ Binja and Lega
○ Central
△ East Bantu, East Africa
× East Bantu, southeast

It is particularly controversial whether Bantu language history was tree-like. In favour of the tree model, Bantu languages are thought to have spread and diversified along with the expansion of farming in this region over the past 2000 years (Huffman 1982). This suggests that a tree model, based on branching and divergence, might be appropriate. Alternatively, the diffusion of Bantu words between neighbouring speech communities is thought to have been widespread (Vansina 1990; Bastin *et al.* 1999; Hinnebusch 1999). It has also been argued that the early stages of Bantu linguistic evolution would be better represented as dialect chains, not clearly bounded, discrete speech communities (Vansina 1995). The level of character fit on the tree was used here to assess how far a tree model accounts for patterns of similarity across the languages in the dataset. Cases of borrowing will appear to be homoplastic on a tree, that is, they will appear as convergent evolution of the same character in different branches, decreasing the level of character fit. A high level of character fit is consistent with a tree model in which words are mainly transmitted by inheritance within speech communities.

2. MATERIAL AND METHODS

Trees were constructed for 73 Bantu and two Bantoid languages. The two Bantoid languages, Tiv and Ejagham, were used as outgroups to root the tree. Tiv is spoken in central Nigeria. Ejagham is spoken by the Ekoi in Cameroon. Both regions are likely locations for the ancestors of Bantu-speaking populations (Vansina 1990, p. 49). The 75 languages in the dataset were also included in the *Ethnographic atlas* of Murdock (1967), a cross-cultural database. This will enable this tree to be used to analyse the evolution of cultural traits. Languages were matched to ethnographic populations by name and geographical location. Further ethnonyms and dialect names were found in Voeghlin & Voeghlin (1977) and Middleton & Rassam (1995). Where more than one language sample was available for an ethno-linguistic group, the language sample that was geographically closest to the focus of ethnographical study was chosen.

The linguistic data comprised of 92 items of basic vocabulary (or meanings) that had previously been coded for cognates by Bastin *et al.* (1999). Cognates are meanings whose form shares a common root in two or more languages. Possibly borrowed forms were also coded as cognate if they shared a common root (Bastin *et al.* 1999, p. 8). For each meaning, different cognate sets were treated as character states. 'Basic vocabulary' refers to meanings that are present in all languages; examples include 'man', 'woman', 'one', 'two', 'tongue' and 'ashes'. The word forms of basic vocabulary are thought to change more slowly than other meanings, being less subject to linguistic innovation or borrowing from neighbouring languages. For this reason, the basic vocabulary of related languages can retain common cognates for hundreds or even thousands of years (Swadesh 1971). Meanings that refer to non-universal aspects of culture, such as 'iron', are not counted as basic vocabulary. The list of 92 items of basic vocabulary was a subset of Swadesh's 100-word standard list of basic vocabulary. Eight meanings from the 100-word list had been previously excluded by Bastin *et al.* (1999) as they were not present in the Bantu languages (e.g 'snow'; see Bastin *et al.* (1983)). Languages with more than 5% missing data were not included in the analysis.

The heuristic search option in Paup* 4.0 (Swofford 1998) was used to search for maximum-parsimony trees. This algorithm does not guarantee finding the shortest tree, but an exhaustive-search algorithm could not be used because the number of languages in the sample was too large. Five hundred replications using tree bisection–reconnection (TBR) with random addition were performed, storing 2000 trees in memory per search. Character states were unordered. Meanings with more than one character state (cognate form) in a particular language were treated as polymorphic.

Characters were then reweighted using the rescaled consistency index (RC), according to their fit on the 88 unweighted trees. Using weighted parsimony gives meanings less prone to homoplasy (from borrowing or convergence) greater weight in building the tree. Another heuristic search for the shortest tree was run using TBR with random addition, with 1000 replications and storing up to 2000 trees in memory.

Bootstrap analysis, which provides a conservative test of the level of support in the data for individual clades on the tree, was performed on the weighted characters. Five hundred bootstrap pseudo-replicates were sampled, using heuristic search with 20 TBR random addition replicates. Clades recovered in the bootstrap analysis are probably supported by several linguistic innovations, and languages in that clade must have few or no conflicting relationships (e.g. resulting from linguistic borrowing) with any language outside that clade.

Maximum-parsimony Bantu trees were compared with the spread of farming in the Neolithic and Early Iron Age (EIA) using archaeological data from Sutton (1972), Schmidt (1975), Huffman (1982, 1989, p. 76), Clist (1987, 1989), Denbow (1990) and Phillipson (1993). The geographical distribution of modern Bantu subgroups, as defined from the parsimony tree, was compared with the spread of early farming traditions in this part of Africa. Archaeological dates associated with early farming traditions were compared with the branching order of clades on the language tree.

3. RESULTS

Using unweighted parsimony, 88 trees were found with a tree length of 2533. The consistency index excluding uninformative characters (CI) was 0.65, the retention index (RI) was 0.59 and the RC index was 0.38. Using RC weighted parsimony, three trees were found with a tree length of 841.208 47. CI was 0.72, RI was 0.68 and RC was 0.49. Variation among the three weighted-parsimony trees was among branch lengths of languages of Bantu zone H. The first RC weighted-parsimony tree is shown in figure 2, in which languages are labelled by name and also by the code of Guthrie (1967–1971), taken from Bastin *et al.* (1999). The parsimony tree can therefore be compared with Guthrie's classification of Bantu languages. The major subgroups on the tree are also indicated and may be compared with their geographical distribution, shown in figure 1.

The following discussion of the topology of parsimony trees refers to all maximum-parsimony trees (unweighted and weighted), unless otherwise stated.

The deepest three splits on the parsimony trees are within languages of the northwest, in Bantu zones A and B.

The West Bantu languages were paraphyletic, that is, they did not share a unique common ancestor. The largest group within the West Bantu languages includes zones C and H, Bira D32 and Kumu D37. Zones K and R are also included in this group on most trees (figure 2). Apart from zones K and R, these languages are spoken in and around the Central African forest. A 'Central' clade of languages spoken in and around modern Zambia was present on all trees. Languages in the Central clade include Lala M52, Lamba M54, Bemba M42, Luba L33, Songe D10, Kaonde L42 and Tonga M64 (figure 2). The West Savannah languages (zones K and R), spoken in southwest Africa, form a clade. Languages in zones K and R remained distinct within this clade. The position of this clade varied across trees. On some unweighted-parsimony trees and on the weighted trees, these languages were most closely related to languages spoken in the rainforest, mainly zones C and H (figure 2). By contrast, on some unweighted-parsimony trees the West Savannah clade was an outlier attached to the clade containing the Central and East Bantu languages.

East Bantu languages were monophyletic. East Bantu languages appear to be descended from a West Bantu-like ancestor. The Central clade is coordinate with East Bantu. Within East Bantu, zone J languages (Lakes Bantu) form a clade. Also within East Bantu, a large clade comprising 16 southeastern languages was present on all trees. This clade included all languages of Bantu zones N (excluding Tumbuka N21) and S. Bantu zones N and S remained distinct within this clade (figure 2).

Bootstrap analysis recovered many clades found in the weighted-parsimony tree, but not all. Many higher nodes on the tree (towards the root) were not recovered in bootstrap analysis. The bootstrap tree is shown in figure 3, in which branch lengths are proportional to the proportion of replicates containing that clade, also shown by node labels.

Four northwestern languages formed a clade in the bootstrap analysis, Bakoko A43, Fang A75, Duala A24 and Kota B25. Among equatorial West Bantu languages, the following two clades were recovered: zone H languages plus Teke B73 formed a clade, and seven zone C languages formed a clade (Kela C75, Mongo C61/3, Nkundo C61/1, Tetela C71, Lele C84 and Sakata C34). The West Savannah languages of zones K and R formed a clade; zones K and R remained distinct within this clade. The Central clade was not recovered in the bootstrap analysis but some subgroups within this group were: a clade including Lala M52, Lamba M54 and Bemba M42 was recovered, in addition to a clade including Luba L33, Songe D10 and Kaonde L42.

The East Bantu clade was not recovered at the 50% level in the bootstrap analysis. The Lakes clade was also not recovered in the bootstrap analysis, but two separate clades within the Lakes languages (zone J) were recovered: Hima J13, Ganda J15, Soga, J16 and Zinza J23; and Rundi J62 and Rwanda J61. The southeastern clade (zone N, excluding Tumbuka N21, and zone S) was recovered in the bootstrap analysis. Zones N and S remained distinct within this clade (figure 3).

The geographical distribution of modern Bantu subgroups, shown in figure 1, may be compared with archaeological traditions associated with the spread of farming in this part of Africa, summarized in figure 4. Comparing figures 1 and 4 reveals that the geographical distribution and temporal order of early farming traditions mirror the distribution and branching order of the modern Bantu subgroups (see § 4).

4. DISCUSSION

(a) *Language transmission: borrowing or inheritance?*

The level of homology or character fit on the tree is indicated by the consistency and retention indices (Swofford 1991). Character fit on the Bantu parsimony trees may be compared with character fit for morphological or genetic data on biological trees (Sanderson & Donoghue 1989). Compared with biological trees with a similar number of taxa, the level of character fit on the Bantu tree was high. This is consistent with a predominantly branching pattern of evolution for basic vocabulary in the Bantu languages, with relatively little borrowing or convergence. Borrowing might be expected to be more common for items of non-basic vocabulary, for example, words for technological innovations such as 'iron' (Vansina 1990, pp. 59–60).

The Bantu tree may also be compared with a parsimony tree of 77 Austronesian languages, spoken in Southeast Asia and the Pacific, constructed by Gray & Jordan (2000). On the Austronesian tree the consistency index was 0.25. The lower level of character fit on the Austronesian language tree is probably partly because Gray & Jordan used a wider lexical dataset to construct their tree, including non-basic items of vocabulary that are more subject to borrowing. Nonetheless, the high consistency and retention indices for the Bantu languages found in the present analysis are striking, suggesting that a tree model of linguistic diversification may fit Bantu better than Austronesian. This is surprising because Bantu languages are spoken across the continental land mass of sub-Saharan Africa, whereas most Austronesian languages are spoken

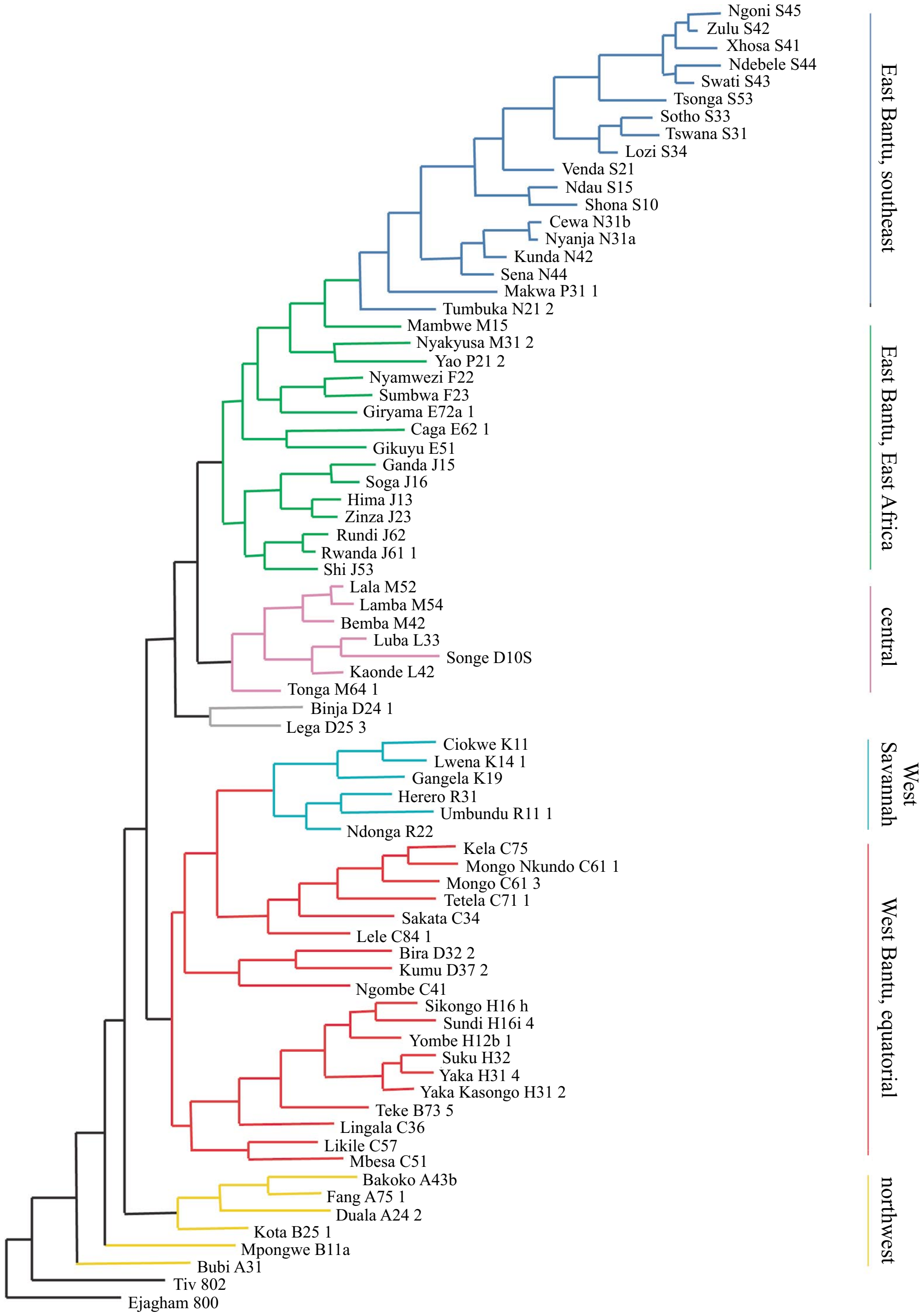

Figure 2. Maximum-parsimony tree of 75 Bantu and Bantoid languages (weighted parsimony). Languages are labelled by name and Guthrie code. Colours show subgroups and labels show the geographical areas where the main subgroups are found (cf. figure 1).

by island populations, which one might expect to be more geographically isolated by the sea. In the Bantu languages, social factors rather than geographical barriers must have maintained distinct speech communities.

Bantu linguistic relationships are closely correlated with the geographical distance between languages: geographically proximate languages also tend to cluster together on the tree (Henrici 1973; Bastin *et al.* 1999; Hinnebusch 1999). Authors have previously argued that a correlation between linguistic and geographical distance implies that borrowing, not descent, is the main process underlying observed patterns of linguistic variation. For example, Bastin *et al.* (1999, p. 1) argued, 'if geography can be reconstructed from comparative linguistic data, geography must have played a large part in shaping those relationships'. Dewar (1995) made a similar argument with respect to Madagascan languages. However, a correlation between geographical and linguistic distance does not necessarily imply extensive borrowing among languages. Closely related populations are also often geographically clustered, because most populations remain near their area of origin. The correlation between the language tree and

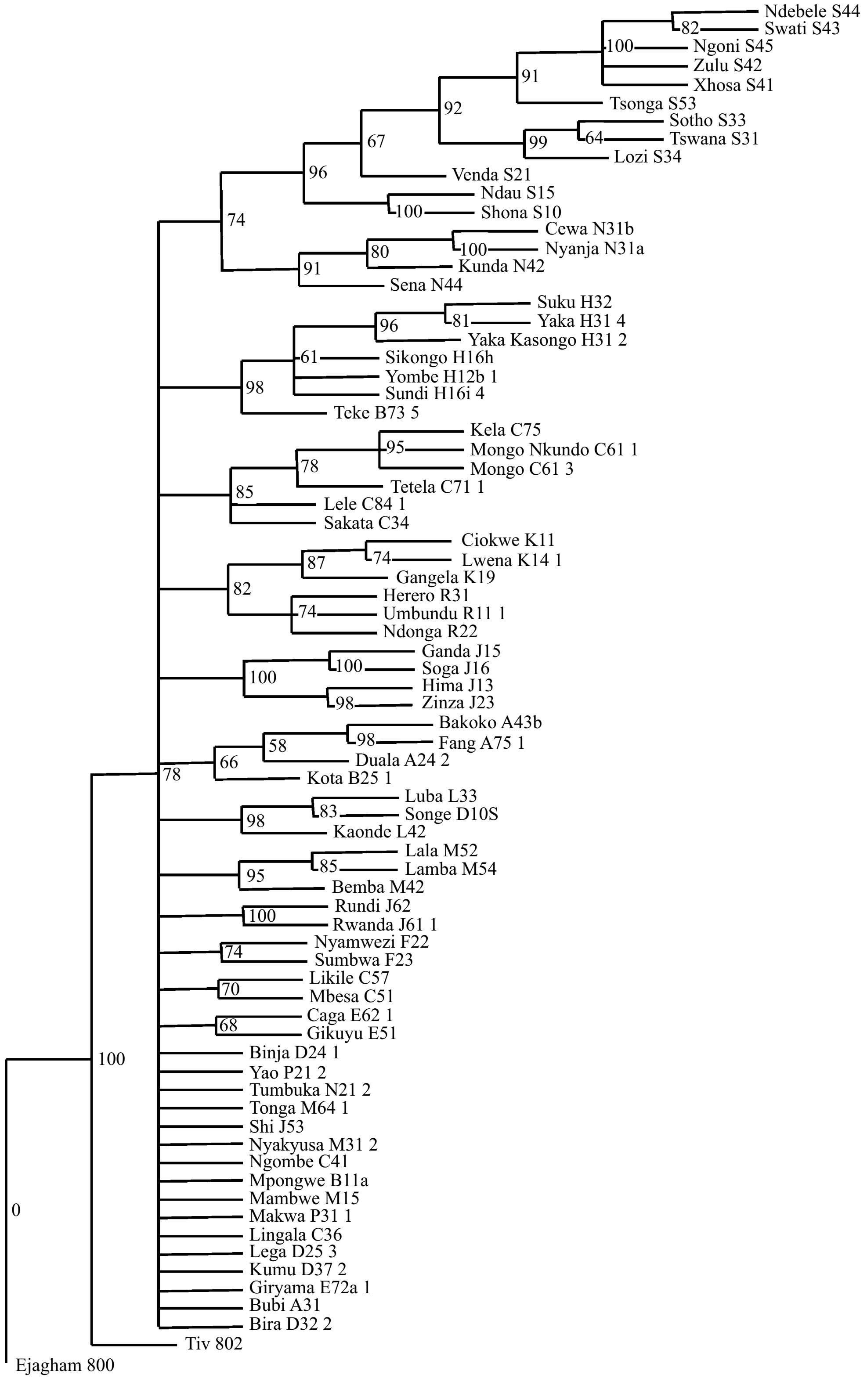

Figure 3. Bootstrap analysis of RC reweighted Bantu lexical characters, showing clades present in 50% or more of bootstrap pseudo-replicates.

independent archaeological evidence for population history in this region suggests that history rather than geography probably underlies many of the observed similarities.

(b) *Comparison with lexicostatistical trees*

In comparison with previously published lexicostatistical trees, maximum-parsimony trees showed the greatest similarity overall to the tree of Heine (1973) and to the branch-average tree of Bastin *et al.* (1999) (cf. Henrici (1973) and other trees in Bastin *et al.* (1999)).

Like parsimony trees, most previously published lexicostatistical trees found the northwestern Bantu languages to be the most divergent. However, there was disagreement among previously published trees as to which languages

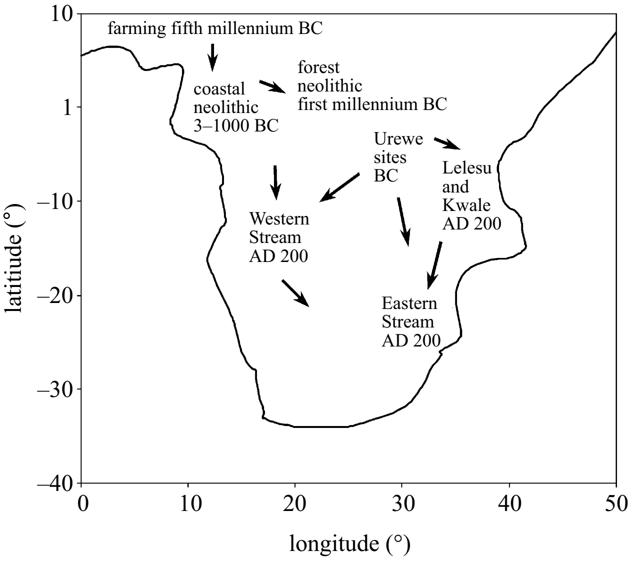

Figure 4. Archaeological traditions associated with the spread of farming across the Bantu-speaking region, following Phillipson (1993) and other sources (see § 4c). Arrows show probable relationships between archaeological traditions as hypothesized by archaeologists from ceramic styles and dates.

were included among these northwestern outliers (cf. Heine 1973; Henrici 1973; Bastin *et al.* 1999). The branch-average tree of Bastin *et al.* (1999) showed the greatest similarity to the parsimony trees presented here, with their northwestern outliers including zone A languages and Mpongwe B11.

On most previously published lexicostatistical trees the West Bantu languages are paraphyletic, as here (Heine 1973; Henrici 1973; cf. Bastin *et al.* 1999). A West Bantu clade containing languages of zones C, H, K and R was present in the group average and branch-average trees of Bastin *et al.* (1999), in agreement with the parsimony tree presented here (figure 2). The ambiguous position of languages in zones R and K has also been found in previous analyses and is thought to result from widespread linguistic borrowing from East Bantu to West Savannah languages (Ehret 1998, pp. 44–45). The internal organization of the East Bantu Lakes languages on parsimony trees agrees with traditional classifications of these languages (Schoenbrun 1998).

Relationships among languages within the East Bantu clade partly agree with the branch-average tree of Bastin *et al.* (1999), on which East Bantu languages were split into two clusters, one including zones E, F and M (Central and East African languages), the other including zones N, P and S (southeastern languages). However, other previously published lexicostatistical trees, including the group average tree of Bastin *et al.* (1999), showed different relationships among East Bantu languages, for example, indicating that zone S languages were the most divergent.

(c) *Bantu language trees and archaeology*

It has long been thought that Bantu languages were introduced to east and southern Africa by the spread of farmers (Huffman 1982; Phillipson 1993) and this has also been argued for West Bantu languages in the Central African forest (Vansina 1984, 1990). Some admixture with local populations, with acculturation of the latter, probably also occurred (Huffman 1982; Vansina 1990).

Maximum-parsimony Bantu language trees mirror closely the spread of farming across this region of Africa in the Neolithic and EIA. The branching order of major subgroups on Bantu parsimony trees and the geographical distribution of those subgroups overlap with distinct archaeological traditions associated with the spread of farming between *ca.* 3000 BC and AD 500 (figures 1, 2 and 4). This overlap is clearer than on previously published Bantu trees (e.g. Heine 1973; Henrici 1973; Bastin *et al.* 1983, 1999).

The earliest Neolithic sites were found in approximately the same area as the northwestern languages of zones A and B ('Northwest' Bantu; figures 1, 2 and 4). Sites were found in coastal Gabon (3000–1000 BC) and at Obobogo near Yaoundé in Cameroon (1000–600 BC) (Clist 1989).

During the first millennium BC, farming spread across and south of the forest to the areas where West Bantu languages of zones C and H and Teke B73 are spoken today ('West Bantu equatorial'; figures 1, 2 and 4). Neolithic (and probable Neolithic) traditions in this region are diverse. They include Tchissanga on the Congo coast (from 580 BC) (Denbow 1990); Okala near Libreville in Gabon (510–320 BC) (Clist 1987, 1989); Imbonga in Equateur, northern Zaire (440–90 BC) (Eggert 1993); Ngovo in Bas Zaire (200 BC to AD 100) (Clist 1989); Batalimo in Centrafrique and the related Maluba tradition in northern Zaire (from 140 BC) (Clist 1989).

By contrast to the northwest and central forest regions, the first farmers in eastern and southern Africa belonged to the EIA. All EIA archaeological traditions in this region are closely related, but three distinct traditions have been identified—in east, southeast and southwest Africa (Phillipson 1993).

The earliest dates for the EIA have been obtained for sites belonging to the Urewe tradition around Lake Victoria (in Uganda, Rwanda, Burundi and eastern Democratic Republic of Congo), probably dating to the last centuries BC (Sutton 1972; Schmidt 1975). Urewe sites are found where Lakes Bantu languages are spoken today. Lega D25 and Binja D24 are also spoken in this area today (cf. Ehret 1998, p. 34). Other EIA traditions in east Africa include Lelesu in Tanzania and Kwale in Kenya, the latter dated to *ca.* AD 200 (Soper 1971; Phillipson 1993; Shillington 1995). In the Lelesu and Kwale areas, modern Bantu languages include Gikuyu E51, Caga E62, Giryama E72, Sumbwa F23 and Nyamwezi F22 (figures 1, 2 and 4).

From the second century AD, the EIA spread into southern Africa in two distinct traditions, called the Eastern Stream in southeast Africa and the Western Stream in southwest Africa (Phillipson 1993). The Eastern Stream correlates with the distribution of modern Bantu languages of zones S and N. Parsimony trees suggest that there was a single long spread of East Bantu languages from around Lake Victoria into the rest of East Africa and then to southeast Africa (figures 1, 2 and 4).

Western Stream archaeological sites correlate with modern Central and West Savannah languages (figures 1, 2 and 4). Both the West Savannah and Central Bantu clades split before the origin of East Bantu, and the earliest populations speaking these languages were probably Neolithic (e.g. Denbow 1990). There was probably horizontal transfer of EIA technology from East Bantu-speaking populations in East Africa to both Central Bantu- and West Savannah-speaking populations (Ehret 1998, pp. 44–45). For example, at Tchissanga on the Congo coast, Neolithic sites have been dated to 580 BC. EIA sites dated to *ca.* second century AD were found nearby at Madingo-Kayes (Denbow 1990). Having acquired EIA ceramics and iron technology, West Savannah-speaking populations probably then spread south from the Congo, forming the Western Stream of the EIA expansions (Huffman 1989, p. 76).

In summary, maximum-parsimony trees of Bantu languages are highly consistent with archaeological evidence for the spread of farming across the whole area of modern Bantu-speaking Africa, although this conclusion must currently remain tentative for clades that receive little bootstrap support, especially the basal splits on the tree (figure 3). The correlation between archaeology and language groups suggests that the major subgroups of modern Bantu stem from the Neolithic and EIA, with little subsequent movement by speech communities. It also suggests that the modern languages within each subgroup evolved *in situ* in these areas.

This research was funded by a Wellcome Trust Fellowship in Bioarchaeology. The author thanks the AHRB-funded Centre for the Evolutionary Analysis of Cultural Behaviour Group at UCL, R. Mace, R. Gray, F. Jordan and two anonymous referees for their comments on earlier versions of the paper.

REFERENCES

Bastin, Y., Coupez, A. & de Halleux, B. 1983 Classification lexicostatistique des langues bantoues (214 releves). *Bull. Seanc. Acad. R. Sci. Outre-Mer Meded Zitt. K. Acad. Overzeese Wet.* **27**, 173–199.

Bastin, Y., Coupez, A. & Mann, M. 1999 Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study. *Annales, Sciences humaines*, **162**.

Bellwood, P. 1996 Phylogeny vs reticulation in prehistory. *Antiquity* **70**, 881–890.

Clist, B. 1987 Early Bantu settlements in West-Central Africa: a review of recent research. *Curr. Anthropol.* **28**, 380–382.

Clist, B. 1989 Archaeology in Gabon, 1886–1988. *Afr. Archaeol. Rev.* **7**, 59–95.

Denbow, J. 1990 Congo to Kalahari: data and hypotheses about the political economy of the western stream of the Early Iron Age. *Afr. Archaeol. Rev.* **8**, 139–176.

Dewar, R. E. 1995 Of nets and trees: untangling the reticulate and dendritic in Madagascar's prehistory. *World Archaeol.* **26**, 301–318.

Eggert, M. K. H. 1993 Central Africa and the archaeology of the equatorial rainforest: reflections on some major topics. In *The archaeology of Africa; food, metals and towns* (ed. T. Shaw, P. Sinclair, B. Andah & A. Okpoko), pp. 289–329. London: Routledge.

Ehret, C. 1998 *An African classical age: eastern and southern Africa in world history 1000 B.C. to A.D. 400*. Oxford: James Currey.

Gray, R. D. & Jordan, F. M. 2000 Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055.

Guthrie, M. 1967–1971 *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages*, vols 1–4. Farnborough, UK: Gregg International.

Heine, B. 1973 Zur genetischen Gliederung der Bantu-Sprachen. *Afr. Übersee* **56**, 164–185.

Henrici, A. 1973 Numerical classification of the Bantu languages. *Afr. Language Stud.* **14**, 82–104.

Hinnebusch, T. J. 1999 Contact and lexicostatistics in comparative Bantu studies. In *Bantu historical linguistics: theoretical and empirical perspectives* (ed. J.-M. Hombert & L. M. Hyman), pp. 173–205. Stanford, CA: Center for the Study of Language and Information.

Huffman, T. N. 1982 Archaeology and ethnohistory of the African Iron Age. *A. Rev. Anthropol.* **11**, 133–150.

Huffman, T. N. 1989 *Iron age migrations*. Johannesburg: Witwatersrand University Press.

Middleton, J. & Rassam, A. (eds) 1995 *Encyclopaedia of world cultures*, IX. Africa and the Middle East. Boston, MA: G. K. Hall & Co.

Moore, J. H. 1994 Putting anthropology back together again: the ethnogenetic critique of cladistic theory. *Am. Anthropol.* **96**, 925–948.

Murdock, G. P. 1967 *Ethnographic atlas*. University of Pittsburgh Press.

Phillipson, D. W. 1993 *African archaeology*, 2nd edn. Cambridge University Press.

Sanderson, M. J. & Donoghue, M. J. 1989 Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795.

Schmidt, P. R. 1975 A new look at interpretations of the Early Iron Age in East Africa. *Hist. Afr.* **2**, 127–136.

Schoenbrun, D. L. 1998 *A green place, a good place: agrarian change, gender, and social identity in the Great Lakes region to the 15th century*. Oxford: James Currey.

Shillington, K. 1995 *History of Africa*, revised edn. London: Macmillan.

Sober, E. 1988 *Reconstructing the past: parsimony, evolution, and inference*, pp. 72–78. Cambridge, MA: MIT.

Soper, R. 1971 A general review of the Early Iron Age of the Southern half of Africa. *Azania* **6**, 5–37.

Sutton, J. E. G. 1972 New radiocarbon dates for eastern and southern Africa. *J. Afr. Hist.* **13**, 1–24.

Swadesh, M. 1971. *The origin and diversification of language* (ed. J. Sherer & D. Hymes). Chicago, IL: Aldine.

Swofford, D. L. 1991 When are phylogeny estimates from molecular and morphological data incongruent? In *Phylogenetic analysis of DNA sequences* (ed. M. M. Miyamoto & J. Cracraft), pp. 295–333. New York: Oxford University Press.

Swofford, D. L. 1998 *Paup* Phylogenetic analysis using parsimony (*and other methods)*, v. 4. Sunderland, MA: Sinauer.

Vansina, J. 1984 Western Bantu expansion. *J. Afr. Hist.* **25**, 129–145.

Vansina, J. 1990 *Paths in the rainforests: toward a history of political tradition in equatorial Africa*. London: Currey.

Vansina, J. 1995 New linguistic evidence and 'The Bantu Expansion'. *J. Afr. Hist.* **36**, 173–195.

Voeghlin, C. F. & Voeghlin, F. M. 1977 *Classification and index of the world's languages*. New York: Elsevier.

Williamson, K. & Blench, R. 2000 Niger-Congo. In *African languages: an introduction* (ed. B. Heine & D. Nurse), pp. 11–42. Cambridge University Press.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.