Naturwissenschaften (2006) 93: 189–194
DOI 10.1007/s00114-006-0088-z

SHORT COMMUNICATION

Kateřina Rexová · Yvonne Bastin · Daniel Frynta

Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data

Received: 6 June 2005 / Accepted: 6 January 2006 / Published online: 15 March 2006

Abstract The phylogeny of the Bantu languages is reconstructed by application of the cladistic methodology to the combined lexical and grammatical data (87 languages, 144 characters). A maximum parsimony tree and Bayesian analysis supported some previously recognized clades, e.g., that of eastern and southern Bantu languages. Moreover, the results revealed that Bantu languages south and east of the equatorial forest are probably monophyletic. It suggests an unorthodox scenario of Bantu expansion including (after initial radiation in their homelands and neighboring territories) just a single passage through rainforest areas followed by a subsequent divergence into major clades. The likely localization of this divergence is in the area west of the Great Lakes. It conforms to the view that demographic expansion and dispersal throughout the dry-forests and savanna regions of subequatorial Africa was associated with the acquisition of new technologies (iron metallurgy and grain cultivation).

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org.10.1007/s00114-006-0088-z

K. Rexová
Department of Philosophy and History of Sciences,
Charles University, Viničná 7,
Praha 2, 128 44, Czech Republic
e-mail: kloskacka@centrum.cz

Y. Bastin
Department of Ethnomusicology and Linguistics,
Musée Royal de l'Afrique Centrale,
Leuvensesteenweg 13,
Tervuren 3080, Belgium

D. Frynta (✉)
Department of Zoology,
Charles University, Viničná 7,
Praha 2, 128 44, Czech Republic
e-mail: frynta@centrum.cz
Tel.: +420-22-1951846
Fax: +420-22-1951841

Introduction

Bantu expansion is an excellent example of prehistoric colonization on a continental scale. A fairly homogenous population of Bantu speaking people had spread from the northwest of the equatorial forest in Cameroon and Nigeria throughout Central, Eastern, and Southern Africa. Although this process was completed in the recent past, our knowledge is based mostly on indirect archeological (e.g., de Maret 1984; Posnansky 1968), anthropological (e.g., Murdock 1959; Vansina 1990), genetic (e.g., Cavalli-Sforza et al. 1994; Pereira et al. 2002), and/or linguistic (e.g., Greenberg 1972; Guthrie 1967–1971; Heine 1973; Nurse 1997) evidence. The linguistic arguments have especially contributed to widely accepted interpretations of the Bantu expansion (Guthrie 1962; Vansina 1995). Nevertheless, traditional comparative linguistics was limited by the huge number of Bantu languages and incompleteness of records. Therefore, most authors have focused on lexical data. The extensive comparative material on proto-Bantu roots collected by Guthrie (1967–1971) was converted to similarity matrices and treated by different tree-building methods (Flight 1988; Henrici 1973). These analyses were, however, limited by the small number of included languages. Consequently, standardized lexical data available in an almost complete set of languages/dialects was introduced (Bastin et al. 1983, 1999; Coupez et al. 1975). Nevertheless, the computational approach used by lexicostatistics, although reasonable (e.g., Embelton 1986), is purely phenetic. The resulting trees were therefore not fully appropriate to the historical reconstruction of language evolution (Hoijer 1956; Nurse 1994–1995).

Thus, an introduction of phylogenetic methodology (Gray and Jordan 2000; Rexová et al. 2003) was needed. This crucial step was performed by Holden (2002) who calculated a maximum parsimony tree for 73 Bantu languages. In spite of its methodological purity, even this study suffers from a low number of characters. Therefore, we searched for additional characters. Grammatical data, although usually considered to be less prone to borrowing

(e.g., Nurse 1994–1995) and therefore, more suitable for the purpose of genetic classification, was overlooked and never used yet for the construction of maximum parsimony trees.

The aim of this study is: (1) to analyze a combined data set consisting of both lexical and grammatical data, (2) to perform a phylogeographic interpretation of the resulting trees, and (3) to discuss the usefulness of phylogenetic methodology in the case of Bantu languages.

The aim of this study is: (1) to analyze a combined data set consisting of both lexical and grammatical data, (2) to perform a phylogeographic interpretation of the resulting trees, and (3) to discuss the usefulness of phylogenetic methodology in the case of Bantu languages.

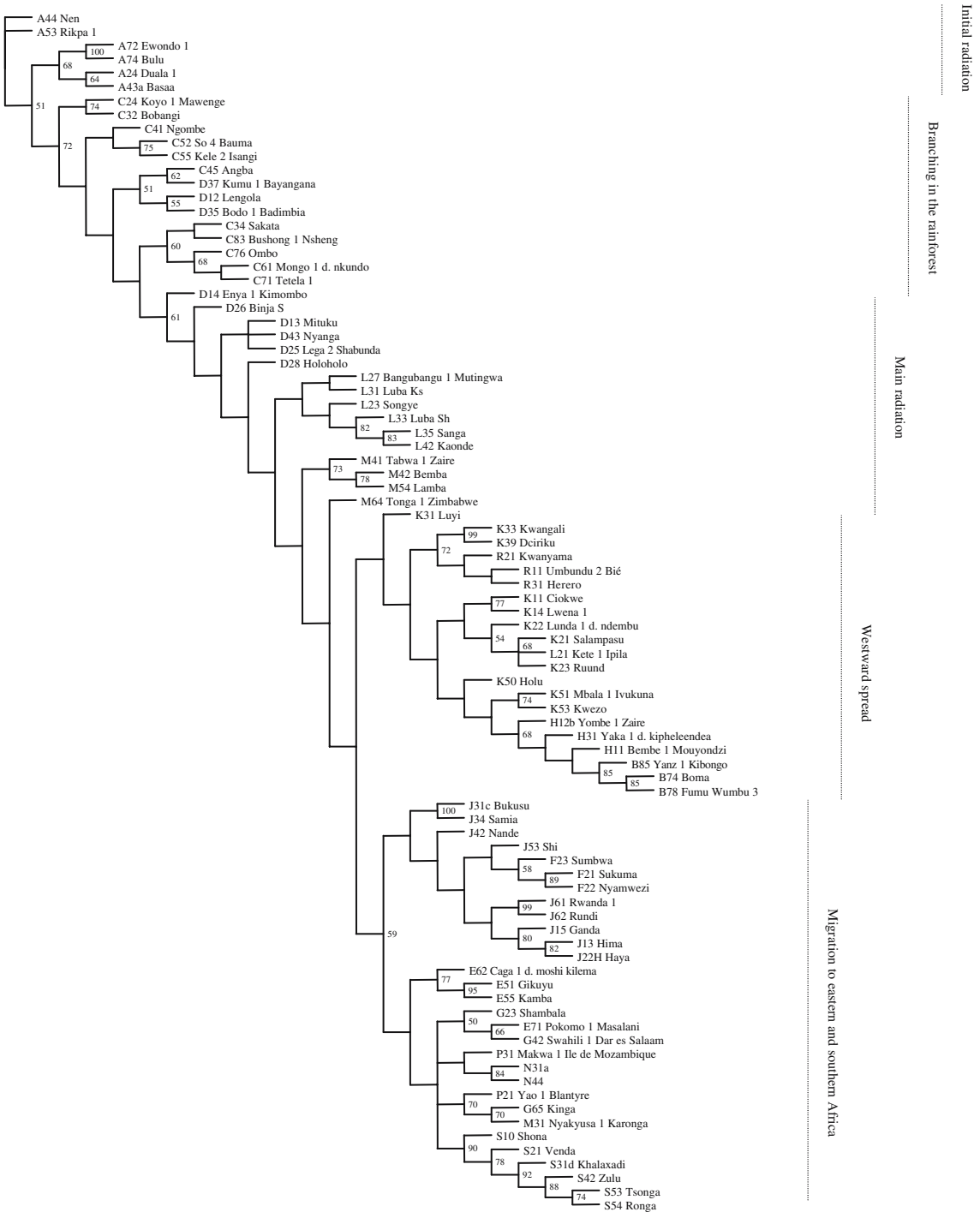

Fig. 1 Maximum parsimony tree of 87 languages based on 92 lexical and 52 grammatical characters. *Numbers* indicate bootstrap values (>50). Language names are adopted from Bastin et al. 1999 and their alphanumerical codes provide information about the geographic location of the language (Guthrie 1967–1971)

Materials and methods

We used the lexical data set collected by Bastin et al. (1999). These data were originally generated for the lexicostatistical analyses of languages, i.e., they are recognized cognate linguistic forms according to the criteria of comparative linguistics. Lexical data were set up on the basis of a list of 100 lexical items (meanings) primarily chosen by Swadesh (1955) and reduced to 92 meanings by Yvonne Bastin. It is widely accepted that words corresponding to meanings on the list are relatively insensitive to borrowing. The grammatical data set (see S1 for description of grammatical characters) consisting of 52 phonological and morphological items was collected by Bastin et al. (1979).

Hence, 87 Bantu languages with both lexical (92) and grammatical (52) items were available for further analyses. The small number of items complicated the separate analyses of lexical and grammatical matrices, especially in the case of grammatical data analysis (see S2 and S3 for separate analyses of grammatical and lexical data set). Although a partition-homogeneity test revealed some statistically significant (P=0.01) incongruence between the lexical and grammatical matrices, we followed the total evidence approach.

The multistate matrix of 144 items was generated and processed by PAUP software (version 4.0b4a (Swofford 2000]). A heuristic search ("addseq = random" and "nreps = 1,000") was performed to find the most parsimonious trees. Items with more than one form in a single language (synonyms) were treated as polymorphic character states. Bremer indices (up to 4) and bootstrap procedure (nreps = 1,000) were computed to indicate support of the individual clades. In addition, weighted parsimony trees were constructed ("reweight") to reduce the effect of possible borrowing. To root the tree of Bantu languages, we used the Nen (A44). This language was selected because it comes from the source region of Bantu expansion. Moreover, in our preliminary analyses of lexical matrices including some Bantoid non-Bantu languages (i.e., the putative sister groups) the Nen language has kept a stable position on the base of the tree.

Next, we recoded the matrix into binary form and further treated it into a MrBayes 3.0B4 (Huelsenbeck and Ronquist 2001). We followed Gray and Atkinson (2003) who performed an analysis of Indo-European languages in adjusting the parameters (samplefreq = 10,000; burnin = 300,000; others: default).

Results

Maximum parsimony analysis generated eight trees. The topology of the consensus tree (length=3,198, CI=0.46, RI=0.55, and RC=0.25; see Fig. 1) was almost the same as that based on three trees resulting from RC weighted-parsimony (length=7,72.53, CI=0.56, RI=0.59, and RC=0.33), therefore we will further discuss the former one.

As expected, all languages in the A zone (Cameroon) are placed on the tree base. The clade consisting of all other Bantu languages (B–S zones) was supported (bootstrap BS=72 and Bremer index Br=2). On the base of this clade, there are offshoots belonging to the C and D zones (N and NE Congo-Kinshasa), three of which are probably monophyletic. The rest of the Bantu languages belong to a well-supported superclade (BS=61 and Br>3). It contains several languages of zone D (E Congo-Kinshasa), L (SE Congo-Kinshasa, Br>3), M (S Congo-Kinshasa, Zambia), and a well-supported clade of Eastern and Southern Bantu languages (BS=59 and Br=2) covering the zones E, F, G, J, M, N, P, and S (from Uganda and Kenya to South Africa). The remaining languages of the superclade form a weakly supported western clade (Br=1) consisting of several clades covering the zones K, L, and R (SW Congo-Kinshasa, Angola, Zambia) and a single well-supported clade (BS=68 and Br>3) of the zones H and B (Congo-Brazzaville, SW Congo-Kinshasa).

Almost all statistically supported clades revealed by maximum parsimony were corroborated by alternative Bayesian methodology (Fig. 2). The differences in the tree topology concern the position of a few languages within C–D zones. The superclade has remained well supported (posterior probability P=95%). The basalmost offshoots of the superclade formed a single basal clade (D zone, P=99%), the remaining ones joined the western clade (zones L, M, and D28, P=93%).

Discussion

Our results suggest a scenario of Bantu expansion consisting of the following steps (see Fig. 3): (1) an initial radiation in Cameroon (A zone); (2) the subsequent branching in the rainforest areas of Congo-Kinshasa (C and D zones); (3) main radiation somewhere in SE Congo-Kinshasa W of the Tanganyika Lake (D and possibly L and M zones); (4) westward spread from this area to K, R, H, and B zones; and (5) migration from the area of main radiation to E and S Africa (J, F, E, G, N, P, S zones).

This scenario disagrees with the results of a previous phylogenetic analysis (Holden 2002) and the current interpretations of Bantu migration (Newman 1995; Vansina 1995) in one important point. We supported the existence of a monophyletic superclade containing all the Bantu languages found in the territories south and east of the rainforest areas of Congo-Kinshasa. Consequently, not one of the western Bantu languages south of the rainforest areas was clustered together with their northern neighbors. If further proved, the scenarios of Bantu migration should be changed considerably. Because the monophyletic group containing both northern and southern Bantu languages from Western Africa was not supported, the early split of the western and eastern branches of Bantu languages in the north rainforest (C zone) areas followed by the migration of the Western group to W Congo-Kinshasa, Congo-Brazzaville, and Angola, as proposed by Curtin et al. (1995) and Vansina (1990), lose its substantiation. The main phylogenetic signal of our data favors the colonization of Angola, SW Congo-Kinshasa and surrounding territories from the more eastern source areas.

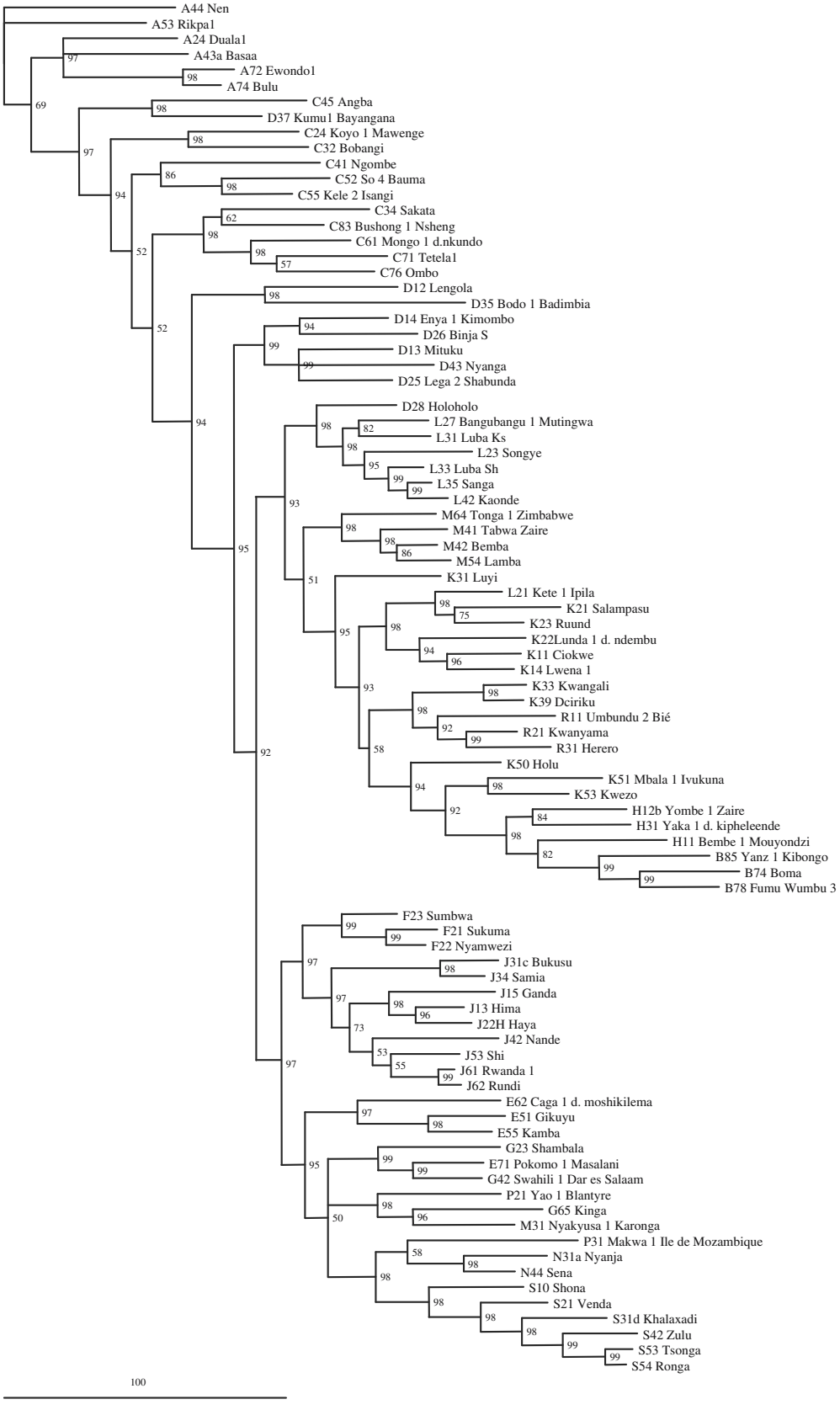

Fig. 2 Bayesian tree of 87 languages. *Numbers* indicate posterior probabilities. Language names are adopted from Bastin et al. (1999) and their alphanumerical codes provide information about the geographic location of the language (Guthrie 1967–1971)

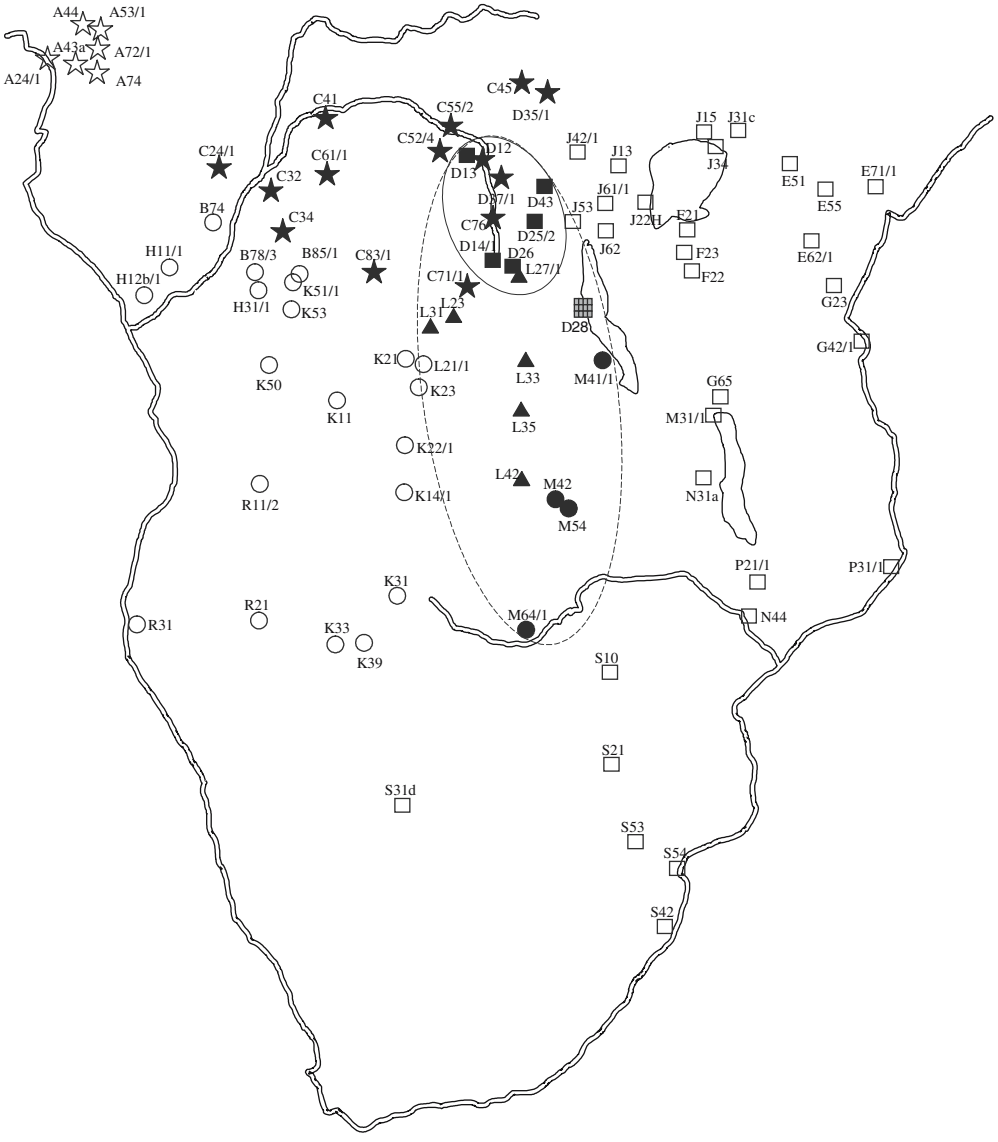

Fig. 3 Geographical distribution of studied languages. The individual clades are visualized by symbols. *Open asterisks* denote the putative initial radiation of Bantu languages; *solid asterisks* the subsequent branching in the rainforest; *solid squares* the basalmost clade of the "main radiation" (*small ellipse*); *solid triangles*, *solid circles*, and *patterned squares* other basal clades of the main radiation (*big ellipse*) placed to the western branch by Bayesian analysis; *open circles* the western clade; and *open squares* the eastern and southern Bantu clade

Our scenario is more intuitive than complex models of Bantu migration proposed by Vansina (1995). It requires just a single passage through the rainforest areas and/or Congo Basin, a single acquisition of the new technologies (iron metallurgy and grain cultivation), possibly somewhere around the Great Lakes, and the consequent demographic expansion followed by dispersal throughout the dry-forests and savanna regions of subequatorial Africa. The time and place of acquisition of the above mentioned technologies are, however, a matter of never ending discussion and disputes.

Our language tree corroborated the results of the recent cross-cultural linguistic comparison carried out by Ehret (1998). Our clade of the Eastern and Southern Bantu languages apparently corresponds to his "Mashariki" group and the four previous offshoots to his "Western savanna," "Botatwe," "Sabi," and "Central savanna" groups, respectively. Moreover, Ehret (1998) provides independent evidence supporting close relationships of all the above-mentioned groups forming a monophylum in our tree.

It should be noted that some ideas of early scholars are congruent with our phylogeographic scenario. Guthrie (1962) in his comprehensive study of Bantu languages concluded that "the bush country to the south of the equatorial forest midway between the two coasts" was a nucleus from where Bantu languages have radiated. Oliver (1966) combined opinions with Greenberg (1966) and Guthrie (1962) and proposed a two-step model including an initial radiation in Cameroon/Nigeria, followed by migration to the "nucleus" and subsequent radiation south of the equatorial forest. Also, Flight (1988) concluded that savanna Bantu formed a single branch.

The rest of the clades do not greatly contradict the views of recent scholars (Bastin et al. 1999). In general, the affinities of languages in local scale are well understood and phylogenetic methodology may not substantially improve the existing classification.

In our previous study (Rexová et al. 2003), we demonstrated a good correspondence between the results of cladistics and traditional comparative linguistics. Never-

theless, there are some specific pitfalls in the case of Bantu languages. Neighboring languages are frequently intelligible and do not behave as fully independent units of evolution analogous to biological species. Borrowing and convergence (Hinnebusch 1999) may obscure the observed pattern of cultural evolution and the analyses suffer from lowered congruence among studied characters. Although these processes may explain lower consistency indices found in phylogenetic analyses of Bantu languages (Holden 2002); this study) when compared to Indo-European ones (Rexová et al. 2003), the methodology used is robust enough to recognize the hidden phylogenetic signal.

References

Bastin Y, Coupez A, De Halleux B (1979) Statistiques lexicale et grammaticale pour la classification historique des langues bantoues. Bull Séances Acad R Sci O-M 23(3):375–387

Bastin Y, Coupez A, De Halleux B (1983) Classification lexicostatistique des langues bantoues (214 relevés). Bull Séances Acad R Sci O-M 27(2):173–199

Bastin Y, Coupez A, Mann M (1999) Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study. Annales, Sciences humaines 162:315–317

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and geography of the human genes. Princeton Univ. Press, Princeton

Coupez A, Evrard E, Vansina J (1975) Classification d'un échantillon de langues bantoues d'après la lexicostatistique. Africana Linguistica 6:133–158

Curtin P, Feierman S, Thompson L, Vansina J (1995) African history. Longmann, New York

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

Embelton SM (1986) Statistics in historical linguistics. Brockmeyer, Bochum

Flight C (1988) Bantu trees and some wider ramifications. Afr Lang Cult 1(1):25–43

Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435–439

Gray RD, Jordan FM (2000) Language tree supports the express-train sequence of Austronesian expansion. Nature 405:1052–1055

Greenberg JH (1966) The languages of Africa, 2nd edn. The Hague, Mouton

Greenberg JH (1972) Linguistic evidence regarding Bantu origins. J Afr Hist 23(2):189–216

Guthrie M (1962) Some developments in the prehistory of the Bantu languages. J Afr Hist 3(2):273–282

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

Heine B (1973) Zur genetischen Gliederung der Bantu-Sprachen. Afr Übersee 56:164–185

Henrici A (1973) Numerical classification of Bantu languages. Afr Lang Stud 14:82–104

Hinnebusch TJ (1999) Contact and lexicostatistics in comparative Bantu studies. In: Hombert J-M, Hayman LM (eds) Bantu historical linguistics: theoretical and empirical perspectives. Centre for the Study of Language and Information, Stanford, pp 173–205

Hoijer H (1956) Lexicostatistics: a critique. Language 32:49–60

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

Huelsenbeck JP, Ronquist F (2001) MrBayes version 3.0B4

de Maret P (1984) L'archéologie en zone bantu jusqu'en 1984. Muntu 1(2):37–60

Murdock GP (1959) Africa: its peoples and their culture history. McGraw-Hill, New York

Newman JL (1995) The peopling of Africa. Yale Univ. Press, New Haven, London

Nurse D (1994–1995) "Historical" classifications of the Bantu languages. Azania 29–30:65–81

Nurse D (1997) The contributions of linguistics to the study of history in Africa. J Afr Hist 38:359–391

Oliver R (1966) The problem of the Bantu expansion. J Afr Hist 7 (3):361–376

Pereira L, Gusmão Alves C, Amorim A, Prata MJ (2002) Bantu and European Y-lineages in sub-Saharan Africa. Ann Hum Genet 66:369–378

Posnansky M (1968) Bantu genesis—archaeological reflexions. J Afr Hist 9(1):1–11

Rexová K, Frynta D, Zrzavý J (2003) Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. Cladistics 19:120–127

Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. Int J Am Linguist 21:121–137

Swofford DL (2000) PAUP, version 4.0b4a. Sinauer, Sunderland

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

Vansina J (1995) New linguistic evidence and "the Bantu expansion". J Afr Hist 36:173–195