Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cladistics 19 (2003) 120–127
Cladistics
www.elsevier.com/locate/yclad

Cladistic analysis of languages: Indo-European classification based on lexicostatistical data

Kateřina Rexová,[a,b,*] Daniel Frynta,[a] and Jan Zrzavý[c]

[a] *Department of Zoology, Charles University, Viničná 7, CZ-128 44 Praha 2, Czech Republic*
[b] *Department of Philosophy and History of Sciences, Charles University, Viničná 7, CZ-128 44 Praha 2, Czech Republic*
[c] *Department of Zoology, Faculty of Biological Sciences, University of South Bohemia, Branišovská 31, CZ-370 05 České Budějovice, Czech Republic*

Accepted 15 July 2002

Abstract

The phylogeny of the Indo-European (IE) language family is reconstructed by application of the cladistic methodology to the lexicostatistical dataset collected by Dyen (about 200 meanings, 84 speech varieties, the Hittite language used as a functional outgroup). Three different methods of character coding provide trees that show: (a) the presence of four groups, viz., Balto-Slavonic clade, Romano-Germano-Celtic clade, Armenian-Greek group, and Indo-Iranian group (the two last groups possibly paraphyletic); (b) the unstable position of the Albanian language; (c) the unstable pattern of the basalmost IE differentiation; but (d) the probable existence of the Balto-Slavonic–Indo-Iranian (''satem'') and the Romano-Germano-Celtic (+Albanian?) superclades. The results are compared with the phenetic approach to lexicostatistical data, the results of which are significantly less informative concerning the basal pattern. The results suggest a predominantly branching pattern of the basic vocabulary phylogeny and little borrowing of individual words. Different scenarios of IE differentiation based on archaeological and genetic information are discussed.

Introduction

The story of language confusion in Babel written in the first book of the Torah (Pentateuch) of the Old Testament is the earliest known text devoted to the diversity of the languages. Starting from the 16th century, the classification of languages has attracted extensive and never-ending research effort. The scientific methodology of classification based on phonetic laws and comparison of closely related languages and/or dialects was introduced by comparative and historical linguists of the 19th century, culminating in the formation of the Neogrammarian school (Brugmann and Delbrück, 1911; Paul, 1880). Many languages are believed to be related in a hierarchical tree-like pattern, and it is therefore possible to apply cladistic methods to infer language trees. Despite this, the barrier between the humanities and the sciences is probably the main reason cladistic methodology has not until very recently been introduced into comparative linguistics for the evaluation of lexical data. The only exceptions are the recent studies by Gray and Jordan (2000) on Austronesian languages and by Holden (2002) on Bantu and Bantoid languages. Also, the phylogenetic algorithm used by Warnow (1997) and Ringe et al. (2002) to assess relationships among ancient Indo-European languages is essentially cladistic. However, the ''comparative'' approach to phonology has been intuitively close to cladistics in emphasizing shared innovations (see Hoenigswald, 1965), but no *explicit* optimality criterion has been used by the comparative linguists. On the contrary, the traditional ''lexicostatistical'' approach to classification of languages is substantially phenetic, based purely on general similarities. The earliest quantitative lexicostatistical method, the glottochronology introduced by Swadesh (1952), may thus serve as a linguistic analogy to the molecular clock in biology (see Gudschinsky, 1964). It is especially the case for studies based on basic vocabularies.

The present study is an attempt to apply the cladistic methodology to the analysis of the basic

* Corresponding author.
E-mail address: kloskacka@centrum.cz (K. Rexová).

0748-3007/02/$ - see front matter © 2002 The Willi Hennig Society. Published by Elsevier Science (USA). All rights reserved.
doi:10.1016/S0748-3007(02)00147-0

vocabulary data. To substantiate this approach, we assume that:

1. *individual* languages are the subjects of cultural evolution (the problem of *general* biological basis of the language can be ignored here) and preserve their continuity throughout long time scales;
2. the evolution of languages is mostly divergent; and
3. the language is transmitted as a whole, and the frequency of borrowing (i.e., horizontal transmission of individual characters) between languages is low.

While the former premise is well substantiated, the latter two appear questionable. However, there is a priori no reason to expect that cases of "word introgression" and language hybridization (Dixon, 1999) are relatively more frequent than the genetic hybridization within some plant and animal taxa, especially at the species level (Arnold, 1997; Harrison, 1993; for the cladistic approach to the phylogenetic reticulations, see Skála and Zrzavý, 1994, and references therein). The large amount of reticulate evolution should be identified by the analysis (e.g., as a conflict between linguistic and genetic trees of the same human populations), not imposed prior to it. Some of the available cladistic analyses of the history of languages are indeed consistent with relatively little word borrowing (see Holden, 2002). Moreover, although there are languages of hybrid origin, they represent an entirely special issue, usually well identifiable prior to phylogenetic analysis, and most languages (including the Indo-European ones) are evidently not hybrids (see Ringe et al., 2002, for a discussion).

For the present study, we have selected the most extensively studied linguistic family, viz., that of the Indo-European ("IE" hereinafter) languages. The origin of IE languages is at the center of the problem of the origin of Europeans (as only three non-IE-speaking populations, viz., relic Basque and lately introduced Uralic and Altaic people, inhabit Europe). Almost all possible hypotheses have been formulated (for a review see Cavalli-Sforza et al., 1994, pp. 263–266) but two possible areas of the IE origin deserve attention: Anatolia (and the spread of IE languages synchronous with the early Neolithic expansion of the agriculture; "Renfrew model" hereinafter) and the Ukraine (and the radiation of pastoral IE tribes during the Eneolithic period; "Gimbutas model" below). Owing to additional historical, archaeological. and biological information which may serve as independent evidence, the IE family is a suitable model for testing the new classification procedures. We adopted the list of 200 basic meanings which had been chosen for lexicostatistical purposes by Swadesh (1952). It is widely accepted that words corresponding to meanings of the list are little sensitive to borrowing (see Dyen et al., 1992).

The aim of this study is to (1) perform a phylogenetic analysis of lexical (basic vocabulary) data using the maximum parsimony approach; (2) evaluate robustness of the trees to different modes of recognition of the primary homologies and to different character-coding strategies; (3) compare the results with earlier analyses of the same dataset carried out by classical lexicostatistical methodology (Dyen et al., 1992); (4) evaluate correspondence between phylogenetic tree and the history of IE expansion as reported by archaeology, history, and genetics; and (5) discuss the applicability of cladistic methodology in the linguistic classification.

Materials and methods

As a source of primary data we used the set collected by Isidore Dyen and presented as file IE-DATA1.f2l at www.ntu.edu.au/education/langs/ielex/IE-DATA1 (cf. Dyen et al., 1992). This file contains the forms (words) used in the 95 modern IE *speech varieties* (i.e., languages, dialects, and creoles) for each of 200 meanings. The forms are classified into *cognate classes* recognized by Dyen et al. (1992) according to the criteria of comparative linguistics. A cognate class includes all words for a given meaning which are presumably homologous by "descent with modification." From the file IE-DATA1.f2l we selected 84 speech varieties (redundant data concerning 11 Slavonic varieties suspect for methodological bias were ultimately excluded, as had been suggested by Dyen et al. (1992)).

Naturally, it is difficult to identify a justifiable functional outgroup since the linguistic family (e.g., IE) is usually defined as a monophyletic group containing *all* the languages that have putatively diverged from a single protolanguage (Crystal, 1985). Therefore, no reliable outgroups outside the studied linguistic group are available prior to global cladistic analysis of the human languages. To root the tree of IE languages, we used the ancient Hittite language (based on the data presented by Tischler (1973)). This decision is supported by Sturtewant (1962), suggesting that Indo-Hittite (= IE *s. lat.*) languages consist of two sister groups, the extinct Anatolian languages (including Hittite, Palaic, Lydian, Luvian, and Lycian) and the IE languages *s. str.* (see also Ringe et al., 2002, for similar decision). The rooted trees presented below should be considered preliminary, and their topology should be compared with the unrooted "trees" based on the analyses of the same IE languages, excluding the Hittite.

The original dataset consists of cards used in data collection. It was therefore necessary to convert it into a matrix useful for further analysis. Using alternative procedures, we recoded the original dataset into three different matrices.

Standard multistate matrix

This is a matrix of 85 speech varieties by 200 characters (meanings). Character state corresponds to

individual cognate classes suggested by the linguistic methods, i.e., to putative primary homologies (*sensu* de Pinna, 1985). Forms that occur simultaneously within a single language (synonyms) and belong to more than one cognate class (e.g., German *Haupt* and *Kopf*, both meaning "head" but only the former being cognate with English *head*) were treated as polymorphic character states. In 25 characters, the number of character states exceeded 32, i.e., the upper limit required by PAUP. Therefore, 141 character states, which have been recorded only in a single speech variety (autapomorphies), were coded as missing values ("?") to avoid this software limitation.

Altered multistate matrix

The standard multistate matrix was done by linguistic methodology, and we checked here the robustness of cognate recognition by comparison of this matrix with another multistate matrix that is based on more superficial similarities. We joined some cognate classes originally reported as distinct but showing obvious superficial similarities each to other. This procedure was done to reduce subjectivity of linguistic classification of forms into the cognate classes (cf. Ringe, 1999), which tends to separate numerous cognate classes when borrowing is presumed or when the homology is not certain. We joined, e.g., North Sardinian *pakos* and Albanian *pak* (both meaning "few") into a single cognate class. Note, however, that we do not attempt to make up a different ("better") classification of the cognate classes but to assess the robustness of the trees derived from the original dataset by Dyen et al. (1992). Polymorphic character states were not allowed.

Binary matrix

Each cognate class of words was treated as a separate character coded as absent/present ("0/1"). The resulting matrix included 2456 characters.

The matrices were then analyzed by PAUP (Version 4.0b4a; Swofford, 2000). First routine "hsearch" ("addseq = random, nreps = 1000") was performed to find the most parsimonious trees. Bremer indices and bootstrap procedure ("nrep = 100" and "nrep = 1000", respectively) were computed to indicate support of the individual clades.

Results

Maximum parsimony analysis of the standard multistate matrix generated 405 trees. The consensus tree (length 4294, CI 0.89, RI 0.93; Fig. 1a) includes all the widely recognized elementary IE linguistic groups, all with a strong support. They include Slavonic (bootstrap support = 100, Bremer index >5), Baltic (97, >5), Indic (99, >5), Iranian (98, >5), Greek (100, >5), Armenian (100, >5), Romance (= Italic; 100, >5), Germanic (100, >5), Albanian (100, >5), and Celtic (100, >5) groups. Some previously suggested but uncertain taxa (subfamilies) of the IE higher classification have received considerable support. They are Balto-Slavonic (90, 1), Indo-Iranian (96, 1), and Romano-Germanic (67, 1) subfamilies. The interesting Balto-Slavonic–Indo-Iranian superclade (59, 1) and the unorthodox Romano-Germano-Albanian-Celtic superclade have appeared in the tree, the latter receiving weak support only (35, 0).

The altered multistate matrix (336 trees, length 4255, CI 0.84, RI 0.90) produced a consensus tree with major clades identical to those derived from the standard multistate matrix (Fig. 1b). The differences concern the branching patterns within the elementary linguistic groups only. As concerns the unorthodox higher taxa, the Romano-Germano-Albanian-Celtic superclade has received marginal bootstrap support (56, 0), but the Balto-Slavonic–Indo-Iranian superclade is not supported here (39, 0).

Parsimony analysis of the binary matrix generated six trees (length 4988, CI 0.49, RI 0.77; Fig. 1c). All the elementary linguistic groups as well as Balto-Slavonic (99, >5) and Romano-Germanic (71, 1) superclades were strongly supported. However, the main branching pattern differs considerably from those produced by the other matrices. The languages that share only a low proportion of the general IE vocabulary (Albanian, Iranian group) have been attracted to the tree basis, and two unorthodox and quite dubious superclades have appeared: the Romano-Germano-Celto-Balto-Slavonic (59, 0) and the group containing all IE languages except Albanian (58, 3).

Discussion

Phylogeny of Indo-European languages

Distant outgroups may lead to spurious relationships based on random similarity (Wheeler, 1990). This phenomenon may apply to the phylogenetic reconstruction of the IE family because the Hittite outgroup (see Ringe et al., 2002) is likely to affect the intra-IE relationships, either because of its phylogenetic distance or because of fragmentary knowledge of the Hittite basic vocabulary. Unrooted trees were therefore constructed for each data matrix and compared with the rooted ones. All the analyses agree that the IE linguistic groups (Indic, Iranian, Baltic, Slavonic, Celtic, Germanic, and Romance) are monophyletic. Moreover, all results indicate that Balto-Slavonic and Romano-Germanic groups are also monophyletic, that Celtic languages are close to the Romano-Germanic superclade (however, with a low

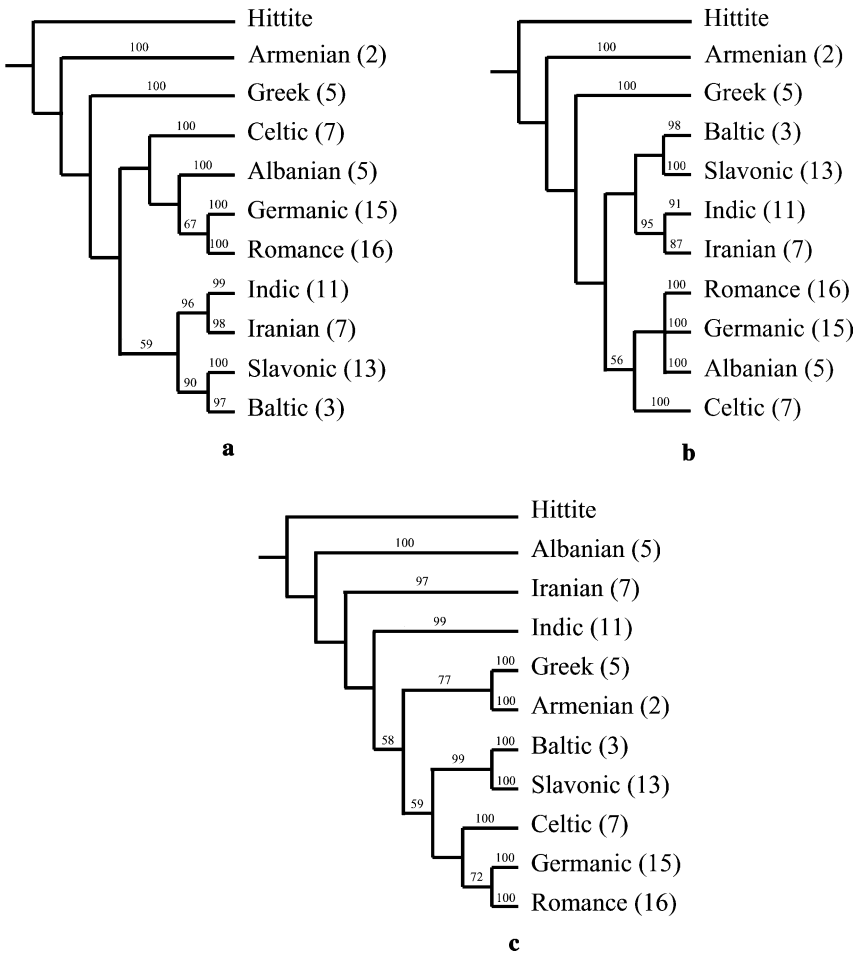

Fig. 1. Phylogeny of Indo-European linguistic groups based on standard multistate matrix (a), altered multistate matrix (b), and binary matrix (c), with bootstrap values indicated above the branches (number of speech varieties per group are in parentheses).

support), and that Armenian and Greek languages are close to one another (however, they may form a paraphyletic assemblage). As concerns higher level relationships, there is a good agreement between the results obtained from the standard and altered multistate matrices—the tree topologies were thus robust to the method used for the identification of the primary homologies. In contrast, different coding of the character states (multistate vs binary) has considerably affected the results. The binary matrix provided a different basal pattern, which is attributable to the behavior of the most aberrant IE languages sharing few linguistic synapomorphies with the other IE languages (and hence numerous 0s in the 0/1 matrix). However, the major difference between both multistate matrices and the binary 1 consists in different tree rooting (near Armenian and Greek languages in the multistate matrices, near to Albanian and Indo-Iranian languages in the binary one).

Within each data matrix, no considerable differences in branching pattern were found in the results of the rooted and unrooted analyses. If the results were presented as unrooted trees, different datasets provide results that are much more compatible than those derived from the rooted analyses, which indicates that the Hittite language groups in a rather chaotic manner with different modern languages in different analyses. All the unrooted trees agree that there are four supergroups of IE languages (Balto-Slavonic, Romano-Germano-Celtic, Armenian-Greek, and Indo-Iranian); however, some of them are likely not to be clades. On the contrary, even the unrooted trees differ in relative position of the above four groups (Indo-Iranian group is close to Balto-Slavonic languages in the multistate matrices, while to Armenian-Greek languages in the binary matrix) and predominantly in position of the "wild card"-behaved Albanian language (close to Romano-Germano-Celtic languages in multistate matrices, close to Indo-Iranian languages in the binary matrix). Because of evident *phenetic* differences between Albanian and Iranian languages on the one hand and the rest of IE on the other hand, we interpret the basalmost position of Albanian and Iranian clades in the rooted binary tree as an "outgroup attraction" artifact and, therefore, restrict

further discussion to results obtained from the multistate matrices. In general, it is possible to conclude that gross topology of the maximum parsimonious trees, constructed exclusively from the basic vocabularies of *modern* languages, is well compatible with the present-day Indoeuropeistic hypotheses.

On the contrary, although based on the same dataset, the phylogenetic analysis presented here disagrees considerably with the earlier lexicostatistical classification (Dyen et al., 1992) and with the tree constructed by Piazza et al. (see Cavalli-Sforza, 2000), both based on the phenetic approach. The Mesoeuropeic *hesion*, a taxon including Romance, German, and Balto-Slavonic subfamilies that has been suggested by Dyen et al. (1992) and has received bootstrap value 95% (Cavalli-Sforza, 2000), is not supported by our results. On the other hand, Indo-Iranian and Balto-Slavonic–Indo-Iranian clades presented here are either missing or poorly supported in the phenetic classifications. The absence of an Indo-Iranian group in the lexicostatistical results is an obvious artifact of the phenetic methodology. As a result of long separation and extensive borrowing from local non-IE linguistic substrates and superstrates (e.g., from the Arabic, a language of religion and power), the Indian and Iranian languages share only a low percentage of characters. Nevertheless, their phylogenetic affinities were recognized by most linguists (e.g., Bloomfield, 1933; Schleicher, 1861–62; Voegelin and Voegelin, 1977), and this view is also supported by other cultural characters. There are, for example, obvious homologies shared by ancient Iranian and Indian religions (Boyce, 1979). The IE classifications obtained by classical comparative linguistic methods (for a review see Ruhlen, 1987) produced less resolved grouping, and except the Indo-Iranian and Balto-Slavonic groups (supported by our results), they usually do not suggest any higher taxa above the well-defined traditional groups.

Cladistics and language phylogeny

The level of character fit on a tree is indicated by consistency and retention indices (CI and RI). The character fit of the presented parsimony trees of IE languages is high (CIs of standard multistate, altered multistate, and binary trees are 0.89, 0.84, and 0.49, respectively), compared with CIs of the biological trees, derived from morphological, molecular, or ecological–behavioral data (Sanderson and Donoghue, 1996). It suggests a predominantly branching pattern of the basic vocabulary phylogeny and little borrowing of *individual* words. The only alternative to the branching phylogeny of languages is a complete acculturation of local population after its admixture with a more aggressive linguistic component—then, the linguistic tree would be incompatible with the biological one, but it would still have high CI.

This conclusion is compatible with the results presented by Holden (2002) on the Bantu and Bantoid African languages (75 languages, 92 items of basic vocabulary, $CI = 0.65$ in the unweighted tree), while the parsimony tree of 77 Austronesian languages (Gray and Jordan, 2000) provided a considerably lower fit ($CI = 0.25$). Although it is difficult to interpret these results, we can conclude that the "word introgression" is less frequent than generally supposed, at least in some linguistic families, and that the amount of reticulate evolution of the languages is not considerably higher than that of the biological species.

Both IE and Bantu languages display a strong correlation between phylogenetic proximity and geographical distance between the languages. In combination with the high consistency indices of both trees, it suggests either that the populations are sedentary, remaining near their areas of origin (see Holden, 2002), or that most of the linguistic differentiation took place *after* migration of the basal populations. Both scenarios are testable by comparison with the archaeological and/or genetic data on the Indo-European history (see below).

In general, the phylogeny of the languages based on lexicostatistical data indicates that language evolution can successfully be reconstructed by the cladistic methods. Naturally, the same rules that we know from the biological cladistics applies to the linguistic cladistics as well—predominantly, the combined analyses are usually better than partial analyses. From this point of view, all the trees derived from basic vocabularies of Austronesian, Bantu, and IE languages should be considered preliminary, and more characters (e.g., grammar) and more taxa (including extinct speech varieties) should be included. On the other hand, lexical items are generally believed to be the least reliable evidence for relationships because they are easily borrowed rather than inherited. This paper is an attempt to test whether lexical data (when containing meanings that seem to show low tendency to be borrowed and when treated by cladistic method) provide results similar to those of classical linguistic classification. The general agreement between our trees and the traditional IE classification suggests that the raw lexical data represent not so labile feature of the language evolution. This conclusion is useful for classification of some non-IE languages for which only lexical data are available.

History of Indo-Europeans

The only IE studies having used a phylogenetic method (Ringe et al., 2002; Warnow, 1997, and methodological discussion and references therein) are devoted to relationships among the oldest known IE languages, supplemented with modern Albanian, Lithuanian, Latvian, and Welsh. The latter study is based on 22 phonological, 15 morphological, and 333 lexical

characters from 24 languages. In the preferred tree, the Hittite–Luvian–Lycian clade and Tocharian were the successive basal branches, followed successively by four superclades of modern IE languages, viz., by Romano-Celtic, Germanic-Albanian, Armenian-Greek, and Balto-Slavonic–Indo-Iranian (''satem'') languages. The rough correspondence to our results is evident; however, it is difficult to judge, as Ringe et al. (2002) provided no formal parameters of the trees nor their support. The position of Germanic languages, which contradicts our findings, is reported as uncertain (Ringe et al., 2002): the authors found that the Germanic languages shared states with disparate language subgroups and believed that Germanic was originally a near sister of Balto-Slavic and Indo-Iranian languages, that at a very early date it lost contact with its more eastern relatives and came into close contact with the western, Romano-Celtic languages. Consequently, Ringe et al. (2002) suggested removing the Germanic languages from analysis. In the best tree with Germanic omitted, Albanian is a sister group of all modern IE languages, while the rest of the tree remains unchanged. Unfortunately, Ringe et al. (2002) did not provide partial phonological, morphological, and lexical trees to show the contributions of different character partitions.

The evolutionary history of IE populations has been extensively studied using genetic methods. The early studies based on protein polymorphism have shown a remarkable correlation between genetic and linguistic similarity matrices (Sokal et al., 1992). This relationship is evident on the level of higher taxa, i.e., linguistic families and/or superfamilies like the Nostratic superfamily (comprising IE, Afro-Asiatic, Kartvelian, Uralo-Altaic, and Dravidian languages). However, the linguistic-genetic accord has become weak at lower hierarchic levels (for a review, see Cavalli-Sforza et al., 1994). Europe is obviously the continent where the increase in genetic dissimilarity per unit of geographic distance is lowest. Most genetic variation among European populations (its first principal component) is arranged along the southeast/northwest cline (Cavalli-Sforza et al., 1994). This cline can be explained by the spread of Neolithic farmers accompanied by demic diffusion (Ammerman and Cavalli-Sforza, 1984; Cavalli-Sforza et al., 1994; Cavalli-Sforza, 1997). This view was supported also by HLA allele frequencies (Sokal and Menozzi, 1982). The putative Neolithic expansion of genes was probably associated exclusively with speakers of the IE family (cf. Renfrew, 1987, 1991). Two other genetic clines from north to southwest and from the Ukrainian steppes west are usually attributed to the impact of the Uralic-speaking populations of the Asian origin and to repeated expansions of steppe nomads westward, respectively (Cavalli-Sforza et al., 1994). Unfortunately, protein data that reflect variation in the recombining nuclear genes are not very useful to solve the question of relationships between European populations and IE-speaking populations in the Middle East and India. The Iranian and Indian populations together with the non-IE-speaking Middle East populations are successive ''sister'' branches of Europeans in the published genetic phenograms, but their position is likely to be affected by hybridization with neighboring populations (Cavalli-Sforza et al., 1994).

Recently, most research effort has been devoted to mitochondrial DNA (mtDNA) variation, which is the best tool for phylogeographic analyses. Mitochondrial DNA has diverged only slightly in non-African human populations (Hedges, 2000; Ingmann et al., 2000; for a review see Avise, 2000), including those from Europe (Malyarchuk and Derenko, 2001; Rando et al., 1998; Simoni et al., 2000). Despite remigration events recorded in the mtDNA pool of extant populations, it is still obvious from the sequence data that Europe was probably colonized from the Middle East and/or the neighboring areas (e.g., Caucasus and/or eastern Europe, see Richards et al., 2000). The majority of the extant mtDNA lineages entered Europe as a few waves during the Upper Paleolithic long before linguistic differentiation. Unfortunately, the immigrant Neolithic component is likely to represent less than one-quarter of the mtDNA pool of modern Europeans (Richards et al., 1998, 2000).

The phylogeography of the non-recombining part of the Y chromosome has shown almost the same results as mtDNA. A considerable part of the sequence variation may be attributed to the Paleolithic migration events (Gibbons, 2000; Semino et al., 2000; Underhill et al., 2001), and the Y-chromosomal diversity within Europe is clinal and influenced primarily by the geography (Rosser et al., 2000). Nevertheless, some Y-chromosome markers clearly suggest the sharp differentiation between European west and east (Malaspina et al., 2000), which may confirm our distinction between western and eastern language clades.

There are two main phylogeographic scenarios explaining the distribution of IE languages. The traditional hypothesis presumes the radiation of pastoral IE tribes immigrating from the ancestral area somewhere in eastern Europe via Russian steppes to western Europe and via Central Asia to Iran and India, during the Eneolithic period (Childe, 1926; Gimbutas, 1985; Mallory, 1989; Sergent, 1995). In contrast, the spread of IE languages from Anatolia westward to the Mediterranean, the Balkans, and Europe, as well as eastward to Central Asia, synchronous with the early Neolithic expansion of the agriculture, is expected by Renfrew (1989). Weng and Sokal (1995) tested the above hypotheses by comparing the phenetic matrix of lexicostatistical distances with geographic distances and with matrices derived from the origin of agriculture according to the Renfrew and Gimbutas models, respectively.

They refuted the Gimbutas "Ukrainian" model and reported that language differentiation and the origin of agriculture are positively correlated within the linguistic subfamilies (particularly among Germanic languages), while the opposite is true when subfamilies of IE languages are compared. They concluded that "differentiation of the major IE branches in Europe seems unrelated to the times of origin of agriculture" and that the "Renfrew hypothesis, if plausible, is far from proven."

Our results suggest a possible existence of the Balto-Slavonic–Indo-Iranian "satem" superclade that occupies most of the eastern territories from east Europe to India and the presence of several well-differentiated clades in Europe and eastern Mediterranean. Due to uncertain rooting of the tree and to the problematic position of Albanian and Germanic languages, it is still premature to give the final phylogeographic interpretation of our IE language cladograms. Nevertheless, the ability of the cladistic methodology to identify several higher taxa within the IE family represents a good prerequisite to reconstruct the history of European and western Asian humankind.

Acknowledgments

We thank Václav Hypša for stimulating discussion, Václav Blažek for consultation in lexicostatistics, and Mike O'Brien and two anonymous reviewers for comments on the drafts of this paper. We are indebted to Martin Rexa who wrote accessory software for data management.

References

Ammerman, A.J., Cavalli-Sforza, L.L., 1984. The Neolithic Transition and the Genetics of Populations in Europe. Princeton University Press, Princeton, NJ.

Arnold, M.L., 1997. Natural Hybridization and Evolution. Oxford University Press, Oxford.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Bloomfield, L., 1933. Language. Holt, New York.

Boyce, M., 1979. Zoroastrians: Their Religious Beliefs and Practices. Routledge & Kegan Paul, London.

Brugmann, K., Delbrück, B., n und Altkirchenslavischen. Grundriss der Vergleichenden Grammatik der Indogermanischen Sprachen: Kurzgefasste Darstellung der Geschichte des Altindischen, Altiranischen (Avestischen u. Altpersischen), Altarmenischen, Altgriechischen, Albanesischen, Lateinischen, Oskisch-Umbrischen, Altirischen, Gotischen, Althochdeutschen, Litauische. Trübner, Strasbourgchenslavischen.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of the Human Gene. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L.L., 1997. Genetic and cultural diversity in Europe. J. Anthropol. Res. 53, 383–404.

Cavalli-Sforza, L.L., 2000. Genes, Peoples, and Languages. North Point Press, New York.

Childe, V.G., 1926. The Aryans: A Study of Indo-European Origins. Knopf, London.

Crystal, D., 1985. A Dictionary of Linguistics and Phonetics. Blackwell, Oxford.

de Pinna, M.C.C., 1985. Concepts and tests of homology in the cladistic paradigm. Cladistics 4, 367–394.

Dixon, R.M.W., 1999. The Rise and Fall of Languages. Cambridge University Press, Cambridge, UK.

Dyen, I., Kruskal, J.B., Black, P., 1992. An Indoeuropean classification: a lexicostatistical experiment. Trans. Am. Philos. Soc. 82, 1–132.

Gibbons, A., 2000. Europeans trace ancestry to Paleolithic people. Science 290, 1080–1081.

Gimbutas, M., 1985. Primary and secondary homeland of the Indo-Europeans. J. Indoeur. Stud. 13, 185–202.

Gray, R.D., Jordan, F.M., 2000. Language tree supports the express-train sequence of Austronesian expansion. Nature 405, 1052–1055.

Gudschinsky, S.C., 1964. The ABC's of lexicostatistics (glottochronology). Hymes 63, 612–623.

Harrison, R.G., 1993. Hybrids and hybrid zones: historical perspective. In: Harrison, R.G. (Ed.), Hybrid Zones and the Evolutionary Process. Oxford University Press, Oxford, pp. 3–12.

Hedges, S.B., 2000. A start for population genomics. Nature 408, 652–653.

Hoenigswald, H.M., 1965. Language Change and Linguistic Reconstruction. University Chicago Press, Chicago.

Holden, C.J., 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. Proc. R. Soc. London 269, 793–799.

Ingmann, M., Kaessmann, H., Pääbo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–712.

Malaspina, P., Cruciani, F., Santolamazza, P., Torroni, A., Pangrazio, A., Akar, N., Bakalli, V., Brdicka, R., Jaruzelska, J., Kozlov, A., Malyarchuk, B., Mehdi, S.Q., Michalodimitrakis, E., Varesi, L., Memmi, M.M., Vona, G., Villems, R., Parik, J., Romano, V., Stefan, M., Stenico, M., Terrenato, L., Novelletto, A., Scozzari, R., 2000. Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa. Ann. Hum. Genet. 64, 395–412.

Mallory, J.P., 1989. In Search of the Indo-Europeans: Language, Archaeology and Myth. Thames & Hudson, London.

Malyarchuk, B.A., Derenko, M.V., 2001. Mitochondrial DNA variability in Russians and Ukrainians: implication to the origin of the Eastern Slavs. Ann. Hum. Genet. 65, 63–78.

Paul, H., 1880. Prinzipien der Sprachgeschichte. Niemeyer, Halle, Germany.

Rando, J.C., Pinto, F., González, M., Hernández, M., Larruga, J.M., Cabrera, V.M., Bandelt, H.-J., 1998. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. Ann. Hum. Genet. 62, 531–550.

Renfrew, C., 1987. Archaeology and Language. The Puzzle of Indo-European Origin. Jonathan Cape, London.

Renfrew, C., 1989. The origins of Indo-European languages. Sci. Am. 261, 82–90.

Renfrew, C., 1991. Before Babel: speculations on the origins of linguistic diversity. Cambridge Archaeol. J. 1, 3–23.

Richards, M.B., Macaulay, V.A., Bandelt, H.J., Sykes, B.C., 1998. Phylogeography of mitochondrial DNA in western Europe. Ann. Hum. Genet. 62, 241–260.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellito, D., Criciani, F., Kivisild, T., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Gölge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A.,

Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Rienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozzari, R., Torroni, A., Bandelt, H., 2000. Tracing European founder lineage in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67, 1251–1276.

Ringe, D., 1999. How hard is it to match CVC-roots? Trans. Philos. Soc. 97, 213–244.

Ringe, D., Warnow, T., Taylor, A., 2002. Indo-European and computational cladistics. Trans. Philos. Soc. 100, 59–129.

Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D.G., Brede, G., Cooper, G., Côrte-Real, H.B.S.M., de Knijff, P., Decorte, R., Dubrova, Y.E., Evgrafov, O., Gilissen, A., Glisic, S., Gölge, M., Hill, E.W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, A., Krumina, A., Kučinskas, V., Lavinha, J., Livshits, A., Malaspina, P., Maria, S., McElreavey, K., Meitinger, A., Mikelsaar, A., Mitchell, R.J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M.J., Previderé, C., Roewer, L., Rootsi, S., Rubinsztein, D.C., Saillard, J., Santos, F.R., Stefanescu, G., Sykes, B.C., Tolun, A., Villems, R., Tyler-Smith, C., Jobling, M.A., 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. 67, 1526–1543.

Ruhlen, M., 1987. A Guide to the World's Languages, vol. 1, Classification. Stanford University Press, Stanford, CA.

Sanderson, M.J., Donoghue, M.J., 1996. The relationships between homoplasy and confidence in a phylogenetic tree. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrency of Similarity in Evolution. Academic Press, San Diego, CA, pp. 67–89.

Schleicher, A., 1861–62. Compendium der Vergleichenden Grammatik der Indogermanischen Sprachen. Böhlau, Weimar, Germany.

Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Afbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A.S., Cavalli-Sforza, L.L., Underhill, P.A., 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y-chromosome perspective. Science 290, 1155–1159.

Sergent, B., 1995. Les Indo-Européens: Histoire, Langues, Mythes. Payot & Rivages, Paris.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., Barbujani, G., 2000. Geographic patterns of mtDNA diversity in Europe. Am. J. Hum. Genet. 66, 262–278.

Skála, Z., Zrzavý, J., 1994. Phylogenetic reticulations and cladistics: discussion of methodological concepts. Cladistics 10, 305–313.

Sokal, R.R., Menozzi, P., 1982. Spatial autocorrelation of HLA frequencies in Europe support demic diffusion of early farmers. Am. Nat. 119, 1–17.

Sokal, R.R., Oden, N.L., Thomson, B.A., 1992. Origins of the Indo-Europeans: genetic evidence. Proc. Natl. Acad. Sci. USA 89, 7669–7673.

Sturtewant, E., 1962. The Indo-Hittite hypothesis. Language 38, 105–110.

Swadesh, M., 1952. Lexico-statistic dating of prehistory ethnic contacts, with special reference to North American Indians and Eskimos. Proc. Am. Philos. Soc. 95, 453–462.

Swofford, D.L., 2000. PAUP, Version 4.0b4a. Sinauer, Sunderland, MA.

Tischler, J., 1973. Glottochronologie und Lexikostatistik. Inst. Sprachwissenchaft, University Innsbruck, Innsbruck.

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Lahr, M.M., Foley, R.A., Oefner, P.J., Cavalli-Sforza, L.L., 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human population. Ann. Hum. Genet. 65, 43–62.

Voegelin, C.F., Voegelin, F.M., 1977. Classification and Index of the World's Languages. Elsevier, New York.

Warnow, T., 1997. Mathematical approaches to comparative linguistics. Proc. Natl. Acad. Sci. USA 94, 6585–6590.

Weng, Z., Sokal, R.R., 1995. Origins of Indo-Europeans and the spread of agriculture in Europe: comparison of lexicostatistical and genetic evidence. Hum. Biol. 67, 577–594.

Wheeler, W.C., 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics 6, 363–367.