Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 6585–6590, June 1997
From the Academy

This paper summarizes a symposium session that was one of the Frontiers of Science symposia, held November 2–4, 1995, at the Arnold and Mabel Beckman Center of the National Academy of Sciences and Engineering in Irvine, CA.

Mathematical approaches to comparative linguistics[†]

TANDY WARNOW

Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6389

ABSTRACT The inference of the evolutionary history of a set of languages is a complex problem. Although some languages are known to be related through descent from common ancestral languages, for other languages determining whether such a relationship holds is itself a difficult problem. In this paper we report on new methods, developed by linguists Johanna Nichols (University of California, Berkeley), Donald Ringe and Ann Taylor (University of Pennsylvania, Philadelphia), and me, for answering some of the most difficult questions in this domain. These methods and the results of the analyses based on these methods were presented in November 1995 at the Symposium on the Frontiers of Science held by the National Academy of Sciences.

Evolutionary Relationships in Linguistics. Evolutionary relatedness of languages is described by observing that the separation of speech communities into distinct and noninteracting subcommunities eventually results in a language developing into new languages in a process quite similar to speciation in biology. Although this is not the only means by which languages change, it is this process which is referred to when we say, for example, "French is a descendent of Latin." This allows us to model the evolution of related languages as a rooted tree in which internal nodes represent the ancestral languages. When a set of languages does not have a common ancestor (as the case may be for a set containing both Dravidian and Indo-European languages), then the evolution of that set is best described by a disjoint collection of rooted trees (i.e., a "forest"). Except in circumstances involving related dialects that continue to have close contact, there is no problem with this model of language evolution.

Careful scholarship over the last century has determined critical features and patterns that, combined with a statistical analysis, can be used to establish that languages share a common ancestor; examples of these features are shared idiosyncracies in the grammars, shared idiosyncratic sound changes, and patterns of sound correspondences. Extending this fundamental statistical analysis, two techniques (the "comparative method" and "subgrouping through shared innovations") have been developed that enable linguists to infer greater information about relatedness and properties of ancestral languages, and—to a limited extent—subgrouping as well. These techniques have established all known linguistic families and subfamilies, and are the basis of historical linguistic scholarship. Known families presently number close to 300, though ongoing comparative work on the languages of New Guinea and of South America—two of the linguistically most diverse and least described places on earth—may reduce this total to as low as 200. Many of these "families" are one-descendent, such as Basque, which is a distinct genetic lineage of its own with no known kin. Although these two techniques provide firm evidence of relatedness between languages, they have so far provided only limited information about subgrouping within sets of related languages. Consequently, linguists have lacked a reliable method for the inference of the full evolutionary history of language families, and the evolutionary histories of many language families remain unresolved, despite decades of debate.

Finally, these techniques are only applicable for comparing well attested languages that are known to be related and whose most recent common ancestor does not lie more than 6,000–8,000 years in the past. At time depths beyond that limit, the critical features upon which the classical techniques are based survive in such small numbers that they cannot reliably be distinguished from chance resemblances (1). Attempts have been made to establish criteria by which such relationships can be inferred for sets of languages with ancestors further back in time than this barrier, but these have been largely unsuccessful and heavily criticized for lacking rigorous statistical foundations. Extending the range of linguistic comparison beyond that critical time depth is therefore a major endeavor within historical linguistics.

In the Frontiers of Science symposium, the panel on Mathematical Approaches to Comparative Linguistics discussed new approaches toward developing methods to accurately infer (*i*) the branching pattern of the evolutionary history of languages known to be related and (*ii*) relationship (whether due to historical contact or to descent from a common ancestor) of languages not already known to be related. The first talk involved a team at the University of Pennsylvania, linguists Donald Ringe and Ann Taylor, and me, in our efforts to develop a methodology for inferring the evolutionary tree for languages known to be related. We formulated a model of evolution based on classical scholarship in historical linguistics, and developed an efficient method that would serve two purposes: first, the model could be tested to see if it fit the data and second, trees that best fit the model could be generated. The application of our methods to the Indo-European family of languages has indicated that the data to a great extent fit the model extremely well, and produced a robust evolutionary tree, potentially settling longstanding controversies in Indo-European studies. In the second talk, Johanna Nichols of the University of California, Berkeley, described her method by which relationships and/or earlier interaction could be reliably inferred between languages not necessarily known to be genealogically related. She described properties of linguistic features that she called "population markers," which would reliably indicate either a genealogical relationship or at least significant and prolonged contact between language communities. Her analysis of the world's languages has implications for our understanding of human migrations and greatly extends the power of comparative linguistic analysis.

e-mail: tandy@central.cis.upenn.edu.

[†] This paper is part of the fourth installment of the new feature, "From the Academy." The first installment appeared in the March 4, 1997 issue, the second in the April 1, 1997 issue. "From the Academy" will be presented occasionally as new NRC reports appear and as essays on the NAS are prepared.

In this report, the basic ideas and results of these two research projects are described, and some of the questions posed by members of the audience at the Symposium are reported.

Evolutionary History Inference of Related Languages. The two fundamental techniques for subgrouping within established families used in historical linguistics are the comparative method, formalized by Henry Hoenigswald (2), and subgrouping through shared innovations. Because the assumptions upon which these two techniques are based are used in the methodology developed by Warnow *et al.* (3), these techniques are described in some detail.

The comparative method. Given a set of languages known to be related, the comparative method has the following steps. Step 1: Observe sound correspondences; that is, compare words for the same (or comparable) meanings and observe patterns of sound correspondences between pairs of languages. Step 2: Infer regular sound change rules. These rules must explain *all* the sound correspondences observed in Step 1. These rules may be context-free or context-dependent, and are specific to each lineage. Step 3: Infer cognation judgments. Two words w and w' from two languages L and L' respectively are said to be *cognate* if it is possible to infer a word $w*$ in some common ancestor of L and L' such that each w and w' can be derived from $w*$ by the sound change rules specific to L and L', respectively. The comparative method distinguishes between words that are similar and those that have a common origin and thus enables linguists to establish that Spanish "mucho" and English "much" are not cognate because applications of the sound change rules do not indicate that they come from a common ancestral word ("mucho" is derived from "multum" in Latin, meaning "much," whereas "much" is derived from "micel" in Old English, meaning "big").

Linguistic characters. The comparative method defines cognate classes so that different words may be considered to be equivalent and thus allows the languages to be defined by a set of equivalence relations, one for each meaning. This is comparable to using morphological features or columns within biomolecular sequences to represent biological taxa; in each case, the primary data are described through the use of partitions of the taxa into equivalence classes. Such partitions are called "characters" in the biological literature.

The comparative method establishes two types of linguistic characters, "lexical" and "phonological." For lexical characters, the character is the semantic slot (e.g., the meaning "hand," with the states of the character defined by cognation judgments). (Were it not for word replacement, which is endemic across all languages, words for the same meaning in related languages would all be cognate and thus all lexical characters would have a single state on any set of related languages. Thus, word replacement is why lexical characters have more than one state.) For phonological characters, the character is a sound change. Languages that share the same outcome (generally, those that undergo the change versus those that do not) exhibit the same state for the character. As a special subtype of lexical characters, morphological characters can also be defined. Here, the character generally is a grammatical feature (e.g., the formation of the future stem, the way the passive is marked, the genitive singular ending of o-stem nouns and adjectives). Languages in which the feature is instantiated in the same way, or by a reflex of the same protomorpheme, exhibit the same state for the character. Because morphological characters resist borrowing, they are especially useful in determining relationships between languages.

Subgrouping through shared innovations. Classical methodology in historical linguistics has used these phonological and morphological characters for subgrouping purposes; when a character has two states in which one is clearly ancestral, then the character defines a linguistic innovation. Linguistic innovations that are useful for subgrouping must be peculiar enough to not be easily repeated and (depending on the particular set of languages examined) should not be too easily lost. When a statistically significant number and quality of innovations are shared, then the set of languages sharing that common set of innovations can be considered to form a linguistic subgroup, such as the Germanic and Italic subfamilies of Indo-European.

Comments. The key observation made by Ringe and myself (see ref. 3) in the fall of 1993 that enabled us to develop a new methodology was that the classical methods in historical linguistics (subgrouping through shared innovations and the comparative method) can be stated as hypothesizing that almost all linguistic characters, if properly encoded, should be compatible with the evolutionary tree for the languages. The term compatible is a technical term from the systematic biology literature, which has the following definition: a character c is compatible with tree T if the nodes in T can be labeled by states of c so that every state of c induces a connected subset of T. An example of a biological character that is compatible is the vertebrate-invertebrate character, whereas the character indicating the presence or absence of wings is not a compatible character on the tree of all animals.

The reason that the hypothesis is stated with the caveat that only *almost all* and not *absolutely all* characters should be compatible is the observation that many phonological characters are based on sound changes that are natural enough to occur repeatedly. By contrast, lexical characters ought to be compatible on the evolutionary tree, provided that borrowing can be detected. Those morphological characters and phonological characters that are based on properties unusual enough to have only arisen once also ought to be compatible on the evolutionary tree. Thus, the hypothesis indicated by the classical methodology is, more precisely, that all lexical characters, and those morphological and phonological characters that represent distinctly unusual traits, should be compatible on the evolutionary tree of a family, provided that the family is well attested and well understood.

Although the linguistic hypothesis is that all properly selected and encoded characters should be compatible on the true evolutionary tree, there are certain specific conditions in which it can be difficult to distinguish between true cognates and words that are borrowed; that is, it may be difficult to distinguish between true and false cognates. Based on these observations, Ringe and I (see ref. 3) formulated the following optimization criterion: find the tree on which it is possible to explain all incompatible character evolution with as simple an explanation as possible, that matches linguistic scholarship as closely as possible.

The optimization problem we formulated is related to a classical problem in biological systematics called the compatibility criterion, in which the tree on which as many characters as possible are compatible is the optimal tree. The compatibility criterion problem caught the interest of the computer science algorithms community because of its combinatorial flavor and interesting graph-theoretic formulation (4). In addition to showing that the compatibility criterion problem is NP-hard (5–7) (and thus unlikely to be solvable in polynomial time; see ref. 8), computer scientists and mathematicians developed polynomial time algorithms for various fixed-parameter formulations of the problem (9–13). Using a program designed by Richa Agarwala (based on ref. 12) to solve the compatibility criterion, Ringe and I decided to test the hypothesis of classical historical linguistics that properly encoded linguistic data should result in highly compatible characters. The program in turn would also permit us to explore all the trees that had optimal and near-optimal scores for the compatibility criterion, and thus select those trees with (hopefully) simple explanations of incompatibility.

Assisted by Libby Levison, then a doctoral candidate at the University of Pennsylvania, we first tested this hypothesis on some small data sets. These preliminary results were very encouraging, and we then turned to the Indo-European (IE) family. Although the IE family is among the best understood of the world's language families, the precise branching pattern of this family had resisted definitive analysis. In particular, we were interested in discovering that the two most heatedly debated hypotheses—the Indo-Hittite and the Italo-Celtic—could be settled by using our methodology. (The Indo-Hittite hypothesis is that the first subfamily to break off from the root of the Indo-European evolutionary tree should be the Anatolian branch, represented by Hittite, and the Italo-Celtic hypothesis is that Italic and Celtic should be sisters within the tree, and without a third sister.)

We selected from each of the subfamilies within IE the oldest, well attested language to represent the subfamily. To reduce the possibility of borrowings among the lexical characters and bias on our part in choosing these characters, we used an existing basic vocabulary list of 212 semantic slots (14).[†] Each semantic slot was treated as a single character and judgments of cognation were made on the basis of the comparative method. An appropriate set of 17 morphological and phonological characters was developed for the IE family.

Over the next 2 years, in collaboration with postdoctoral researcher Ann Taylor, Ringe and I studied the Indo-European family of languages. We discovered that a phenomenon termed "polymorphism" in which, for example, more than one word is available in a particular semantic slot (consider "big" and "large"). Polymorphism creates significant difficulties for reconstructing the evolutionary history in Indo-European, and there was no rigorous methodology in place for handling polymorphic characters. In collaboration with other computer scientists, I developed algorithms to handle polymorphic character data (15), which were then used to analyze the Indo-European data. Because rooted trees are desirable, directionality constraints implied by some of the linguistic data were encoded as characters by using techniques already in use by systematic biologists, and these characters were included in the data set.

These algorithms were then applied to the entire data set for Indo-European, and all the trees with optimal or near-optimal compatibility scores were examined. The two best trees had 12 and 13 incompatible characters, respectively, but were remarkably similar except for the placement of Germanic. When Germanic was removed from the data set, however, a tree was obtained on which every character was compatible! Such a tree is called a perfect phylogeny and indicates that the data (minus Germanic) fit the model proposed by us exactly. We then examined whether the deletion of any other single language would result in a comparable situation, but the removal of any other single language resulted in many incompatible characters. This suggested that Germanic might be a singular problem for the Indo-European family and suggests that the correct tree for the Indo-European family would be obtained by placing Germanic within one of the optimal or near-optimal trees obtained when Germanic is removed.

Assisted by postdoctoral researcher Libby Levison and Alexander Michailov, we then considered the near-optimal trees to establish the degree of confidence for each of the features of the optimal tree. Although our original data set contained 229 characters, only 61 of these were informative, because the remaining 148 characters fit every possible tree on the family. The subgroups Balto-Slavic and Indo-Iranian are strongly supported, as is the subgrouping together of these two subgroups to comprise the Satem Core; however, these subgroupings had already been suggested by traditional methods and have generally not been argued about by the historical linguistic community. On the other hand, many hotly contested subgroupings are supported by this analysis to various degrees. The Indo-Hittite hypothesis is supported by only one character, but it is difficult to impugn that character. Should that character be impugned, a subgrouping of Hittite and Tocharian is possible, but moving the root below the Italo-Celtic subgroup seems less likely than the present rooting due to geographic constraints. Tocharian can move only slightly within the tree without causing a significant decrease in the compatibility score; hence it is reasonable to consider its placement to be relatively well constrained. The Italo-Celtic subgroup was supported by three characters, indicating relatively strong support. The Greco-Armenian subgroup was supported by five characters, and thus is strongly supported by the data. Each of these three subgroupings had been debated significantly over the past many decades, and the strong support of some of these subgroups through this analysis was surprising. The only features that remained somewhat unclear through this analysis were the exact placement of Tocharian within the tree (which, as we noted, was nevertheless fairly constrained), the exact placement of the root (Proto-Indo-European), and where Albanian fits in the tree. These questions require further data before a definitive answer can be obtained.

We then sought to reintroduce Germanic into the optimal and near-optimal trees to consider whether there was a reasonable explanation for the incompatible characters that were obtained. The result was that there were two reasonable locations for Germanic; the first, and best, was to place Germanic within the Satem Core, as a sister to the Balto-Slavic subgroup. In this placement, the pattern of incompatibility has a simple explanation: it appears to point to a situation in which Germanic began to develop within the Satem Core (as evidenced by its morphology) but moved away before the final satem innovations. It then moved into close contact with the "western" languages (Celtic and Italic) and borrowed much of its distinctive vocabulary from them at a period early enough that these borrowings cannot be distinguished from true cognates. Because statements of cognation depend on unbroken descent from a common ancestor through genetic inheritance, and not from borrowing, this hypothesis implies that words in Germanic borrowed from pre-proto-Italic and pre-proto-Celtic are not cognate with the corresponding words in Italic and Celtic. If this relatively simple hypothesis is accepted, then all the characters are compatible on the tree. The second placement for Germanic that produces a reasonable fit is just outside the Satem Core. This placement avoids the need to posit an early geographic move for Germanic, but does not provide a simple explanation for all the incompatible characters. Hence, the best location for Germanic seems to be obtained by taking the best tree for the family with Germanic removed and introducing Germanic as a sister to Balto-Slavic. This tree is given in Fig. 1.

We concluded by noting that although our method has produced what seems to be a likely solution for the evolutionary history of the Indo-European family, the major point of our research is the model of language evolution, which seems to be well supported by the data (as evidenced by the existence of a perfect phylogeny when Germanic is removed). Our method then permits linguists to infer whether their judgments are consistent with the model, and to obtain a tree which best fits their judgments and the model. However, because the data supporting the tree are somewhat limited, ongoing research is likely to modify the results obtained over time. In fact, the analysis given here differs somewhat from what was presented at the Symposium on the Frontiers of Science, due to the continuing data collection and analysis, and because this

[†] The list has more items than Tischler's (14) because we split some items that indicated more than one semantic slot into several items. For example, Tischler's list includes *day* as one item, and this item was split into two items, *period of 24 hours* and *period of daylight.*

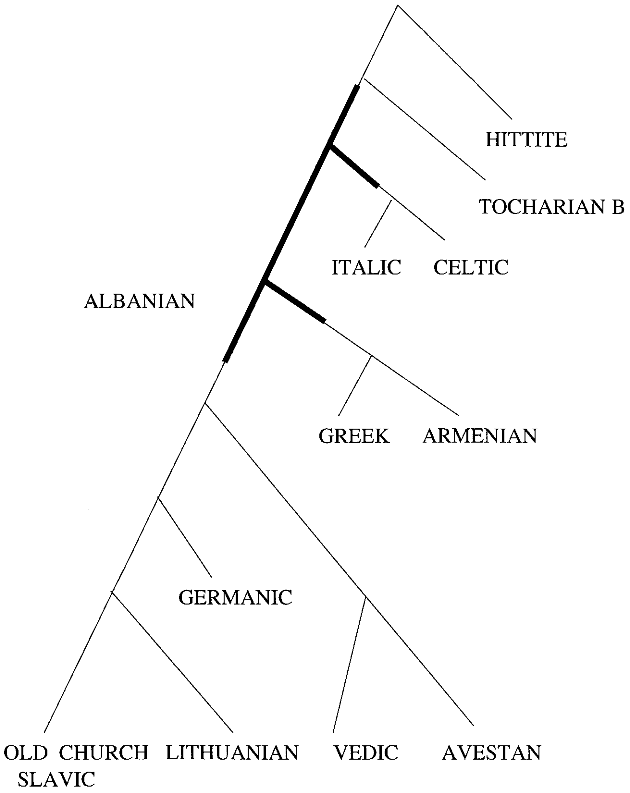

FIG. 1. The topology of the rooted evolutionary tree for Indo-European. The tree is not drawn to scale—the only indication of time that can be inferred is through ancestry. Albanian can be attached to this tree along any thick edge.

project is ongoing, there is a possibility that continued analysis will change the solution obtained to some degree.

Questions. Several questions were asked at the end of the talk, with the most significant question addressing the issues of whether languages really do form trees. In particular, the audience wanted to know how borrowing was handled, whether creoles and pidgins contradict the use of trees to model language evolution, and if there are problems with this method when the family has related dialects. The speakers said that analyzing dialects is definitely a problem for this method, because in those cases trees may not be appropriate models of evolution; however, when the problem is restricted to languages rather than dialects, the process of speech communities separating indicates that trees are appropriate. Furthermore, creoles and pidgins do not cause any problem for this method because creoles and pidgins can be detected as such.

Comparing Languages not Known To Be Related. The previous section described how the evolution of a set of languages sharing a common origin can be inferred from the features of the languages when properly encoded and analyzed. Johanna Nichols' work studies the case of languages that are either unrelated, or that have diverged to the point where the diagnostic features used to infer genetic relatedness between languages have been largely lost. Thus, rather than attempting to establish a genetic relationship between languages, Nichols' work endeavors to establish techniques by which similarity due to common origin *or* prolonged and intimate contact can be established. She proposes specific features, which she calls "population markers" or "historical markers," whose distribution can be used to formulate hypotheses about linguistic prehistory. Nichols suggests that her results can be used in conjunction with archeological evidence to develop better theories about early human migrations. Her findings, applied to a database of the world's populations, have the potential to greatly extend current knowledge of human migrations and relationships between languages.

Genetic vs. historical markers. Genetic markers are features that indicate a genetic relationship between languages, and thus indicate that languages sharing the genetic marker have a common ancestor. By contrast, historical markers (also called population markers) indicate a nonaccidental relationship, although they cannot tell us whether that relationship is specifically genetic; it could have been significant before contact between speech communities or before contact with a now-defunct third party. There are essentially three mechanisms by which languages can share features: (*i*) through inheritance from a common ancestor, indicating a genetic relationship; (*ii*) through borrowing (whether direct or indirect) between neighboring speech communities, indicating a historical (but not necessarily genetic) relationship; and (*iii*) through spontaneous reappearance of the same trait in different lineages.

For any feature to be useful for detecting genetic or historical relationships, the feature must be unlikely to evolve spontaneously; otherwise, spurious relationships will be posited. To establish a specifically genetic relationship (as opposed to the more general historical relationship), it must be possible for the linguist to distinguish between acquisition through borrowing and acquisition through inheritance. Features that are difficult to borrow are appropriate for use as genetic markers, but borrowable features also can be analyzed correctly in genetic terms provided that borrowing can be detected. Essentially, genetic markers must have the following traits: (*i*) the feature must be extremely unlikely to arise twice—for lexical characters, the comparative method establishes this strong probability; and (*ii*) borrowing of the feature must either be extremely unlikely, or it must be possible to detect such borrowing.

As Ringe and I observed (and subsequent research with Taylor supported; refs. 3 and 15), it follows that genetic markers should define characters that are compatible on the evolutionary tree for the language family. This observation allows a linguist to posit that some set of features is inherited genetically and this hypothesis, in turn, can be tested [by using the methodology of Warnow *et al.* (3)] as described in the previous section.

Historical markers must also have certain properties that enable a historical relationship to be detected, although these properties are somewhat different from those required for genetic markers. Although the trait should not be likely to arise twice, the condition that borrowing should either be unlikely or detectable need not hold. If a historical marker is based on a trait that is never borrowed, then it cannot be used to provide evidence of contact between different languages otherwise not known to be related. On the other hand, if the trait is too easily borrowed, or too easily lost, then there will be no pattern of relationship that permits nontrivial observations. Thus, historical markers, to be useful, must be capable of being borrowed, but must not be lost too easily once acquired.

Each type of marker (genetic or historical) enables the detection of a relationship of some sort, either through descent from a common ancestor or through contact, and the best markers (whether genetic or historical) are low-frequency features that form a single frequency peak or cluster, resulting in a frequency asymmetry that is statistically significant. Genetic markers such as these permit subgrouping at a fine-grained level, while historical markers of this type provide greater insight into the history of early human migrations, because the findings can be compared with archeological evidence.

Nichols proposes a method by which historical markers can be selected and analyzed. She shows how the geographical distribution of a candidate historical marker among the world's languages can provide evidence for common histories between languages and, in particular, can lead to hypotheses about early migrations, which can then be tested against archeological evidence.

Nichols' research. Nichols selected 14 different traits that had the specified properties required for historical markers, and that, in addition, were believed to be independent of each other (J. Nichols, unpublished results). These were morphological ergativity, morphological complexity, head-marking morphology, inclusive/exclusive oppositions in first-person pronouns, genders or other noun classes, numeral classifiers, tones, possessive affixes, regular transitivization in verbal derivation, identical stems in "I/me" and "we/us", m as root consonant in first-person singular pronouns ("I/me"), m as root consonant in the second-person singular pronoun ("you"), verb-initial word order, and secondary glottal articulation.

A selection of the world's languages was then studied to determine the incidence of these traits throughout the world. Of the 200–300 different language families that have been established, some of these families are very well understood and others are less so. Because some languages are only recently attested (and not as well studied as others), there is a distinct possibility that in time, linguistic scholarship will be able to identify genetic relationships between certain families. Thus, the number of linguistic families may eventually be reduced to about 200; that is, languages that now seem to be unrelated genetically may in time be established as having a common origin. In developing a database of the world's languages, Nichols selected, at most, one language from each major branch within each linguistic family to obtain her sample of languages. The sample she has obtained (more than 200 languages and still growing) has the property that no two languages within the sample are likely to be more closely related than two distantly related Indo-European languages (like French and Armenian).

Geographical distribution of markers. Nichols discovered striking patterns in the geographic distribution of these historical markers around the world. All findings point in the same direction: strong affinities between Australia and the western Old World and different but also strong affinities all around the Pacific Rim. The linguistic distributions point to coastal spread around the Pacific beginning in very early times and to an earlier expansion from Africa via southern Asia to Australasia. Both expansions are widely assumed by archeologists and human geneticists, but the linguistic distributions seem to provide the clearest evidence of them.

For example, some markers are most frequent in Europe, Africa, or both; least frequent in Australia; and of middling frequency in Asia and the Americas. This geographical distribution correlates with archeological research that establishes that the Americas were settled by people migrating from Siberia (i.e., from Asia). Other markers are densely clustered in Australia, well represented around the Pacific and in the Americas, but rare in the Old World (Europe, Africa, central Asia), implying that the distribution of these markers must have taken place before the colonization of the Pacific Islands and the New World. The pattern also suggests that the impetus for expansion came from the west, ultimately from Africa.

A similar pattern occurs within Australia and New Guinea, where the frequencies of population markers show that a subset of the Australian languages defined by specific geographic boundaries closely resembles a subset of the languages of New Guinea, again defined by specific geographic boundaries. Other interesting correlations between Australia and New Guinea are present in this analysis, showing generally an east–west trend in the frequencies of the different markers. Australia and New Guinea were originally (during the Ice Age) parts of the same continent, which was split by a postglacial sea-level rise. Human colonization of these two lands emanated from Southeast Asia, and the landfall point for this colonization was the northwest coast of the continent. The patterns between these two lands actually indicate multiple linguistic colonizations and support the previous research indicating that human colonization occurred when the two lands were in a single continent.

There are many striking patterns that can be observed when the frequencies of these population markers are compared with geography, and these patterns, when combined with archeological evidence, provide significantly more detailed information (or at least better hypotheses!) about early migrations.

Questions. The questions posed by the audience mostly focused on two issues: Nichols' selection of markers and whether all languages are assumed to be genetically related, so that the issue is not whether there is a tree, but what the tree is. Nichols agreed that the selection of markers would influence the findings, but indicated that she used markers from all over the typological literature and in consultation with specialists on various language families. Nichols' answer to the question of whether all the world's languages are related through "genetic descent" was that this is a hard question. She noted that the arguments proposed in favor of the theory that all languages being related are generally considered discounted by serious linguists for statistical reasons (they are based on similarities between words that could have arisen due to chance), and that rigorous methods by which such questions might be answered have yet to be discovered.

Further Reading. A good introduction to phylogenetic tree construction methodology in biology can be found in ref. 16. The methodology of the Warnow *et al.* research is described in refs. 3 and 15. More detailed information about the mathematics of the compatibility criterion problem can be obtained in refs. 10 and 17. Additional material on historical linguistic methodology can be obtained in refs. 2, 18, and 19. Johanna Nichols' work is described in greater detail in refs. 20–22. Discussions of the Indo-Hittite and Italo-Celtic hypotheses (and other controversies in Indo-European studies) can be found in refs. 23–27. A discussion of the archeological evidence related to the discussion of migrations in Australia and New Guinea can be found in refs. 28 and 29. These research projects have been ongoing; the result is that some of the findings have been clarified.

I thank Dr. Libby Levison and Alexander Michailov, who assisted with the further analysis of the Indo-European data. This work was supported in part by National Science Foundation Grants CCR-9457800 and SBR-9512092, by an Army Research Office Grant DAAL0389-C0031, by the Institute for Research in Cognitive Science at the University of Pennsylvania, and by financial support from Paul Angello.

1. Ringe, D., Jr. (1992) *Trans. Am. Philos. Soc.* **82(1)**.
2. Hoenigswald, H. M. (1960) *Language Change and Linguistic Reconstruction* (Univ. of Chicago Press, Chicago).
3. Warnow, T., Ringe, D. & Taylor, A. (1996) *Proceedings of the Seventh Annual ACM/SIAM Symposium on Discrete Algorithms* (Assoc. Comput. Machinery, New York, and Soc. Industr. Appl. Math., Philadelphia), pp. 314–322.
4. Buneman, P. (1974) *Discrete Math.* **9,** 205–212.
5. Bodlaender, H., Fellows, M. & Warnow, T. (1992) *Proc. Int. Colloq. Autom. Language Processing* **510,** 373–383.
6. Day, W. H. E. & Sankoff, D. (1986) *Syst. Zool.* **35(2),** 224–229.
7. Steel, M. (1992) *J. Classification* **9,** 91–116.
8. Garey, D. & Johnson, D. (1979) *Computers and Intractibility* (Freeman, New York).
9. Kannan, S. & Warnow, T. (1994) *SIAM J. Computing* **23(4),** 713–737.
10. Kannan, S. & Warnow, T. (1997) *SIAM J. Computing*, in press.
11. McMorris, F. R., Warnow, T. & Wimer, T. (1994) *SIAM J. Discrete Math.* **7(2),** 296–306.

12. Agarwala, R. & Fernández-Baca, D. (1994) *SIAM J. Computing* **23(6),** 1216–1224.
13. Agarwala, R. & Fernández-Baca, D. (1996) *Int. J. Foundations of Computer Science* **7,** 11–21.
14. Tischler, J. (1973) *Glottochronologie und Lexikostatistik* (Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck, Austria).
15. Bonet, M., Phillips, C., Warnow, T. & Yooseph, S. (1995) Proceedings of the Twenty-eighth annual ACM Symposium on the Theory of Computing, May 22–24, 1996, Philadelphia, pp. 220–229.
16. Felsenstein, J. (1982) *Q. Rev. Biol.* **57(4),** 379–404.
17. Warnow, T. (1993) *New Zealand J. Bot.* **31,** 239–248.
18. Greenberg, J. H. (1957) in *Essays in Linguistics*, ed. Greenberg, J. H. (Univ. of Chicago Press, Chicago), pp. 35–45.
19. Meillet, A. (1925) *La Méthode Comparative en Linguistique Historique* (H. Aschehoug & Co., Oslo).
20. Nichols, J. (1990) *Language* **66,** 475–521.
21. Nichols, J. (1992) *Linguistic Diversity in Space and Time* (Univ. of Chicago Press, Chicago).
22. Nichols, J. (1995) in *Proceedings of the 11th International Congress of Historical Linguists*, ed. Andersen, H. (John Benjamins, Amsterdam), pp. 337–355.
23. Cowgill, W. (1970) in *Indo-European and Indo-Europeans*, eds. Cardona, G., Hoenigswald, H. M. & Senn, A. (Univ. of Pennsylvania Press, Philadelphia), pp. 113–153.
24. Cowgill, W. (1975) in *Proceedings of the 11th International Congress of Linguists*, ed. Heilmann, L. (Mulino, Bologna), pp. 557–570.
25. Cowgill, W. (1979) in *Hethitisch und Indogermanisch*, eds. Neu, E. & Meid, W. (Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck, Austria), pp. 25–39.
26. Ringe, D., Jr. (1991) *Die Sprache* **34,** 59–123.
27. Sturtevant, E. (1933) *A Comparative Grammar of the Hittite Language* (Linguistic Society of America, Philadelphia).
28. White, J. P. & O'Connell, J. F. (1982) *A Prehistory of Australia, New Guinea, and Sahul* (Academic, New York).
29. Roberts, R. G., Jones, R. & Smith, M. A. (1990) *Nature (London)* **345,** 153–156.