Multilateral Comparison and Significance Testing of the Indo-Uralic Question

Brett Kessler and Annukka Lehtonen

Washington University in St. Louis

Unofficial prepublication draft of chapter 3 (p. 33–42) of:
Forster, Peter, & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages.* Cambridge, England: McDonald Institute for Archaeological Research, 2006. ISBN 1902937333.

Among Joseph Greenberg's many contributions to linguistics (Croft 2001, 2002), the one he may be best remembered for is his advocacy and prolific use of a methodology he called *multilateral comparison*. Using that technique, he claimed to demonstrate genetic relationship between many languages: four families in Africa (Greenberg 1963), previously unclassified languages of Papua and vicinity (the Indo-Pacific hypothesis, 1971), most of the native languages of the Americas (the Amerind hypothesis, 1987), and, most recently, a huge number of languages ranging across Eurasia and into North America (the Eurasiatic hypothesis, 2000, 2002). Indeed, Greenberg clearly believed that the technique was capable of demonstrating relationships among all languages. His last book presented preliminary evidence to support the notion that the Eurasiatic group was related to the Amerind languages (2002, 2–3), and he and his colleagues have often spoken of etyma purported to descend from a hypothesized Proto-World, the original human language (Bengtson & Ruhlen 1994; Ruhlen 1994, 101–24).

The prospect of making such great progress in uncovering the phylogeny of human language has excited many people and inspired them to apply multilateral comparison techniques to demonstrate the existence of very large genetic groups. At the same time, the reactions of the overwhelming majority of academic historical linguists have ranged from dismissive to hostile. Especially pursuant to the publication of the Amerind book (1987), several prominent linguists published detailed rebuttals of Greenberg's findings and of his methodology in general (e.g. Campbell 1988; Matisoff 1990; Ringe 1996; Salmons, 1992).

In contrast, adherents of multilateral comparison have not presented a very rigorous explanation of their methodology. What has been offered does not proceed much beyond the most cursory geometrical sketch. One makes a tableau of words,

where the columns represent many different languages, and the rows represent many different concepts; the data cells contain words expressing those concepts in the given languages. One then looks for columns that are more similar in corresponding rows than are other columns. Those similar columns are considered to be languages that are related, or more closely related to each other than to the other languages in the tableau (e.g. Greenberg 1993). Such a description immediately raises concerns because it seems essentially identical to a prescientific methodology known to have performed badly in the past (Poser & Campbell 1992). Occasional attempts to elaborate the methodology mathematically have only hurt its case. For instance, Greenberg and Ruhlen (1992) once claimed to show that a huge number of the world's language families are related because it is possible to find in each family some member language that has a word that has some connection to the concept 'swallow' (e.g. 'suck', 'neck', 'breast') and that has the consonants /m/, /l/, /k/, or most of them, or consonants that are similar. They argued that the probability of such a constellation of facts is vanishingly small; therefore the languages must be related. The computation was wrong on many levels (Hock 1993), but the most damning was the (admittedly widespread) assumption that a single low-probability event suffices to prove a hypothesis. Greenberg repeatedly expressed the idea that multilateral comparison is meant to be effective specifically because the huge data tableaux afford many opportunities to find interesting low-probability patterns. At the same time, he believed that the large amount of data afforded statistical protection against errors. But as many reviewers have pointed out, a strong case can be made for exactly the opposite assertion: that the more data one looks at, the more likely it is that one will find a sizable number of interesting patterns that are simply coincidental, chance occurrences.

In light of such problems, it is perhaps understandable that there has been a backlash against multilateral comparison, with theorists sometimes insisting on procedures that are in all points its exact opposite. Instead of comparing lexemes, the traditional comparative method is meant to require comparing grammatical morphemes and the patterns of their use; instead of looking for similarities, one must specifically look for recurring sound correspondences; instead of being satisfied with matches in a small part of words, one must insist on matches involving at least a CVC sequence; instead of comparing many languages simultaneously, one must look at two at a time (e.g. Nichols 1996; Poser & Campbell 1992).

We concur that such recommendations are likely to result in a much smaller and more manageable set of interesting data that are much less likely to lead researchers

into finding spurious connections between languages. However, we question whether such restrictions are either completely necessary or completely sufficient to solve the problem. Not sufficient, in that even the most rigorous classically trained linguists often remain unsure as to whether the number of recurrent sound correspondences they or a colleague have uncovered is enough to prove that languages are related. Not absolutely necessary, in that perhaps a convincing linguistic study can be performed without radically rejecting all of the principles of multilateral comparison.

In this paper we attempt to elaborate multilateral comparison into something that produces valid, convincing, and perhaps even useful results, while retaining as many of the properties of the methodology as possible. The tack we take is to graft on principles of statistical hypothesis testing so that one can evaluate whether the amount of similarities detected is significantly more than one would expect to see by chance. We take multilateral comparison as point of departure in part because it is an interesting case from a science-theoretic point of view: being diametrically opposed to mainstream techniques, it would not seem at first blush to be a promising candidate for rehabilitation. But in fact, we will show that certain aspects of the methodology make it especially useful for hypothesis testing.

Significance Testing in Historical Linguistics

By now, several studies have addressed the general question of significance testing in historical linguistics; see, for example, the overview in Kessler (2001). Beginning with Ross (1950), all such methodologies exploit the idea that the connection between sound and meaning is arbitrary in a natural language. If one is given the words for ‘black’ and ‘white’ in a language but not told which one is which, there should be no way of solving the puzzle without knowledge of the language or perhaps of a related language. Due to this *arbitraire du signe* (Saussure 1916), if one can show that words for the same concept across two languages share some phonetic property significantly more than do words for different concepts, then that amounts to showing that the words and therefore the languages as a whole are historically connected. Further, if one can exclude the possibility of loans, then one is demonstrating genetic relationship. Of course, not all words are completely unmotivated; ‘mother’ tends to be similar across languages, and if one knows the words for ‘black’ and ‘white’ it is much easier to identify the words for ‘blackness’ and ‘whiteness’. But it is generally believed that such cases can be identified with a little hard work.

Crucial to all significance testing is the proviso that all data be collected in a way unbiased with respect to the research hypothesis. All work after Ross (1950) addresses this by stipulating the use of a specific, predefined list of concepts. On encountering the concept 'black', the researcher is expected to objectively determine the single most usual word for 'black' in the two languages and enter them into the data set, totally without considering whether they constitute good evidence for the language relationship.

Significance-testing methodologies then require something to measure: What property of words should one look at in order to see if it is more abundant between words for the same concept than between words for different concepts? There have been two main threads of research here. One thread (e.g. Guy 1980; Kessler 1999, 2001; Ringe 1992, 1993, 1995; Ross 1950; Villemin 1983), following the traditional comparative method's emphasis on recurrent sound correspondences, has counted the number of times the same pairs of phonemes were found in words expressing the same concept. For example, in Ringe (1992), part of the evidence connecting English and Latin was the fact that an unusually high number of concepts (six) is expressed in Latin by words starting in /k/ and in English by words starting in /h/ (e.g. *cor, heart*; *cornu*, *horn*); any phonetic similarity between the two segments was deemed completely irrelevant. Another thread of research (e.g. Baxter & Manaster Ramer 2000; Oswalt 1970, 1998) has computed the phonetic similarity of the words. For example, Baxter and Manaster Ramer counted how many words begin with segments that are similar to each other. Latin /k/ couldn't match /h/ in their scheme, no matter how many words have that correspondence. But even a single pairing of phonetically similar sounds would be counted, such as /k/~/tʃ/.

The final step in significance testing is to determine whether the measure so computed is significantly greater than expected by chance. Various methods have been proposed, some simpler than others, all differing more or less in accuracy and general validity. Given the wide availability of fast computers, we now believe that the best solution in terms of reliability, accuracy, and applicability across a wide variety of methodologies is the Monte Carlo test of significance; see, for example, Kessler (1999, 2001) for its application to recurrence measures, and Baxter & Manaster Ramer (2000) for its use with phonetic similarity measures, all of which sources explain the theory. In a nutshell, the idea is that if we want to see what the relationship between words of the same meaning would look like across languages if only chance were involved, all we need to do is randomly rearrange the associations between the words and their meanings.

Of course, any particular rearrangement is only one possible chance outcome; what we really want to do is to try all possible rearrangements and see what percentage of them has a measure better than or equal to the measure we actually found before rearranging. If five per cent of the rearrangements has such a high measure, we say that there is a five per cent probability that the relationship between the words is due to chance; or, more technically, we say that the results are significant at $p = .05$. Of course, even with high-speed computers it is impossible to do every rearrangement (a full permutation test) when large amounts of data are involved, but for all practical purposes, sampling a large number of random rearrangements (a Monte Carlo test) is just as good (Good 1994).

The significance testing techniques we have briefly described here may raise two broad classes of objections. First, they ignore interesting classes of data, such as morphology. This is undeniably true, but no one has suggested that historical linguists abandon other methodologies when adding significance testing to their arsenal. It is simply the case that this particular tool requires unbiased, consistent collection of data for which mutual independence of form can be reasonably assumed, and in our present state of knowledge lexical lists afford the most reliable means to that end. Second, word lists have scary associations with glottochronology, which gained some disrepute due in part to inflated expectations for its precision (Embleton 2000). More damningly, virtually every amateur attempt to show connections between languages takes the form of pointing out phonetic similarities between words in a word list. However, problems with such approaches lie not in word lists themselves but in almost every other aspect of the methodology, such as collecting unbiased samples, matching up words by strictly defined criteria of semantic equality, and showing that the similarity is really greater than expected by chance. If such problems are corrected, Saussure's arbitrariness hypothesis actually makes word lists a very good choice for statistical analysis.

Multilateral Significance Testing

With this general background in mind, let us examine how such techniques may be applied to multilateral comparison. Some properties that characterize that methodology are:

- Construction of flexible lexical word lists forming tableaux of words for the same concept across languages;
- use of similarity criteria rather than recurrent sound correspondences;
- simultaneous comparison across many languages.

To what extent can we apply these characteristics to hypothesis testing?

Lexical Lists

As we have seen, the first property—working with lists of words that express the same concept in different languages—is already a staple in significance testing research, and so we adopt it without hesitation. While Greenberg certainly studied grammatical elements, his work showed that lexical comparison proper can be treated as a separable methodology; for example, in his Eurasiatic work he presented lexical analysis in a separate monograph (2002). We propose therefore to restrict our purview here to lexical morphemes.

In one respect, however, we have chosen to depart somewhat from Greenbergian practice. We eschew the great length of the word lists reported in Greenberg's studies, which can number several hundred words. In general, it is true that increasing the amount of data increases the accuracy and significance of tests. Yet empirical studies show that increasing the size of word lists does not always help very much (Kessler 2001; Ringe 1992) and may in fact hurt the analysis. Some words are simply more probative than others because they are less subject to replacement. There is a point past which adding more words just waters down the data and makes it harder to uncover true relationships between languages. It is not clear just how small the lists can be. Although it is intriguing that Baxter and Manaster Ramer (2000) reported success using just 33 words from the Yakhontov list, the bulk of research has used one of the two Swadesh lists—200 or 100 words (Swadesh 1952, 1955, respectively)—and in this study we stay within that conservative range.

In addition to being long, another property of Greenberg's word concept list is that it differed from study to study. These differences must be due in part to the convention against reporting negative data—words for which no cognates are found in a particular study are simply not mentioned—but in part to some real flexibility. We have undertaken to model such flexibility while bowing to the exigencies of controlled hypothesis testing: in particular, we must always guard against selecting words on the basis of what favours the research hypothesis.

We started with the Swadesh 200 list (1952) but omitted the concepts that are typically not fully lexical: 'and', 'at', 'because', 'few', 'he', 'here', 'how', 'I', 'if', 'in', 'not', 'some', 'there', 'they', 'this', 'we', 'what', 'when', 'where', 'who', 'with', 'ye'. We then introduced a process whereby the list would be reduced for specific comparisons, based on how suitable the remaining concepts were for the languages

being tested. First, concepts were discarded outright if a language had no attested word for it (e.g. ‘swim’ in Gothic), or if all relevant words were sound symbolic (e.g. ‘mother’ in many languages), or loanwords (e.g. Latin *petra* ‘stone’, from Greek), or repeat a root that is used more typically elsewhere in the word list (e.g. ‘dig’ in Latin, *fodere*, is also used for ‘stab’).

Next, we assigned the remaining words a suitability factor, essentially an estimate of how likely it is that the word was truly old. This was a two-step process. First, for each individual language, we assigned each word a derivation factor, an estimate of how likely it was to have been derived from a different meaning. For example, the root of Latin *intestina* ‘guts’ clearly means ‘in’, and so it is given a high derivation factor. The other step of this process was to mark each concept by an estimate of its long-term retention rate. The score averages values reported in Swadesh (1955), Oswalt (1971), and values from three studies reported in Kruskal, Dyen, and Black (1973). The values given in the last are actually replacement rates, and were converted to retention scores using Oswalt’s inverse power function transformation (1975). All values were given equal weight. A sixth term in the average indicated whether the semantic concept was used or advocated by a variety of language researchers: Swadesh (1955, where a refined, 100-item list was introduced), O’Grady, Black, and Hale (Alpher & Nash 1999), Yakhontov (Baxter & Manaster Ramer 2000), and Dolgopolsky (1986). For a given concept in a particular language, the suitability factor was its retention rate, proportionally reduced by the best (lowest) derivation rate of any of the words for that concept.

Finally, in any given comparison between languages, we ranked the concepts by the product of their suitability factor in each language, then used the top-ranking 100 words in the analysis. This process combines the best of two traditions: the Swadeshian tradition of using the universally most stable concepts for comparison—though enhanced by using a good deal of research unavailable to Swadesh himself—and the Greenbergian tradition of adapting the word lists to the study at hand. At the same time, nothing in the process biases the selection either in favour of or against the research hypothesis. The only disadvantage is that it requires a bit more thoughtful linguistic analysis than blindly following a word list, and it can be disappointing that one ends up discarding about half of one’s painstakingly collected data. But in the end one does not wish a demonstration of linguistic relationship to fail because one has relied heavily on young words like *intestina* or on concepts subject to constant lexical replacement, like ‘dirty’, which has a retention rate of 3 on a scale of 0 to 100.

There is another way in which Greenberg's use of word lists was unusually flexible. In significance testing as well as almost all other uses of word lists, such as glottochronology, standard procedure has always insisted on selecting exactly one word to represent a concept in each language. Greenberg, however, never shied from using multiple words. In his Eurasiatic comparison (2002), for example, under the concept 'fire' he selected the Old Japanese word *pi* to compare with words like Greek *pyr*, but *atu-* 'hot' to compare with words like Old Irish *áith* (which he glossed as 'furnace'). Although it is hard to defend the imprecision of many of Greenberg's semantic matches, we must concede that often words have synonyms that are hard to choose between. Is there a way to incorporate into a significance test multiple words for individual concepts?

Multiple words would be difficult to handle in many standard statistical frameworks, but they turn out to be very tractable in Monte Carlo tests, because significance testing is done in exactly the same way as the initial measurements—except for having to rearrange and repeat thousands of times. We propose that when measuring the similarity or difference between two languages for a given concept that may be expressed by multiple words, one simply does all pairwise comparisons between the words, and takes their average. For example, if for the concept 'back' Old English offers both *hrycg* and *bæc* and Latin offers both *tergum* and *dorsum*, we would take four measurements: those for *hrycg* vs. *tergum*, *hrycg* vs. *dorsum*, *bæc* vs. *tergum*, and *bæc* vs. *dorsum*, then use the average of those four measurements. As long as exactly the same technique is performed during the rearrangements, the significance testing will be correct and unbiased.

Phonetic Similarity

We have also seen that the second principle—the use of phonetic similarity measures—also has a precedent in significance testing. Its linguistic motivation is unassailable: words that have the same origin are, on average, more likely to sound alike than two randomly selected words that are not cognate; sounds do not always change, and when they do, they are more likely to change by small amounts than to become completely different. The only real problem is that one must define similarity precisely if one wishes to use it in a significance test.

As far as we can determine, Greenberg never published an explicit algorithm for quantifying the similarity of two words. Our proposal is to report the distance between the places of articulation of the first consonants of the word. More precisely, we suggest that one compute the phonetic difference between a pair of words by first isolating the

first consonants of their roots; for all-vowel roots, the first vowel is used. Each of those two segments is scored according to its relative distance from the front of the mouth, in broad terms (labial, 0; dental to prepalatal, 4; palatal, 6; velar, 9; postvelar, 10); sounds with two places of articulation, like /w/, are given two scores (0, 9). The difference between two segments—and therefore between the two roots—is defined as the smallest absolute difference between all crosswise pairings of those scores. For example, /j/ would get a score of 6; when compared to /w/, |6 – 9| is smaller than |6 – 0|, so the distance between /j/ and /w/ is 3. In addition, half a point is added if the segments are not identical to each other.

This choice of similarity function is based on the observation that the place of articulation is the phonetic property most likely to remain constant over great amounts of time. It may appear objectionable that such a function throws away a lot of important data. However, as was the case with long word lists, it turns out that it actually does harm to include additional data that is weaker than the strongest datum, because counterevidence counts against the hypothesis just as surely as positive evidence counts for it. If place of articulation is more durable on average than voice, then requiring one to incorporate voice in a distance metric simply waters down the stronger evidence that place of articulation provides. In the end, counting any evidence beyond the most probative ends up weakening the case, leading one to falsely conclude that related languages are not related (see Kessler 2001 for some empirical tests).

Multilateral Comparison

As far as we know, all previous work involving significance testing has been inherently bilateral, comparing two languages at a time. For example, even though Kessler (2001) performed many tests on eight different languages, all of the comparisons were bilateral; his discussion of multilateral comparison was small and disappointing. When looking for connections between the Indo-European languages and the Uralic languages, Oswalt (1998) did many bilateral comparisons, and Ringe (1998) used protolanguage reconstructions for the two families, thereby reducing a potentially very multilateral analysis to a bilateral analysis again. But Greenberg and his colleagues have repeatedly stressed that multilateralism is the essential nonnegotiable feature of their methodology (e.g. Greenberg 1993, 2000).

There are many interesting things one can do with multi-way interactions between multiple factors, many of which have an elegant mathematical description but pose a challenge for real-world interpretation. For example, there is a straightforward

extrapolation of the Monte Carlo test for bilateral comparisons, where, instead of scrambling the connections between two columns (languages), one scrambles more columns. In the end, such a test may tell whether there is some connection between the languages that were entered. But what is the point of that? If one entered English, Basque, and Sumerian into a three-column table and found out only that there was some connection there somewhere, certainly the very next thing one would do would be to run bilateral tests to find out whether the relationship was restricted to a specific pair of those languages. So why not just do the bilateral comparisons in the first place?

It seems to us that the most immediately fruitful way to deploy multilaterality is not vertically but horizontally. If one knows that a set of languages is related, it is imaginable that that set of languages may prove a more useful comparandum against an unknown candidate language than would any of those languages singly. Even if it may not be immediately clear how that would work exactly, the intuition is probably clear. A language such as Albanian might be very difficult to identify with certainty as an Indo-European language if one could only compare it with a single other Indo-European language, and indeed it was accepted comparatively late into the fold (Bopp 1854). Only in the larger context, when one compares it with the pattern that emerges from considering the more easily grouped Indo-European languages, does the membership of Albanian become more evident.

We propose to emulate that multilateral capability in precisely the same way we have proposed treating multiple words for a concept in the same language. For example, if we know that Latin, Greek, and Gothic belong to the same group (call it Indo-European), we could build a data file where each concept lists words from all those languages. When we got to the concept ‘five’, we would have the set {*quinque*, *pente*, *fimf*}, precisely as if they were synonymous terms in one language. Then if we wanted to see whether Albanian belonged to this group, we would proceed exactly as if comparing Albanian to any single language. When we got to ‘five’, we would end up comparing the phonetic similarity of Albanian *pesë* to the set {*quinque*, *pente*, *fimf*}. By our proposed procedure, we would in effect measure the difference between *pesë* and each of those three words, and take the average. In general, as long as the selection of languages is unbiased, the larger the set, the more likely we will include a similar cognate which will bring up the average score for the comparison.

So, if languages are known to be related, they can be grouped together and treated as an entity in our multilateral comparisons. What if more than two groups remain after we have grouped known siblings? We propose a methodology analogous to

a nearest-neighbour hierarchical clustering: Perform comparisons between all pairs of groups and see which of the pairs is the most strongly connected, then group them together. Then repeat. More precisely, for each of the language comparisons that is significant at the .05 level, we compute the magnitude of the effect by first computing the total dissimilarity across all matching concepts (call that *m*), then computing what that would be by chance (*c*), then reporting the proportional improvement, $(c - m) / c$. The chance dissimilarity *c* can be computed while doing the Monte Carlo significance test: it is simply the average of the total dissimilarity measures across all the rearrangements of the data.

This clustering technique gradually builds up a group that has at its core the most certainly related items. As it grows, it becomes easier to bring in outlying languages whose relationship is harder to establish. Crucially, though, while we believe that the power of the test grows, its bias does not. Just because a cluster might become more attractive as a partner to other Indo-European languages does not imply that it will be more likely to be chosen to partner with an unrelated language.

Test Case: Indo-Uralic

Data and Procedure

The Indo-European and Uralic families are useful test cases of a methodology because each of them comprises many languages whose relationship with each other is now considered completely secure, but which vary widely in how obvious the connection is on the surface. While each family contains language groups whose interrelatedness must always have been obvious (e.g. the Germanic languages, the Balto-Finnic languages), the bulk of the connections weren't discovered until the late eighteenth century, and several others weren't acknowledged as members of the family until later (e.g. Albanian into Indo-European, Samoyedic languages into the Uralic group). A methodology's performance when confronted with these language families would give a good idea of its power.

The possibility of a genetic relationship between Indo-European and Uralic provides a test of a different kind. Such a relationship, nicknamed Indo-Uralic, has long been expected and often been claimed. It is a linchpin of most variants of the Nostratic hypothesis, and is an important element in Greenberg's Eurasiatic hypothesis. If a link between the two families can be demonstrated, it is happy news for almost everybody. If, on the other hand, a statistically rigorous version of multilateral comparison fails to

uncover the connection between Indo-European and the family that most people consider to be its most likely neighbour, then the results of prior multilateral comparisons may be called into question.

Detailed information about the test can be found at our web site, http://BrettKessler.com/multilat. Here follows a synopsis.

Eleven Indo-European languages were chosen. Abundantly attested older languages (Old Church Slavic, Old English, Gothic, Classical Greek, Old High German, Old Irish, Classical Latin, Old Norse, and Sanskrit) were favoured because they should be closer to any putative Proto-Indo-Uralic ancestor and therefore make the connection easier to find. Albanian and Lithuanian, which are only attested relatively recently, were added to flesh out the range of languages considered. In the set of 11 languages, Old English, Gothic, Old High German, and Old Norse form a relatively recently diverging group whose interconnectedness is patent: the Germanic group. Lithuanian and Old Church Slavic form a branch, Balto-Slavic, that is less obvious but nowadays accepted by the vast majority of linguists. For Uralic, four mutually divergent languages were chosen: Finnish, Hungarian, Mari, and the Samoyedic language Nenets. XML files (W3C 2004) were built to store information about the words expressing each of the Swadesh 200 concepts for each of the languages. The words were gathered from a variety of sources without regard to the possibility of genetic cognacy with words from other languages in the study (Balg 1889; Buck 1949; Collinder 1955, 1957; Drizari 1957; Glare 1982; Kessler 2001; Köbler 2003a,b; Kulonen 2000; Lehtisalo 1956; Liddell & Scott 1889; Miklosich 1963; Moisio, Galkin & Vasil′ev 1995; Monier-Williams 1899; Országh 1959; Pewtress & Gerikas n.d.; Quin 1990; Ringe 1992; Sadeniemi 1966). An attempt was made to exclude all loanwords, motivated (sound symbolic) words, and words that share the same root with another word in the same language. For the Uralic languages, words that appear to be loans from Germanic or Balto-Slavic were excluded (principally words restricted to the Balto-Finnic languages), but not those that are merely suspected of being early Indo-European loans into Uralic as a whole (e.g. words for ‘name’ and ‘water’; in Finnish, *nimi* and *vesi*), since these could also be interpreted as evidence for common descent from a shared ancestor. Suitability assignments were made for each concept, as described earlier.

Clustering multilateral analysis of the 15 languages was performed from the bottom up, as if pretending that we did not know of any existing relationships between any of the languages. We started with all pairwise comparisons among the 15 languages, took the pair that had the greatest significant magnitude at p .05 or lower, combined

them into a new group, then repeated with another round of pairwise comparisons, thenceforth treating that combined group exactly like a language. In each of the significance tests, the data was rearranged 100,000 times. This iterative process stopped when no language (or group) was found to be significantly related to any other language or group.

Results

In the initial iteration of the method, when all comparisons were bilateral comparisons between individual languages, the test found statistically significant evidence that 79 per cent of the pairings of intrafamily languages (an Indo-European language vs. an Indo-European language, or a Uralic one vs. a Uralic one) were related to each other. The most difficult Indo-European case was Albanian. While a connection was found between it and each of the Balto-Slavic languages, Old Irish, Greek and Sanskrit, the program reported insufficient evidence for linking it with Latin or any of the four Germanic languages. Some of the Germanic languages had trouble in other pairings: Norse was not connected with Church Slavic or Irish, and Gothic was not connected with Church Slavic or Greek. Greek did not get connected to the Balto-Slavic languages either. Within the Uralic family, the most trouble was caused by Nenets: while the program connected it with Mari, it did not report a connection between Nenets and either Hungarian or Finnish. None of the pairwise matches between Indo-European and Uralic languages were reported as significant.

High German and Gothic were found to be related at a significance level of $p = .00000$, with a magnitude of 78 per cent: The phonetic dissimilarity measure m was only 84, much less than the chance measure c of 382. Therefore that pair was pulled out and treated in the next iteration as a single language. The clusters formed by this process in successive clustering cycles were as follows:

Magnitude	Grouping
78%	High German with Gothic
75%	English with the High German–Gothic group
65%	Norse with English–High German–Gothic
43%	Church Slavic with Lithuanian
34%	Latin with Norse–English–High German–Gothic
31%	Hungarian with Finnish
24%	Albanian with Church Slavic–Lithuanian

23% Mari with Hungarian–Finnish
21% Sanskrit with Latin–Norse–English–High German–Gothic
20% Irish with Albanian–Church Slavic–Lithuanian
14% Greek with Sanskrit–Latin–Norse–English–High German–Gothic
12% Irish–Albanian–Church Slavic–Lithuanian with Greek–Sanskrit–Latin–Norse–English–High German–Gothic
9% Nenets with Mari–Hungarian–Finnish

After this last iteration, the remaining two groups, representing the Indo-European and the Uralic languages respectively, were not found to be connected; the significance level was $p = .45$.

General Discussion

A traditional bilateral approach to significance testing would have been essentially our first clustering cycle, which by itself yielded some paradoxical results that cannot easily be interpreted. The finding, for example, that Albanian is related to Greek, and Greek is related to Latin, but Albanian is not related to Latin has no real-world interpretation, at least not in terms of traditional Stammbaum phylogenetics. The result may be due to some experimental error—perhaps we were overly quick to reject suspected loans from Latin into Albanian but not perceptive enough to catch enough real loans from Greek—but most likely the result reflects uncertainty. A negative result can mean that two languages, although related, changed their vocabulary so fast that they simply do not any longer look much like each other. A positive result could occur simply because, if we are willing to accept matches at a five per cent significance level, then by definition we should expect five per cent of our tests to return a false positive. In view of these uncertainties, it is not clear how conflicting results could be meaningfully resolved at a bilateral level.

The multilateral approach successfully addressed this problem. By iteratively building up increasingly large groups starting with islands of certainty such as the Germanic languages and the Balto-Slavic languages, it eventually built up clusters that had sufficient useful comparanda to enable all the Indo-European languages to be identified and all the Uralic languages as well.

The success in grouping these languages may be taken as a vindication of multilateral comparison. It supports what Greenberg always declared to be the central

tenet of his methodology: Comparing many languages synoptically can uncover evidence of relatedness that is not discernible bilaterally.

In two respects, however, this experiment fails to wholly vindicate Greenberg. Our methodology is not an exact duplication of multilateral comparison as it has generally been practiced. We have introduced several techniques of experimental design that multilateralists rarely if ever discuss or, by implication, practice. Foremost among these are unbiased selection of comparanda, use of strict criteria for determining phonetic similarity, and a method of significance testing to see whether the evidence that turns up is more than expected by chance. To be fair, the majority of historical work lacks these qualities as well, and linguists of all stripes might profitably avail themselves of significance-testing techniques such as those presented here, when the amount of evidence they proffer for historical connections fails to immediately convince colleagues.

The other piece of bad news for the multilateralist research program is that this more rigorous version of the methodology failed to turn up any connection between Indo-European and Uralic. None of the pairwise tests were significant, and the p value of .45 that was obtained when doing the last, great multilateral comparison is weaker by far than any value that would ever be considered to even hint at possible significance. Now we will be the first to admit that the failure to find positive results in any single test set is by no means definitive, and it is imaginable that our results would have been more positive if any number of parameters had been changed. But until new evidence emerges, it is difficult to avoid the conclusion that the Indo-Uralic hypothesis is not well supported by the sort of data afforded by multilateral lexical comparison. Given that multilateralists generally believe that Uralic is one of the closest neighbours of Indo-European and therefore lies at the foundational level of much of their work, the validity of long-range multilateral analyses in general is called into question.

References

Alpher, B. & D. Nash., 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics* 19, 5–56.

Balg, G.H., 1889. *A Comparative Dictionary of the Gothic Language.* New York (NY): Westermann.

Baxter, W.H. & A. Manaster Ramer, 2000. Beyond lumping and splitting: Probabilistic issues in historical linguistics, in Renfrew, McMahon & Trask, 167–88.

Bengtson, J.D. & M. Ruhlen, 1994. Global etymologies, in *On the Origin of Languages*, ed. M. Ruhlen. Stanford (CA): Stanford University Press, 277–336.

Bopp, F., 1854. *Über das Albanesische in seinen verwandtschaftlichen Beziehungen.* Berlin.

Buck, C.D., 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages.* Chicago (IL): University of Chicago Press.

Campbell, L., 1988. Review of Greenberg (1987). *Language* 64, 591–615.

Collinder, B., 1955. *Fenno-Ugric Vocabulary.* Stockholm: Almqvist and Wiksell.

Collinder, B., 1957. *Survey of the Uralic Languages.* Stockholm: Almqvist and Wiksell.

Croft, W., 2001. Joseph Harold Greenberg [obituary]. *Language* 77, 815–30.

Croft, W., 2002. Correction to Greenberg obituary. *Language* 78, 560–4.

Dolgopolsky, A.B., 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia, in *Typology, Relationship, and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*, eds. V.V. Shevoroshkin & T.L. Markey. Ann Arbor (MI): Karoma, 27–50.

Drizari, N., 1957. *Albanian–English and English–Albanian Dictionary.* New York (NY): Ungar.

Embleton, S., 2000. Lexicostatistics/Glottochronology: from Swadesh to Sankoff to Starostin to future horizons, in Renfrew, McMahon & Trask, 143–65.

Glare, P.G.W., (ed.), 1982. *Oxford Latin Dictionary.* Oxford: Clarendon Press.

Good, P., 1994. *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses.* New York (NY): Springer.

Greenberg, J.H., 1963. The languages of Africa. *International Journal of American Linguistics, supplement* 29(1), pt. 2.

Greenberg, J.H., 1971. The Indo-Pacific hypothesis, in *Linguistics in Oceania*, ed. T.A. Sebeok. (Current Trends in Linguistics, 8, vol. 1). The Hague: Mouton, 807–71.

Greenberg, J.H., 1987. *Language in the Americas.* Stanford (CA): Stanford University Press.

Greenberg, J.H., 1993. Observations concerning Ringe's *Calculating the Factor of Chance in Language Comparison. Proceedings of the American Philosophical Society* 137, 79–89.

Greenberg, J.H., 2000. *Indo-European and its Closest Relatives: the Eurasiatic Language Family: Grammar.* Stanford (CA): Stanford University Press.

Greenberg, J.H., 2002. *Indo-European and its Closest Relatives: the Eurasiatic Language Family: Lexicon.* Stanford (CA): Stanford University Press.

Greenberg, J.H. & M. Ruhlen. (1992). Linguistic origins of Native Americans. *Scientific American* 267, 94–9.

Guy, J.B.M., 1980. *Glottochronology Without Cognate Recognition.* Canberra: Australian National University, Department of Linguistics.

Hock, H.H., 1993. Swallow tales: chance and the "world etymology" MALIQ'A 'swallow, throat', in *CLS 29: Papers from the 29th Regional Meeting of the Chicago Linguistic Society, vol. 1, The main session*, eds. K. Beals, et al. Chicago (IL): Chicago Linguistic Society, 215–38.

Kessler, B., 1999. *Estimating the Probability of Historical Connections Between Languages.* Ph.D. dissertation, Stanford University.

Kessler, B., 2001. *The Significance of Word Lists.* Stanford (CA): Center for the Study of Language and Information.

Köbler, G., 2003a. *Altnordisches Wörterbuch.* http://www.koeblergerhard.de/anwbhinw.html.

Köbler, G., 2003b. *Neuhochdeutsch–althochdeutsches Wörterbuch*. http://www.koeblergerhard.de/germanistischewoerterbuecher/althochdeutscheswoerterbuch/nhd-ahd.pdf.

Kruskal, J.B., I. Dyen & P. Black, 1973. Some results from the vocabulary method of reconstructing language trees, in *Lexicostatistics in Genetic Linguistics,* ed. I. Dyen. The Hague: Mouton, 30–55.

Kulonen, U.-M. (ed.), 2000. *Suomen sanojen alkuperä: etymologinen sanakirja.* Jyväskylä: Gummerus Kirjapaino.

Lehtisalo, T., 1956. *Juraksamojedisches Wörterbuch.* Helsinki: Suomalais-Ugrilainen Seura.

Liddell, H.G. & Scott, R., 1889. *An Intermediate Greek–English Lexicon.* Oxford: Clarendon Press.

Matisoff, J.A., 1990. On megalocomparison. *Language* 66, 106–20.

Miklosich, F. von, 1963. *Lexicon Palaeoslovenico-Graeco-Latinum.* Neudruck der Ausg. Wien 1862--65. Aalen: Scientia.

Moisio, A., I.S. Galkin & V.N. Vasil′ev, 1995. *Suomalais-marilainen sanakirja*. Turku: Turun yliopisto.

Monier-Williams, M., 1899. *A Sanskrit–English Dictionary.* Delhi: Banarsidass.

Nichols, Johanna. 1996. The comparative method as heuristic, in *The Comparative Method Reviewed*, eds. M. Durie & M. Ross. New York (NY): Oxford University Press, 39–71.

Országh, L., 1959. *Magyar-angol kéziszótár.* Budapest: Akadémiai Kiadó.

Oswalt, R.L., 1970. The detection of remote linguistic relationships. *Computer Studies* 3, 117–29.

Oswalt, R.L., 1971. Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics* 13, 421–34.

Oswalt, R.L., 1975. The relative stability of some syntactic and semantic categories. *Working Papers on Language Universals* 19, 1–19.

Oswalt, R.L., 1998. A probabilistic evaluation of North Eurasiatic Nostratic, in Salmons & Joseph, 199–216.

Pewtress, H.H. & T. Gerikas (eds.), n.d. *Marlborough's English–Lithuanian and Lithuanian–English Dictionary.* London: Marlborough.

Poser, W.J. & L. Campbell, 1992. Indo-European practice and historical methodology, in *Proceedings of the Eighteenth Annual Meeting of the Berkeley Linguistics Society,* eds. L.A. Buszard-Welcher, L. Wee & W. Weigel. Berkeley (CA): Berkeley Linguistics Society, 214–36.

Quin, E.G. (ed.), 1990. *Dictionary of the Irish Language, Based Mainly on Old and Middle Irish Materials*. Dublin: Royal Irish Academy.

Renfrew, C.; A. McMahon & L. Trask (eds.), 2000. *Time Depth in Historical Linguistics.* Cambridge, England: McDonald Institute for Archaeological Research.

Ringe, D.A., Jr., 1992. *On Calculating the Factor of Chance in Language Comparison.* Philadelphia: American Philosophical Society.

Ringe, D.A., Jr., 1993. A reply to Professor Greenberg. *Proceedings of the American Philosophical Society* 137, 91–109.

Ringe, D.A., Jr., 1995. The "mana" languages and the three-language problem. *Oceanic Linguistics* 34, 99–122.

Ringe, D.A., Jr., 1996. The mathematics of 'Amerind'. *Diachronica* 13, 135–54.

Ringe, D.A., Jr., 1998. Probabilistic evidence for Indo-Uralic, in Salmons & Joseph, 153–97.

Ross, A.S.C., 1950. Philological probability problems. *Journal of the Royal Statistical Society, Series B (Methodological)* 12(1), 19–59.

Ruhlen, M., 1994. *The Origin of Language: Tracing the Evolution of the Mother Tongue.* New York (NY): Wiley.

Sadeniemi, M. (ed.), 1966. *Nykysuomen sanakirja.* Porvoo: WSOY.

Salmons, J., 1992. A look at the data for a global etymology: **Tik* 'finger', in *Explanation in Historical Linguistics*, eds. G.W. Davis & G.K. Iverson. Amsterdam: Benjamins, 207–28.

Salmons, J.C. & B.D. Joseph (eds.), 1998. *Nostratic: Sifting the Evidence*. Amsterdam: Benjamins.

Saussure, F. de, 1916. *Cours de linguistique générale*. Paris: Payot.

Swadesh, M., 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 452–63.

Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–37.

Villemin, F. (1983). Un essai de détection des origines du japonais à partir de deux méthodes statistiques, in *Historical Linguistics*, ed. B. Brainerd. Bochum: N. Brockmeyer, 116–135.

W3C, 2004. *Extensible Markup Language (XML).* http://www.w3.org/XML.