Word Similarity Metrics and Multilateral Comparison

Brett Kessler
Washington University in St. Louis
bkessler@wustl.edu

Abstract

Phylogenetic analyses of languages need to explicitly address whether the languages under consideration are related to each other at all. Recently developed permutation tests allow this question to be explored by testing whether words in one set of languages are significantly more similar to those in another set of languages when paired up by semantics than when paired up at random. Seven different phonetic similarity metrics are implemented and evaluated on their effectiveness within such multilateral comparison systems when deployed to detect genetic relations among the Indo-European and Uralic language families.

1 Introduction

Because the historical development of languages is analogous to the evolution of organisms, linguists and biologists have been able to share much of their cladistic theory and practice. But in at least one respect, linguists are at a disadvantage. While all cellular organisms on Earth are patently related to each other, no such assumption can be made for languages. It is possible that languages were invented multiple times, so that the proper cladistic analysis of all human languages comprises a forest rather than a single tree. Therefore historical linguists undertaking a cladistic analysis – more often referred to as *subgrouping* – have to ask a question that rarely arises at all in biology: Are the entities for which I am undertaking to draw a family tree related to each other in the first place?

The question of whether two or more languages are related is addressed by looking at characters that differ between languages and asking whether observed similarities in those characters are so great as to lead to the conclusion that the languages have a common ancestor. Researchers have investigated many types of characters for this purpose, including fairly abstract ones such as the structure of paradigms, but the most commonly used characters have been the individual morphemes of the language. Morphemes are associations between strings of phones and specific language functions such as lexical meanings or more general grammatical properties. Crucially, those associations are arbitrary to a very great extent. Knowing that a 'tree' is /strom/ in Czech will not help one figure out that it is /ets/ in Hebrew; nor should Hebrew speakers confronted with two Czech lexical morphemes, such as /strom/ vs /firad/, be able to guess which one means 'tree' and which one means 'castle'. An implication of this arbitrariness is that if one pairs morphemes by meaning between two languages, that set of pairs should not have any systematic phonetic property that would not be obtained if morphemes were paired up without regard to meaning. Thus, if one does observe some systematic phonetic property across the semantically paired morphemes, one can conclude that there is some historical contingency that gave those languages that property. Namely, one can conclude that at one time the languages shared the same morpheme for at least some of the meanings, either because of borrowing or because of descent from a common ancestor.

The most straightforward application of this prin-

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 6–14,
Prague, June 2007. ©2007 Association for Computational Linguistics

ciple is to see whether the morphemes for the same concept in two different languages appear unusually similar to each other. Anyone seeing that the morpheme for 'all' was /æɑlː/ in Old English and /alː/ in Old High German, that 'animal' was /deːor/ and /tior/, respectively, and that 'back' was /hrydːʒ/ vs. /hrukː/, and so forth, might well conclude that the languages were related to each other, as indeed they were. Unfortunately, the universal properties of language mean that even unrelated morphemes have something in common; it is not always obvious whether the amount of similarity between semantically matched morphemes is significantly greater than that between semantically mismatched morphemes. For nearly two centuries now, the standard recourse in case of doubt has been the comparative method. One counts how many times the same pair of sounds match up in semantically matched morphemes; for example, Old English /d/ often corresponds to Old High German /t/. A large number of recurrent sound correspondences appearing in several positions in a large number of different words has been considered proof that languages are related. This method is more sophisticated than eyeballing similarities, not least because it recognizes the effect of phonetic apomorphies – sound changes – such as the change of /d/ to /t/ in Old High German. The standard methodology gives no concrete guidance as to how many recurrent sound correspondences constitute proof. However, there have been attempts to recast the comparative method in terms of modern statistical theory and experimental methodology, providing clearcut quantification of the magnitude and significance of the evidence that languages are related (see Kessler, 2001, for recent developments and a summary of earlier work).

One drawback to recent statistical adaptations of the comparative method is that they have been limited to comparing two languages at a time. It has been claimed, however, most prominently by Greenberg (e.g., 1993), that when one wishes to test whether a large set of languages are related, conducting a series of bilateral tests loses power: there may be information contained in a pattern of relations across three or more languages that is not manifest in the bilateral partitioning of the set of languages. Greenberg's approach to multilateral comparison was a step backward to the days before the development of the comparative method (Poser & Campbell, 1992). By his own account, he simply eyed the data and apparently never failed to conclude that languages were related.

Most linguists have rejected Greenberg's approach and many have written detailed refutations (e.g., Campbell, 1988; Matisoff, 1990; Ringe, 1996; Salmons, 1992). But Kessler and Lehtonen (2006) believed that multilateral comparison could be valid and advantageous if applied with some statistical rigour. Adapting Greenberg's basic approach, they developed a methodology that involved computing phonetic similarity between semantically matched morphemes across several languages at a time. This was different from the comparative method, because recurrent sound correspondences were not sought: large numbers of recurrences are not typically found across large numbers of languages. However, it is conceptually straightforward to aggregate similarity measures across morphemes in many languages. Crucially, the similarity across semantically matched morphemes was compared to that obtained across semantically mismatched morphemes. Thus, this application of multilateral comparison is based on the same principles about sound−meaning arbitrariness on which the comparative method was based. Because the similarity computations were completely algorithmic and applied to data collected in an unbiased fashion, the new methodology provided a way to reliably quantify and test the significance of phonetic similarity as evidence for historical connections between two sets of multiple languages. Kessler and Lehtonen demonstrated that the method was powerful enough to detect the relationship between 11 Indo-European languages and that between 4 Uralic languages, but it did not detect any connection between those two families.

The core of the multilateral comparison methodology is the phonetic similary metric. To my knowledge, Greenberg never specified any particular metric. However, many different phonetic comparison algoriths have been proposed for many purposes, including this task of looking for similarities between words (reviewed in Kessler, 2005); in particular, Baxter and Manaster Ramer (2000) and Oswalt (1998) developed algorithms expressly for investigating language relatedness, though only in bilateral tests. In this paper I explore several different phonetic comparison algorithms and evaluate how well they perform in Kessler and Lehtonen's (2006) multilateral comparison task for Indo-European and Uralic.

2 Multilateral Comparison

The basic multilateral algorithm is described in Kessler and Lehtonen (2006); here I give a summary of the relevant facts. For each of 15 languages, we collected all of the words expressing concepts in the Swadesh (1952) list of 200 concepts. However, words were discarded if they violated the key assumptions discussed in the introduction. For example, onomatopoeia and sound symbolism would violate the assumption of arbitrariness: languages could easily come up with similar words for the same concept if they both resorted to natural associations between sounds and their meanings. Grammatical words were rejected because they tend to have certain phonetic properties in common across languages, such as shortness; this also violates arbitrariness. Loanwords were discarded in order to focus on genetic relationships rather than other types of historical connection.

In addition to rejecting some words outright, we tagged others for their relative suitability for a historical analysis. The concepts themselves were scored for how much confidence other researchers have placed in their suitability for glottochronological studies. Some of the contribution to this score was quite subjective; other parts of it were derived from studies of retention rates: how long words expressing the concept tend to survive before being replaced by other words. The words were stripped down to their root morpheme, and then tagged for how concordant that root meaning is with the target concept; for example, if a word for 'dirty' literally means 'unclean', the root 'clean' does not express the concept 'dirty' very well. None of the conditions indicated by these suitability measures invalidates the use of a word, but low retention rates and complex semantic composition mean the word has a lower chance of being truly old and consequently of being a very good datum in a comparison of languages suspected of being only distantly related. These suitability scores were combined for each word in each language. Then, in any given comparison between languages, the suitability scores for each concept were aggregated across words, and the 100 concepts with the best rankings were selected for actual comparison. This technique both ensures the availability of a reasonably large amount of data and also attempts to ensure that the words themselves will be reasonably probative without biasing the test in either direction.

In any single multilateral test, it is assumed that we have a single specific hypothesis: whether one group of one or more languages is related to another group of one or more languages. The approach taken therefore is to determine for each concept how different the words in one group are to the words in the other group. If there are more than one word in each group, then all crosspairs are computed and their average is taken. This approach applies both to the situation where there are multiple languages in a group and multiple words for a given language. These averages are then summed across all 100 concepts, giving a single distance measure: a score of how different the two groups of languages are from each other.

It is important to note, however, that this distance measure is not meaningful in itself. Sets of languages could get relatively low distance measures just because their phonological inventories and phonotactics are very similar to each other's; such typological similarity is not, however, strong evidence for historical connectedness between languages. Rather, what is needed is a relative comparison: how dissimilar would the words be across the two sets of languages if they were not matched by semantics? This is computed by randomly matching concepts in one set of languages with concepts in another set of languages and recomputing the sum of the dissimilarity measures. Each such rearrangement may give a different total distance, which may not be representative, so this procedure is done 100,000 times and the distance is averaged across all those iterations, yielding a very close estimate of the phonetic difference between words that are not matched on semantics. From this one can compute the proportion by which the semantically matched distance is less than the semantically mismatched distance. This proportion is the magnitude m of the evidence in favour of the hypothesis that sets of languages are related to each other. At the same time that the magnitude is computed, one can also compute the significance level of the hypothesis, by counting what proportion of the 100,000 rearrangements has a total distance score that is at least as small as that between the semantically matched words. That number estimates how likely it is that the attested amount of evidence would have occurred by chance, given the phonology of the sets of languages. This paper follows the usual convention in the social sciences of considering significance levels, p, below .05 as being reasonably comfortable.

While each individual test can tell the probability that two sets of languages are related, specific studies may seek to find out which of three or more sets of languages are related. To investigate that, a nearest-neighbour hierarchical clustering is used. In each cycle of the procedure, comparisons are made between all pairs of sets of languages to see which pairs have significant evidence ($p < .05$) of being related. Of those, the pair with the highest magnitude m are combined to form a new, larger, set of languages. The cycles repeat until all languages are grouped into one large set, or no pair of sets have sufficiently significant evidence of being related.

3 Phonetic Distance Metrics

Phonetic distance metrics can be evaluated on several different principles. The ultimate goal is that they should result in p values that are very low when languages are related and high when they are not related. Unfortunately, that goal is only partly evaluable. There are no two languages known for sure not to be related; otherwise there would be no monogeneticists. The best one can test for is m values that correlate well with our incomplete knowledge of the degree of relatedness between languages.

Beyond basic engineering goals of simplicity and efficiency, therefore, a good algorithm should give a relatively low distance score for words or languages known to be related. To the extent possible, it should take minimal account of phonetic features that change quickly over time, and weight more heavily features that tend to be stable over time.

It is perhaps less obvious that a phonetic distance metric should be based on features that are widespread, both across the languages of the world and within individual languages. To take a clearly absurd example, a bad metric would give a distance of 0 if two words agree in whether or not they contained a click, and 1 otherwise. For the vast majority of languages, all word pairs would be assigned a distance 0, because neither word has a click. Such a metric would find no evidence that any pair of clickless languages are related, because the distance of the semantically matched pairs would be no less than the distance of the mismatched pairs. Similarly, even if a feature is found in both languages, it should be neither too common nor too rare. For example, many languages have a contrast between lateral and central sounds, but lateral sounds tend to be vastly less common than central sounds. A metric that compares sounds based on central/lateral distinctions may again end up finding little probative evidence. This observation may seem commonplace for statisticians, but is worth pointing out because the tradition in historical linguistics has always been to look for pieces of evidence that are individually spectacular for their rarity, such as a pair of words whose first five sounds are all identical. It is great to report such evidence when it is found, but bad to demand such evidence in advance, because typically any specific type of spectacular evidence will not show up even for related languages. In a statistical analysis it is much better to look for common pieces of evidence to ensure that their distribution in any particular study will be typical and therefore reasonably conducive to a reliable quantitative analysis.

A much more subtle danger is that a poorly chosen phonetic distance metric might be influenced by parts of the phonology that are not as completely arbitrary as one might like them to be. Because the arbitrariness hypothesis is almost always observed to be applicable in practice, and because it has attained the status of dogma, linguists do not know all there is to know about conditions in which the association between sound and meaning may not be entirely arbitrary and the ways in which that non-arbitrariness may repeat across languages, spuriously indicating that languages are related. However, one strong contender for non-arbitrariness is word length. It appears to be true that words that are longer in one language tend to be longer in another. If a phonetic distance metric is sensitive to word length, it could indicate that semantically matched words are more or less similar than mismatched words, just because their length is similar. This study attempts to minimize that effect by discarding grammatical words, which tend to be systematically shorter than lexical words. It also reduces words to their root morpheme, in part because crosslinguistic tendencies favouring longer words are probably due largely to a tendency to use more morphemes when building lower-frequency concepts. Nevertheless, even these steps are not proof against matching-length effects, and so it would be better for phonetic distance metrics not to be sensitive to word length.

3.1 Candidate Metrics

Seven different phonetic distance metrics were evaluated for this study.

C_1-place. The phonetic distance metric used by Kessler and Lehtonen (2006) was based on the observations that in language change, consonants tend to be more stable than vowels, the front of the word tends to be more stable than the end of the word, and place of articulation tends to be more stable than other features. Consequently it is based on the place feature of the first consonants (C_1) found in the comparanda; only if a comparandum has no consonant at all is its first vowel used instead. Places of articulation are assigned integer values from 0 (lips) to 10 (postvelar), and candidate phones are assigned a list of these values, which allows for secondary and double articulation. The phonetic distance between two sounds is the smallest absolute difference between the crosswise pairings of those place values. In addition, half a point is added if the two sounds are not identical. For example, when comparing the Old English word for 'child', /tʃild/, with the corresponding Old High German word, /kind/, the algorithm would extract the first consonants, /tʃ/ and /k/; assign the postalveolar /tʃ/ a place value of 4 and the velar /k/ a value of 9; and report the difference plus an extra 0.5 for being non-identical: 5.5.

P_1-Dolg. Baxter and Manaster Ramer (2000), in a demonstration of bilateral comparison, used a phonetic distance metric adapted fom Dolgopolsky (1986). Dolgopolsky grouped sounds into 10 classes, which were defined by a combination of place and manner of articulation. Two sounds were considered to have a distance of 0 between them if they fell in the same class; otherwise the distance was 1. Instead of using the first consonant in the word, the first phoneme (P_1) is used instead, but all vowels are put in the same class. Dolgopolsky's idea was to group together sounds that tend to change into each other over time; thus one class contains both velar stops and postalveolar affricates, because the sound change [k] → [tʃ] is common. Thus in the example of /tʃild/ vs. /kind/, the reported distance would be 0.

C_1-Dolg and P_1-place. These metrics were introduced in order to factor apart the two main differences between C_1-place and P_1-Dolg. C_1-Dolg uses Dolgopolsky classes but operates on the first consonant, if any, rather than on an initial vowel. P_1-place uses the place comparison metrics of C_1-place, but always operates on the first phoneme, even if it is a vowel. So many morphemes begin with a consonant that this is often a distinction without a difference, as in the 'child' example. But note how in comparing Old English /æːɣ/ with Latin /oːw/, both 'egg', the P_1 versions would compare /æː/ with /oː/, for a distance of 3.5 by the P_1-place metric (palatal vs. velar vowels) and 0 by the P_1-Dolg metric (all vowels are in the same class); whereas the C_1 metrics would compare /ɣ/ with /w/, for a distance of 0.5 by C_1-place (both sounds have velar components) and 1 by C_1-Dolg.

P_1-voice. This metric is designed to be as simple as possible. Two words have a distance of 0 if their first phones agree in voicing, 1 if they disagree. Breathy voice was counted as voiced. The idea here is that phonation contrast is reasonably universal, and it is a relatively simple matter to partition all known phones into two sets.

C*-DolgSeq. In the comparative method, the best evidence for genetic relatedness is considered to be the presence of several words that contain multiple sounds that all evince recurrent sound correspondences. In particular, multiple consonant matches between words are often sought as particularly probative evidence. This metric implements this desideratum by lining up all the consonants (C*) in the words sequentially (hence Seq). Each such pair of aligned consonants contributes a distance of 1 to the cumulative distance between the words if the consonants are not in the same Dolgopolsky class. If the one word has more consonants than the other word, alignment begins at the beginning of the word, and the extra consonants at the end are ignored. To avoid making this metric sensitive to word length, the total distance is divided by the number of consonant pairs. Continuing the 'child' example, /tʃ/ and /k/ contribute 0 because they are in the same Dolgopolsky class; /l/ and /n/ contribute 1 because they are in different classes; and /d/ and /d/ contribute 0; the sum 1 is averaged across 3 comparisons to give a score of 0.33.

C*-DolgCross. Although the C*-DolgSeq metric attempts to exploit information from multiple consonants in each pair of words, it fails to exploit all possible information. The extra consonants at the end of the longer word are ignored. Further, there is the possibility that the sequential alignment would fail under some fairly common situations. For example, if in one language a consonant is deleted or vocalized, the later consonants will not be aligned correctly. To address this issue, this metric examines all crosswise pairs of consonants and reports their average Dolgopolsky metric. In the example, /tʃ/ is compared to /k/ (0), /n/ (1), and /d/ (1); /l/ is compared to /k/ (1), /n/ (1), and /d/ (1); and /d/ is compared to /k/ (1), /n/ (1), and /d/ (0). Thus the metric is $7/9$, or 0.78.

3.2 Test

Data from 15 languages were used. These languages were selected to give a reasonably wide range of variation in their relatedness to each other. Eleven of the languages were Indo-European, and four were Uralic. Within both of those families there are subclades that are noticeably more closely related to each other than to other languages in the same family. The Indo-European set contains four Germanic languages (Old English, Old High German, Gothic and Old Norse) and two Balto-Slavic languages (Lithuanian and Old Church Slavonic); all the other languages are traditionally considered as belonging to separate branches of Indo-European: Latin, Albanian, Greek, Latin, Old Irish, and Sanskrit. The Uralic set contains three languages that subgroup in a clade called Finno-Ugric (Finnish, Hungarian, and Mari), which is rather distinct from the Samoyedic branch, which contains Nenets. Several linguists believe that the Indo-European and Uralic languages are related to each other (e.g., Bomhard, 1996; Greenberg, 2000; Kortlandt, 2002), though this hypothesis is far from being universally accepted. For each of the 15 languages, translation equivalents were found for each of the Swadesh 200 concepts, as described in Kessler and Lehtonen (2006).

The multilateral comparison algorithm described above was performed once with each of the above-described phonetic distance metrics. Each of the analyses comprised a complete hierarchical clustering of all 15 languages. For each metric, the main concern was whether a multilateral analysis performed with it would group together languages known to be related, however remotely. A second question was what similarity magnitudes would be reported for languages known to be related. In general one would expect a good phonetic distance metric to yield high magnitudes and low p values for languages known to be related, and that, all things being equal, magnitudes should increase the more closely related the languages are.

A large amount of information is available about each run of the program. The algorithm begins by performing bilateral comparisons for each pair of languages, and it might be somewhat interesting to compare those 105 data points across each of the seven metrics. Perhaps more interesting and decidedly more succinct is to focus on the numbers for each of the major clades described above (Table 1). Because almost all of the runs of the program created clusters that contained exactly the languages in each of the clades named in the column headers, it was possible to show the m value reported by the program when that cluster was formed: the degree of similarity between the two subsclusters that were joined to form the cluster in question. For example, when the algorithm using the C_1-place metric joined Old Norse up with a cluster containing Old English, Old High German, and Gothic, it reported an m value of .65 between those two groups. Because of the nature of the clustering algorithm, this represents the weakest link within the clade: in general, the similarity between languages in each of those two subclades will be higher than this number.

A striking feature of Table 1 is the stability of

Metric	Germanic	Balto-Slavic	Indo-European	Finno-Ugric	Uralic	Indo-Uralic
C_1-place	.65**	.43**	.12**	.23**	.09*	.00
C_1-Dolg	.65**	.42**	.12**	.26**	.09**	.02*
C*-DolgCross	.22**	.14**	.05**	.10**	.05**	.01
C*-DolgSeq	.57**	.37**	.09**	.22**	.07**	.02*
P_1-Dolg	.66**	.41**	.13**	.25**	.10**	.02
P_1-place	.66**	.45**	.13**	.31**	.09*	-.01
P_1-voice	.68**	.57**	(.19)	.37**	(.05)	(.05)

Table 1: Similarity Magnitudes Reported for Each Linguistic Clade. *$p < .05$. **$p < .001$. Numbers are the m values reported when the clade is constructed via clustering. If the algorithm does not posit the clade as a cluster, table reports in parentheses the average m reported for each pair of languages in the clade.

the algorithm across different phonetic distance metrics. All of them constructed the relatively easy subclades (Germanic, Balto-Slavic, and Finno-Ugric), reporting very strong significance values. All of them except P_1-voice constructed Indo-European and Uralic, which are both fairly difficult to identify; in fact P_1-voice nearly did so, except that it misclassed Nenets with the Indo-European languages. All of them assigned very low similarity magnitudes to a proposed Indo-Uralic grouping: that is, they found very little similarity between Indo-European and Uralic words for the same concept. Furthermore, the magnitudes for the various clades are all ranked in the same order. As one would hope, the subclades within each family are given much higher m values than the families themselves.

In direct comparisons between comparable version of the place metric and the Dolgopolsky metric (C_1-place vs. C_1-Dolg and P_1-place vs. P_1-Dolg), no very consistent patterns emerge. But the Dolgopolsky metrics tend to reveal the Uralic family with much higher significance levels than do the other measures, and they are also the only metrics that ever posit an Indo-Uralic clade at acceptable significance levels (C_1-Dolg at $p = .04$; C*-DolgSeq at $p = .02$). An optimistic explanation is that the Dolgopolsky classes are better at finding subtle evidence of language relatedness, and that this may be due to their being constructed eclectically. Sounds were claimed to have been grouped into classes based on the frequency with which they are known to develop into each other in the course of language change (Dolgopolsky, 1986:35), not based on any a priori principle; place of articulation clearly is a consideration, but there are many other factors involved. For example, one group comprises the coronal obstruents, except that sibilant fricatives are in a separate group of their own, and sibilant affricates are grouped with the velars. One might expect a system based on empirical data to perform better than one based on a monothetic property such as place of particulation. However, it must also be cautioned that Dolgopolsky did not explain how he gathered the statistics upon which his classes are based. Since the classes were introduced in a paper designed to show that Indo-European and Uralic, among other families, are related to each other, it is possible that the statistics were informed at least in part by patterns he perceived between those language families. There is therefore some small cause to be concerned that Dolgopolsky classes may be, if only inadvertently, somewhat tuned to the Indo-Uralic data and therefore not completely unbiased with respect to the research question.

A more consistent trend in the table is that the metrics that attempt to incorporate more information about the comparanda return lower similarity magnitudes. The C*-DolgSeq metric, which aligns the consonants and reports the average distance across all the pairs. gave substantially lower numbers than the metrics that analyze single phonemes. This observation applies even more strongly to the C*-DolgCross metric, which reported magnitudes a third the size of other measures. The result is not unexpected. It is common knowledge that initial consonants tend to be more stable than other consonants in the word; incorporating non-initial consonants into the metric means that a higher proportion of the data the metric looks at will be more dissimilar. This being the case, it may seem surprising that C*-DolgSeq and C*-DolgCross showed essentially the same connections between languages as did the other metrics, and at strong significance levels. Even though the similarity levels are close to background levels (those of semantically unmatched pairs), they are still measurably above background levels; the p values are only concerned with whether the matched data is more similar than the unmatched data, not by how much they are different.

P_1-voice was introduced to experiment with a metric that takes the other approach: instead of incorporating more material into the measure, it incorporates less. Being based on a single binary phonetic feature, P_1-voice is arguably the most minimal metric possible. Perhaps not unexpectedly, it has the opposite effect of that of C*-DolgSeq and C*-DolgCross: m measures are raised. At the same time, this metric too appears to reveal the known relations between languages. The several gaps in the table are due to a single odd choice that the algorithm made: it concluded that the Uralic language Nenets was quite similar to the Germanic languages, at least with respect to whether the first sound is voiced in semantically matched words. Presumably this connection was just a chance accident; indeed, saying that one is working with significance levels of .05 is another way of saying that one is willing to tolerate such errors about 5% of the time.

4 Conclusions

The evaluation of the methodology across 15 languages did not provide overwhelming evidence favouring one type of phonetic distance metric over another. Perhaps, by a small margin, the strongest results are obtained by comparing what Dolgopolsky classes the first consonants – or, equally well, the first phonemes – of the words fall into, but nothing seriously warns the researcher away from other approaches.

Conceivably further experiments with other data sets will reveal strengths and weaknesses of different metrics more convincingly. Until such time, however, it may be most useful to choose phonetic distance metrics primarily on theoretical, if not philosophical, criteria. Metrics that look at many parts of the word have the advantage of not missing information, even if it turns up in unusual places. It is not unknown for a branch of a language family to do something unusual like drop all initial consonants; in such an event, all the single-phoneme metrics explored here would fail entirely. One does not really wish to change one's metric for different sets of languages, because if one has the freedom to fish for different metrics until a test succeeds, one can almsot certainly – and spuriously – prove that almost all languages are related. So there is some advantage to having a metric that covers all the bases. But the similarity measures returned under such circumstances do tend to be small, and although such reduction in m did not seem to have any deleterious effect in the present experiment, it is not unreasonable to worry that weak similarity measures may cause problems in some data sets. Further, the more of a word one is looking at, the more likely it is that one will inadvertently encode length information into the metric.

The main conclusion to be drawn from this study is that the basic methodology is very hospitable to a variety of phonetic distance metrics and performs adequately and stably with any reasonable metric. Unlike parametric methods, this randomization-based methodology does not require the researcher to develop new formulas to compute strength and significance values for each new distance metric. The simple expedient of randomly rearranging the data a large number of times and recomputing the distance metric for each rearrangement provides the most literal and straightforward way of applying the key insight of the arbitrariness hypothesis: the phonetic similarity of semantically matched words will be no greater than that of semantically mismatched ones, unless some historical contingency such as descent from a common language is involved.

References

William Baxter and Alexis Manaster Ramer. 2000. Beyond lumping and splitting: probabilistic issues in historical linguistics. In *Time Depth in Historical Linguistics*, eds. C. Renfrew, A. McMahon., and L. Trask. McDonald Institute for Archaeological Research, Cambridge, England. 167–188.

Allan R. Bomhard. 1996. *Indo-European and the Nostratic Hypothesis.* SIGNUM Desktop Publishing, Charleston, SC.

Lyle Campbell. 1988. Review of Greenberg (1987). *Language* 64:591–615.

Aaron B. Dolgopolsky. 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 and T. L. Markey. Karoma, Ann Arbor, MI. 27–50.

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

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

Brett Kessler. 2001. *The Significance of Word Lists.* Center for the Study of Language and Information, Stanford, CA.

Brett Kessler. 2005. Phonetic comparison algorithms. *Transactions of the Philological Society* 103:243–260.

Brett Kessler and Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. *Phylogenetic Methods and the Prehistory of Languages*, eds. P. Forster and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge, England. 33–42.

Frederik Kortlandt. 2002. The Indo-Uralic verb. In *Finno-Ugrians and Indo-Europeans: Linguistic and Literary Contacts.* Shaker, Maastricht, 217–227.

James A. Matisoff. 1990. On megalocomparison. *Language* 66:106–120.

Robert L. Oswalt. 1998. A probabilistic evaluation of North Eurasiatic Nostratic. In *Nostratic: Sifting the Evidence.*, eds. J. C. Salmons and B.D. Joseph. Benjamins, Amsterdam. 199–216.

William J. Poser and Lyle 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, and W. Weigel. Berkeley Linguistics Society, Berkeley, CA. 214–236.

Donald A. Ringe. 1996. The mathematics of 'Amerind'. *Diachronica* 13:135–154.

Joseph Salmons. 1992. A look at the data for a global etymology: *Tik 'finger'. In *Explanation in Historical Linguistics*, eds. G.W. Davis and G.K. Iverson. Benjamins, Amsterdam, 207–228.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96:452–463.