Language and Linguistics Compass 2/5 (2008): 821–839, 10.1111/j.1749-818x.2008.00083.x

The Mathematical Assessment of Long-Range Linguistic Relationships

Brett Kessler*
Washington University in St. Louis

Abstract

Language classification differs from biological cladistics in that monogenesis cannot be assumed. Before a cladogram or family tree can be accepted, linguists must be convinced that the languages are related at all. Morpheme tables, or word lists, provide a good framework for investigating relatedness, but methodologies for quantifying and assessing the data statistically are still being developed. The comparative method furnished a viable statistic, recurrent sound correspondences, but by no means to see whether they exceeded levels expected by chance. Organizing correspondences into contingency tables permitted hypothesis testing, with Monte Carlo resampling methods providing the flexibility to support a wide variety of test statistics, including different ways of computing sound recurrences and phonetic similarity. Thus, techniques from both the comparative method and multilateral comparison can be deployed with rigorous numeric assessment. Experiments seek to increase the power of the tests to explore new hypotheses and verify long-range language relationships.

Of the many fascinating things linguists do, one might think that the hoary field of comparative linguistics would be one of the least likely to capture the attention of the rest of the world. Yet the public, or at least the journalists who serve them (e.g. Nova 1997; Wade 2000), are fascinated to learn of prehistoric connections between languages. Perhaps more surprising still, comparative linguistics has caught the attention of scientists in many other fields. Computer scientists, mathematicians, geneticists, and biological taxonomists are now some of the most visible practitioners of comparative linguistics (Forster and Renfrew 2006; McMahon and McMahon 2006).

Much of the appeal of comparative linguistics for biologists stems from the fact that a key issue in both disciplines is essentially the same. Given a set of entities (languages, organisms) known to have evolved from a common ancestor, scientists seek to recover the series of divergences that led to the current state of affairs. Biologists have been at the forefront of the effort to systematize this field, which is now generally called cladistics, and to develop algorithms for performing automatic cladistic analysis.

Because linguists have been slower to develop similar approaches to the analogous cladistic problem in their field – subgrouping – biologists and their colleagues have stepped up to apply their tools to language. The result has been a new era of mathematical approaches to comparative linguistics, which has been admirably surveyed by McMahon and McMahon (2005). These new approaches are fascinating not only to linguists, but also to archaeologists and geneticists, who increasingly treat subgroupings between languages as evidence for how human populations dispersed (Cavalli Sforza 2006).

Applying biological tools to linguistics depends crucially on respecting the relevant differences between biology and language. An important difference that is often overlooked is the different status of the theory of monogenesis in the two fields. In biology, it goes without saying that all cellular organisms have a common origin; in linguistics, monogenesis of languages is one good theory, but so is polygenesis. Therefore, when comparative linguists are confronted with a set of languages to analyse cladistically, they must undertake a task that biologists do not confront: they must assess whether there is sufficient evidence for grouping the languages together at all. Whereas a cladistic tool designed for biology might valiantly produce the best possible phylogenetic tree for any set of data, a linguist must acknowledge the theoretical possibility of polygenesis, and therefore in some cases accept that the better analysis might be the one that posits a forest of unrelated trees. This article surveys how linguists have approached the problem of demonstrating whether languages are related, with emphasis on mathematical or computational methods and the mathematical justifications for methods that are not explicitly quantificational. The reader is warned that this narrow purview purposely excludes some of the most exciting recent topics in modern comparative linguistics. The article does not deal with subgrouping, but restricts itself to means of deciding whether languages are demonstrably related and therefore eligible for subgrouping in the first place. Nor does this article examine the role of external evidence for language relatedness, such as similarities in the genetics of populations speaking the languages in question. Convergence of data between linguistics, archaeology and genetics is an exciting trend in the study of human prehistory (see, for example, the papers presented at the Workshop on Alternative Approaches to Language Classification 2007 – http://aalc07.psu.edu), but an important prerequisite for convergence is that each discipline makes its own independent contribution.

Morpheme Lists and Phonetic Similarity

The simplest method of demonstrating that languages are related is simple inspection. No one seriously proposes that linguists have an obligation to mathematically prove obvious relationships, such as that of English to Dutch. Thus, the use of the word 'long-range' in the title of this article,

a long-range relationship being any proposed family relationship that encounters some serious skepticism.

One early refinement of the simple inspection technique was to organize the language data into morpheme tables. Jones (2007), when he famously declared in 1786 the relatedness of what are now universally recognized as several branches of the Indo-European language family, did so by invoking similarity between 'the roots of verbs and the forms of grammar'. That is, he drew up a list of concepts and the morphemes that expressed them in Sanskrit, Latin, Greek, and so forth. Then he observed that morphemes expressing the same meaning or grammatical function tended to be especially similar across these languages. Below is a fragment of such a table, including three languages that Jones discussed plus Hebrew, a non-Indo-European language.

Meaning	Sanskrit	Latin	Greek	Hebrew
'carry'	*bhar*	*fer*	*pher*	*naśa*
'stand'	*sthā*	*stā*	*stē*	*'amad*

Most of this article will discuss methodologies that either use such tables directly or can be conceptualized in such terms. Such tables will be called 'morpheme tables' for the sake of clarity; 'word lists' is the more common terminology. Concepts or functions along with their translation equivalents in various languages will frequently be referred to as rows, languages as columns. Accordingly, comparing translation equivalents to each other may be referred to as comparisons within rows.

As obvious as Jones's procedure might be, Jones already had a major methodological insight. His comparison is strictly between morphemes that have the same meaning or function. He did not compare the overall appearance of the languages, an approach that is not likely to be very productive. One may suspect that languages are related if they share some common typological feature that impresses us as strange: perhaps a super-abundance of vowels, or the presence of contrastive tones, or an exotic word order. But such typological features have continually proven to be disappointing (Campbell 2005: 277), because so much is unknown about how to objectively characterize such features, how often they would be expected to appear by chance in unrelated languages, how likely they are to be borrowed from other languages, how rapidly they might change within a given language, whether there is a natural direction for such change, and what the joint probabilities are for different typological features to co-occur. Although the growing availability of large-scale typology databases (Haspelmath et al. 2005; Bickel and Nichols 2007) is beginning to provide clues to some of these questions, it would be premature to use typological congruence within a specific set of languages as a definitive demonstration of their relatedness.

The other insight worth pointing out here is that Jones compared morphemes, not words per se: roots and inflections were treated as two separate pieces of information. One might conceptualize this as having an additional row in the table extract that contains the concept 'first person singular agreement marker' with forms like *-mi*, *-m*, *-ō*, etc. across the various languages. If instead one used a full-word list, so that these markers appeared dozens or hundreds of times on many different verbs, the apparent similarities between languages would be vastly increased. Any similarity in this inflectional ending across languages is just one similarity, worthy of exactly one row in the table.

One thing that is not particularly clear in Jones's explanation – it is just one paragraph in an essay that immediately shifts topics – is what standard of comparison he intended. Are *bhar*, *fer* and *pher* remarkably similar to each other compared to their similarity to non-Indo-European words like *naśa*? Or, is it that Sanskrit *bhar* is more like Latin *fer*, a morpheme in the same row, than it is like Latin *stā*, a morpheme in a different row? These can be distinguished as between-columns and between-rows contrasts. A reader's first intuition is typically that between-columns contrasts are intended. There is a lot to be said for this intuition: it is true, on average, that a column will be more similar row-by-row to a column for a related language than to a column for an unrelated language. Yet, there is no fundamental, theoretical reason to think that between-columns differences will suffice to prove whether languages are related or not. It is possible to have a wide range of similarities in a world where all languages are related, or all unrelated. It is also easy to imagine similarity measures that would be thrown off when the languages in consideration are typologically similar. Languages with similar sounds in their inventories will have words that are more similar than other languages. For example, two languages that just happen to use *k* a lot at the beginning of words, such as Finnish and Hawai'ian, might have a lot of similarity by some measures, but this does not make Hawai'ian related to Finnish and not to, say, Tahitian, a language totally lacking *k*.

If we accept one of the primary axioms of modern linguistics, Saussure's claim that the association between sound and meaning in words is arbitrary (1916), then it becomes immediately clear that what is called for is between-rows comparison: Sanskrit *bhar* is more like Latin *fer* than it is like Latin *stā*. If Saussure was right, then there would be no reason for within-row similarities to be greater than between-row similarities for the same languages, other than some sort of historical connection between the languages, such as the descent from a common ancestor that one is trying to prove.

Of course, whatever comparisons Jones was thinking of, he no doubt relied entirely on human intuition, unaided by mathematics, to decide whether the magnitude of those similarities was significant, whether exceptions were few enough to be disregarded, and so forth. But his

technique was arguably much more powerful than inspecting the general appearance of languages, and it laid the groundwork for more rigorous statistical evaluations a couple of centuries on.

It is curious that a procedure very similar to Jones's technique was introduced in the latter half of the twentieth century as a new cutting-edge methodology. Greenberg's method of multilateral comparison consisted of visual inspection of word tables, with the proviso that the table should be very large, particularly in the horizontal dimension: many languages should be compared at a time (Greenberg 1993). No explicit quantification or significance testing was added to the procedure, despite major advances in statistics and in computational science since the time of Jones. However, Greenberg did justify his procedure on mathematical grounds. He demonstrated how much less likely, and therefore much more probative, it is to find a similarity across many languages than across few languages. That is obviously true, but many critics (e.g. Ringe 1993) have pointed out that Greenberg rarely if ever reported similarities across all the languages that are meant to be related. If one is free to proffer similarities across any subset of a large number of languages, then it actually becomes hugely more likely, and therefore infinitesimally probative, to adduce such similarities.

Greenberg primarily compared similarities between columns rather than between rows, because his main concern was to perform a subgrouping of languages, based on the assumption that all languages are related just like all biological organisms are (Greenberg 1987). He would say that *bhar*, *fer* and *pher* are more like each other than any are like *naśa*, and therefore this is a piece of evidence grouping Sanskrit, Latin, and Greek together to the exclusion of Hebrew. Because of this profound difference in goals, it may be wrong to think of multilateral comparison as a methodology for assessing whether languages are related. Perhaps some of the difference in reception Greenberg's analyses found among linguists compared with other scientists such as geneticists is due to programmatic differences resulting from different understandings of the status of monogenesis. When Greenberg claimed that there is a Eurasiatic grouping of languages (2000), biologists may have largely understood his claim to mean simply that of all languages, which are all assumed to be related anyway, the preponderance of evidence is that the languages in that proposed group are more closely connected to each other than to the other languages he considered. Historical linguists, who must labour under the possibility of polygenesis, took it as an unconvincing attempt to demonstrate that Indo-European is definitively related to Uralic, Japanese, Eskimo, and many others (e.g. Georg and Vovin 2003).

Comparative Method

The development of the comparative method in the nineteenth century put comparative linguistics on a more objective footing. The core of this method starts off, at least conceptually, with morpheme tables. But instead

of simply affirming that the words look similar across the rows, the special connection is measured in terms of recurrent sound correspondences (Campbell 2005). The linguist first aligns each of the sounds in words in the same row. Aligning words is an ill-defined and difficult task, but the process can be kick-started by looking first at the initial sounds of the words, which, in most languages, are more stable over the millennia than other sounds in words. In the data shown in the table above, for example, *s* ~ *s* ~ *s* is one sound correspondence. If the same correspondence is found in another row of the table, it graduates to being a recurrent sound correspondence. There is no requirement that the sounds in the correspondences should be identical or even similar: *bh* ~ *f* ~ *ph* is also a recurrent sound correspondence that is just as highly valued as *s* ~ *s* ~ *s*.

Recurrent sound correspondences were emphasized, because the comparative method has other goals beyond just proving languages related. Sound correspondences help recover the specific history of the languages. But for present purposes, they also help demonstrate that languages are related by providing evidence that can be precisely defined and quantified. One can report how many different recurrent sound correspondences were uncovered; how many times each of them appeared in words; and how many rows have words that agree with each other in having two, three, or more different sounds in them that participate in recurrent sound correspondences. To illustrate this last point: in Indo-European, the correspondences *bh* ~ *f* ~ *ph*, *a* ~ *e* ~ *e*, and *r* ~ *r* ~ *r*, among many others, are all found in many words. The fact that the roots for 'carry' agree in having a recurrent sound correspondence for each phoneme is a compelling piece of evidence that the languages are truly related.

People who work with the comparative method use other evidence for genetic relatedness as well. Some claim that comparing the grammars of languages is the best or only reliable evidence (Poser and Campbell 1992). The strongest motivation for such a claim is that grammatical morphemes are less likely to be borrowed than are whole words. However, some claims for the superiority of grammar over lexicon are more debatable. Often it is claimed that similarities between paradigms, such as similarities in several case endings, are especially probative. This is true as far as it goes, but there is no clear reason for believing that similarities between three morphemes are any more probative when people think of them as belonging to a paradigm than are similarities between any other set of three morphemes, except to the extent that they constitute a more restricted search space (see below). There are also good arguments for discounting grammatical information. Grammatical morphemes tend to be short and somewhat similar between languages. These factors make it more likely that any given pair of grammatical morphemes across languages will be more similar to each other than will a pair of content words, so one needs to amass more morpheme matches when working with grammatical elements. Languages often have very different grammatical structures, which can make it impossible

to find morphemes that have the same function in two different languages. This can lead the researcher to do more abstract grammatical comparisons, which easily devolves into typological comparison with all of its concomitant analytical problems. In practice, invoking grammatical information typically means that the researcher has found some striking similarity between languages, such as Hrozný's (1917) demonstration that Hittite, like some other Indo-European languages, has a noun class with an *r*/*n* alternation. Such discoveries can be impressive and convincing from an informal perspective. But from a statistical perspective, the idea that one should look through the grammar until one sees something impressive is difficult to operationalize, not least because it is virtually impossible to delineate the search space: exactly how many other possible candidates for impressive similarities failed to materialize? To take an artificial example: imagine you know that a certain type of similarity has only a 1% chance of happening by coincidence, and you have found five such similarities when comparing two languages. If you had looked at 100 independent candidates for such similarity, the probability of finding at least five or more similarities would be 0.003, an impressively low number. If you had looked at 1,000 candidates, the probability of finding five or more similarities would be 0.971, close to a dead certainty. Although the binomial distribution makes this easy to calculate, in practice it is difficult to know what numbers to plug in when one has intensively dug through two languages looking for any and all striking similarities.

The comparative method has been a huge success and is still actively being used to develop and evaluate long-range proposals; recent examples include an expansion of the Austronesian language family (Blevins 2007) and a link between languages of North America and Siberia (http://www.uaf.edu/anlc/dy2008.html). At the same time, acceptance of the comparative method has cut down on the number of long-range analyses that linguists feel obliged to take seriously, because the requirement to adduce numerous well-supported recurrent sound correspondences introduced a quantifiable component into comparative linguistics. But there was a glaring omission in the method: it gave no guidance as to what quantity is needed before one can confidently claim that the number of recurrences found are more than expected by chance.

Rules of thumb emerged. Meillet (1925) wrote approvingly of basing evidence on rows of words that all contain, in the same order, four sounds in recurrent correspondence sets; three sounds make for a proof that is 'moins forte' (less strong); two sounds, 'fragile'; and one sound, 'à peu près nulle' (1925: 36). This is, of course, scant guidance based on no hard data at all, but arguably it is impossible to do much better, considering all the factors that can influence the reliability of the proof. Certainly, the more sounds that match, the better, but also the more rows of words that match, the better: might a large score in one dimension make up for a deficiency in the other?

Ross (1950) developed a promising and remarkably simple way of quantifying the role of chance. He proposed using a contingency table to keep track of correspondences. The rows are labelled with the sounds of the one language and the columns with the sounds of the other language. Every time a correspondence is encountered between sounds in the two languages (e.g. it is observed that in Sanskrit the word for 'carry' starts with *bh* and in Latin, *f*), a tally is added to the cell for that sound pair. Such a table can, in principle, be subjected to statistical tests of significance. However, Ross only sketched how such a test might be developed, and did not report deploying a finished version.

Probabilistic Models

There have been several attempts to quantify the evidence for language relatedness and present the result as a single number. These usually involve basic probability theory. Collinder (1947) argued for a large Uralic-Altaic family by adducing 13 similarities between the member languages. He claimed that if he were to work out the probability of 13 features agreeing by chance, the number would be infinitely small, and therefore a historical connection between the languages was infinitely likely. Hymes (1956) made a similar argument for linking Tlingit with the Athapaskan languages by computing that the odds against their having the same order of verb affixes was 1,216,189,440,000 to 1. Dolgopolsky (1986) found similarities between words for 13 different concepts, and decided that the probability of chance occurrence would be 10^{-20}, which proves the existence of a broad Sibero-European family. Nichols (1996) demonstrated that any language that had an Indo-European gender system would be, in fact, Indo-European. She did this by computing probabilities, many of them derived from frequencies observed across large numbers of languages, that a language would have genders, that it would have at least three genders, that one of the gender markers would be *-s*, and so forth. When she finished with all the probabilities, she multiplied them together to get 0.00000057, a very small number.

Several researchers have addressed the question of how many CVC (consonant–vowel–consonant) matches one could expect to find when comparing words (Swadesh 1954; Cowan 1962; Bender 1969; Ringe 1999). The answer depends on how similar sounds must be in order to be considered a match, and how many languages are being compared. In the bilateral case, Swadesh (1954) used probability theory to demonstrate that only about four matches would be expected per hundred words. Using different formulas, Cowan (1962) concluded that no more than three matches would turn up in 95% of chance cases. Bender (1969) took the empirical tack of counting the number of matches between pairs of 21 unrelated languages and lowered the number to two per hundred. Ringe (1999) used probability theory to emphasize that a very large

number of matches are expected by chance if one reports any CVC match between any pair of languages out of a large set of languages, as is done in multilateral comparison.

Probabilistic analysis and the language modelling it entails are worthy topics of research, but linguists have rightfully been wary of claims of language relatedness that are based primarily on probabilities. If nothing else, skepticism is aroused when one is informed that a potential long-range relationship whose validity is unclear to experts suddenly becomes a trillion-to-one sure bet when a few equations are brought to bear on the task. Attempts to use probabilistic reasoning to assess specific long-range hypotheses almost always run into one or more of the following problems.

Defining the problem as finding the probability of a specific observation. In a complex system like language, one expects that it will be fairly easy to find coincidental similarities that would have very low probability of occurring. It is true that, all things being equal, the lower the probability of an observed similarity the less like it is to be coincidental, and most probabilistic analyses strive to present a very low number. The problem is that there is no straightforward way to decide whether a number is low enough to be convincing in itself. One sometimes sees the number 0.05 adduced, which is used in the context of statistical significance testing in social science experiments. But significance testing is based on several conditions that must be fulfilled before that numeric cut-off can be meaningfully applied. One requirement is that the analysis should be based on a substantial number of observations made in a fashion unbiased by one's hypothesis; one magnificent anecdote is not sufficient. Another requirement is that one does not compute the probability that the measure in question – here, similarity – would be exactly the observed value by chance. One calculates the probability that the measure would have at least the observed value; that is, one must calculate the cumulative tail of the distribution. The difference can be illustrated by considering a man who claims that he is unusually wealthy because he has $4,746.29 in the bank. Suppose it comes to light that only eight people in the USA have exactly that sum in their bank account. Considering the population of the USA, the probability of having that bank balance would be 0.00000003. That is an impressively small number, but hardly enlightening. The question really calls for computing the proportion of people who have $4,746.29 or more, which is likely to be a much larger number.

Failure to precisely specify the search space. Five good correspondence sets obtained from studying 25 words might be strong evidence; finding only five good correspondence sets after studying 10,000 words might not be so impressive. It is understandable that linguists want to uncover and utilize all available evidence. It is also understandable that publishers are rarely interested in publishing thousands of pieces of negative evidence.

The result, though, is a collection of data that is not susceptible to statistical control. A breakthrough came when Swadesh (1950) and others began developing standardized lists of concepts to be used in statistical studies. Basing an analysis on all the concepts on the list and no others provides the sort of control that is needed. However, even with an explicit list, it is shockingly easy to dramatically expand the search space in an unquantifiable way by using wide latitude in selecting translation equivalents. A linguist who has just supplied *Hund* as the German word for 'dog' might feel tempted to select the word *hound* for English. In a non-statistical analysis, this is arguably the right thing to do: one wants to find the cognates, and one can imagine how *hound* could easily acquire its present meaning if it originally meant 'dog' in general. From the standpoint of statistical assessment, such freedom means disaster, because there is no way to quantify the space from which the linguist has drawn the observation.

Using simplifying models of morphemes. It is not at all uncommon for researchers to model what a morpheme can look like in a language by defining it as a string of sounds, each different sound having an equal chance of occurring in each position of the morpheme. Some approximations get better, or add slop factors (Swadesh 1954; Cowan 1962), but rarely does one see a model that takes into account the real inventories, phoneme frequencies and phonotactics of the languages under consideration, not to mention morpheme level constraints: whether the language is likely to repeat the same consonant in a root, how likely it is to tolerate two roots that are homophonous, and so forth.

Assuming that similarity itself is probative. Assuming that similarity itself is probative in the absence of any theory as to why that should be the case. Because closely related languages tend to be more similar to each other than languages not known to be related, it is tempting to think that a relatively high degree of similarity between two languages is proof of similarity. But a simple thought experiment refutes that idea. If all languages were unrelated, they would still differ in how similar they were to each other. Some language pairs would be more similar to each other than would be the case for typical pairs of languages, but that would not mean that they were related. Thus, no proof can be based on demonstrations of relative degree of similarity. Another tack might be to measure the degrees of similarity between all related languages and between all unrelated languages; if there turned out to be little or no overlap between these two ranges of similarity, one could classify new pairs of languages by seeing which range their similarity measure fell into. Unfortunately, there are no pairs of languages that are known to be unrelated, so there can be no calibration of their similarity scale. There appears to be no way to say whether any particular similarity measurement in itself is probative evidence

of a relationship between two languages, or whether it is likely to occur in a pair of unrelated languages.

Arbitrariness Testing

More recently a paradigm has emerged under which the problems inherent in the probabilistic approach are largely overcome. It may be called arbitrariness testing, because the statistical tests performed are motivated explicitly by Saussure's doctrine that the connection between sound and meaning is arbitrary. The chosen statistic, whether it is a count of recurrent sound correspondences or some measure of phonetic similarity, is first computed across columns in the same rows; Oswalt (1970) called this the 'gross' score. The same score is then computed in the same way, but between rows; Oswalt called this the 'background' score. Neither the gross nor the background score is of interest in itself, but rather the relation between them. Of all ways of computing the background score by matching words across different rows, does the gross score end up looking pretty typical, or is it higher than the great majority of background scores? Only in this latter case can one conclude that the similarity between words in the same row is significantly greater than one would expect by chance. That would mean that some historical contingency must be responsible for an apparent invalidation of Saussure's doctrine.

In precise terms, one wants the tail of the distribution: the proportion of all background scores that are at least as high as the gross score. That would constitute the probability that the gross score would occur by chance if the vocabularies of the two languages are not historically connected. This value is often called 'the significance level' of an analysis. Most social scientists are satisfied if that level is 0.05 or lower.

Because similarity between rows is an important factor, negative data obviously must be retained and quantified. This requirement is satisfied by building up unbiased morpheme tables. 'Unbiased' in this context means that the selection of data is not influenced in one way or the other by the research hypothesis.

ARBITRARINESS TESTING USING RECURRENT SOUND CORRESPONDENCES

The contingency tables introduced by Ross (1950) present both positive and negative data from the standpoint of the comparative method, and so are suitable for arbitrariness testing, although it took several decades to work out exactly how to do that. One might imagine doing a χ^2 test over the whole table. However, the standard technique for interpreting the significance of the χ^2 statistic invokes the χ^2 distribution, which does not give reliable answers when the numbers in most cells are very low, as is almost always the case. Villemin (1983) attempted to finesse the problem by running a separate χ^2 test on each cell of the contingency table, and

claimed that the languages should be considered related if any cell had a very low significance level. Unfortunately, picking the one test out of hundreds that gives the desired answer virtually guarantees that any analysis will support one's hypothesis.

Ringe, in an important series of papers (1992, 1993, 1995, 1998), worked to refine the statistics to avoid that fallacy. The first step of his procedure was to count how many cells in the contingency table have a count significantly higher than expected by the hypergeometric distribution at a chosen significance level. The next step was to run a second significance test, using the binomial distribution to see how many statistically significant cells make for a statistically significant table, given the chosen significance level and, as the number of trials, the average number of different correspondences the two morpheme lists would have if the morphemes were paired at random.

Ringe's methodology gave believable answers; in particular, it is reassuring that it does not share an unfortunate tendency of many numerical methods in linguistics, which is to find that virtually every tested relationship is proven correct with almost infinite odds, such as Hymes's (1956) figure of 1,216,189,440,000 to 1. Ringe's believable results suggested that his approach to significance testing was, in most cases, a decent approximation to reality. On the other hand, testing a contingency table by doing significance tests on the results of running significance tests on dozens of parts of the table is complicated enough that it necessitates simplifying assumptions that can lead to wrong results in certain cases (Baxter 1998). If researchers do not fully comprehend the mathematical properties of a methodology, they may shy away from it, use it inappropriately, or place too little or too much credence in the results it outputs.

Kessler (2001) radically simplified the method by using Monte Carlo tests of significance. In this procedure, a computer literally does exactly what is called for by the definition of arbitrariness testing. After computing the gross score, the words are rearranged, so that the words are no longer necessarily paired by meaning, and a background score is computed for this rearranged data. The goal is to find what fraction of such rearrangements have a background score at least as high as the gross score; that fraction is the significance level. Ideally, one might wish to do this over all possible re-arrangements – a permutation test – but because that would take approximately forever, only a random sampling of all possible re-arrangements is used – a Monte Carlo test. The result can be made as precise as desired by increasing the number of random re-arrangements: typically numbers between a thousand and a million are used (Good 1994).

Monte Carlo tests make very few assumptions about the data; unlike χ^2 tests, for example, they do not give false positives when the amount of data is small. The researcher is also free to define virtually any type of test measure that seems linguistically appropriate, because there is no need to

find an existing statistical test that is mathematically appropriate for the measure. This flexibility leaves the researcher free to explore counting recurrences in creative ways. Kessler (2001), for example, tested several possibilities and decided it was best to square the recurrence counts before summing them, in order to capture the intuition that one sound correspondence that recurs many times is more probative than two sound correspondences that each recurs half as many times. Other variations in the procedure explore issues that are more linguistic than statistical. For example, the cited research has focussed on finding recurrences for the initial sound or the first consonant of the morpheme, presenting some analyses that confirm the observation that that position tends to be the most stable over time; but see Goh (2000) for a strong demurral.

ARBITRARINESS TESTING USING SIMILARITY JUDGEMENTS

Randomization tests are not limited to using recurrent sound correspondences as the test statistic. They can be applied just as well to the sort of similarity judgements used by Jones and Greenberg. They are even more important in such analyses, because unrelated languages with similar sound inventories can have high gross similarity scores, whereas a language that has undergone several sound changes may appear on the surface quite dissimilar from its relatives. Randomization tests compensate for typological influences, because they are included as part of the background score.

Oswalt (1970, 1991, 1998) developed the first fully algorithmic method of this type. Using a list of 100 basic concepts (Swadesh 1955), translation equivalents were found for the two languages under consideration. The test statistic was to count how many words match, where matching means that a predetermined number of consonants in the two words are similar to each other in a predetermined number of articulatory features. After measuring the gross score, Oswalt systematically shifted all the translations in one language by one row, to get a background score. Background scores were computed for all 99 shifts. This distribution of background scores was fitted to a normal curve in order to infer the statistical significance of the gross score. This procedure was made more straightforward and precise by Baxter, who converted it into a true Monte Carlo test (Baxter and Manaster Ramer 2000).

Quite a bit of exploration has gone into tweaking the parameters of this method. Oswalt (1998) explored possible relations between candidate Nostratic languages by running his tests five times, with different criteria for matches. This illustrates both the flexibility of this methodology and a concomitant problem: it is not always easy to interpret contradictory results from multiple partially dependent tests. In another variation of the test, Baxter and Manaster Ramer (2000) only considered word-initial consonants and used a different similarity statistic, recognizing a match when the two consonants are members of the same Dolgopolsky class – sets

of consonants that Dolgopolsky (1986) claimed descend particularly often from a common ancestral sound. Kessler (2007) compared this Dolgopolsky statistic to six other candidates and found it to be at least as reliable as the others. But incorporating information from other sounds in the word did not do much damage when results were verified against two secure families, Indo-European and Uralic. Kessler and Lehtonen (2006) further demonstrated the flexibility of the Monte Carlo approach by experimenting with several other innovations, including comparing multiple languages simultaneously and translating the same concept with multiple synonyms in the same language.

Current methodologies that use arbitrariness testing appear to be substantially correct from a statistical point of view, but not infrequently certain linguistic errors creep in when they are put to use. Most of these constitute failure to allow for exceptions to the arbitrariness doctrine, which may cause false positives:

- *Loanwords*. These are likely to be more similar within rows than between rows. All linguists understand this, but they rarely do anything about it when running statistical tests. For example, Villemin (1983) proved that Japanese and Korean were related, without considering the near certainty of there being loans in his morpheme table.
- *Onomatopoeia* and other words that tend to have similar forms across languages, such as 'mother'.
- *Morpheme repetitions*. If the translation of two different concepts includes the same morpheme in one language, even an unrelated language has an elevated chance of doing the same thing.

Apparently, researchers fear that removing such items would be a biased tinkering with the lists. But there would seem to be no point in running the tests at all when known loans appear in the list.

In addition to finding good, unbiased ways to deal with such exceptions to arbitrariness, many other open questions remain:

- *What is the optimal concept list?* Although quite a bit of work has been done on finding optimal concept lists for lexicostatistic work in general (Embleton 1986; McMahon and McMahon 2005), most has been within the context of subgrouping and glottochronology; relatively few studies have focused specifically on what list works best for uncovering long-range linguistic relationships. Ringe (1992), Oswalt (1998), and Kessler (2001) all reported that the Swadesh 100 concept list works at least as well as Swadesh's list of 200 words. Baxter and Manaster Ramer (2000) reported that their test found a connection between English and Hindi even when using a 33-concept list. These findings would seem to fly in the face of the statistical commonplace that more data makes for more accuracy, but apparently that statistical fact is counterbalanced by the linguistic fact that there are only a few concepts for which the

words tend to be stable across long stretches of time (Lohr 1999); adding in unstable concepts, whose fate after several millennia is likely to be indistinguishable from chance, just introduces more noise and error into the exercise.

- *Which is the best type of statistic?* Similarity measures and recurrent sound correspondences are both based on secure linguistic principles. Sounds that descend from the same sound in a parent language are likely to be especially similar, and they are also likely to show up in recurrent correspondences. Because it is not obvious which of those two properties would dominate at the time depths needed for exploring possible long-range relationships, the choice between the two may have to be decided on methodological rather than linguistic grounds. The trend in recent years has been towards working with ever shorter concept lists, which reduces considerably the opportunities for recurrences to emerge. In such a situation, similarity statistics may be the only viable option.
- *How much of the word should be used?* Research suggests that most power can be had by using just the first sound, or the first consonant; additional sounds are more likely to have been lost or to have changed in complicated ways that will appear to the algorithm as indistinguishable from chance, therefore weakening the power to recognize true relationships between languages. On the other hand, some language situations may be different, particularly languages in which prefixing or aphaeresis has played a role.
- *What if Saussure was wrong?* An unfavourable answer to this question could dismantle the entire arbitrariness testing enterprise. If words for the same concept across unrelated languages are even slightly more similar to each other than words for different concepts, then a powerful test might be able to pick that up and falsely conclude that unrelated languages appear related.

Assessment of the Assessment Methods

Current methodologies are generally well-founded from both the statistical and the linguistic point of view. But empirical validation is tricky. If linguists do not agree that languages are related, how can any answer be scored as right or wrong? Even for languages known to be related, it is not always obvious how damning an occasional false negative is. If knowledge of the status of difficult languages is based largely on information not available to the test, is the test to be faulted for its limited purview, or is to be excused for doing the best it can with limited information? To this needs to be added the consideration that tests only tell us how likely it is that chance is the cause of observed similarities between languages. If researchers wish to accept a cut-off such as 5% as meaning that chance is not involved, then, by definition, that interpretation is going to be wrong in one test out of 20, even when the test is doing exactly what it is designed to do.

With all those quibbles, it does appear that the assessments are believable. Researchers have been conscientious about running and publishing comprehensive validations. The usefulness of such validations is diminished only by the fact that the above-mentioned exceptions to arbitrariness are often not dealt with at the outset. This may have led to a bias towards testing positive for long-range relationships. In general, though, the tests appear to be good, if not perfect, at verifying relations within known families, such as Indo-European and Uralic. Such performance is fairly impressive, considering how divergent some of the languages are within those families.

Because the underlying theory and attested results of these assessment tools are both sound, further development is to be encouraged, in an effort to increase the power and reliability of the tests and to solve some of the open issues already addressed. But there may be some cultural impediments to wide adoption of these techniques. Some of these attitudes, along with possible rebuttals, are:

- *Believable results are boring*. So far, it appears that the power of these tests is similar to the power of the comparative method. If the tests tell what was already known, why run the tests? In reply, one may argue that independent verification of a theory is always good, and that well-validated tools may be valuable when debating controversial hypotheses or exploring new language sets.
- *Lexicostatistics is invalid*. In the past, linguists have confronted many methodologies that use explicit word tables and for which extravagant claims have been made. Multilateral comparison and glottochronology are two prominent examples. Of course, that does not mean that any linguistic technique that uses word or morpheme tables is wrong.
- *Sound similarity is an invalid statistic*. Any favouring of similarity over sound recurrence may come as a surprise, if not an affront, to practicing historical linguists, who recognize that the comparative method has in the past given consistently more reliable results than methods that just invoke sound similarity. But when both approaches are subjected to precise quantification and statistical significance testing, they become more equally credible. Besides, no one has asked linguists to abandon the comparative method. Finding sound correspondences is still an essential part of reconstructing the protolanguage and documenting its history.
- *New techniques do not consider all the data*. This is completely true. It is unlikely that in the foreseeable future, any fully algorithmic technique amenable to statistic significance testing will be able to operate over the full panoply of data that comparative linguists are able to reason with. This is why everyone involved proposes the new numerical methods only as a supplement to non-statistical techniques, and by no means as a replacement.

Short Biography

Brett Kessler's research concentrates on computational, statistical and experimental approaches to the study of language. Currently, his two main threads of research are mathematical approaches to comparative linguistics and, with his colleagues in psychology, experimentation in how readers and spellers cope with inconsistency in phoneme–grapheme mappings. Recent papers include 'Word similarity metrics and multilateral comparison' (ACL 2007) and 'Feedback consistency effects in single-word reading' (Psychology Press). Kessler holds a BA in Linguistics, an MS in Computer Science, and an MLS from Indiana University, and a PhD in Linguistics from Stanford. He has worked in human–computer interaction and natural language processing at Xerox and Hewlett Packard Labs and in reading research at Wayne State University. He is now at Washington University in St. Louis, where he teaches linguistics in the Psychology Department.

Note

* Correspondence address: Brett Kessler, Psychology Department, Washington University in St. Louis, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899, USA. E-mail: bkessler@wustl.edu.

Works Cited

Baxter, William H. 1998. Response to Oswalt and Ringe. Nostratic: sifting the evidence, ed. by Joseph C. Salmons and Brian D. Joseph, 217–36. Amsterdam, The Netherlands: Benjamins.

Baxter, William H., and Alexis Manaster Ramer. 2000. Beyond lumping and splitting: probabilistic issues in historical linguistics. Time depth in historicl linguistics, ed. by Colin Renfrew, April McMahon and Larry Trask, 167–88. Cambridge, UK: McDonald Institute for Archaeological Research.

Bender, Marvin L. 1969. Chance CVC correspondences in unrelated languages. Language 45.519–31.

Bickel, Balthasar, and Johanna Nichols. 2007. Autotyp <http://www.uni-leipzig.de/~autotyp>.

Blevins, Juliette. 2007. A long lost sister of Proto-Austronesian? Proto-Ongan, mother of Jarawa and Onge of the Andaman Islands. Oceanic Linguistics 46(1).154–98.

Campbell, Lyle. 2005. How to show languages are related. The handbook of historical linguistics, ed. by Brian D. Joseph and Richard D. Janda, 262–82. Malden, MA: Blackwell.

Cavalli Sforza, Luigi Luca. 2006. Genes and languages. Marges Linguistiques 11.261–80.

Collinder, Björn. 1947. La parenté linguistique et le calcul de probabilités. Uppsala, Sweden: Almqvist & Wiksells Boktruckeri.

Cowan, H. J. K. 1962. Statistical determination of linguistic relationship. Studia Linguistica 16.57–96.

Dolgopolsky, Aaron B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. Typology, relationship, and time, ed. by Vitalij V. Shevoroshkin and T. L. Markey, 27–50. Ann Arbor, MI: Karoma.

Embleton, Sheila M. 1986. Statistics in historical linguistics. Bochum, Germany: Brockmeyer.

Forster, Peter, and Colin Renfrew (eds.). 2006. Phylogenetic methods and the prehistory of languages. Cambridge, UK: McDonald Institute for Archaeological Research.

Georg, Stefan, and Alexander Vovin. 2003. Review of Greenberg 2000. Diachronica 20(2).331–62.

Goh, Gwang-Yoon. 2000. Probabilistic meaning of multiple matchings for language relationship. Journal of Quantitative Linguistics 7(1).53–64.
Good, Phillip. 1994. Permutation tests. New York, NY: Springer.
Greenberg, Joseph H. 1987. Language in the Americas. Stanford, CA: Stanford University Press.
——. 1993. Observations concerning Ringe's calculating the factor of chance in language comparison. Proceedings of the American Philosophical Society 137.79–89.
——. 2000. Indo-European and its closest relatives: the Eurasiatic language family. Volume 1, Grammar. Stanford, CA: Stanford University Press.
Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds) 2005. The world atlas of language structures. Oxford, UK: Oxford University Press.
Hrozný, Bedřich. 1917. Die Sprache der Hethiter: Ihr Bau und ihre Zugehörigkeit zum indogermanischen Sprachstamm. Leipzig, Germany: Hinrichs.
Hymes, Dell H. 1956. Na-Déné and positional analysis of categories. American Anthropologist 58.624–38.
Jones, William. 2007. The third anniversary discourse: On the Hindus. A reader in nineteenth century historical Indo-Europen linguistics, ed. by Winfred P. Lehmann, Chapter 1 <http://www.utexas.edu/cola/centers/lrc/books/read00.html>.
Kessler, Brett. 2001. The significance of word lists. Stanford, CA: CSLI Publications.
——. 2007. Word similarity metrics and multilateral comparison. Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, 6–14. Stroudsburg PA: Association for Computational Linguistics <http://www.aclweb.org/anthology-new/W/W07/W07-1302.pdf>.
Kessler, Brett, and Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. Phylogenetic methods and the prehistory of languages, ed. by Peter Forster and Colin Renfrew, 33–42. Cambridge, UK: McDonald Institute for Archaeological Research.
Lohr, Marisa. 1999. Methods for the genetic classification of languages, dissertation. Cambridge, UK: University of Cambridge.
McMahon, April, and Robert McMahon. 2005. Language classification by numbers. Oxford, UK: Oxford University Press.
Meillet, Antoine. 1925. La méthode comparative en linguistique historique. Oslo, Norway: Aschehoug.
Nichols, Johanna. 1996. The comparative method as heuristic. The comparative method reviewed: Regularity and irregularity in language change, ed. by Mark Durie and Malcolm Ross, 39–71. New York, NY: Oxford University Press.
Nova. 1997. In search of the first language <http://www.pbs.org/wgbh/nova/transcripts/2120glang.html>.
Oswalt, Robert L. 1970. The detection of remote linguistic relationships. Computer Studies 3.117–29.
——. 1991. A method for assessing distant linguistic relationships. Sprung from some common source, ed. by Sydney M. Lamb and E. Douglas Mitchell, 389–404. Stanford, CA: Stanford University Press.
——. 1998. A probabilistic evaluation of North Eurasiatic Nostratic. Nostratic: sifting the evidence, ed. by Joseph C. Salmons and Brian D. Joseph, 199–216. Amsterdam, The Netherlands: Benjamins.
Poser, William J., and Lyle Campbell. 1992. Indo-European practice and historical methodology. Berkeley Linguistics Society 18.214–36.
Ringe, Donald A., Jr. 1992. On calculating the factor of chance in language comparison. Transactions of the American Philosophical Society 82(1).1–110.
——. 1993. A reply to Professor Greenberg. Proceedings of the American Philosophical Society 127(1).91–109.
——. 1995. The 'Mana' languages and the three-language problem. Oceanic Linguistics 34.99–122.
——. 1998. Probabilistic evidence for Indo-Uralic. Nostratic: sifting the evidence, ed. by Joseph C. Salmons and Brian D. Joseph, 153–97. Amsterdam, The Netherlands: Benjamins.

——. 1999. How hard is it to match CVC-roots? Transactions of the Philological Society 97(2).213–44.
Ross, Alan S. C. 1950. Philological probability problems. Journal of the Royal Statistical Society, Series B (Methodological) 12.19–59.
Saussure, Ferdinand de. 1916. Cours de linguistique générale. Paris, France: Payot.
Swadesh, Morris. 1950. Salish internal relationships. International Journal of American Linguistics 16.157–67.
——. 1954. Perspectives and problems of Amerindian comparative linguistics. Word 10.306–32.
——. 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. Historical linguistics, ed. by B. Brainerd, 116–35. Bochum, Germany: Brockmeyer.
Wade, Nicholas. 2000. Joseph H. Greenberg: What we all spoke when the world was young (Scientist at Work). New York Times, 1 February 2000 <http://query.nytimes.com/gst/fullpage.html?res=9406E0D8163FF932A35751C0A9669C8B63 >.

