Book Reviews

The Significance of Word Lists

Brett Kessler
(Wayne State University)

Stanford, CA: CSLI Publications
(Dissertations in linguistics, edited by
Joan Bresnan, Sharon Inkelas, and Peter
Sells), 2001, x+277 pp; distributed by
the University of Chicago Press;
hardbound, ISBN 1-57586-299-9, $65.00,
£41.00; paperbound, ISBN
1-57586-300-6, $23.00, £14.50

Reviewed by
Grzegorz Kondrak
University of Toronto

The idea that statistical significance tests can be applied to the task of determining relatedness of languages is known to provoke rather emotional reactions. "Fallacious," "specious," "circular," "superficially plausible but in fact utterly unreliable," "exhibiting innumeracy to a fatal degree"—these are epithets actually used by researchers to describe one another's work in this area. What is generally agreed upon is that languages of common origin normally exhibit more common traits than unrelated languages. The main difficulty is finding an objective and reliable measure of the probability that the regularities are not due to chance.

A number of phenomena make such statistical comparisons difficult. First, lexical replacement steadily decreases the number of cognates shared by related languages. Second, with the passage of time, words can radically change their phonetic form and sometimes also their meaning. Finally, lexical transfer of words between unrelated languages that come into contact further obscures the true nature of their relationship. Nevertheless, a number of proposals have been put forward, the most prominent being the method of Ringe, introduced in his 1992 monograph and refined in his subsequent papers.

The new book by Kessler, based on his Ph.D. dissertation, is the most comprehensive work on the subject so far. It critically analyzes the previous approaches, points out a number of possible pitfalls of statistical testing, and proposes a novel solution to the problem. Although its main target audience is linguists interested in statistical argumentation, the book is also of computational interest. Kessler's approach crucially depends on computerized simulations, which are used in lieu of deduction in order to arrive at nontrivial results. Moreover, proper application of statistical reasoning is a topic of concern for many computational linguists. No deep background in historical linguistics is necessary for understanding the problems discussed in the book.

In the opening chapters, Kessler clearly introduces the problem by means of several illuminating examples. He explains in detail the important difference between measures based on phonetic similarities and those focused on sound recurrences, and his reasons for preferring the latter. The previously proposed methods are carefully analyzed, and their linguistic biases and mathematical flaws are pointed out. The

author then proposes his solution, which is based on permutation tests. In this technique, a measure of relatedness is computed for a pair of word lists representing two languages and then compared against the results calculated for a large number of randomly generated permutations of the lists. The probability of a historical connection between languages is estimated by counting the percentage of random orderings that produce a higher value of the relatedness measure.

In the second part of the book, the author discusses various phenomena that can compromise the validity of statistical testing. These include onomatopoeic words, borrowings, and language-internal cognates and regularities. Each analysis is concluded with an outline of practical procedures that should be taken to ensure the correctness of testing.

In the remainder of the book, the author proposes other measures of relatedness that can be used besides the χ^2 deviation statistic. These measures are designed to more closely correspond to intuitive techniques used by historical linguists. In order to evaluate the performance of the measures, Kessler employs a test suite of eight 200-word lists representing languages that exhibit various degrees of mutual relatedness: English, French, German, Latin, Albanian, Turkish, Hawaiian, and Navajo. (The complete data are included in the appendix together with some etymological information.) A measure is considered to perform well if it finds a statistically significant correlation between languages if and only if they are known to be related. The experiments conducted by the author indicate that his method is powerful enough to establish medium-range relationships such as French–English, which are unlikely to be demonstrable by traditional methods from the surface data alone.

Kessler wisely refrains from making excessive claims about the power of significance testing. He suggests that statistical assessment should complement rather than replace the venerable comparative method. It may be particularly useful as a preliminary test in cases where an in-depth analysis has not yet been applied. Therefore, I find quite disappointing the author's refusal to use his method to evaluate any of the controversial hypotheses of relatedness. After investing a considerable effort in developing and testing the measures, it would seem logical to try them on some of the hotly disputed cases, such as the Nostratic or Amerind families. Kessler prefers to leave that task to others, and provides a detailed description of steps to be followed "for those wishing to try this method for themselves." One can only hope that the comparative linguists will take up the suggestion.

In my opinion, one of the book's main accomplishments is in showing the limits of statistical techniques. Unlike the traditional comparative method, statistical testing is more about *sampling* rather than *analyzing* the data. A linguist attempting to perform statistical testing is advised to discard a lot of potentially valuable data. For example, increasing the size of word lists and considering more than a single phonotactic position are likely to weaken the power of a statistical test. Morphological information, which is considered particularly valuable by comparative linguists, is also better left out because its significance is difficult to evaluate statistically. Paradoxically, the more the experimenter knows about the tested languages, the more lexical items he is likely to disqualify, thus lowering the chances of obtaining a positive result.

Great care has to be exercised in interpreting the result of a statistical test. If the result is negative, it must not be considered as evidence against a genetic relationship (cf. Baxter and Manaster Ramer 2000). Kessler points out that it is not even theoretically possible to show that two languages are unrelated. A positive result, on the other hand, can neither demonstrate nor decide historical connection between languages; it can merely state that there is some statistically significant correlation between two word lists.

The lack of a single universally accepted measure of relatedness and the ease with which new measures can be constructed is somewhat disturbing. Since measures are typically validated by applying them to a handful of language pairs, it is not surprising that so many of them have been shown to be flawed (both by Kessler and by other authors). Kessler's favorite R^2 metric does not look impregnable either: it seems to me that in some cases accidental recurrences between frequent phonemes may easily overwhelm true recurrences between infrequent phonemes.

A close look at the results produced by various measures proposed in the book reveals that most of the tested language pairs that have no known cognates nevertheless exhibit statistically significant correlation according to *some* measure. This is not entirely surprising because for any fixed level of significance, a certain percentage of false positives is to be expected. Kessler exhorts everybody who does statistical testing to report all unsuccessful tests together with the successful ones. However, he also notes that it is common in the literature on long-range comparisons to present only the most striking results. Unfortunately, the reliability of a statistical test depends to a large degree on the experimenter's zeal to eliminate various kinds of subtle biases, even when that may lead to a result contrary to his linguistic intuition.

There is no simple procedure to verify the correctness of statistical testing. Due to the nondeterministic nature of the method, even rewriting the program and running it on the same data does not guarantee replicating the result. In general, Kessler is content with providing a single value reflecting the likelihood that the correlation between languages is statistically significant. Only rarely does he attempt to find out which pairs of words contribute a significant result. In my opinion, significance testing is only the first step toward showing the relatedness. After discovering some kind of regularity in the data, the next logical step is to follow the trail and analyze the regularities on an individual basis. Kessler makes one step in this direction when he takes a closer look at the results of his W metric and enumerates the cognates that are implied by the experiment. This seems to be the right way to make the method more transparent and verifiable.

Two important papers that are relevant to the subject are not mentioned in the book. Ringe (1998) "largely supersedes" Ringe's earlier work, which is discussed in great detail throughout the book. Baxter (1995) employs a method of multiple scrambling of word lists that is very similar to Kessler's approach.

Overall, the book exceeds in quality and detail most of what has previously been published on the subject. It is well written and coherent as a whole, and it covers all important aspects of the problem. Moreover, since none of the author's findings have been reported in article form, reading the book is the only way to get to know this interesting work.

References

Baxter, William H. 1995. 'A stronger affinity . . . than could have been produced by accident': A probabilistic comparison of Old Chinese and Tibeto-Burman. In W. S.-Y. Wang, editor, *The Ancestry of Chinese Language*. Project on Linguistic Analysis, Berkeley, pages 1–39.

Baxter, William H. and Alexis Manaster Ramer. 2000. Beyond lumping and splitting: Probabilistic issues in historical linguistics. In Colin Renfrew, April McMahon, and Larry Trask, editors, *Time Depth in Historical Linguistics*. The McDonald Institute for Archeological Research, Cambridge, UK, pages 167–188.

Ringe, Don. 1992. On calculating the factor of chance in language comparison, *Transactions of the American Philosophical Society*, 82(1).

Ringe, Don. 1998. Probabilistic evidence for Indo-Uralic. In Joseph C. Salmons and Brian D. Joseph, editors, *Nostratic: Sifting the Evidence*. John Benjamins, Amsterdam, pages 153–197.

Grzegorz Kondrak is interested in computational approaches to historical linguistics. He has published papers on the identification and alignment of cognates. Kondrak's address is: Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4; e-mail: kondrak@cs.toronto.edu; URL: http://www.cs.toronto.edu/∼kondrak.