The Mathematics of Glottochronology Revisited
Author(s): Annette J. Dobson, Joseph B. Kruskal, David Sankoff and Leonard J. Savage
Source: *Anthropological Linguistics*, Vol. 14, No. 6 (Sep., 1972), pp. 205-212
Published by: The Trustees of Indiana University on behalf of Anthropological Linguistics
Stable URL: http://www.jstor.org/stable/30029322
Accessed: 23-11-2015 20:26 UTC

THE MATHEMATICS OF GLOTTOCHRONOLOGY REVISITED

Annette J. Dobson
James Cook University
of North Queensland

Joseph B. Kruskal
Bell Telephone Laboratories

David Sankoff
Université de Montréal

†Leonard J. Savage
Yale University

0. The widespread and well-established use of lexicostatistics by field linguists and anthropologists as an aid in subgrouping languages within a language family, as an adjunct to comparative analysis, and in assessing time depths of linguistic divergence, is attested to by the sheer volume of recent publications, especially on Oceanic, African, American, and Australian languages.[1] On the theoretical level, however, the usefulness and validity of lexicostatistical techniques are relatively less appreciated, for two types of reasons. First, it is frequently claimed that lexicostatistics and glottochronology don't work when compared with historical and linguistic facts. It is not our intention to review the empirical evidence for and against these methods; suffice it to say that in subgrouping (without time estimates), lexicostatistics produces classifications which are generally quite compatible with those produced by comparative analysis, and the cases in which glottochronology produces poor time estimates[2] are balanced by cases in which the results are very realistic.[3] No statistically meaningful comparison of large numbers of historical and glottochronological dates has been carried out since the early efforts of Lees (1953).

The second type of criticism is that the theory behind lexicostatistics and glottochronology is wrong. It is a simple theory; nobody disputes the fact that it ignores the causes and mechanisms of lexical change. This does not mean it is wrong, however; it does not try to account for, predict, or even describe individual lexical changes. It only claims that these complex and multi-faceted processes result in a certain regularity at the statistical or aggregate level.[4] A more serious criticism frequently leveled at glottochronology is that its mathematical basis is faulty. For ten years now, students of linguistics have been referred to Chrétien's The Mathematical Models of Glottochronology

(1962) as an invalidation of the theoretical foundations of lexicostatistics. It is widely read, frequently cited, and has been reprinted in a popular series of linguistics readings, though it is known by many mathematicians and statisticians to be mathematically unsound.

At a recent conference on lexicostatistics at Yale, there was discussion of the article and its impact on the use of statistics in historical linguistics. This note, signed by four mathematicians interested in lexicostatistics, is the outcome of that discussion.

Chrétien's calculations and arguments suffer from three major errors. We emphasize that these are not mere details of mathematical rigor, but are serious conceptual and logical errors which go to the very heart of his criticisms, and completely invalidate them. We discuss each of these errors in turn.

1. The fundamental formula, the keystone of his whole discussion, is Chrétien's formula (5):

$$K = Nr^t,$$

where N = the number of basic meanings in the test-list being used,
t = the number of millennia which have elapsed,
K = the number of words for the basic meanings which have been retained over this period,
r = a constant between 0 and 1 which must be determined empirically for any given list of basic meanings, and which measures the rate at which words for list meanings are replaced by non-cognates.

While this statement was used and intended only as an approximation in the fundamental papers of Swadesh (1952, 1955), Lees (1953), and Gudschinsky (1956), Chrétien uses it as a precise statement of the relationship between K and t. The correct statement of the precise relationship, which was implicit in the early papers, is[5]

$$\text{expected value of } K = Nr^t.$$

Unfortunately, the early literature in lexicostatistics never explicitly states this precise equation.

The early papers contain many mathematical statements which are not expressed as precisely as a mathematician would express them, even though their substance was correct. Among these are equations essentially the same as Chrétien's formula (5). Thus Chrétien's misinterpretation is understandable.

Nevertheless, Chrétien did misinterpret when he assumed a precise relationship between t and K. In each of the early papers taken as a whole, it is abundantly clear to any mathematician or statistician that the equation $K = Nr^t$ is meant to be understood as correct only in a statistical sense, that is, the equation only holds on the average. The papers all explicitly discuss the statistical variation which will inevitably occur, thus indicating that a precise relationship only exists between t and the expected value of K.

A continuing source of misunderstanding concerning the relationship between t and K may result from the superficial similarity between radioactive decay and lexical replacement. Equations essentially the same as $K = Nr^t$ are fre-

quently found in discussions of radioactive decay, even though experts in that field all know that this equation is merely an approximation. The reason this approximation is so commonly used for radioactive decay is that the value of N is typically enormous there: $N = 10^9$ = one billion would not be unusual in many applications to radioactivity. The larger the value of N, the better the approximation between K/N and 'expected value of K'/N.

We note in passing that the equation

$$\text{expected value of } K = Nr^t$$

is based on the simplifying assumption that r is the same for all meanings in the test-list. In some recent papers, this assumption has been replaced by more general and accurate assumptions, so that this equation gives way to a more refined and complicated successor.

By itself, Chrétien's first error would not be too serious, but combined with his other arguments it causes real trouble.

2. Chrétien's Tables IV and V show probabilities for various values of K and of t, given that C has been observed and its value known, where:

t = the number of millennia since two cognate languages separated,
$K = Nr^t$,
C = the number of meanings (from among the given list of basic meanings) for which both languages have retained the word used at the time of separation,

and where N and r (as defined above) are taken to be 100 and 0.86 respectively. Tables IV and V are based on the calculations underlying Tables I, II, and III. Due to the first error described above, the earlier tables are incorrect. However, because their foundation is correct in a rough and ready sense (that is, on the average), they also have a certain rough and ready validity. The values they contain are approximately correct if properly interpreted.

However, in going on to Tables IV and V, Chrétien makes a further error which completely invalidates his calculations. The latter tables are drastically in error.

As Dyen (1972) points out, the reason for this error is that Chrétien, in making his calculations (see page 18 and page 26), assumes that each of his 'joint sets' is equally likely. In the calculation on page 18, the joint sets being compared all involve the same value of K, so this assumption is correct in this context. In the calculation on page 26, joint sets for different values of K are involved in a single calculation, and the assumption is thoroughly incorrect.[6]

3. Chrétien points out that for a given value of K, lexicostatistical theory permits a whole range of possible values of t which could have produced it. He finds this objectionable. Similarly he objects that for a given value of either t or C, there is a whole range of the other variable which corresponds to it. He suggests that this is a fundamental weakness in the glottochronological method. This suggestion is in error.

In the following discussion, N is of course fixed, and we assume that r has already been determined. Suppose t is given, and we ask how many words (corresponding to the list of basic meanings) will be retained over the period.

This number K is not fixed precisely: rather it has a probability distribution. Many values of K are possible (specifically, any K for which $N \geqslant K \geqslant 0$), but only a limited range of values of K would have probabilities large enough to be plausible. In fact, the most likely values of K (that is, the values of K with highest probabilities) will be integers which are close to Nr^t. (Notice that Nr^t is generally not an integer, and hence K cannot be exactly equal to Nr^t in most situations.) As we consider integers more and more distant from Nr^t, the probabilities get smaller and smaller. In fact, the probabilities follow a well-known formula called the binomial distribution.[7]

Likewise, suppose we observe K and wish to deduce t. The observed value of K permits a range of plausible values of t, though any $t > 0$ is possible. Since the equation $K = Nr^t$ is not justified, neither is its rearrangement $t = \log(K/N)/\log r$. It is true, however, that the

maximum likelihood estimate[8] of $t = \log(K/N)/\log r$

and the most plausible values of t will lie near this estimate. As we move to values of t away from this region, the plausibility drops off. The details are more complicated in this case than in the previous paragraph, because they involve not only probability theory but also questions of statistical inference.

The relationship[9] between C and t precisely parallels the relationship between K and t. If we are given t, then the lexicostatistical model does not assert that $C = Nr^{2t}$, but it does assert that

expected value of $C = Nr^{2t}$.

The most likely values of C are integers close by Nr^{2t} (itself seldom an integer), and more distant integers are increasingly unlikely. However, every integer C with $N \geqslant C \geqslant 0$ is possible. The distribution of C also satisfies a binomial formula. Conversely, given C, it is not true that $t = \log(C/N)/2\log(r)$, which is a simple rearrangement of $C = Nr^{2t}$. However, the lexicostatistical model does assert that the maximum likelihood estimate of t is $\log(C/N)/2\log(r)$. The most plausible values of t are close to this value. Plausibility drops off as we move away from this region. However any $t > 0$ is possible.

To sum up, Chrétien's third error consists of repeatedly criticizing glottochronology for a phenomenon that is unavoidable anywhere in science. This phenomenon is that estimates of key quantities — however arrived at — are subject to random fluctuation.

4. It is understandable that linguists are not primarily interested in probability and statistics as tools in the analysis of their data. Nonetheless, it seems to us that this does not justify the continued acceptance of Chrétien's mathematically incorrect and misleading treatment of glottochronology as the definitive article on the subject of lexicostatistics. It is deplorable that such a paper is in large part responsible for the retarded development of this field until very recently.

Finally, we would stress that lexicostatistics is not a theory about the details of lexical change in particular languages, nor about causal and systematic relationships. Rather, it is a theory about aggregates of lexical changes, usually in an aggregate of languages. As such it is not falsifiable by selected

examples drawn from one language or another. It is falsifiable only in the statistical sense, by aggregates of data. Even when a particular lexicostatistical model is shown to be untenable in some respect, it may be more appropriate to improve and modify the model than to abandon it completely. In fact, one of the important contributions of Dyen, James and Cole (1967) was to demonstrate statistically that the parameter r differs substantially among the meanings of the Swadesh list, and to show how improved assumptions reflecting this fact could be incorporated into lexicostatistical theory. This has led to the establishment of a firmer mathematical and statistical basis for quantitative historical linguistics,[10] which will in turn, we hope, encourage a new interest in the methods of collection, analysis, and interpretation of lexicostatistical data.

WORKS CITED

Bender, M. L. (1971). The Languages of Ethiopia: A New Lexicostatistical Classification and some Problems of Diffusion, AL 13.165-288.

Bergsland, Knut and Hans Vogt (1962). On the Validity of Glottochronology, Current Anthropology 3.115-53.

Blount, Ben, and Richard T. Curley (1970). The Southern Luo Languages: A Glottochronological Reconstruction, Journal of African Languages 9.1-18.

Brainerd, Barron (1970). A Stochastic Process Related to Language Change, Journal of Applied Probability 7.69-78.

Chowning, Ann (1969). The Austronesian Languages of New Britain, Pacific Linguistics A 21.17-43.

Chrétien, C. Douglas (1962). The Mathematical Models of Glottochronology, Lg 38.11-37. Reprinted as Bobbs-Merrill Language and Linguistics Series 16.

Clark, David J. (1971). Three "Kua" Languages of Eastern Nigeria, Journal of West African Linguistics 1.27-36.

Dobson, Annette J. (1969). Lexicostatistical Grouping, AL 11.216-21.

Dobson, Annette J. (1970). Interval Estimates of Language Divergence Times, Journal of Applied Probability 7.781-4.

Dobson, Annette J. (1972a). Estimating Time Separations for Languages. Proceedings of the Yale Conference on Lexicostatistics, to be published, Mouton & Co.

Dobson, Annette J. (1972b). Estimating Language Divergence Times, Journal of the Royal Statistical Society, to appear.

Dutton, T. E. (1969). The Peopling of Central Papua, Pacific Linguistics B 9.

Dyen, Isidore (1972). The Impact of Lexicostatistics on Comparative Linguistics. Proceedings of the Yale Conference on Lexicostatistics, to be published, Mouton & Co.

Dyen, Isidore, A. T. James and J. W. L. Cole (1967). Language Divergence and Estimated Word Retention Rate, Lg 43.150-71.

Elmendorf, William W. (1969). Geographic Ordering, Subgrouping, and Olympic Salish, IJAL 35.220-5.

Freund, John E. (1962). Mathematical Statistics. Prentice-Hall.

Grimes, James L. (1971). A Reclassification of the Quichean and Kekchian (Mayan) Languages, IJAL 37.15-9.

Gudschinsky, Sarah C. (1956). The ABC's of Lexicostatistics, Word 12.175-210.

Gursky, Karl-Heinz (1969). A Lexical Comparison of the Atakapa, Chitimacha, and Tunica Languages, IJAL 35.83-107.

Hooley, Bruce A. (1970). Mapos Buang, Territory of New Guinea. Ph. D. dissertation, University of Pennsylvania.

Kruskal, Joseph B., Isidore Dyen, and Paul Black (1971). The Vocabulary Method of Reconstructing Family Trees: Innovations and Large-Scale Applications. Mathematics in the Archaeological and Historical Sciences. (edited by Hodson, Kendall and Tautu) Edinburgh University Press. pp. 360-80.

Kruskal, Joseph B., Isidore Dyen, and Paul Black (1972). Some Results from the Vocabulary Method of Reconstructing Family Trees. Proceedings of the Yale Conference on Lexicostatistics, to be published, Mouton & Co.

Lees, R. B. (1953). The Basis of Glottochronology, Lg 29.113-27.

McElhanon, K. A. (1970). Lexicostatistics and the Classification of Huon Peninsula Languages, Oceania 40.214-31.

McElhanon, K. A. (1971). Classifying New Guinea Languages, Anthropos 66. 120-44.

McNamara, Leo F. (1961). Morpheme Retention in Irish, AL 3:9.23-30.

Miller, Wick R., James L. Tanner, and Lawrence P. Foley (1971). A Lexicostatistic Study of Shoshoni Dialects, AL 13.142-64.

O'Grady, Geoffrey N. and Terry J. Klokeid (1969). Australian Linguistic Classification: A Plea for Coordination of Effort, Oceania 39.298-311.

Oswalt, Robert L. (1971). Towards the Construction of a Standard Lexicostatistic List, AL 13.421-34.

Sankoff, David (1970). On the Rate of Replacement of Word-Meaning Relationships, Lg 46. 564-69.

Sankoff, David (1972). Mathematical Developments in Lexicostatistic Theory. Current Trends in Linguistics 11 (edited by T. Sebeok) to appear.

Satterthwait, Arnold C. (1960). Rate of Morphemic Decay in Meccan Arabic, IJAL 26. 256-61.

Spitzbardt, Harry (1969). Kritische Bemerkungen zur Glottochronologie, Ethnographische-Archaologische Zeitschrift 10. 41-50.

Swadesh, Morris (1952). Lexicostatistic Dating of Prehistoric Ethnic Contacts, APS-P 96. 452-63.

Swadesh, Morris (1955). Towards Greater Accuracy in Lexicostatistic Dating, IJAL 21. 121-37.

Troike, Rudolf C. (1969). The Glottochronology of Six Turkic Languages, IJAL 35. 183-91.

Vérin, Pierre, Conrad P. Kottak and Peter Gorlin (1969). The Glottochronology of Malagasy Speech Communities, Oceanic Linguistics 8. 26-83.

Webb, Nancy M. (1971). A Statement of some Phonological Correspondences among the Pomo Languages. IJAL, Memoir 26.

Wilson, D. B. (1969). The Binandere Language Family, Pacific Linguistics A 18. 65-86.

Zakar, Andras (1971). Sumerian-Ural-Altaic Affinities, Current Anthropology 12. 215-26.

Z'graggen, J. A. (1969). Classification and Typological Studies in Languages of the Western Madang District, New Guinea. Ph. D. dissertation, Australian National University.

NOTES

1. Since 1969, this includes, for Oceania, Chowning (1969), Dutton (1969), Hooley (1970), O'Grady and Klokeid (1969), Vérin et al (1969), Wilson (1969), Z'graggen (1969) and McElhanon (1970, 1971) — the latter reviews the work in New Guinea, including the contributions of S. A. Wurm; for Africa, Bender (1971), Blount and Curley (1970) and Clark (1971); for American, Elmendorf (1969), Grimes (1971), Gursky (1969), Miller et al (1971) and Webb (1971). For other areas, see Troike (1969), Zakar (1971), Oswalt (1971) and Spitzbardt (1969).

2. See, for example, Bergsland and Vogt (1962) or Spitzbardt (1969).

3. For example, Satterthwait (1960), McNamara (1961) and Troike (1969).

4. The particular statistical model used to describe this regularity might well be inappropriate, in which case the theory should be criticized and improved. This is precisely what has been happening in recent lexicostatistical theory, resulting in increasingly realistic models and sophisticated techniques (cf. Note 10 below).

5. Lexicostatistics is a probabilistic theory, specifying only the probability of each possible value of K. The 'expected value of K' is the mean, or center of gravity, of this probability distribution. See Freund (1962: 90-1).

6. Suppose the common source of languages 1 and 2 is language 0. Let A_1 be the set of meanings with cognate words in language 0 and 1; and similarly with A_2. Let A_{12} be the intersection of A_1 and A_2. Chrétien calls the pair (A_1, A_2) a <u>joint set</u>. On p. 18, he calculates the probability that there are C meanings in A_{12}, given that there are K in each of A_1 and A_2. He assumes that all joint sets where A_1 and A_2 each contain K meanings, are equally likely. This is perfectly correct in the context of the Swadesh theory. On p. 26, Chrétien wishes to calculate the probability that both A_1 and A_2 contain K meanings, given that A_{12} contains C (and, implicitly, that both A_1 and A_2 contain the same number). He assumes that all joint sets in which A_1 and A_2 have equal numbers of meanings and C meanings in common (i. e. C meanings in A_{12}) are equally likely. These joint sets, however, involve many different values of K; and as K varies, the probability of joint sets varies greatly.

7. The probability that K words will be retained for given values of N, r, and t is $\binom{N}{K} r^{tK}(1 - r^t)^{N-K}$ for $0 \leqq K \leqq N$, where $\binom{N}{K} = N!/[K!(N-K)!]$ is the binomial coefficient.

8. The maximum likelihood estimate of t is the value of the parameter t which, inserted into the formula of note 7 above, maximizes the probability of the observed value of K. See Freund (1962: 223).

9. Chrétien's discussion of this relationship is confused by his misunderstanding of the nature of statistical independence. He postulates that the selection of words discarded in one language is statistically independent of the selection in a cognate language. This is a well-understood and frequently invoked assumption. Its mathematical formulation is in terms of the product law of probability. On p. 13 Chrétien explains statistical independence in nonmathematical terms which nonetheless clearly imply the product law. He actually uses the product law on p. 14 because it is a consequence of independence, but on pp. 15 and 31 he implies that the product law is an assumption about probabilities which is not part of the statistical independence postulate.

10. For example, see the papers of Brainerd (1970), Dobson (1969, 1970, 1972a, 1972b), Kruskal, Dyen and Black (1971, 1972) and Sankoff (1970, 1972).

† Professor Savage died November 1, 1971. In agreeing to be a co-signer of this note, he had suggested several revisions, almost all of which were incorporated.