Lexicostatistical Tree Reconstruction Incorporating Borrowing*

Sheila M. Embleton

0. Abstract

A persistent concern of lexicostatistics has been to eliminate all meanings susceptible to borrowing from the test-list used in family-tree reconstruction for a group of related languages. However, it is not clear that any meanings exist which are completely immune to borrowing, especially in situations of large-scale language contact. There are also considerable statistical advantages in the use of a longer test-list. Thus lexicostatistical tree reconstruction methods can be improved by incorporating borrowing rates as well as replacement rates in the reconstruction process. The system of differential equations necessary for this reconstruction (where the replacement rate \( r \) and the borrowing rate \( b \) are known) will be presented, together with a computer simulation of the divergence of a language family in which both lexical replacement and borrowing are involved. The results of comparisons of the reconstructed trees with the computer-generated trees for different lengths of test-list and for different values of \( r \) and of \( b \) show that a high degree of accuracy, both in the topology of the tree and in its relative branch lengths, can be obtained, even for values of \( r \) and \( b \) considerably higher than those found in natural language.

1. Introduction

Swadesh (1950, 1952) introduced a model of language change based on a vocabulary turnover process analogous to radioactive decay. It is claimed that a test-list of \( N \) meanings can be constructed which are likely to be found in all cultures. Such a list is now generally known as a Swadesh-list, and contains meanings referring to such supposed cultural universals as body parts, numerals, topographical terms, certain kinship terms, and simple activities. The construction of this list in itself is problematic (Hoijer 1956), as we can see from the progressive reduction in length of the Swadesh-list over the years and from the slight variation in the actual list used (compare, for example, Swadesh 1955 with Gudschinsky 1956).

In applying Swadesh's glottochronological technique to the reconstruction of the family tree for a group of related languages, we take a Swadesh-list (generally \( 100 < N < 215 \)), and find a single gloss for the meaning in each of the languages in the group. Over time, morpheme decay occurs. That is, in each language, some morphemes representing meanings on the list are replaced by other morphemes, for a variety of reasons (e.g. taboo, adoption
of forms from a superstratum or substratum, simple borrowing, gradual semantic shifts). Assume that this rate of morpheme decay, \( R \), is constant over time. Consider any one language, evolving from time \( t_0 \). Let \( m_1, m_2, \ldots, m_N \) be the set of meanings, \( w_i(t) \) be the word at time \( t \) which represents \( m_i \), and \( n(t) \) be the number of \( w_i(t) \) still cognate to \( w_i(t_0) \). Then the rate of change of \( n(t) \) over time can be expressed as

\[
\frac{dn(t)}{dt} = Rn(t) \, ,
\]

with the boundary condition

\[
n(t_0) = N \, ,
\]

since all \( N \) meanings are cognate to \( w_i(t_0) \) at \( t_0 \).

From Equation (1), by separation of variables,

\[
\frac{dn(t)}{n(t)} - R dt = 0 \, .
\]

Integrating,

\[
\ln|n(t)| - Rt = c. 
\]

Since \( n(t) \geq 0 \), \( \ln|n(t)| = \ln n(t) \).

Thus,

\[
n(t) = e^{c}e^{Rt} \, .
\]

From the boundary condition (Equation (2)),

\[
e^{c} = Ne^{-Rt} \, .
\]

Thus

\[
n(t) = Ne^{-R(t-t_0)} \, .
\]

We are interested in determining the time depth, which is the value of \( t - t_0 \). From Equation (3),

\[
t - t_0 = \frac{1}{R} \ln \frac{n(t)}{N} \, .
\]

Suppose we now consider two languages, both evolving independently after \( t_0 \). The number of shared items at time \( t \) is given by the product \( Ne^{-R(t-t_0)} \cdot e^{-R(t-t_0)} \), \( e^{-R(t-t_0)} \), or \( Ne^{-2R(t-t_0)} \), giving a time depth of \( \frac{1}{2R} \ln \frac{n(t)}{N} \).

The use of this method for family tree reconstruction presupposes that we already have a value for the rate-constant \( R \). This constant can be determined by using the same equations to solve for \( R \) where the time depth is known (e.g. most Indo-European
languages, Chinese). Lees (1953:118-119) uses this process to determine that R is approximately 19% per 1000 years for the 200-word Swadesh-list 'for all languages, at all times'. This is obviously a very sweeping statement, and failure to understand that it represents no more than a simplifying assumption for the model has led to much abuse.

Many studies were undertaken on the basis of this simplified model (Kroeber 1955, Arndt 1959, Taylor and House 1955, Gudschinsky 1955, etc.). Expectedly, the results sometimes diverged considerably from known historical fact. Unfortunately, the fact that Swadesh's model is stochastic was only implicit in the original articles (Swadesh 1950, 1952; Lees 1953). A mathematician's reaction would be to realize that 'any degree of deviation is predicted to occur in a certain proportion of cases by [a] stochastic model' (Sankoff 1973) and to try to revise and improve the model to minimize the deviation. Linguists treated the model as a deterministic one, and used empirical deviations as evidence against it (Bergsland and Vogt 1962). The controversy over glottochronology generated a huge amount of literature. Swadesh (1955) reduced the size of the list in an attempt to make it less culturally-oriented, but Hoijer (1956) and Levin (1964) still had problems with it. Teeter (1963:642) concluded pessimistically: 'The closer we get to universal validity, the fewer items we have. My own opinion is that there will be no items at all on the perfect list.' To avoid problems of cognate determination, Householder (1964:326) suggested replacing cognateship by an arbitrarily-set limit (e.g. 75%) of phonetic similarity. Chrétien (1962) purported to disprove the mathematics of the model. This article was widely read, accepted, and cited among linguists as an invalidation of the theoretical foundations of glottochronology. Chrétien (1962) was also supported by further such articles (Lunt 1964; Fodor 1965; Chrétien 1966). Unfortunately, much of this literature simply rehashes the same old arguments time and time again. As Hymes (in van der Merwe 1966:492) succinctly states:'It is striking how critics of glottochronology
continue to discover the same criticisms, but not the constructive attempts to deal with them.' These criticisms were known by mathematicians and statisticians to be unsound, but these scholars were hampered by a lack of linguistic knowledge, and were therefore understandably reluctant to become involved in what seemed to be a very controversial field in linguistics. In 1971, the First Conference on Genetic Lexicostatistics was held at Yale University, a gathering involving statisticians as well as linguists and anthropologists, with a view to clearing glottochronology of its bad name. Chrétien 1962 and its impact on the use of statistics in historical linguistics were discussed extensively. This discussion was then embodied in Dobson, Kruskal, Sankoff, and Savage 1972, in which Chrétien's major conceptual and logical errors were pointed out one by one and in which it was shown how these errors completely invalidated his criticisms. A second such conference, Lexicostatistics in Genetic Linguistics II, was held at the Centre de Recherches Mathématiques at the Université de Montréal in 1973. The continuing involvement of statisticians in co-operation with linguists augurs well for the future of glottochronology, as many of the numerous pitfalls which have so plagued the earlier investigations can now be avoided, as both mathematical and linguistic expertise will be readily available.

2. Improvements to the Basic Model

Swadesh's basic model was a very simple one, and various refinements have been made subsequently. It was discovered that there was non-homogeneity in the test-list (i.e. R is different for different meanings). Van der Merwe (1966) deals with this problem by splitting the test-list into classes, where each class has its own replacement rate, thus improving the original Swadesh model. This modification had earlier been suggested in an abstract by Joos (1964), but he never actually incorporated it into a model. Dyen (1964) showed that not only does the mean replacement rate depend on the meaning involved, but the dependence is similar in different language families. This culminated in an important contribution
by Dyen, James, and Cole (1967), in which each meaning is associated with its own mean replacement rate. They showed how to use analysis of variance techniques to simultaneously estimate these rates and the divergence times for the family. Simultaneous estimation of divergence times and mean replacement rates was further researched in Kruskal, Dyen, and Black (1971).

The next major improvement (Brainerd 1970) generalized the Swadesh model to allow for *recurrent cognition* (where the word elicited at the end of the time interval is cognate to the word for the same meaning at the beginning of the interval, but there have been one or more replacements involving non-cognation in the interval) and *chance cognition* (words may coincidentally be phonologically similar). Brainerd also allows the rate parameter \( R \) and the probability that a replacement word is cognate to the original word for a meaning to have different values in different languages. In addition Sankoff (1973:100) points out that 'in Brainerd's theory any type of word feature may be used as a criterion of word relationship, for example whether or not both have Latinate stems.'

Another improvement deals with the non-independence of evolution of related languages. After separation of two languages, there is an initial period of 'inertia', where a degree of parallelism between the changes in one language and those in the other results. If this drift effect is significant, it can also be introduced into the model (see Gleason 1959). If there is additional contact between the two languages, Sankoff (1973:100) suggests introducing a borrowing rate parameter, but notes that 'the mathematics become considerably more complicated.'

The final improvement concerns Swadesh's assumptions that we can elicit exactly one word for each test-list meaning and that the replacement process for any given word must be instantaneous. Brainerd's version of the model would also easily have generalized to allow multiple synonyms for any meaning, although he did not actually incorporate this. Sankoff (1970,1971,1973) allows synonyms for each test-list meaning. There is a constant
probability per unit time that a new word may be introduced (initially at a low frequency) for a meaning. During each small interval of time, the usage frequency for a word has equal probabilities of increasing or decreasing by a small increment (modelled by a random-walk process). Occasionally a word drops to zero frequency, thus becoming permanently obsolete. To compare a set of synonyms at one time or in one language with a partially different set at another time or in another language, we can use either of two strategies:

Strategy (1): Always choose only the most frequently used synonym (Sankoff does not deal with a 'tie' situation; presumably one of the synonyms is chosen arbitrarily).

Strategy (2): Compute a metric distance between the frequency distributions of the two sets.

The important outcomes are that on the average Strategy (1) works as well as Strategy (2), and that the results are not very dependent on the exact nature assumed for the random walk (only the parameter for the rate of introduction of new words is important).

A convenient summary of the improvements to Swadesh's basic model is provided by Sankoff's *fully parametrized lexicostatistics* (1973:102-103). The parameters are:

- $\alpha, \beta$ parameters of the $\Gamma$-distribution of replacement rates (for details, see Sankoff 1973:98)
- $\gamma$ the probability of chance recurrent cognition (Brainerd)
- $\kappa$ the drift time constant (Gleason)
- $\theta$ the borrowing probability

For times greater than $\kappa$, the expected value for the proportion of cognates, $C/N$, is

$$[1 + \gamma \theta - \gamma - \theta][1 + 2\beta(t - \kappa)]^{-\alpha} + \gamma + \theta - \gamma \theta$$

and the maximum likelihood estimator of $t$ is

$$\left(\frac{1 - \gamma - \theta + \gamma \theta}{C/N - \gamma - \theta + \gamma \theta}\right)^{1/\alpha} - 1 + \kappa.$$

Examination of limiting cases as parameters approach zero reveals the earlier simple cases and, as $t$ becomes indefinitely large,
'the lexicostatistic relationship approaches an equilibrium value \( \gamma + \theta - \gamma\theta \), dependent only on borrowing and chance cognation. Thus the genetic aspects of linguistic relationship gradually become less important than the diffusion aspects.' (Sankoff 1973:103).

An important and perhaps surprising result is that these more complex models behave remarkably similarly to the simpler ones, partially justifying some of the earlier (and often implicit) assumptions, which should no longer be challenged as being unrealistic. Sankoff (1973:104) also shows that longer test-lists (e.g. \( N \geq 1000 \)) give results comparable to the shorter lists. This has the advantage of minimizing the effect of a few bad cognate judgments and of semantic or cultural problems. We can also have a different test-list for different language families, provided we keep the same list throughout each entire experiment, enabling us to enjoy the statistical benefits of the longer test-lists.

A persistent concern in the construction of a test-list has been the elimination or at least the reduction in number of meanings susceptible to borrowing. It is by no means clear that any meanings exist which are completely immune from borrowing, especially if we try to apply lexicostatistic techniques to situations of large-scale language contact (at the extreme, processes such as pidginization and creolization). These facts, coupled with the statistical advantages of a longer list, show that lexicostatistics must find a way to incorporate borrowing rates, despite the 'considerably more complicated' mathematics which will become necessary. It is to this end that the rest of this paper is directed.

3. Incorporating Borrowing Rates with N Indefinitely Large

Sankoff (1972), working within a biological framework, develops a model of genetic divergence of populations which could be used to reconstruct the history of an evolutionary tree from information derived from the currently existing populations alone. He assumes a fixed set of genetic sites (corresponding to the linguist's test-list), sufficiently large 'that we can neglect statistical fluctuation in the dynamic models' (i.e. \( N \) must approach \( \infty \)), and ignores 'the
small proportion of gene sites for which there may be different types within a single population' (i.e. synonymy and dialect divergence within a language) (Sankoff 1972:597). In my exposition of Sankoff's model, I replace the biological terms with their linguistic equivalents and omit Sankoff's complete mathematical proofs for each proposition.

Let $X_t$ denote a language at time $t$ and $S_{X_t Y_u}$ the similarity between languages $X$ and $Y$ at times $t$ and $u$ respectively, measured by the proportion of cognates. $S_{X_t Y_u} = S_{Y_u X_t} > 0$, and $S_{X_t X_t} = 1$. Assuming that each meaning has constant probability $r$ per unit time of having its word replaced with a new word, and that recurrent cognation does not occur, then, for $X_t$ ancestral to $X_u$,

$$\frac{dS_{X_t X_u}}{du} = -rS_{X_t X_u}$$

With the initial condition

$$S_{X_t X_t} = 1,$$

we have the solution

$$S_{X_t X_u} = e^{-r(u-t)}.$$  

For two independently evolving languages, we have

$$S_{X_t Y_u} = e^{-r(v-t)}e^{-r(v-u)},$$

where $v$ is the time at which $X$ and $Y$ become independent from each other. To reconstruct the evolutionary tree from present-day data alone (i.e. when $u = t = 0$), we merely solve

$$S_{XY} = e^{-2rv},$$

for each $(X,Y)$ pair.

The time subscripts have been suppressed for present-day languages without danger of ambiguity.

Sankoff then proceeds to the more complicated (and linguistically more realistic) case where languages can influence one another after their separation. This involves considering the geography as well as the history of the tree. Whenever a language splits into two, Sankoff (1972:600) splits the country corresponding to that language as follows. Two distinct borders of the country are chosen, and the new border is drawn to connect the midpoints of these two borders. If the country has an exterior border, two points
on this border may be chosen as endpoints for the new border. The original split of one country into two must also be accomplished in this latter fashion.

There are various ways of modelling the borrowing process, but 'mathematically speaking, these all lead to the same type of problem' (Sankoff 1972:601). Sankoff chooses to allow borrowing between neighbouring languages only. As before the replacement rate is r, and there is also a borrowing rate \( b/k_x \) of borrowings into \( X \) from each of its \( k_x \) neighbours.

Suppose \( X \) is a neighbour of \( Y \) (write \( X \in N_Y \)). \( S_{XY} \) changes due to three factors: change due to replacement in \( X \) and in \( Y \), change due to borrowings between \( X \) and \( Y \), and change due to borrowings into \( X \) and into \( Y \) from their other neighbours. Change due to replacement in \( X \) and in \( Y \) is \(-2rS_{XY}\). Total borrowings between \( X \) and \( Y \) are \( b \left( \frac{1}{k_x} + \frac{1}{k_y} \right) \), but a proportion \( S_{XY} \) of the words borrowed are already cognate in the two languages. Thus change due to borrowings between \( X \) and \( Y \) is

\[
(1 - S_{XY}) b \left( \frac{1}{k_x} + \frac{1}{k_y} \right).
\]

The effect of borrowings into \( X \) from \( Z \), where \( Z \) is any of the \( k_x - 1 \) remaining neighbours of \( X \), is more complicated and merits a detailed derivation here (Sankoff's derivation (1972:601) is in error). Borrowings from \( Z \) into \( X \) increase \( S_{XY} \) if \( X \) and \( Y \) had non-cognate words but \( X \) borrows a word from \( Z \) where \( Y \) and \( Z \) have cognate words.

Borrowings from \( Z \) into \( X \) decrease \( S_{XY} \) if \( X \) and \( Y \) had cognate words but \( X \) borrows a word from \( Z \) where \( Y \) and \( Z \) have non-cognate words. Thus borrowings from \( Z \) into \( X \) change \( S_{XY} \) at the rate

\[
(10) \frac{b}{k_x} \left[ (1 - S_{XY}) S_{YZ} - S_{XY} (1 - S_{YZ}) \right] = \frac{b}{k_x} \left[ S_{YZ} - S_{XY} \right]
\]

and change due to borrowings into \( X \) from all its neighbours other than \( Y \) is

\[
(11) \frac{b}{k_x} \sum_{Z \in N_X \setminus Y} \left[ S_{YZ} - S_{XY} \right].
\]

Similarly, borrowings into \( Y \) from all its neighbours other than \( X \) is

\[
(12) \frac{b}{k_y} \sum_{Z \in N_Y \setminus X} \left[ S_{XZ} - S_{XY} \right].
\]

Collecting terms,
\[
\frac{dS_{XY}}{dt} = -2rS_{XY} + (1-S_{XY})b\left(1 + \frac{1}{k_X}\right) + b \sum_{Z} e_{N_X} \left[S_{YZ} - S_{XY}\right]
+ b \sum_{Y} e_{N_Y} \left[S_{XZ} - S_{XY}\right].
\]

Since \(S_{XX} = S_{YY} = 1\), Equation (13) may be further simplified to

\[
\frac{dS_{XY}}{dt} = -2rS_{XY} + b \frac{\sum_{Z} e_{N_X}}{k_X} \left(S_{YZ} - S_{XY}\right) + b \frac{\sum_{Y} e_{N_Y}}{k_Y} \left(S_{XZ} - S_{XY}\right).
\]

The above derivation follows Sankoff's method, omitting his errors, and resulting in a much simpler expression for \(dS_{XY}/dt\). Equation (14) can however be derived more directly. There are essentially five cases to consider when X borrows from a neighbour Z when determining the change in \(S_{XY}\).

<table>
<thead>
<tr>
<th>Case</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>per unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>word 1</td>
<td>word 1</td>
<td>word 1</td>
<td>α</td>
</tr>
<tr>
<td>2</td>
<td>word 1</td>
<td>word 1</td>
<td>word 2</td>
<td>β</td>
</tr>
<tr>
<td>3</td>
<td>word 1</td>
<td>word 2</td>
<td>word 1</td>
<td>γ</td>
</tr>
<tr>
<td>4</td>
<td>word 1</td>
<td>word 2</td>
<td>word 2</td>
<td>δ</td>
</tr>
<tr>
<td>5</td>
<td>word 1</td>
<td>word 2</td>
<td>word 3</td>
<td>ε</td>
</tr>
</tbody>
</table>

Only Cases 2 and 4 result in a change in \(S_{XY}\). Thus the rate of change due to borrowings from Z into X is given by

\[
\frac{dS_{XY}}{dt} = \frac{b}{k_X} (+ \delta - \beta).
\]

Now \(\alpha + \beta = S_{XY}\)

\(\alpha + \delta = S_{YZ}\)

\(\delta - \beta = S_{YZ} - S_{XY}\).

Borrowings from Z into Y are analogous. Thus the total rate of change of \(S_{XY}\) is given by Equation (14).

To reconstruct the evolutionary tree, we use the system of equations implied by Equation (14) together with the present-day similarities as initial conditions. We trace the similarities back through time, until one of them equals 1. The two populations involved are amalgamated by deleting the border between their two countries. The process is repeated until only one country remains. Note also that this method of reconstruction not only yields the topology of the evolutionary tree, but also the times of the splits. For proof of the uniqueness of the solution, see Sankoff (1972:602-603).
Sankoff's solution obviously applies to an idealized case. In linguistic application, practical problems arise. N is not indefinitely large, and hence statistical fluctuation cannot be ignored. The constants r and b are unknown and must be determined. These constants may also change between populations and between different meanings on the test-list. The next section addresses the first of these practical problems.

4. Incorporating Borrowing Rates with N not Indefinitely Large

To test Sankoff's solution (i.e. the corrected version given by Equation (14)) to the reconstruction problem in a linguistically more real case (N<500), it was decided to construct a computer simulation of the language change process. In order to be of manageable size but still large enough to be an interesting problem, the simulation is terminated when 10 present-day countries exist. Reconstruction of the tree is then attempted from the present-day similarities and geography alone, using the same values of r and b as in the generation of the data. By comparing the generated tree with the reconstructed tree, we can determine the accuracy of the reconstruction. Only an outline of the computer simulation will be given here; the author is happy to enter into private correspondence on any details.

The simulation begins with one country which splits at time 0. After a random time interval (exponential with parameter 2), one of these splits between any two points on its borders. This is more general than Sankoff's process, where two distinct borders must be chosen. After another random time interval (exponential with parameter 3), one of these splits, and so on, until we have 10 countries. Meanwhile, linguistic divergence along the branches is simulated. Each meaning in the parent at the time of the first split is represented by a word labelled 1. After a random time interval, exponential with parameter nN(r+b), where n is the number of countries in existence, a change event
occurs. To decide whether this change is to be a borrowing or a replacement, choose an x in the interval (0,1). If 
\[ x < r/(r+b) \], a replacement occurs; if \[ x > r/(r+b) \], a borrowing occurs. For a replacement event, we choose one of the n languages at random, then choose one of the N meanings at random, and replace the word for that meaning by the next highest available label for that meaning. For a borrowing event, we choose one of the n languages at random to be the borrowing language, then choose one of its neighbours at random to be the 'lending' language, and then copy the word for that meaning. These processes continue until immediately before the eleventh split is due to occur, when we terminate the data generation and construct a 10 x 10 similarity matrix. The values of r and b, the generated tree, the times of split, and the similarity matrix are all printed out. All the cognate data and the final geographical configuration can each be printed out as an option to the program.

Reconstruction is then attempted, using the same values for r and b as in the data generation, and using Equation (14) with the 10 x 10 similarity matrix as the initial conditions. Each of the similarities is traced back through time, using a Runge-Kutta fourth order approximation to integrate S with respect to t, until one of these similarities equals 1. The two countries are then amalgamated, by deleting the border between them and coalescing the two appropriate rows and columns in the similarity matrix (the two values which must coalesce are simply averaged). The procedure is continued until only one country remains. The reconstructed tree and the reconstructed times of split are printed out. Occasionally, we find that we must coalesce two countries which in actual fact do not have a common border. This represents no complication for the algorithm, but a message that the coalescing countries do not border is printed. As an option to the program, countries can be prevented from coalescing if they do not share a border. In this case, we continue integrating S with
respect to t until the similarity becomes 1 for a pair of countries which do have a common border.

The next step is to compare the generated and reconstructed trees. The most obvious way is simply to count how many of the reconstructed trees agree perfectly with the corresponding generated tree. In a similar experiment, but with a less general splitting process and with reconstruction by the method of simple hierarchical cluster analysis (which cannot incorporate b into the reconstruction), Sankoff and Dobson (1971) used this measure. For a 200-word list, they found a maximum of 55% topologically correct trees when b=0, and this percentage rapidly approached 0% as b increased. My own results (with 20 simulation runs for each (r,b) pair) show a maximum of 65% topologically correct trees when b=0, with this percentage decreasing to an average of approximately 30% when b=30 (see Table 1). However, this simple measure is also simplistic; it fails to distinguish cases in which trees differ in some minor respect from trees with major topological differences. In other words, it cannot distinguish a mistake in reconstruction involving a node low in the tree from a mistake involving a node high in the tree; both mistakes are considered to be equally drastic.

A major improvement is the introduction of a coefficient which can distinguish the degree of error in the reconstruction. For this purpose, I have adapted a measure first proposed by Sokal and Rohlf (1962) in order to compare dendrograms in numerical taxonomy research in biology. The adaptation is slight; thus, I retain their name cophenetic correlation coefficient (CCC) for this measure. Assume there are n present-day countries. For the reconstructed tree and for the generated tree, we must calculate a derived similarity matrix. This is an upper triangular matrix minus the diagonal (i.e. a triangular matrix with \( \binom{n}{2} \) entries), where the entry in the (X,Y) cell is the actual time of split between countries X and Y. In Sokal and Rohlf's calculation, the actual times of split themselves are not used; instead the time axis is divided into a number of class intervals, each interval is assigned an ordinal value, and these interval values are entered into the derived similarity matrix.
The CCC is now computed simply by calculating the ordinary (Pearson) product-moment correlation coefficient between corresponding elements of the two derived similarity matrices to be compared. In our particular case then, the CCC will measure the degree of accord between the relative branch lengths and the topologies of the reconstructed and the generated trees. Note that relative branch lengths and not absolute branch lengths are relevant; this is because the correlation between aU+b and cV+d is the same as the correlation between U and V, where a and c are any positive constants and b and d are any constants. For N=200 and n=10, the CCC ranges from 0.99 for b=0 to 0.91 for b=30, with the values decreasing roughly exponentially in between (see Table 2). For a correlation coefficient, these are remarkably high values, indicating that even in cases where the reconstructed tree is not perfectly correct, the topological errors are relatively minor.

Even the CCC is partially dependent on the accuracy of our reconstruction of relative branch lengths. In order to evaluate only the topological accuracy of our reconstructions, we must find yet another measure, which must have two basic properties: it should have the value 1 if and only if the reconstruction is topologically perfect and its value should decrease in proportion to the severity of the topological error of the reconstruction. Such a measure (call it \textit{topological similarity coefficient (TSC)}) can be constructed in the following way. For each of the n countries, work backwards in time up the tree, labelling each node encountered 1, 2, 3, \ldots until the top of the tree is reached. Then, for each node, retain only the maximum label assigned to it. Next, as for the CCC, calculate a derived similarity matrix, where the entry in the (X,Y) cell is the label assigned to the lowest node dominating both X and Y (i.e. the node assigned to the most recent common ancestor of X and Y). The TSC is calculated as the product-moment correlation coefficient between corresponding elements of the derived similarity matrices. Note that the TSC is totally independent
of the time-scale and satisfies the two basic properties required above. Two sample calculations are included as Table 4. For N=200 and n=10, the TSC ranges from 0.965 to 0.85, again decreasing as b increases (see Table 3). As for the CCC, the TSC values are high for a correlation coefficient, reflecting a high degree of accuracy in the topology of our reconstructions. The greater range of values for the TSC as compared to the CCC is expected because the TSC is designed to be particularly sensitive to the severity of the topological error encountered.

An examination of Tables 1-3 shows that the reconstruction method discussed in this paper can produce highly accurate lexicostatistical tree reconstructions for families of 10 languages using a 200-word list, for values of r ranging from 15 to 30 and for values of b ranging from 0 to 30. Lees (1953) determined that r was approximately 19% per 1000 years for the 200-word Swadesh-list; values for b are, except in extreme circumstances, much lower. Thus our reconstruction method shows a high degree of accuracy for values of r and b both within and beyond the range of values found from actual historical language data. The method is also robust in that the topology and relative branch lengths of the reconstructed tree depend only on the ratio of the values of r and b, and not on their absolute values.

In order to investigate the effect of the length of the test-list on the accuracy of the reconstruction, the simulation experiment was repeated with N=500 (see Tables 5-7) and with N=100 (see Tables 8-10). Since r is approximately 14% per 1000 years for the 100-word Swadesh-list (Swadesh 1955:127), the range of values for r has been extended to include r=10 and r=14 when N=100. Comparing the results for N=200 with those for N=500, it can be seen that accuracy, whether measured by percentage topologically correct, by CCC, or by TSC, is slightly improved for N=500 over N=200. The improvement is sufficiently small, however, that, when faced with the practical problems of real language data, it is doubtful
whether the researcher will find it worth the time and
trouble to work with a 500-word list in preference to a
200-word list. On the contrary, comparison of results for
N=200 with those for N=100 shows that accuracy, by all
three measures, is considerably decreased by using a
100-word list. Hence the conscientious researcher must prefer
a 200-word list over a 100-word list. Note also that,
since we are now able to take account of borrowings in the
reconstruction process, we are no longer constrained to
use a Swadesh-list, specially chosen for its resistance to
borrowing; we may use any word list, subject only to the
practical constraint of the availability of our chosen
meanings in the language family with which we are working.

It was pointed out earlier, in the description of the
reconstruction algorithm, that sometimes we should coalesce
two countries which do not share a border, but that as an
option to the computer program we can prevent these countries
from coalescing and instead continue the algorithm. This
option can be used to increase further the accuracy of the
reconstructions. If we are about to coalesce two countries
which do not border, we are obviously about to make a
mistake, as we will be reconstructing a situation which
could not possibly have occurred in the generation process.
This particular mistake can thus be prevented. The percentage
topologically correct can be increased up to 15%, and the
CCC and the TSC can be increased up to 3% (typically,
slightly less than 1%).

5. Working with Real Language Data

The theoretical results discussed above all have the same
value of \( r \) for each language and the same value of \( b \) throughout
the language family. This represents a simplification which
should be abandoned when working with real language data.
The replacement rate \( r_i \) should be allowed to vary for each
language \( L_i \), and the borrowing rate \( b_{ij} \) should be allowed
to vary for each language pair \( (L_i, L_j) \). Note that \( b_{ij} \) need
not necessarily equal \( b_{ji} \); language history offers many
instances of non-equal borrowing rates between a pair of languages. Neither of these refinements constitutes a problem for our reconstruction algorithm. A matrix of values is simply substituted for the single constant.

Reconstruction for a real language family of n languages involves entering the cognateship data (for an N-word list) as an N x n matrix, entering a neighbour matrix (an n x n matrix with a 1 in cells representing a pair of countries which border, 0 otherwise), and entering values for $r_i$ and $b_{ij}$. The same algorithm is used as described in Section 4 above.

The author is currently testing the reconstruction algorithm for the Germanic family, using 14 present-day Germanic languages in the actual reconstruction and 9 historically-attested Germanic languages to aid in the estimation of replacement and borrowing rates. Preliminary results indicate that the reconstruction algorithm is successful in reconstructing a tree in accord with general linguistic views of the history of the Germanic language family.

* I would like to thank Barron Brainerd for comments on an earlier draft of this paper, and J.S.A. Hepburn for help with the mathematical formulae and the computer programming.
Table 1  Percentage Topologically Correct Trees (N=200)

<table>
<thead>
<tr>
<th>r</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>65</td>
<td>40</td>
<td>50</td>
<td>30</td>
<td>25</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>25</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>50</td>
<td>40</td>
<td>45</td>
<td>30</td>
<td>30</td>
<td>25</td>
<td>10</td>
<td>15</td>
<td>5</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>45</td>
<td>40</td>
<td>55</td>
<td>60</td>
<td>50</td>
<td>50</td>
<td>30</td>
<td>35</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>25</td>
<td>55</td>
<td>50</td>
<td>65</td>
<td>40</td>
<td>50</td>
<td>40</td>
<td>55</td>
<td>60</td>
<td>30</td>
<td>20</td>
<td>45</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>50</td>
<td>75</td>
<td>65</td>
<td>35</td>
<td>30</td>
<td>55</td>
<td>25</td>
<td>20</td>
<td>35</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 2  CCC (N=200)

<table>
<thead>
<tr>
<th>r</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>.9895</td>
<td>.9844</td>
<td>.9904</td>
<td>.9697</td>
<td>.9679</td>
<td>.9517</td>
<td>.9567</td>
<td>.9414</td>
<td>.9500</td>
<td>.9529</td>
<td>.9456</td>
<td>.9290</td>
</tr>
<tr>
<td>19</td>
<td>.9918</td>
<td>.9806</td>
<td>.9861</td>
<td>.9777</td>
<td>.9746</td>
<td>.9653</td>
<td>.9716</td>
<td>.9523</td>
<td>.9440</td>
<td>.9095</td>
<td>.9489</td>
<td>.9110</td>
</tr>
<tr>
<td>20</td>
<td>.9883</td>
<td>.9808</td>
<td>.9878</td>
<td>.9752</td>
<td>.9767</td>
<td>.9757</td>
<td>.9669</td>
<td>.9646</td>
<td>.9554</td>
<td>.9457</td>
<td>.9229</td>
<td>.9084</td>
</tr>
<tr>
<td>25</td>
<td>.9858</td>
<td>.9935</td>
<td>.9936</td>
<td>.9811</td>
<td>.9793</td>
<td>.9760</td>
<td>.9702</td>
<td>.9562</td>
<td>.9385</td>
<td>.9439</td>
<td>.9455</td>
<td>.9354</td>
</tr>
<tr>
<td>30</td>
<td>.9913</td>
<td>.9905</td>
<td>.9923</td>
<td>.9893</td>
<td>.9815</td>
<td>.9743</td>
<td>.9715</td>
<td>.9531</td>
<td>.9673</td>
<td>.9579</td>
<td>.9453</td>
<td>.9298</td>
</tr>
</tbody>
</table>

Table 3  TSC (N=200)

<table>
<thead>
<tr>
<th>r</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>.9656</td>
<td>.9463</td>
<td>.9598</td>
<td>.8915</td>
<td>.8308</td>
<td>.9129</td>
<td>.9499</td>
<td>.9207</td>
<td>.8684</td>
<td>.8822</td>
<td>.9007</td>
<td>.9161</td>
</tr>
<tr>
<td>19</td>
<td>.9669</td>
<td>.9252</td>
<td>.9166</td>
<td>.9119</td>
<td>.9092</td>
<td>.8842</td>
<td>.9025</td>
<td>.8637</td>
<td>.8895</td>
<td>.8494</td>
<td>.8606</td>
<td>.8710</td>
</tr>
<tr>
<td>20</td>
<td>.9632</td>
<td>.9247</td>
<td>.9604</td>
<td>.9426</td>
<td>.9434</td>
<td>.9167</td>
<td>.8955</td>
<td>.8996</td>
<td>.8731</td>
<td>.8493</td>
<td>.8492</td>
<td>.8799</td>
</tr>
<tr>
<td>25</td>
<td>.9314</td>
<td>.9687</td>
<td>.9328</td>
<td>.9377</td>
<td>.9309</td>
<td>.9520</td>
<td>.9387</td>
<td>.9529</td>
<td>.8632</td>
<td>.8950</td>
<td>.8874</td>
<td>.8864</td>
</tr>
<tr>
<td>30</td>
<td>.9558</td>
<td>.9440</td>
<td>.9764</td>
<td>.9585</td>
<td>.9151</td>
<td>.9096</td>
<td>.9341</td>
<td>.9021</td>
<td>.8635</td>
<td>.9008</td>
<td>.8514</td>
<td>.8966</td>
</tr>
</tbody>
</table>
Table 4 Sample Calculations of TSC
A. for two topologically similar trees.

**Generated Tree**

```
L_1 L_9 L_8 L_4 L_7 L_5 L_6 L_{10} L_2 L_3
```

**Reconstructed Tree**

```
L_1 L_9 L_8 L_4 L_7 L_5 L_6 L_{10} L_2 L_3
```

**Trees with Maximum Node Labels**

**Generated Tree**

```
  5
 / \
 4   1
 /   /\
3   1 2
 /   /   /\
1 1 1 1
```

**Reconstructed Tree**

```
  5
 / \
 4   1
 /   /\
3   1 2
 /   /   /\
1 1 1 1
```

**Derived Similarity Matrices**

**Generated Tree**

<table>
<thead>
<tr>
<th></th>
<th>L_2</th>
<th>L_3</th>
<th>L_4</th>
<th>L_5</th>
<th>L_6</th>
<th>L_7</th>
<th>L_8</th>
<th>L_9</th>
<th>L_10</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_1</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>L_2</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>L_3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>L_4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_9</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Reconstructed Tree**

<table>
<thead>
<tr>
<th></th>
<th>L_2</th>
<th>L_3</th>
<th>L_4</th>
<th>L_5</th>
<th>L_6</th>
<th>L_7</th>
<th>L_8</th>
<th>L_9</th>
<th>L_10</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_1</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>L_2</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>L_3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>L_4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L_9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TSC = 0.9861
B. for two trees less topologically similar than in Table 4A.

Trees with Maximum Node Labels

Derived Similarity Matrices

<table>
<thead>
<tr>
<th>Generated Tree</th>
<th>Reconstructed Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_2</td>
<td>L_3</td>
</tr>
<tr>
<td>L_1</td>
<td>5</td>
</tr>
<tr>
<td>L_2</td>
<td>5</td>
</tr>
<tr>
<td>L_3</td>
<td>4</td>
</tr>
<tr>
<td>L_4</td>
<td>2</td>
</tr>
<tr>
<td>L_5</td>
<td>5</td>
</tr>
<tr>
<td>L_6</td>
<td>5</td>
</tr>
<tr>
<td>L_7</td>
<td>4</td>
</tr>
<tr>
<td>L_8</td>
<td>5</td>
</tr>
<tr>
<td>L_9</td>
<td>2</td>
</tr>
</tbody>
</table>

TSC = 0.9306
Table 5 Percentage Topologically Correct Trees (N=500)

<table>
<thead>
<tr>
<th>r/b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>55</td>
<td>40</td>
<td>40</td>
<td>45</td>
<td>40</td>
<td>50</td>
<td>30</td>
<td>60</td>
<td>40</td>
<td>30</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>19</td>
<td>70</td>
<td>85</td>
<td>75</td>
<td>60</td>
<td>55</td>
<td>40</td>
<td>35</td>
<td>50</td>
<td>55</td>
<td>35</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>65</td>
<td>70</td>
<td>35</td>
<td>45</td>
<td>60</td>
<td>25</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>25</td>
<td>65</td>
<td>70</td>
<td>70</td>
<td>75</td>
<td>55</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>65</td>
<td>30</td>
<td>55</td>
<td>35</td>
</tr>
<tr>
<td>30</td>
<td>65</td>
<td>25</td>
<td>75</td>
<td>80</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td>50</td>
<td>45</td>
<td>60</td>
<td>25</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 6 CCC (N=500)

<table>
<thead>
<tr>
<th>r/b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>.9942</td>
<td>.9958</td>
<td>.9899</td>
<td>.9851</td>
<td>.9818</td>
<td>.9797</td>
<td>.9692</td>
<td>.9687</td>
<td>.9467</td>
<td>.9541</td>
<td>.9410</td>
<td>.9369</td>
</tr>
<tr>
<td>19</td>
<td>.9955</td>
<td>.9967</td>
<td>.9956</td>
<td>.9857</td>
<td>.9808</td>
<td>.9756</td>
<td>.9672</td>
<td>.9583</td>
<td>.9669</td>
<td>.9542</td>
<td>.9368</td>
<td>.9146</td>
</tr>
<tr>
<td>20</td>
<td>.9983</td>
<td>.9926</td>
<td>.9911</td>
<td>.9917</td>
<td>.9841</td>
<td>.9805</td>
<td>.9695</td>
<td>.9630</td>
<td>.9763</td>
<td>.9494</td>
<td>.9303</td>
<td>.9320</td>
</tr>
<tr>
<td>25</td>
<td>.9964</td>
<td>.9962</td>
<td>.9952</td>
<td>.9863</td>
<td>.9871</td>
<td>.9793</td>
<td>.9779</td>
<td>.9634</td>
<td>.9674</td>
<td>.9645</td>
<td>.9513</td>
<td>.9449</td>
</tr>
<tr>
<td>30</td>
<td>.9965</td>
<td>.9938</td>
<td>.9946</td>
<td>.9860</td>
<td>.9916</td>
<td>.9790</td>
<td>.9869</td>
<td>.9710</td>
<td>.9489</td>
<td>.9642</td>
<td>.9180</td>
<td>.9514</td>
</tr>
</tbody>
</table>

Table 7 TSC (N=500)

<table>
<thead>
<tr>
<th>r/b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>.9538</td>
<td>.9467</td>
<td>.9323</td>
<td>.9326</td>
<td>.9528</td>
<td>.9591</td>
<td>.8921</td>
<td>.9636</td>
<td>.9090</td>
<td>.8708</td>
<td>.8998</td>
<td>.8448</td>
</tr>
<tr>
<td>19</td>
<td>.9721</td>
<td>.9923</td>
<td>.9909</td>
<td>.9487</td>
<td>.9600</td>
<td>.9219</td>
<td>.8979</td>
<td>.9014</td>
<td>.8950</td>
<td>.9106</td>
<td>.9187</td>
<td>.8635</td>
</tr>
<tr>
<td>20</td>
<td>.9819</td>
<td>.9646</td>
<td>.9387</td>
<td>.9413</td>
<td>.9852</td>
<td>.9614</td>
<td>.8910</td>
<td>.8911</td>
<td>.9511</td>
<td>.8856</td>
<td>.8684</td>
<td>.8817</td>
</tr>
<tr>
<td>25</td>
<td>.9816</td>
<td>.9795</td>
<td>.9766</td>
<td>.9682</td>
<td>.9493</td>
<td>.9320</td>
<td>.9604</td>
<td>.9516</td>
<td>.9650</td>
<td>.8865</td>
<td>.9356</td>
<td>.8892</td>
</tr>
<tr>
<td>30</td>
<td>.9696</td>
<td>.9337</td>
<td>.9770</td>
<td>.9759</td>
<td>.9303</td>
<td>.9568</td>
<td>.9690</td>
<td>.9353</td>
<td>.8968</td>
<td>.9401</td>
<td>.8593</td>
<td>.9075</td>
</tr>
</tbody>
</table>
### Table 8  Percentage Topologically Correct Trees (N=100)

<table>
<thead>
<tr>
<th>r</th>
<th>b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td>30</td>
<td>15</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>30</td>
<td>30</td>
<td>45</td>
<td>35</td>
<td>60</td>
<td>30</td>
<td>25</td>
<td>40</td>
<td>25</td>
<td>30</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>35</td>
<td>10</td>
<td>30</td>
<td>40</td>
<td>15</td>
<td>25</td>
<td>20</td>
<td>35</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>40</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>30</td>
<td>10</td>
<td>30</td>
<td>45</td>
<td>30</td>
<td>25</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>40</td>
<td>30</td>
<td>30</td>
<td>25</td>
<td>40</td>
<td>20</td>
<td>30</td>
<td>30</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>45</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>30</td>
<td>35</td>
<td>30</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>35</td>
<td>45</td>
<td>60</td>
<td>40</td>
<td>35</td>
<td>10</td>
<td>35</td>
<td>10</td>
<td>30</td>
<td>15</td>
<td>25</td>
<td>20</td>
</tr>
</tbody>
</table>

### Table 9  CCC (N=100)

<table>
<thead>
<tr>
<th>r</th>
<th>b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td>.9620</td>
<td>.9539</td>
<td>.9028</td>
<td>.9381</td>
<td>.9180</td>
<td>.9458</td>
<td>.8912</td>
<td>.9417</td>
<td>.8648</td>
<td>.8615</td>
<td>.8825</td>
<td>.8359</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>.9746</td>
<td>.9669</td>
<td>.9849</td>
<td>.9605</td>
<td>.9727</td>
<td>.9512</td>
<td>.9084</td>
<td>.9433</td>
<td>.9291</td>
<td>.9270</td>
<td>.8744</td>
<td>.8812</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>.9745</td>
<td>.9618</td>
<td>.9715</td>
<td>.9613</td>
<td>.9565</td>
<td>.9543</td>
<td>.9396</td>
<td>.9474</td>
<td>.9389</td>
<td>.9496</td>
<td>.8971</td>
<td>.8919</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>.9784</td>
<td>.9636</td>
<td>.9796</td>
<td>.9754</td>
<td>.9710</td>
<td>.9642</td>
<td>.9669</td>
<td>.9516</td>
<td>.9266</td>
<td>.9104</td>
<td>.9174</td>
<td>.8737</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>.9868</td>
<td>.9812</td>
<td>.9583</td>
<td>.9716</td>
<td>.9664</td>
<td>.9567</td>
<td>.9606</td>
<td>.9347</td>
<td>.9080</td>
<td>.9042</td>
<td>.9113</td>
<td>.9274</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>.9759</td>
<td>.9854</td>
<td>.9694</td>
<td>.9838</td>
<td>.9680</td>
<td>.9621</td>
<td>.9556</td>
<td>.9276</td>
<td>.9286</td>
<td>.9439</td>
<td>.9301</td>
<td>.9142</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>.9795</td>
<td>.9804</td>
<td>.9762</td>
<td>.9769</td>
<td>.9482</td>
<td>.9554</td>
<td>.9669</td>
<td>.9326</td>
<td>.9282</td>
<td>.9329</td>
<td>.9086</td>
<td>.9454</td>
</tr>
</tbody>
</table>

### Table 10  TSC (N=100)

<table>
<thead>
<tr>
<th>r</th>
<th>b</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td>.8839</td>
<td>.8970</td>
<td>.8819</td>
<td>.8641</td>
<td>.8613</td>
<td>.8738</td>
<td>.8539</td>
<td>.8753</td>
<td>.8097</td>
<td>.8243</td>
<td>.8721</td>
<td>.8173</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>.9007</td>
<td>.9144</td>
<td>.9649</td>
<td>.9162</td>
<td>.9763</td>
<td>.8908</td>
<td>.8313</td>
<td>.8888</td>
<td>.8736</td>
<td>.8553</td>
<td>.8018</td>
<td>.8703</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>.9128</td>
<td>.9240</td>
<td>.9363</td>
<td>.9304</td>
<td>.9128</td>
<td>.8876</td>
<td>.8485</td>
<td>.9157</td>
<td>.8873</td>
<td>.8949</td>
<td>.8362</td>
<td>.8003</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>.9090</td>
<td>.8792</td>
<td>.9284</td>
<td>.9246</td>
<td>.9333</td>
<td>.8621</td>
<td>.8993</td>
<td>.9124</td>
<td>.8805</td>
<td>.8616</td>
<td>.8835</td>
<td>.8101</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>.9494</td>
<td>.9441</td>
<td>.9048</td>
<td>.8987</td>
<td>.8671</td>
<td>.8372</td>
<td>.8661</td>
<td>.8653</td>
<td>.8388</td>
<td>.8297</td>
<td>.8582</td>
<td>.8600</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>.9065</td>
<td>.9419</td>
<td>.9328</td>
<td>.9329</td>
<td>.9138</td>
<td>.9085</td>
<td>.9033</td>
<td>.8566</td>
<td>.8934</td>
<td>.8867</td>
<td>.8453</td>
<td>.8894</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>.9096</td>
<td>.9295</td>
<td>.9349</td>
<td>.9170</td>
<td>.8847</td>
<td>.8593</td>
<td>.8692</td>
<td>.8518</td>
<td>.8725</td>
<td>.8606</td>
<td>.8737</td>
<td>.8909</td>
</tr>
</tbody>
</table>
References


Householder, Fred. 1964. a comment in Current Anthropology 5:361.

Joos, Martin. 1964. Glottochronology with Retention-Rate Inhomogeneity.


