

On detecting borrowing:
Distance-based and character-based approaches

James W. Minett and William S.-Y. Wang[†]

Language Engineering Laboratory,
Department of Electronic Engineering,
City University of Hong Kong,
Kowloon, Hong Kong, P.R.C.

[PRE-PRINT OF AN ARTICLE TO APPEAR IN *DIACHRONICA* 20.2]

Email: minett@ee.cityu.edu.hk (Minett),

wsyw@ee.cityu.edu.hk (Wang)

Tel: (+852) 2788 7187

Fax: (+852) 2788 7791

[†] We would like to thank the three anonymous reviewers for their helpful comments on an earlier draft of this paper. This work has been supported in part by grants from the City University of Hong Kong (9010001) and the Research Grants Council of Hong Kong (CityU 1224/02H).

1. Introduction

The comparative method is the most widely used method for deriving the classification of a family of genetically related languages. It prescribes a procedure for reconstructing the proto-language of a language family by determining regular correspondences among putative cognate sets, usually of lexical items. Innovations shared by groups of languages are then established, by which a classification of the family is derived (Durie and Ross 1996). However, when a language family has diverged from its proto-language at such great time depth that too few cognates remain to establish regular sound correspondences, the comparative method can no longer be applied. Even for families with shallower time depths, sound changes can obscure cognates to such a degree that the classification cannot be accurately constructed. It is therefore essential to distinguish true cognacy from other types of lexical similarity such as chance resemblance, sound symbolism, and, in particular, borrowing, all of which reduce the accuracy of the reconstruction and hence the classification. For example, although the lower-order groupings within each of the sub-families of Indo-European are very well attested, there is still much doubt about the higher-order groupings among them. (Warshaw et al. (1995) appear to have made some progress toward solving this problem, although their work is yet to be generally accepted by Indo-Europeanists.) Correctly distinguishing undetected borrowings from cognates may help to resolve the classification of both Indo-European and other, less well documented language families back to greater time depths.

Ross (1996) explains that the comparative method can perform poorly when confronted with borrowing, particularly among languages having no nuclear zone that has avoided

contact with neighboring languages. Putative cognates for which no regular correspondences can be established are typically excluded from the analysis, thereby potentially introducing bias into the reconstructed proto-language and, hence, the genetic classification. A number of alternative methods for dealing with borrowing have been put forward. Hinnebusch (1996), for example, has proposed a lexicostatistical method by which language contact is inferred from skewing in the lexicostatistical data. Wang (1989) has suggested another method in which systematic deviations between the input lexical distances and those calculated from the reconstructed optimal tree are used to infer borrowing. With cladistic analysis becoming more commonly used in linguistic classification, character-based methods for detecting borrowing are now also being explored. For example, Warnow, Ringe and Taylor (1995) have proposed a classification of Indo-European based on a cladistic analysis of both lexical and morphological characters in which they resolve the position of the Germanic branch by distinguishing cognates from loans. Methods have also been suggested in other fields. For example, in their study of human genetic history, Cavalli-Sforza, Menozzi, and Piazza (1994) have suggested that bootstrapping can be used to detect admixture among populations — it should be possible to translate this concept to the task of detecting linguistic borrowing. None of the above methods and ideas has yet been developed to the stage that borrowing can be detected reliably.

Our aim is to develop a computational method to reliably detect borrowing among a family of genetically related languages, to determine which linguistic features have been borrowed, and, ideally, to identify the donor and recipient languages. In this paper, we focus on the borrowing of lexical items. We propose two methods: a distance-based method, derived from lexicostatistics, and a character-based method. We begin by de-

scribing the distance-based method for detecting lexical borrowing in Section 2. We test the hypothesis that the presence of a negative branch length in the lexicostatistical tree for any group of three genetically related languages is indicative of borrowing among them. We present simple mathematical formulae for determining the presence of a negative branch length. However, we are able to show, by means of simple examples, that the presence of a negative branch length is not a sufficient condition for borrowing. We therefore examine whether the probability of occurrence of a negative branch length when there *is* borrowing exceeds that when there *is no* borrowing — Monte-Carlo simulations indicate that the increase in probability is too small to be of practical benefit, causing us to reject the proposed method.

In Section 3, we go on to present a character-based method for detecting borrowing. In this method, we make the assumption that linguistic characters arise independently only once; all further apparent instances of a particular innovation are assumed to be borrowings. Although this assumption is clearly inappropriate for phonological characters, for which parallel convergence is rather common, it may well be appropriate for lexical characters provided that obvious instances of chance resemblance, sound symbolism, and parallel convergence are excised. After isolating the borrowed characters, it is a relatively simple matter to determine the sets of languages among which the borrowing may have occurred. We test this method by application to data for the seven main dialects of Chinese. We show that the method can determine for each informative character a set of languages among which borrowing may have taken place, and can, in some cases, identify the most likely donor and recipient languages.

A discussion and suggestions for future work on detecting borrowing follow in Section

4.

2. A Distance-Based Method for Detecting Borrowing

2.1 Lexicostatistics

Lexicostatistics is a general term for the analysis of languages based on statistical properties of their lexicons. Often, however, the term ‘lexicostatistics’ is used to refer specifically to Swadesh’s method (1950, 1951, 1952, 1955) for classifying a family of genetically related languages based on a statistical comparison of their basic vocabularies. Basic vocabularies are prepared by selecting a list of meanings that are considered to be relatively culture-free and resistant to replacement; the Swadesh 100- (Swadesh 1955) and 200- (Swadesh 1951) word lists are two such commonly used lists. For each language in the family, each meaning is translated into the word that most naturally represents it (Campbell 1998). The similarity of two languages is then measured by counting the proportion of meanings for which the words in the basic vocabularies of those languages *appear* to be cognate¹; we shall refer to such words as ‘apparent cognates’ to reflect the possibility that some such pairs may turn out not to be cognates. Ideally, the apparent cognate count should include only true cognates, although in practice pseudo-cognates, i.e. chance resemblances, and undetected borrowings may sometimes be in-

¹ As Trask (1996) discusses, many linguists argue that cognates can *only* be identified after application of the comparative method. Nevertheless, lexicostatistics is sometimes applied without establishing cognacy, although when, for example, loans are consistently classed as cognates, the resultant classifications tend to be very poor (Campbell 1998).

correctly counted as cognates. Languages that have a greater proportion of apparent cognates are deemed to have split from the proto-language more recently and, consequently, to be more closely related than those having fewer apparent cognates. The inclusion of false cognates such as undetected borrowings in the cognate count skews the relationships among the languages, making them appear to be more closely related than genetically is the case.

A particularly controversial aspect of Swadesh’s method, ‘glottochronology’, is the estimation of the time depths of language splits. Lees (1953) developed a formula that relates the proportion of cognates shared by two genetically related languages, denoted by C , and the rate at which words in each language are retained, r , to the time depth since the languages split from the proto-language, t ,

$$\langle t \rangle = \frac{\log C}{2 \log r} \quad (1)$$

— this formula should be interpreted as specifying the ‘expected value’ of the time depth, hence the use here of the symbol ‘ $\langle \rangle$ ’ to denote statistical expectation. The use of this formula to estimate time depth has been widely criticized; a summary is provided in (Embleton 1986). Arguably the weakest aspect of glottochronology and other related lexicostatistical methods is their assumption that the retention rate is homogeneous across all languages and meanings, and across time. Retention rates certainly are not homogeneous, across languages, across meanings, or across time. For example, Bergsland and Vogt (1962) have commented on the high retention rate of Icelandic compared to that of closely related Faroese (which itself has a higher retention rate than other Germanic languages), while Blust (2000) has recently demonstrated how inhomogenei-

ties in the retention rate can give rise to erroneous lexicostatistical classifications. Some progress has been made, however, in modeling inhomogeneous retention rates. For example, Cavalli-Sforza and Wang (1986) have shown that the proportions of cognates shared among meanings with similar replacement rates tend to be distributed according to the spatial separation of the respective languages. Soon after, in his *Language in the Americas*, Greenberg (1987) put forward a generalization of glottochronology for an arbitrary number of languages with inhomogeneous retention rate. However, the practical implementation of this and other similar methods is problematic.

The modeling of borrowing has been addressed by lexicostatistical methods with some success. Embleton (1981, 1986) has developed a lexicostatistical model for languages that have come into contact with their geographic neighbors. Embleton (1986) models change in the lexical similarity between any two neighboring languages due to three factors: parallel lexical replacement, lexical borrowing between the two languages, and borrowing into each language from other neighbors, deriving a system of differential equations that relates these factors. To use the method, both the retention rates and the borrowing rates must be known. Pagel (2000) has recently shown that retention rates for both extant languages and proto-languages can be estimated by maximum-likelihood estimation, although this approach has not yet been extended to estimating borrowing rates. Another recent method whose aim is to detect language contact has been developed by Hinnebusch (1996), based on ideas that are apparently common practice in African linguistics (see Heine 1974, Nurse 1979, for example). Hinnebusch notes that borrowing between languages very often co-occurs with skewing in the lexicostatistical data. We will discuss this method in more detail in Section 2.5 in relation to our own

lexicostatistical method for detecting borrowing.

2.2 Lexical Distance

Lees obtained formula (1) for the time depth by assuming that any two genetically related languages retain their basic vocabulary at a constant and equal rate, r , after splitting from the proto-language. After some period of time, t , and in the absence of non-genetic processes, each language would be expected to have retained a proportion r^t of its basic vocabulary. Consequently, the expected proportion of shared basic vocabulary, or ‘lexical similarity’, would be

$$\langle C \rangle = r^{2t} \quad (2)$$

Taking the natural logarithm of (2) gives

$$\log \langle C \rangle = 2t \log r . \quad (3)$$

The value of $\log \langle C \rangle$ approaches zero as either the time depth approaches zero (that is, the languages have only just split from the proto-language) or the retention rate approaches 100% (that is, the languages have undergone negligible independent change since splitting from the proto-language).

Given the lexical similarity, C , of two genetically related languages, then by (3) we would expect $\log C$ to be proportional to the time depth, t . We can measure the dissimilarity of a pair of languages by the ‘lexical distance’ between them, defined by

$$D = -\log C . \quad (4)$$

Note that the minus sign in (4) ensures that the lexical distance is positive. An alterna-

tive, more transparent approach is to calculate the dissimilarity of language pairs using the measure

$$D' = 1 - C, \tag{5}$$

or else to work directly with the similarity, C . Indeed, Dyen et al. (1994) work with C directly in their lexicostatistical classification of 84 Indo-European languages. Taking the logarithm in the definition of lexical distance, however, has the advantage that when the retention rates are approximately constant, lexical distance gives a rough indication of time depth (although we make no attempt to infer time depth from lexical distance in this paper). It therefore provides a measure of the degree to which two genetically related languages appear to have diverged from the proto-language. Furthermore, the use of the logarithm renders the measure additive. For example, consider two genetically related languages that have developed independently since splitting from the proto-language such that each retains 80% of the proto-language basic vocabulary. They would be expected to have 64% of their basic vocabularies in common — $80\% \times 80\% = 64\%$. However, taking the logarithm of the lexical similarities makes the relationship additive — $\log(0.80) + \log(0.80) = \log(0.64)$. This property is particularly useful as it allows the construction of metric trees whose branch lengths are calculated to indicate the degrees of divergence of the languages; ideally, the sum of the branch lengths connecting any two nodes in the tree should equal the lexical distance between them. Note that the lexical distance between two arbitrary languages is not impacted by multiple replacements of a particular meaning; no matter how many times the word for a particular meaning is replaced, its effect on the lexical distance remains the same.

The approach taken in distance-based classification is to determine the topology and

branch lengths of the additive tree that in some sense best fit the linguistic dissimilarities among the set of languages being classified. One such method by Qiao and Wang (1998) is to examine each possible topology for the branch lengths that fit the lexical distance data with the smallest mean-squared error — of the trees so generated, the one with the smallest error is considered the best tree. The problem with this exhaustive-search method is that the runtime for even as few as 20 languages can exceed 24 hours on the fastest personal computers currently available. Branch-and-bound algorithms can reduce runtime somewhat by selectively discarding groups of sub-optimal trees based on the analysis of previously examined trees. Alternative methods, such as the Fitch-Margoliash (1967) algorithm, and Saitou and Nei's (1987) Neighbor-Joining algorithm, use heuristic algorithms to estimate the best tree, reducing runtimes by several orders of magnitude. There is no guarantee, however, that they converge to the best fitting tree.

2.3 Negative Branch Lengths as Potential Indicators of Borrowing

In order to introduce the distance-based method for detecting borrowing described here, it is instructive to consider the case of classifying a set of three genetically related languages by constructing an 'unrooted' tree. In the context of linguistic classification, an 'unrooted' tree represents the branching relationships among a group of languages, but without explicitly indicating ancestry. An unrooted linguistic tree constructed to represent the branching relationships among a group of genetically related languages would therefore not include a node for the proto-language. A 'rooted' linguistic tree, however, would include such a node, allowing the order in which various sub-groups of languages split off from the proto-language to be explicitly indicated.

Denoting the set of three languages by $\{L_i, L_j, L_k\}$, the optimal unrooted tree and its branch lengths can be calculated simply and, above all, uniquely from the pairwise lexical distances. Only a single unrooted tree topology is possible: that shown in Fig. 1. The sum of the lengths of the branches that connect any pair of languages should equal the lexical distance between that pair. Thus, the branch lengths, denoted by l_i , l_j and l_k , can be calculated by solving the simultaneous equations

$$D_{ij} = l_i + l_j, \tag{6a}$$

$$D_{jk} = l_j + l_k, \tag{6a}$$

$$D_{ki} = l_k + l_i, \tag{6a}$$

where, for example, D_{ij} denotes the lexical distance between languages L_i and L_j .

[Fig. 1 here ...]

The system of simultaneous equations (6) has the unique solution

$$l_i = \frac{1}{2} (D_{ij} + D_{ki} - D_{jk}), \tag{7a}$$

$$l_j = \frac{1}{2} (D_{jk} + D_{ij} - D_{ki}), \tag{7b}$$

$$l_k = \frac{1}{2} (D_{ki} + D_{jk} - D_{ij}). \tag{7c}$$

Notice that although the lexical distances must, by definition, be positive, there is no such constraint on the branch lengths. From (7a), we observe that the branch length l_i , for example, is negative when

$$D_{jk} > D_{ij} + D_{ki}. \tag{8}$$

Substituting this expression into (7b), we have

$$l_j > D_{ij} > 0. \quad (9)$$

Similarly, it can be shown that $l_k > 0$. Thus, when l_i is negative, l_j and l_k must both be positive. Since the labeling of the languages is arbitrary, corresponding results must hold when either l_j or l_k are negative. Hence at most one branch length in the lexicostatistical tree for three languages may be negative. A negative branch length therefore occurs when

$$D_{jk} > D_{ij} + D_{ki} \quad (10)$$

for any distinct i, j and k . An equivalent condition, stated in terms of the lexical similarities, is

$$C_{jk} < C_{ij} \times C_{ki} \quad (11)$$

for any distinct i, j and k .

Our aim in this section is to test the hypothesis that the presence of a negative branch length in a unrooted lexicostatistical tree for three languages (which can be reliably detected by condition (10)) is indicative of borrowing. If this hypothesis is found to be valid, we can expand upon the method to detect borrowing among an arbitrary number of related languages. We continue by examining two simple scenarios in which a negative branch length is induced in the lexicostatistical tree for three languages.

In the first scenario, we consider three hypothetical, genetically related languages, L_i , L_j and L_k , that descend from a common proto-language, PL . We use lexicostatistics to classify these languages based on a basic vocabulary consisting of ten meanings, $M_1 \dots$

M_{10} , that have the states shown in Table 1.

[Table 1 here ...]

In Table 1, cognates are marked with the same state: "0" represents the archaic state of a meaning retained from the proto-language PL , while other states represent innovations. In practice, the archaic state of each meaning would typically be unknown; glosses for meanings M_7 and M_{10} , with the states given in Table 1, would therefore provide the same classification information — we introduce archaic states in the table simply so that we can specify the retention rate (also unknown to the linguist) of each language since splitting from the proto-language. In this example, each language has retained 40% of the basic vocabulary of the proto-language. Clearly languages L_j and L_k are more closely related to one another than to L_i , and hence, together with their intermediate proto-language, PL_{jk} , form a ‘monophyletic group’ — to paraphrase Kitching (1998), a ‘monophyletic group’ of languages is one that consists of a proto-language plus all and only all languages which derive from that proto-language. The lexical distances among L_i , L_j and L_k are calculated from the observed lexical similarities, both shown in Table 2. The lexicostatistical tree, shown as a rooted tree in Fig. 2, is then constructed by calculating the branch lengths from the lexical distances using (7). All branch lengths are positive.

[Table 2 here ...]

[Fig. 2 here ...]

Suppose that languages L_i and L_j now come into contact such that L_i donates some of its basic vocabulary to L_j . If these loans were included in the cognate count for L_i and L_j ,

these two languages would appear to be *more* closely genetically related than is actually the case. Conversely, it is possible that some words in L_j that were previously cognate with words in L_k would be replaced by loan words from L_i , resulting in languages L_j and L_k appearing to be *less* closely genetically related than is the case. For example, if L_i were to donate words for meanings M_3 and M_4 to L_j , languages L_i and L_j would then appear to be more closely related than is the case. However, the lexical distance between L_j and L_k would be unchanged due to the replaced words in L_j not having previously been cognate with words in L_k . The apparent lexical similarities and the resultant lexical distances after such contact between L_i and L_j are summarized in Table 3.

[Table 3 here ...]

Constructing the lexicostatistical tree using these lexical distances, we observe that the resultant tree, shown in Fig. 3, has the same topology as that in Fig. 2 but has different branch lengths, given in Table 4. In particular, the length of the branch connecting L_j to the body of the tree is negative, demonstrating that borrowing can indeed induce a lexicostatistical tree having a branch with negative length. The detection of negative branch lengths in lexicostatistical trees is therefore a candidate method for detecting borrowing.

[Fig. 3 here ...]

[Table 4 here ...]

In the second scenario, however, we demonstrate that a negative branch length can also appear in the lexicostatistical tree when there is no borrowing among the languages. Consider now three genetically related languages, L_i , L_j and L_k , for which the basic vocabularies for ten meanings have the states shown in Table 5.

[Table 5 here ...]

Each language has retained 50% of the basic vocabulary of the proto-language. It is clear that, together with their intermediate proto-language, languages L_j and L_k form a monophyletic group. The observed lexical similarities and the resultant lexical distances among L_i , L_j and L_k are shown in Table 6.

[Table 6 here ...]

The lexicostatistical tree, shown in Fig. 4, is constructed using the same method as before. The branch lengths are given in Table 7 — the length of the branch connecting L_k to the body of the tree is negative. We must therefore conclude that the presence of a negative branch length in the lexicostatistical tree is not a sufficient condition for detecting borrowing.

[Fig. 4 here ...]

[Table 7 here ...]

2.4 Simulation

Having determined that the presence of a negative branch length in the lexicostatistical tree for three languages is not a sufficient condition for detecting borrowing, we seek to determine the probability that the lexicostatistical tree contains a negative branch length (a) when there is no borrowing among the languages, and (b) when there is borrowing among the languages. If it can be shown that a negative branch length is significantly more likely to occur when there is borrowing, we can treat its occurrence as evidence (but not proof) of borrowing. We treat condition (10) as the decision rule of the binary

hypothesis test

$$D_{jk} \begin{matrix} > \\ < \end{matrix} \begin{matrix} H_1 \\ H_0 \end{matrix} D_{ij} + D_{ki} \quad (12)$$

for any distinct i, j and k , where H_0 and H_1 are the hypotheses:

- H_0 : the ‘null hypothesis’ — no borrowing among L_i, L_j and L_k ,
- H_1 : the ‘alternative hypothesis’ — borrowing among L_i, L_j and L_k .

A ‘decision rule’ specifies the conditions under which the null hypothesis, H_0 , is either accepted or else rejected in favor of the alternative hypothesis, H_1 . A binary hypothesis test has four possible outcomes:

- H_0 accepted, H_0 true — correct decision
- H_0 rejected, H_1 true — correct decision
- H_0 rejected, H_0 true — Type I Error, “false alarm”
- H_0 accepted, H_1 true — Type II Error, “miss”

and is usually characterized by the probabilities of Type I and Type II error, also called the ‘false alarm rate’ and ‘miss rate’, respectively. For example, our method generates a false alarm whenever condition (10) indicates the presence of a negative branch length in the lexicostatistical tree for languages among which there has been no borrowing. Ideally, we would want the probabilities of both types of error to be zero. Typically, however, adjusting the decision rule so as to reduce the probability of one type of error causes the other to increase (DeGroot 1986). We must therefore accept that the false

alarm rate and miss rate cannot both be made arbitrarily small.

We proceed to estimate the false alarm rate and miss rate for decision rule (12) using Monte-Carlo simulation, a general method for estimating solutions to various types of system using random number generation and sampling — a comprehensive, mathematical summary of the Monte Carlo method can be found in (Fishman 1996). The generation of random data of any type is problematic (although true random number generators can be implemented by tapping a non-deterministic source, for example white noise in an electrical circuit). A more tractable approach is to use ‘pseudo-random’ data constructed from chaotic sequences of integers that, while deterministic, share many of the requisite properties of random numbers. In order to implement the simulation, we have developed two algorithms running in the Matlab programming environment (MathWorks 2000) to generate lexicostatistical data for sets of three hypothetical languages, one for each hypothesis. Our algorithms make use of the pseudo-random number generator built into Matlab.

2.4.1 H_0 : No Borrowing

The algorithm we use to construct pseudo-random lexicostatistical data for three languages among which there is no borrowing is described in Appendix A. The parameters of the algorithm are the number of meanings, N , the time depth, t , the retention rates of the extant languages, r_i , r_j and r_k , and the retention rate of the intermediate proto-language, r_{jk} . Fig. 5 summarizes the data generated from one typical run of the algorithm.

[Fig. 5 here ...]

Algorithm 1 can be run repeatedly for a particular set of parameter values to generate a number of pseudo-random lexicostatistical trees — the relative frequency of false alarms provides an estimate of the false alarm rate for those parameter values. For example, for a set of 100 runs with the number of meanings set to 100, the time depth set to 6,000 years, and the retention rate of each language set to 75% per millennium, the false alarm rate was estimated to be about 7%.

Fig. 6 shows a plot of the estimated false alarm rate as a function of retention rate and time depth for 100 meanings based on 5000 runs. The plot shows clearly that the false alarm rate is negligible except within two regions: at high retention rate and low time depth, and along a band of values of retention rate and time depth, such as retention rate 70% per millennium and time depth about 7,000 years. The first region corresponds to languages that have high retention rate and which have split from the proto-language relatively recently; two such languages would be very closely genetically related, in which case there would be a high probability that they would both have retained the proto-form for any particular meaning.

[Fig. 6 here ...]

For example, consider three genetically related languages L_i , L_j and L_k having the character states shown in table 8.

[Table 8 here ...]

The lexical distances among L_i , L_j and L_k are

$$D_{ij} = 0.01005,$$

$$D_{jk} = 0.02020,$$

$$D_{ki} = 0.01005.$$

The value of $D_{ki} + D_{ij}$, which equals 0.02010, is less than the value of D_{jk} , 0.02020, implying, by (10), that the length of the branch connecting language L_i to the tree is negative — the single innovation of meaning M_1 in language L_k is sufficient to induce this negative branch length.

The second region of non-negligible false alarm rate appears to depend on some functional relationship between the retention rate and the time depth. Plotting the estimated false alarm rate as a function of the retention rate raised to the power of the time depth, r^t , shown in Fig. 7, confirms this. The plot comprises 36 lines, one line for each iteration of the time depth. The lines are obviously equivalent, with only slight fluctuations from the mean due to the finite number of samples. We conclude that the probability of false alarm depends functionally on the parameter r^t . The peak false alarm rate of about 32% occurs at approximately $r^t = 0.095$. For example, at a time depth of 6,000 years, this corresponds to a retention rate of about 68% per millennium. We have not yet determined the reason for the relationship between the peak value of the false alarm rate and the corresponding value of r^t .

[Fig. 7 here ...]

The property that the false alarm rate is negligible across many combinations of values of the retention rate and time depth is desirable. The peak in the false alarm rate at high retention rate and low time depth will be of little concern in many practical applications. For example, a retention rate of 86% and time depth of 200 years imply an estimated false alarm rate of about 5%, perhaps an acceptable level of error. Of far more concern, however, is the second peak for which the false alarm rate rises to about 32%. Such a

high false alarm rate will probably only be acceptable if the miss rate is very low. As we discuss in the following section, this is not the case.

2.4.2 H_1 : Borrowing

In order to estimate the miss rate, we require an algorithm to generate pseudo-random lexicostatistical data for languages among which borrowing has taken place. There are many possible scenarios in which borrowing could take place among three languages: more than one pair of languages could come into contact; contact could occur several times, introducing several strata of loans; borrowing could be bi-directional. We focus on just the simplest model of borrowing: a single instance of borrowing from a single donor language to a single recipient language. Algorithm 1 is amended by pseudo-randomly selecting a time instant at which to inject the borrowing. The donor and recipient languages are chosen pseudo-randomly and non-cognate words are copied from the donor to the recipient at a specified borrowing rate, b . This algorithm, Algorithm 2, is described in Appendix B. It has one more parameter than Algorithm 1, the borrowing rate, b . Fig. 8 summarizes the data generated from one typical run of the algorithm; note that this tree would be classed as a miss since it contains no branch with negative length despite borrowing having been injected.

[Fig. 8 here ...]

We use Algorithm 2 to estimate the miss rate in much the same way that we have used Algorithm 1 to estimate the false alarm rate. In the following simulation we set the borrowing rate to 20%. All other parameter values are set to the same values as in the false alarm rate simulation. Fig. 9 shows a plot of the estimated miss rate for 100 meanings based on 2000 runs. The weakness of using decision rule (12) to detect borrowing

is now apparent. Across much of the range of values of retention rate and time depth, the miss rate approaches 100% — the algorithm fails to indicate borrowing. This region corresponds very closely to the region in which the false alarm rate is negligible. A negative branch length is approximately equally likely to occur whether there has been contact among the languages or not. The miss rate exhibits two troughs, corresponding closely to the regions of peak false alarm rate. Fig. 10 shows a plot of the miss rate against the retention rate raised to the power of the time depth, r^t . Again, we observe that the probability of error is functionally related to r^t . The minimum value of the miss rate occurs at about $r^t = 0.075$. At a time depth of 6,000 years, this corresponds to a retention rate of about 65% per millennium.

[Fig. 9 here ...]

[Fig. 10 here ...]

For this method to have any practical use in detecting borrowing, the probability that borrowing is indicated when there has been contact among the languages should exceed that when there has been no contact. In other words, one minus the miss rate, i.e. the ‘detection rate’, should exceed the false alarm rate. Fig. 11 shows a plot of the detection rate minus the false alarm rate against r^t . The detection rate exceeds the false alarm rate significantly only across a small range of values of r^t , centered at about 0.04, at which point the detection rate exceeds the false alarm rate by about 22%. At a time depth of 6,000 years, this corresponds to a retention rate of about 58% per millennium. For a retention rate of 80%, the peak performance occurs for a time depth of about 14½ millennia. This level of performance would probably be unsatisfactory for most practical attempts to detect borrowing. The proposed distance-based method for detecting

borrowing must therefore be discarded.

[Fig. 11 here ...]

In Section 3, we describe an alternative method for detecting borrowing using a character-based approach. First, however, we relate the distance-based method just discussed to the method of Hinnebusch (1996), which is based on detecting skewing in lexicostatistical data.

2.5 Lexicostatistical Skewing as an Indicator of Borrowing

Hinnebusch (1996) has suggested a method for detecting language contact that relates skewing in lexicostatistical data to non-genetic effects, particularly borrowing. He quotes an example cited by Heine (1974), which we describe here briefly to illustrate the basic concept of the method. Two Nilotic languages, Samburu and Nandi, share 9.9% of a 200-word basic vocabulary. Masai and Nandi, however, share 15.7% of the basic vocabulary. Since Masai and Nandi are known to have recently come into contact, the 5.8% difference in the lexical similarities is assumed to be due to borrowing between Masai and Nandi. Hinnebusch (1996) uses this same basic idea to identify lexical borrowing among several sub-families of Bantu. Among a group of genetically related languages, borrowing is inferred between two languages in different clusters when the lexicostatistical similarity between them significantly exceeds the average similarity between two arbitrary languages, one chosen from each cluster.

Although the concept underlying the Hinnebusch method differs somewhat from that underlying the distance-based method which we have proposed, the two methods turn out to have a similar structure. To simplify the comparison of the two methods, we

focus on detecting borrowing among three genetically related languages only. We assume that the genetic relationships among three languages, L_i , L_j and L_k , are known, as shown in Fig. 12 and that the lexical similarities between each pair of languages, C_{ij} , etc., are available. Although it is generally the case that the genetic classification of a group of related languages is unknown a priori, there are occasions when a particular classification of a small subset of languages is implied by independent linguistic data — such information may be sufficient to classify certain subsets of languages, even though their classification with respect to the entire set of languages may be unknown. For example, Hinnebusch (1996) relies on the testimony of native speakers and inferences drawn from the application of other linguistic techniques to assess the sub-families to which the languages being studied belong. Note that Hinnebusch’s method only requires that the languages be classified into genetically related sub-families; the groupings within each of the sub-families need not be known.

Borrowing between L_j and L_k is indicated when the ‘skewing’, given by $C_{jk} - C_{ik}$, is large (although Hinnebusch does not specify just how large the skewing must be for borrowing to be indicated). In other words, borrowing is indicated when

$$C_{ik} \ll C_{jk} . \tag{13}$$

Under the same scenario, our method would indicate possible borrowing between L_j and L_k if, for example, the length of the branch in the lexicostatistical tree connecting L_j to the body of the tree were found to be negative. In terms of the lexical similarities, borrowing would be indicated (by (11)) when

$$C_{ik} < C_{ij} \times C_{jk} . \tag{14}$$

Both methods indicate borrowing when the lexical similarity between L_i and L_k is small, but whereas we set a specific threshold for C_{ik} , Hinnebusch does not. We have shown that our proposed method for detecting borrowing offers poor performance; in terms of lexical skewing, the threshold value induced by decision rule (12) is simply not sensitive to borrowing. A worthwhile line of research therefore would be to determine appropriate threshold values for lexical skewing that reliably indicate borrowing.

[Fig. 12 here ...]

3. A Character-Based Method for Detecting Borrowing

3.1 Cladistics

Cladistics, also known as phylogenetic systematics, is a method for performing hierarchical classification of taxa, first derived formally by the entomologist Willi Hennig (1950) (although similar methods had been used in historical linguistics more than 60 years earlier by Brugmann (1884)). Hennig's aim was to devise a method for discovering the ancestor-descendant relationships among biological species implied by Darwin's theory of evolution. In a cladistic analysis, each taxon is described in terms of the states of certain characters. For example, *hyperoartia* (lampreys) can be distinguished from other vertebrates by the absence of jaws (Kitching et al. 1998): in this case, the character is "jaws" and the character state is "absent". One attempts to distinguish the characters or states that have been retained by direct descent from ancestor species, 'plesiomorphies', from the characters or states that are innovations, 'synapomorphies' (Kitching et al. 1998). The taxa are then organized into a nested hierarchy based on the shared

innovations.

A common criterion for determining the optimal classification is the ‘maximum parsimony criterion’, which selects the hierarchy “that accounts for the greatest number of characters in the simplest way” (Kitching et al. 1998). A simple example, after Kitching et al. (1998), illustrates the concept. Table 9 shows the character states for four hypothetical taxa having six binary characters: "0" represents the absence of a character, "1" its presence.

[Table 9 here ...]

Three putative classifications are considered, the cladograms for which are shown in Fig. 13. Marked on each cladogram are state changes that produce the observed character states. The most parsimonious classification is that represented by Cladogram A, which requires 7 character state changes, fewer than Cladograms B and C, which require 9 and 8 character state changes, respectively. The minimum number of character state changes required on a cladogram is called the ‘length’ of the cladogram. We calculate length using ‘Fitch parsimony’ (Fitch 1971) whereby characters are assumed to be ‘unordered’; that is, the sequence of permissible character state changes is unknown or else all permissible character state changes are considered to be equally likely. Another commonly used measure of parsimony, ‘Wagner parsimony’ (Wagner 1961; Kluge & Farris 1969), works with ‘ordered’ characters for which the sequence of permissible character state changes is assumed known — we consider Wagner parsimony to be less appropriate for use in linguistic cladistic analysis as there is no fundamental order to the sequence of innovations that may occur to any lexical item.

[Fig. 13 here ...]

An alternative criterion for determining the optimal classification is the ‘compatibility criterion’. Warnow et al. (1995) define a character as ‘compatible’, or ‘convex’, when each of its states has arisen only once. The optimal classification according to the compatibility criterion is that having the fewest number of non-compatible characters. Going back to the putative classifications shown in Fig. 13, Cladogram A has only 1 non-compatible character, Character 6. Cladograms B and C, however, have 3 and 2 non-compatible characters, respectively. The classification associated with Cladogram A is therefore also optimal according to the compatibility criterion.

Both of the above criteria are used in the phylogenetic classification of biological taxa. Felsenstein (1982) has argued that there are no general grounds for preferring either criterion over the other. Nevertheless, Warnow and colleagues have published a body of work in which they apply cladistic methods to the classification of languages based on the compatibility criterion (see Warnow et al. 1995, Warnow 1997, etc.). In their words, “the fact that a character is not [compatible] on the tree under consideration is much more significant than the precise number of extra evolutionary steps required by that character on the tree” (Warnow et al. 1995). By adopting the compatibility criterion, they have been able to construct a phylogeny for the Indo-European family in which several of the most contentious sub-groupings of the sub-families of Indo-European appear to be resolved. In particular, they propose that, genetically, Germanic is a sister sub-family to Balto-Slavic and that it later came into contact with the Italic and Celtic sub-families to which it now shares a strong lexical affinity. They assume that the states of appropriately chosen linguistic characters, such as lexical items (but not sound changes), have arisen only once. Implicit in this assumption is the belief

that non-compatible characters are indicative of borrowing or other non-genetic effects. We now explore the use of this relationship as a method for detecting borrowing.

3.2 The Method

The proposed method for character-based detection of borrowing among a group of genetically related languages is based on the assumption that no state of any character emerges independently more than once — second and further occurrences of a particular innovation in a cladistic analysis are assumed to be due to borrowing. Note that we allow characters to revert to the states of earlier, possibly extinct, languages, such as has been observed, for example, in the revival of word-final *-d* in Swedish (Janson 1977). We treat such character reversion as a special case of borrowing.

We begin the proposed method by following a standard cladistic analysis for reconstructing genetic relationships, much like that used by Warnow et al. (1995). A gloss for a set of meanings (the characters) should be prepared for each language. Words that are judged to be cognate (preferably by application of the comparative method) are assigned the same character state. Of course, some of these judgments may be incorrect — the aim of the method is to detect these errors, which we ascribe to borrowing. Other types of characters, such as morphological features, can also be used provided that each character state can reasonably be expected to have emerged independently only once. Having prepared the character state data, the length of each possible cladogram for the group of languages should then be calculated — we follow the procedure described by Fitch (1971). The most parsimonious trees, that is, those having the shortest length, can then be determined, allowing a set of candidate genetic classifications of the languages

to be reconstructed.

We continue by identifying the characters that are incompatible — such characters have states that have arisen more than once. Three examples will illuminate how we relate compatibility to borrowing. Consider the two cladograms shown in Fig. 14, which have the same topology and terminal states of a single character but which assign the state changes of that character differently. In both cases, the character is incompatible. In Cladogram A, the plesiomorphic state is 1. The character state 2 arises at four points on the cladogram, which we interpret as indicating one instance of innovation plus three instances of borrowing among the taxa *B*, *C*, *D*, and *E*. In Cladogram B, however, the plesiomorphic state is 2. Since character state 1 arises at only two points, this cladogram is the more parsimonious and indicates one instance of borrowing between the taxa *A* and *F*. We therefore select Cladogram B and infer that borrowing has occurred between *A* and *F*. In this case, the direction of the borrowing cannot be determined.

[Fig. 14 here ...]

Fig. 15 shows a situation in which multiple cladograms are equally parsimonious. The two cladograms have the same topology and terminal states for a single character but assign the state changes of that character differently. In Cladogram A, the plesiomorphic state is 1. Character state 2 arises at three points on the cladogram, indicating two instances of borrowing among the taxa *B*, *C*, and *F*. In Cladogram B, the plesiomorphic state is 2. Character state 1 arises at three points, again indicating two borrowings, in this case among the taxa *A*, *D*, and *E*. Since the two cladograms are equally parsimonious, we are unable to resolve the borrowing, inferring only that borrowing has occurred twice among *B*, *C*, and *F*, or twice among *A*, *D*, and *E*.

[Fig. 15 here ...]

The cladogram shown in Fig. 16 demonstrates how the direction of borrowing can sometimes be inferred. The plesiomorphic state is 1, retained by taxa A, B and C. Character state 2, observed at taxa D, E and F, arises most parsimoniously at a single point, as indicated on the cladogram. The character state 1 is also observed at taxon G — according to our assumption that a particular innovation occurs only once, this cannot be an innovation since the character state 1 was already present at the root. For the same reason, taxa A, B and C cannot be recipients of borrowing. Hence borrowing is indicated either from AB to G or from C to G; in either case, the recipient taxon is G.

[Fig. 16 here ...]

In this way, borrowing hypotheses can be constructed for each of the most parsimonious trees. In general there will be more than one most parsimonious tree. In the next section, we apply the method to detect lexical borrowing among seven Chinese dialects.

3.3 Case Study — Chinese

We have used the above method to detect lexical borrowing among seven Chinese dialects. The dialects we have considered are Mandarin (as spoken in Beijing), Xiang (as spoken in Changsha), Yue (as spoken in Guangzhou), Gan (as spoken in Nanchang), Wu (as spoken in Suzhou), Hakka (as spoken in Meixian), and Min (as spoken in Xiamen) — a conventional classification of these seven Chinese dialects is shown in Fig. 17 (You 2000). A gloss for each of the Swadesh 100 basic meanings was prepared for

each dialect by Prof. Xu Tong-Qiang² (refer to Appendix C for details). Having determined the shared basic vocabulary, we constructed the character state matrix; for each meaning, apparent cognates were assigned the same character state.

[Fig. 17 here ...]

In cladistics, a character that contains no grouping information relevant to a particular classification analysis is termed ‘uninformative’ — a character is uninformative, for example, when all taxa are assigned the same state or when a character state is restricted to a single terminal taxon (Kitching et al. 1998). Of the 100 lexical characters examined for the seven Chinese dialects, 85 were found to be uninformative. The classification of the dialects was therefore carried out using the remaining 15 informative characters. Ideally, we would prefer to carry out the classification using more than just 15 informative characters. Nevertheless, this case study will serve to illustrate our methodology. The states of the 15 informative characters, among which borrowing can potentially be detected, are shown in the character state matrix in Table 10.

[Table 10 here ...]

To see how the algorithm works, consider Topology 1, shown in Fig. 18(a). On this topology, nine characters are compatible, together requiring a minimum of 29 state changes. Six characters are incompatible, "feather", "grease", "say", "small", "sun", and

² Prof. Xu, of the Department of Chinese and Literature at Peking University, very kindly provided us with unpublished glosses for the seven Chinese dialects, originally collected for (Xu 1991). Mr. Wang Feng, of the Language Engineering Laboratory at the City University of Hong Kong, then identified apparent cognates for each pair of dialects. All further processing of the data is our own.

"what", requiring at least 24 state changes. For example, the most parsimonious assignment of state changes to the character "feather" on Topology 1 requires 4 state changes, as shown in Fig. 19. Since two states are observed for the character "feather" and since we have assumed that each observed character state arises independently only once, these 4 state changes would imply two instances of innovation plus 2 instances of borrowing among Mandarin, Xiang and Min, as marked on the figure. Topology 1 has length 53, indicating 46 instances of innovation plus 7 instances of borrowing. For a second topology, Topology 2, shown in Fig. 18(b), all characters except "what" are incompatible, requiring a minimum of 64 state changes (46 instances of innovation plus 18 instances of borrowing). Topology 2 is therefore less parsimonious than Topology 1.

[Fig. 18 here ...]

[Fig. 19 here ...]

We find a total of 55 most parsimonious topologies for the seven Chinese dialects, each of length 53 — the next most parsimonious topologies, of which there are 77, each have length 54. The most parsimonious topologies can be classified into five types, each type characterized by a set of characters that are indicated as subject to borrowing. Rather than display each of the most parsimonious topologies, we show only the consensus tree for each type — a ‘consensus tree’ is a diagram that combines a set of topologies into a single topology by retaining components that occur sufficiently often, for example in more than 50% of the topologies being combined (majority-rule consensus tree (Swoford 1991)). Fig. 20 shows the consensus tree for the topologies of each type as well as for the entire set of most parsimonious topologies. The borrowings indicated for each type are summarized in Table 11. Borrowing of the characters "feather", "small", and

"what" is indicated for each type. "know" is indicated as borrowed in 4 of the 5 types. However, borrowing of the remaining characters, "grease" (3 types), "say" (3 types), "sun" (3 types), "give" (2 types) and "who" (2 types), appears to be less strongly supported.

[Fig. 20 here ...]

[Table 11 here ...]

How are we to assess in an objective manner which of the indicated borrowings are significant? We deal with this issue here by using simple statistical significance testing procedures to test whether the number of topologies for which borrowing of a particular character is indicated is greater than would be expected by chance if we were to have selected the 55 topologies at random — if so, we accept the hypothesis that borrowing for that meaning has occurred. We know from probability theory that if we draw a random sample of n objects without replacement from a population of N objects of which X possess a certain attribute, then the number of objects in the sample that possess that attribute, x , has a ‘hypergeometric distribution’ with probability function (Evans et al. 1993)

$$\Pr(x) = \frac{\binom{X}{x} \binom{N-X}{n-x}}{\binom{N}{n}}. \quad (15)$$

Thus the number of topologies picked at random that indicate borrowing for any particular meaning has a hypergeometric distribution. Table 12 lists the number of topologies for which borrowing of each character is indicated among the 55 most parsimonious topologies and among all possible topologies; there are $3 \times 5 \times \dots \times 11$ (i.e. 10395)

possible rooted binary topologies relating the 7 taxa.

[Table 12 here ...]

The probabilities that fewer than the observed numbers of topologies indicate borrowing are also shown in the table, calculated using (15). At the 1% significance level, we conclude that meanings "feather", "small" and "what" have each undergone borrowing. According to this criterion, however, there is insufficient evidence to conclude that meanings "give", "grease", "know", "say", "sun" and "who" have undergone borrowing.

We now examine the consensus trees for topologies of each type to determine among which dialects each instance of borrowing may have occurred. For example, Fig. 21 shows all the most parsimonious assignments of borrowing of the character "feather" with respect to Types I–V. Types I–III admit one instance of innovation plus two instances of borrowing among Mandarin, Xiang, and Min. Type III also admits borrowing of the character "feather" between Wu and either Hakka or Yue, and between Min and either Mandarin or Xiang. Types IV & V, however, admit only a single instance of borrowing between Min and Mandarin or, perhaps, Xiang. For those borrowing hypotheses involving Mandarin, we consider Mandarin the more likely donor because of its status as a prestige language, Beijing having been the political and cultural capital of China for almost 800 years.

[Fig. 21 here ...]

The dialects among which borrowing of "small" and "what" may have occurred can be determined in a similar manner. (Borrowing hypotheses could also be determined similarly for "give", "grease", "know", "say", "sun" and "who", although we would have to

treat these hypotheses more tentatively.) Table 13 summarizes the most parsimonious assignments of borrowing among the seven Chinese dialects implied by our method — for each hypothesis, we indicate the direction of borrowing with the strongest support. Note that in no case is a unique set of donor and receiver languages indicated. Nevertheless, the method does offer the linguist a set of borrowing hypotheses that can be tested in more detail using other methods.

[Table 13 here ...]

4. Discussion

We have examined two methods for detecting borrowing among a family of genetically related languages, one method using distance-based techniques, the other using character-based techniques. The proposed distance-based method, discussed in Section 2, detects branches with negative length in the lexicostatistical tree for any three languages in the family, the hypothesis being that such negative branch lengths are indicative of borrowing. We have shown, however, that the presence of a negative branch length in the tree for three particular languages is not a sufficient condition for borrowing to have occurred among them. Nor is there a significant increase in the probability of occurrence of a negative branch length when there is borrowing above that when there is not. Rather, the probability of occurrence of a negative branch length depends on the proportion of the basic vocabulary retained by each of the three languages from the proto-language. The method must therefore be rejected.

The proposed character-based method for detecting borrowing, discussed in Section 3, follows a similar approach to genetic classification as that adopted by Warnow et al.

(1995), although we determine the genetic classification using the maximum parsimony criterion rather than the compatibility criterion. We assume that no innovation arises independently more than once — all instances of apparent parallel innovation are assumed to be due to borrowing. The method is implemented in two steps: first, genetic classification is performed by ranking all possible cladograms according to their length; then, the most parsimonious cladograms are examined for borrowing. Applying the method to seven Chinese dialects, we have shown that the method indicates for each informative character a set of languages among which borrowing may have taken place, in some cases identifying the most likely direction of the borrowing.

4.1 Future Directions

The failure of the distance-based method to detect borrowing has led us to reject the hypothesis that the presence of negative branch lengths in lexicostatistical trees is indicative of borrowing. As we have discussed in Section 2.5, however, the lexicostatistical skewing method of Hinnebusch (1996) is a far more promising line of research. We are currently undertaking a statistical study of the performance of this method under various scenarios of contact.

In Section 1, we remarked that Cavalli-Sforza et al. (1994) have suggested that bootstrapping may serve to detect admixture (equivalent to borrowing) among biological taxa. Bootstrapping is the process of sampling with replacement from a data set in order to simulate a random sample drawn from the population containing that data set. It is frequently used in cladistic analysis to determine the support for a particular clade (Kitching et al. 1998). Cavalli-Sforza et al. note that when bootstrapping is applied to a metric tree, “mixed populations often tend to be attached to different clusters in differ-

ent bootstrap trees” — in other words, taxa that have undergone borrowing tend to shift to different positions in the reconstructed topology depending on which taxa have contributed to the borrowing. The potential of applying this method to detect borrowing among languages should be explored. The idea suggested by Wang (1989), whereby borrowing can be inferred from differences in the input and output lexical distance matrices, should also be examined in more detail; we are not aware of any recent application or analysis of this method.

The character-based method, which we have applied to detect borrowing among seven Chinese dialects, appears to be able to identify characters for which borrowing has occurred, also specifying sets of putative donor and recipient languages. It must now be established that the method, which should be regarded only as indicating the most parsimonious borrowing hypotheses, performs robustly when applied to a language family for which both the vertical classification and the borrowings among a particular set of lexical items are well attested — Indo-European, for example, would be an ideal choice. If the method can be shown to recover a reasonable proportion of the known borrowings without too many false alarms, we can have greater confidence that the method is indeed robust and that it can be used to reliably detect borrowing among less-well studied groups of languages.

Appendix A — Algorithm for Pseudo-Random Lexicostatistical Data; No Borrowing

The following algorithm constructs pseudo-random lexicostatistical data for three languages among which there is no borrowing. The input parameters are: the number of meanings, N ; the time depth, t ; the retention rates of the extant languages, r_i , r_j and r_k ; and the retention rate of the intermediate proto-language, r_{jk} .

Algorithm 1 (Pseudo-Random Data — H_0 : No Borrowing)

1. Specify the time depth, t , of the proto-language and the number of meanings in the basic vocabulary, N .
2. Pseudo-randomly select one language, L_i , to be the sister to the other two languages, L_j & L_k , which have the intermediate proto-language PL_{jk} .
3. Pseudo-randomly select the time depth, t_{jk} , of the intermediate proto-language — this should be some fraction of the total time depth, t .
4. Specify the retention rate for each extant language, r_i , r_j & r_k , and for the intermediate proto-language, r_{jk} .
5. Pseudo-randomly select the basic vocabulary retained by L_i from the proto-language — the probability that each word is retained is r_i^t .
6. Pseudo-randomly select the vocabulary retained by the intermediate proto-language from the proto-language — the probability that each word is retained is $r_{jk}^{t-t_{jk}}$.
7. Pseudo-randomly select the vocabulary retained by L_j and by L_k from the intermediate proto-language — the probability that each word is retained is $r_j^{t_{jk}}$ and $r_k^{t_{jk}}$, respectively.
8. Count the proportion of basic vocabulary shared by each pair of languages and thereby determine the lexical distances.

Appendix B — Algorithm for Pseudo-Random Lexicostatistical Data; Borrowing

The following algorithm constructs pseudo-random lexicostatistical data for three languages among which there is borrowing between a single pair of languages. This algorithm has the input parameter b , the borrowing rate, in addition to the input parameters of Algorithm 1.

Algorithm 2 (Pseudo-Random Data — H_1 : Borrowing)

1. Specify the time depth, t , of the proto-language, the number of meanings in the basic vocabulary, N , and the borrowing rate, b .
2. Pseudo-randomly select one language, L_i , to be the sister to the other two languages, L_j & L_k , which have the intermediate proto-language PL_{jk} .
3. Pseudo-randomly select the time depth t_{jk} of the intermediate proto-language.
4. Specify the retention rate for each extant language, r_i , r_j & r_k , and for the intermediate proto-language, r_{jk} .
5. Pseudo-randomly select the donor language, the recipient language, and the time depth, t_b , of the borrowing — the donor and recipient are selected from among the extant languages *and* the intermediate proto-language.
6. There are now *three* time intervals within which to determine retentions: the interval *between* the borrowing and the splitting of the intermediate proto-language, the time interval *before* that, and the time interval *after* that. Retentions are selected pseudo-randomly within each interval following the same procedure as in steps 5–7 of Algorithm 1.
7. Count the proportion of basic vocabulary shared by each pair of languages and thereby determine the lexical distances.

Appendix C — Lexical Data for Seven Chinese Dialects

Glosses for the Swadesh 100-word list were prepared for each of the following seven Chinese dialects by Prof. Xu Tong-Qiang of the Department of Chinese and Literature at Peking University: Xiang (Xi), Gan (Ga), Wu (Wu), Mandarin (Ma), Hakka (Ha), Min (Mi), and Yue (Yu). As discussed in Section 3.3 in the text, only 15 characters were found to be informative, those listed below. IPA Transcriptions for each character were obtained from (Beijingdaxue 1989, 1995). The symbol ‘·’ refers to the neutral tone; the symbol ‘*’ refers to a character whose etymology is unknown.

Character	Xi	Ga	Wu	Ma	Ha	Mi	Yu
"eat"	吃 tɕ ^h ia 24	吃 tɕ ^h iak 55	吃 tɕ ^h iɿ? 44	吃 tɕ ^h ɿ 55	食 sət 55	食 tsia? 55	吃 ʃɿk 22
"egg"	蛋 tan 31	蛋 t ^h an 31	蛋 dɛ 31	鸡子儿 tɕi 55 tser 214	卵 lɔn 31	卵 nɿŋ 33	蛋 tan 22
"eye"	眼睛 ŋan 31 · tɕin	眼睛 ŋan 214 · tɕiaŋ	眼睛 ŋɛ 31 tsin 44	眼 iɛn 214	目珠 muk 11 tsu 44	目矐 bat 55 tsiu 55	眼睛 ŋan 13
"feather"	毛 mau 33	羽 y 213	羽 jy 31	毛儿 maur 35	羽 i 44	毛 mɔ̃ 55	羽 jy 22
"give"	把 pa 31	把 pa 214	界 pa? 44	给 kei 214	分 pun 44	互 hɔ 33	界 pei 35
"grease"	油 iəu 13	肥 fəi 45	油 jɿy 24	肥 fei 35	肥 p ^h i 44	肥 pui 24	肥 fei 21
"know"	晓得 ɕiau 31 tɕ 24	晓得 ɕieɯ 214 · tet	晓得 ɕiæ 52 tɕ? 44	知道 tɕɿ 55 · tau	知得 ti 44 tet 11	知 tsai 55	知 tɕi 53
"say"	讲 kan 31	话 ua 21	说 sɿ? 44	说 suo 55	讲 kɔŋ 31	讲 kɔŋ 51	讲 kɔŋ 35

Character	Xi	Ga	Wu	Ma	Ha	Mi	Yu
"small"	细 ɕi 35	小 ɕieʊ 214	小 siæ 52	小 ɕiau 214	细 sɛ 52	细 sue 11	细 ʃei 33
"stand"	站 tsan 35	站 tsan 35	立 liiʔ 23	站 tʃan 51	倚 kʰi 44	倚 kʰia 33	倚 kʰei 23
"sun"	太阳 tʰai 35 ian 13	日头 nit 55 • tʰeu	日头 nitʔ 23 dy 24	太阳 tʰai 51 • iaŋ	日头 nit 11 tʰeu 11	日 lit 55	热头 jit 22 tʰeu 21
"swim"	洗冷水澡 ɕi 31 lən 31 ɕeyi 31 tsau 31	玩水 uan 35 sui 214	游水 jiy 24 sɥ 42	鬼水 fu 35 ʃui 214	洒水 siu 11 sui 31	洒水 siu 24 tsui 51	游水 jɛu 21 ʃøy 35
"walk"	走 tsəu 31	走 tseu 214	走 tsy 52	走 tsou 214	行 haŋ 11	行 kiã 24	行 haŋ 21
"what"	么子 mo31 • tsɿ	什里 sət 55 li 35	啥 sɔ 412	什么 ʃen 35 • mə	嘢个 mak 11 kɛ 52	什物 sim 51 mĩʔ 55	嘢 mɛt 55
"who"	哪个 la 31 ko 35	哪个 la 214 • ko	啥人 sɔ 412 nɪn 24	谁 sei 35	瞞*人 man 31 nɪn 11	啥人 sĩa 31 laŋ 24	边*个 pin 55 kɔ 33

SUMMARY

Two computational methods for detecting borrowing among a family of genetically related languages are proposed. One method, based on the detection of branches with negative length in lexicostatistical trees, is shown to work poorly. As we demonstrate, this method is similar to another recently proposed method for detecting borrowing based on skewing in lexicostatistical data. A second method, using character-based classification techniques in common use in the classification of biological taxa, is shown to be more effective. This method allows borrowed characters and the languages among which the borrowing may have taken place to be identified — in some cases, the most likely direction of the borrowing can also be specified.

RÉSUMÉ

Cet article présente deux approches computationnelles pour détecter les emprunts dans une famille de langues génétiquement reliées. Nous mettons en évidence les mauvaises performances de la première méthode, basée sur la détection de branches de longueurs négatives dans les arbres lexicostatistiques. Nous démontrons également qu'elle est similaire à une approche récemment proposée pour détecter les emprunts, et basée sur les biais des données lexicostatistiques. Une seconde approche, qui repose sur les techniques cladistiques couramment utilisées en biologie pour la classification des taxons, se révèle plus efficace. Elle permet d'identifier les caractères empruntés, ainsi que les langues dans lesquels l'emprunt aurait pu avoir lieu — dans certains cas, la direction du changement peut également être spécifiée.

1. ZUSAMMENFASSUNG

Zwei computergestützte Methoden zur Ermittlung von Entlehnungen innerhalb einer Familie genetisch verwandter Sprachen werden vorgeschlagen. Eine Methode, basierend auf der Ermittlung von Zweigen negativer Länge in lexikostatistischen Diagrammen, ist, wie gezeigt wird, ineffektiv. Wir zeigen, dass diese Methode einer anderen, kürzlich vorgeschlagenen ähnelt, welche auf dem Skewing in lexikostatistischen Daten beruht. Eine zweite Methode, die kladistische Techniken nutzt, die gewöhnlich zur Klassifikation biologischer Taxa angewandt werden, ist, so wird gezeigt, effektiver. Diese Methode erlaubt es, entlehnte Merkmale und die Sprache, zwischen denen die Entlehnung stattgefunden haben könnte, zu identifizieren — in einigen Fällen kann auch die Richtung der Entlehnung spezifiziert werden.

REFERENCES

- Beijingdaxue Zhongguoyuyanwenxuexi Yuyanxue Jiaoyanshi. 1989. *Hanyu Fangyan Zihui*, 2nd ed. Beijing: Wenzhi Gaige Chubanshe.
- Beijingdaxue Zhongguoyuyanwenxuexi Yuyanxue Jiaoyanshi. 1995. *Hanyu Fangyan Cihui*, 2nd ed. Beijing: Yuwen Chubanshe.
- Bergsland, Knut & Hans Vogt. 1962. "On the validity of glottochronology". *Current Anthropology* 3.115–53.
- Blust, Robert. 2000. "Why lexicostatistics doesn't work: the 'universal' constant hypothesis and the Austronesian languages". *Time Depth in Historical Linguistics*, Vol. 2 ed. by Colin Renfrew, April McMahon and Larry Trask, 311–31. Cambridge: The McDonald Institute for Archaeological Research.
- Brugmann, Karl. 1884. "Zur Frage nach den Verwandtschaftsverhältnissen der Indogermanischen Sprachen". *Internationale Zeitschrift für allgemeine Sprachwissenschaft* 1.226–56.
- Campbell, Lyle. 1998. *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Cavalli-Sforza, Luigi Luca & William S.-Y. Wang. 1986. "Spatial distance and lexical replacement". *Language* 62.38–55.
- Cavalli-Sforza, Luigi Luca, Paolo Menozzi & Alberto Piazza. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.
- DeGroot, Morris H. 1986. *Probability and Statistics*. 2nd ed. Reading, MA: Addison-Wesley.
- Durie, Mark & Malcolm Ross, eds. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. New York: Oxford University Press.
- Dyen, Isidore, Joseph B. Kruskal & Paul Black. 1992. "An Indoeuropean Classification: A Lexicostatistical Experiment". *Transactions of the American Philosophical Society*

ety 82:5.

- Embleton, Sheila. 1981. *Incorporating Borrowing Rates in Lexicostatistical Tree Reconstruction*. Unpublished Ph.D. thesis, Department of Linguistics, University of Toronto.
- Embleton, Sheila. 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.
- Evans, Merran, Nicholas Hastings & Brian Peacock. 1993. *Statistical Distributions*. 2nd ed. New York: John Wiley.
- Farris, James S. 1972. "Estimating phylogenetic trees from distance matrices". *American Naturalist* 106.645–68.
- Felsenstein, Joseph. 1982. "Numerical methods for inferring evolutionary trees". *The Quarterly Review of Biology* 57:4.379–404.
- Fishman, George S. 1996. *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer-Verlag.
- Fitch, Walter M. and Emanuel Margoliash. 1967. "Construction of phylogenetic trees". *Science* 155.279–84.
- Fitch, Walter M. 1971. "Toward defining the course of evolution: minimum change for a specific tree topology". *Systematic Zoology* 20:4.406–16.
- Greenberg, Joseph H. 1987. *Language in the Americas*. Stanford: Stanford University Press.
- Heine, Bernd. 1974. "Historical linguistics and lexicostatistics in Africa". *Journal of African Linguistics* 11:3.7–20.
- Hennig, Willi. 1950. *Grundzüge einer Theorie der phylogenetischen Systematik*. Berlin: Deutsche Zentralverlag.
- Hinnebusch, Thomas J. 1996. "Skewing in lexicostatistic tables as an indicator of contact". Paper presented at the *Round Table on Bantu Historical Linguistics*, Université Lumière 2, Lyon, France, May 30–June 1, 1996.
- Janson, T. 1977. "Reversed lexical diffusion and lexical split: loss of -d in Stockholm".

- The Lexicon in Phonological Change* ed. by William S.-Y. Wang. 252–266. The Hague: Mouton.
- Kitching, Ian J., Peter L. Forey, Christopher J. Humphries & David M. Williams. 1998. *Cladistics: The Theory and Practice of Parsimony Analysis*. 2nd ed. New York: Oxford University Press.
- Kluge, Arnold G. & James S. Farris. 1969. “Quantitative phyletics and the evolution of anurans”. *Systematic Zoology* 18.1–32.
- Lees, Robert B. 1953. “The basis of glottochronology”. *Language* 29.113–27.
- Nurse, Derek. 1979. *Classification of the Chaga Dialects*. Hamburg: Helmut Berske.
- Pagel, Mark. 2000. “Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies”. *Time Depth in Historical Linguistics*, Vol. 1 ed. by Colin Renfrew, April McMahon and Larry Trask, 189–207. Cambridge: The McDonald Institute for Archaeological Research.
- Qiao, Sanzheng & William S.-Y. Wang. 1998. “Evaluating phylogenetic trees by matrix decomposition”. *Anthropological Science* 106:1.1–22.
- Ross, Malcolm D. 1996. “Contact-induced change and the comparative method: cases from Papua New Guinea”. *The Comparative Method Reviewed* ed. by Mark Durie and Malcolm Ross, 180–217. New York: Oxford University Press.
- Saitou, Naruya & Masatoshi Nei. 1987. “The neighbor-joining method: a new method for reconstructing phylogenetic trees”. *Molecular and Biological Evolution* 4:4.406–25.
- Swadesh, Morris. 1950. “Salish internal relationships”. *International Journal of American Linguistics* 16.157–67.
- Swadesh, Morris. 1951. “Diffusional cumulation and archaic residue as historical explanations”. *Southwestern Journal of Anthropology* 7.339–46.
- Swadesh, Morris. 1952. “Lexico-statistic dating of prehistoric ethnic contacts”. *Proceedings of the American Philosophical Society* 96.452–63.

- Swadesh, Morris. 1955. "Towards greater accuracy in lexico-statistic dating". *International Journal of American linguistics* 18.121–37.
- Swofford, David L. 1991. "When are phylogeny estimates from molecular and morphological data incongruent?". *Phylogenetic Analysis of DNA Sequences* ed. by Michael M. Miyamoto and Joel Cracraft, 295–333. Oxford: Oxford University Press.
- The MathWorks, Inc. 2000. *Matlab* computer software.
- Trask, Robert L. 1996. *Historical linguistics*. London: Arnold.
- Wagner, Warren H. Jr. 1961. "Problems in the classification of ferns". *Recent Advances in Botany* 1.841–4.
- Wang, William S.-Y. 1989. "The migration of the Chinese people and the settlement of Taiwan". *Anthropological Studies of the Taiwan Area: Accomplishments and Prospects* ed. by Kwang-chih Chang, Kuang-chou Li, Arthur P. Wolf and Alexander Chien-chung Yin. Taipei: Department of Anthropology, National Taiwan University.
- Warnow, Tandy. 1997. "Mathematical approaches to comparative linguistics". *Proceedings of the National Academy of Science* 94.6585–90.
- Warnow, Tandy, Donald Ringe & Ann Taylor. 1995. "Reconstructing the evolutionary history of natural languages". Paper presented at the *Workshop on Historical Linguistics*, University of Pennsylvania, 1995.
- Xu, Tong-Qiang. 1991. "Lishi yuyanxue". *Shangwu Yinshuguan*.
- You, Yu-Jie. 2000. *Hanyu Fangyanxue*. Shanghai: Shanghai Jiaoyu Chubanshe.

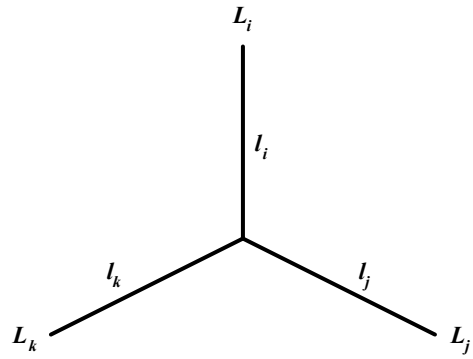


Fig. 1. The unique unrooted lexicostatistical tree for three languages, L_i , L_j and L_k .
Also marked on the tree are the branch lengths, l_i , l_j and l_k .

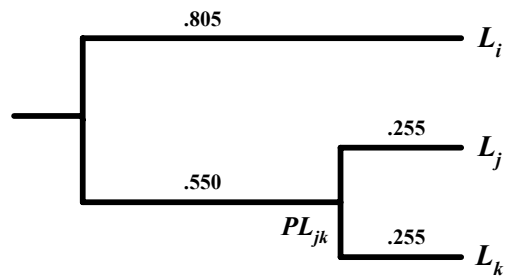


Fig. 2. The lexicostatistical tree for languages L_i , L_j and L_k *before* contact between L_i and L_j . This and all subsequent rooted trees displayed in Section 2 have been rooted using midpoint rooting (Farris 1972).

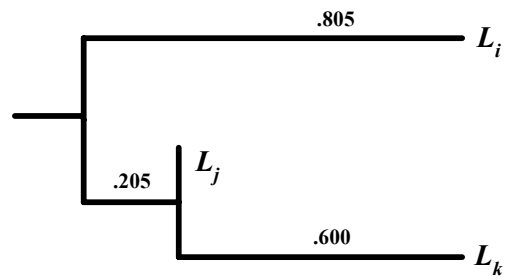


Fig. 3. The lexicostatistical tree for languages L_i , L_j and L_k *after* contact between L_i and L_j . The length of the branch connecting L_j to the body of the tree is negative (-0.090).

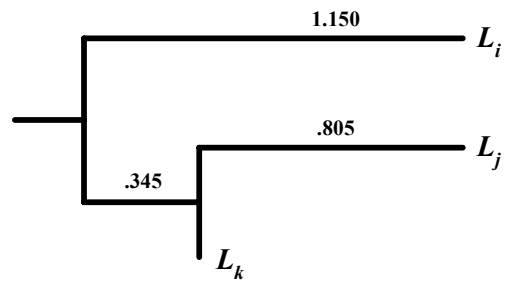


Fig. 4. The lexicostatistical tree for languages L_i , L_j and L_k — no contact. The length of the branch connecting L_k to the body of the tree is negative (-0.295).

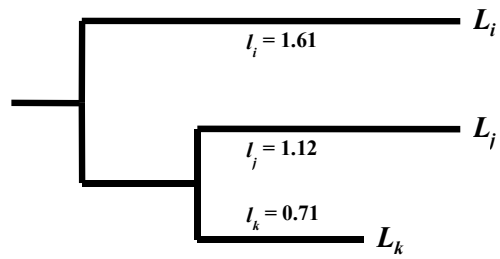


Fig. 5. The lexicostatistical tree and data generated by one typical run of Algorithm 1.
 ($N = 100$, $r_i = r_j = r_k = r_{jk} = 80\%$, $t = 5$ millennia)

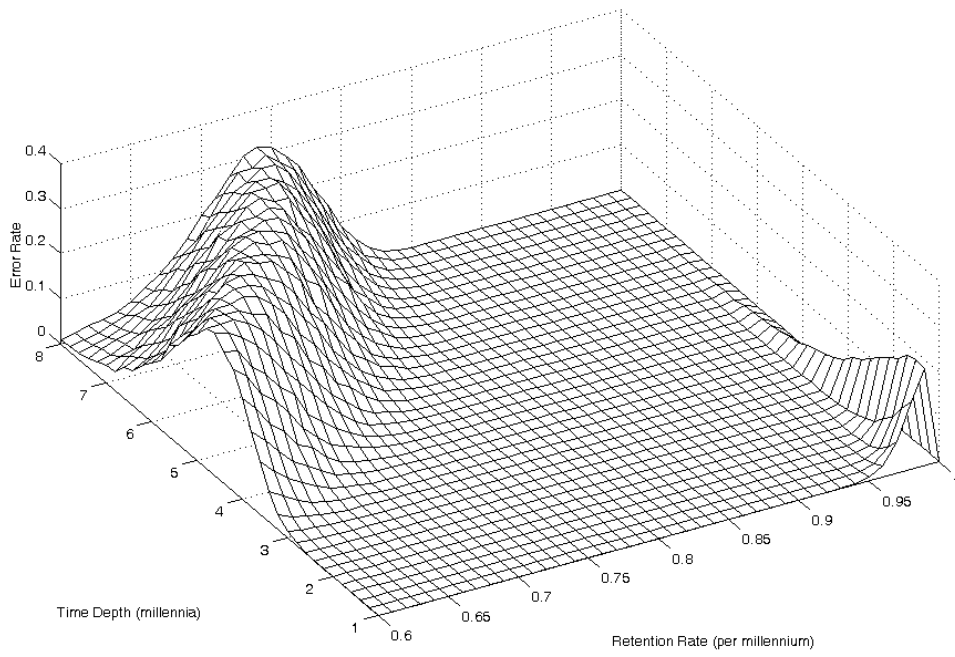


Fig. 6. Estimated false alarm rate for 100 meanings (5000 runs).

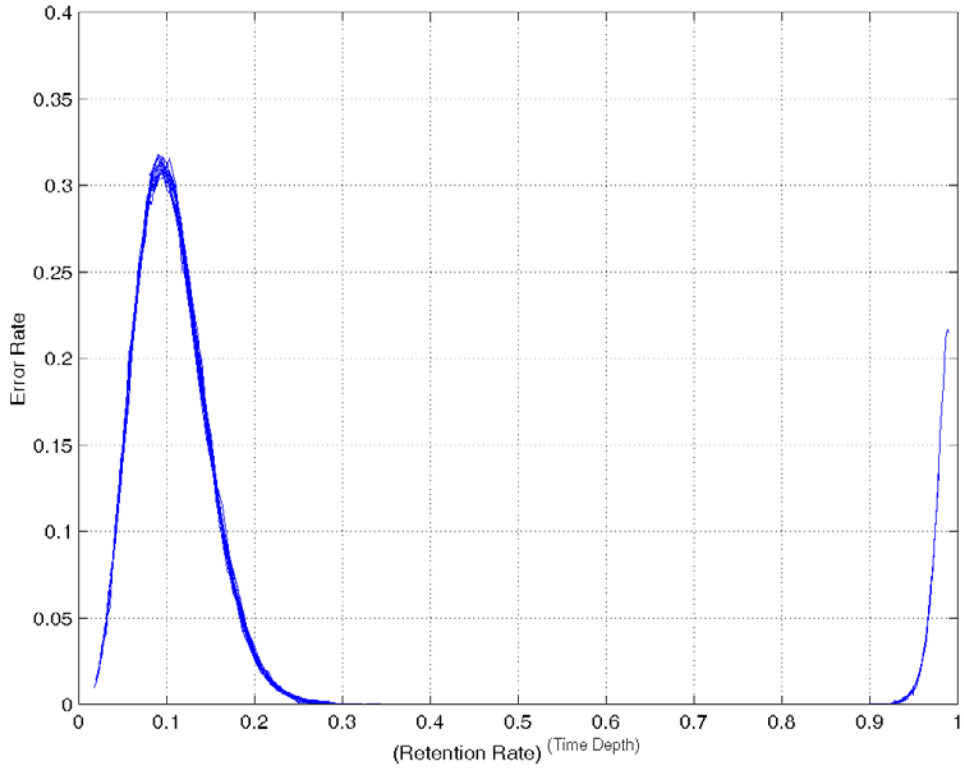


Fig. 7. Estimated false alarm rate for 100 meanings as a function of r^t .

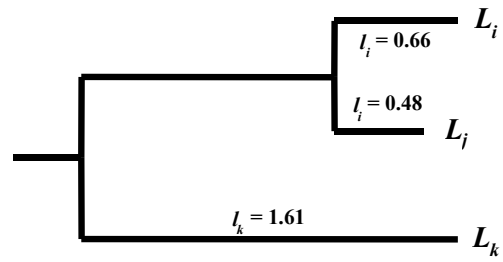


Fig. 8. The lexicostatistical tree and data generated by one typical run of Algorithm 2.

($N = 100$, $r_i = r_j = r_k = r_{jk} = 80\%$, $t = 5$ millennia)

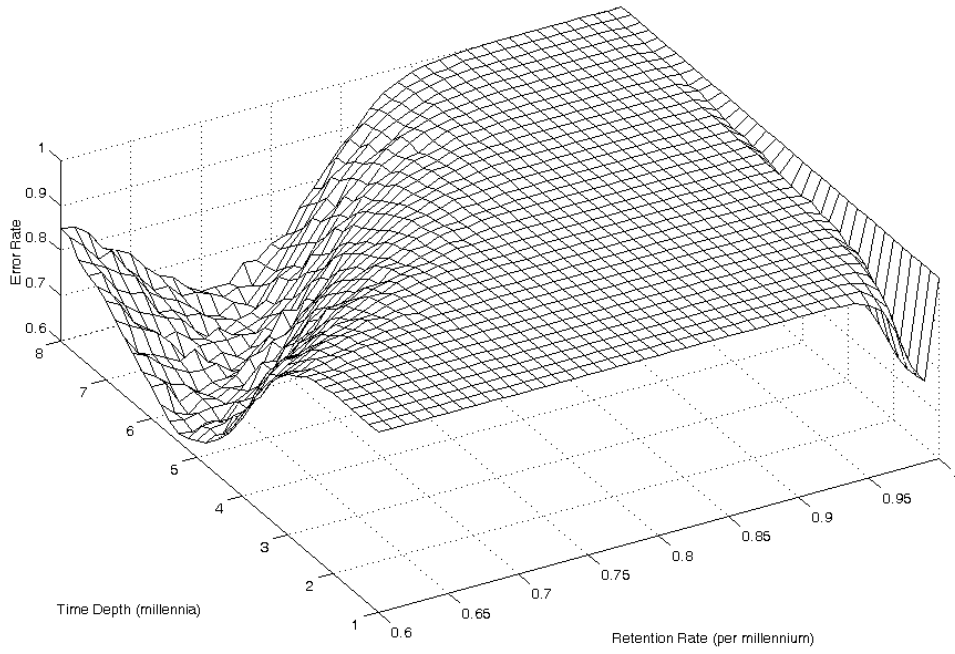


Fig. 9. Estimated miss rate for 100 meanings — 20% borrowing (2000 runs).

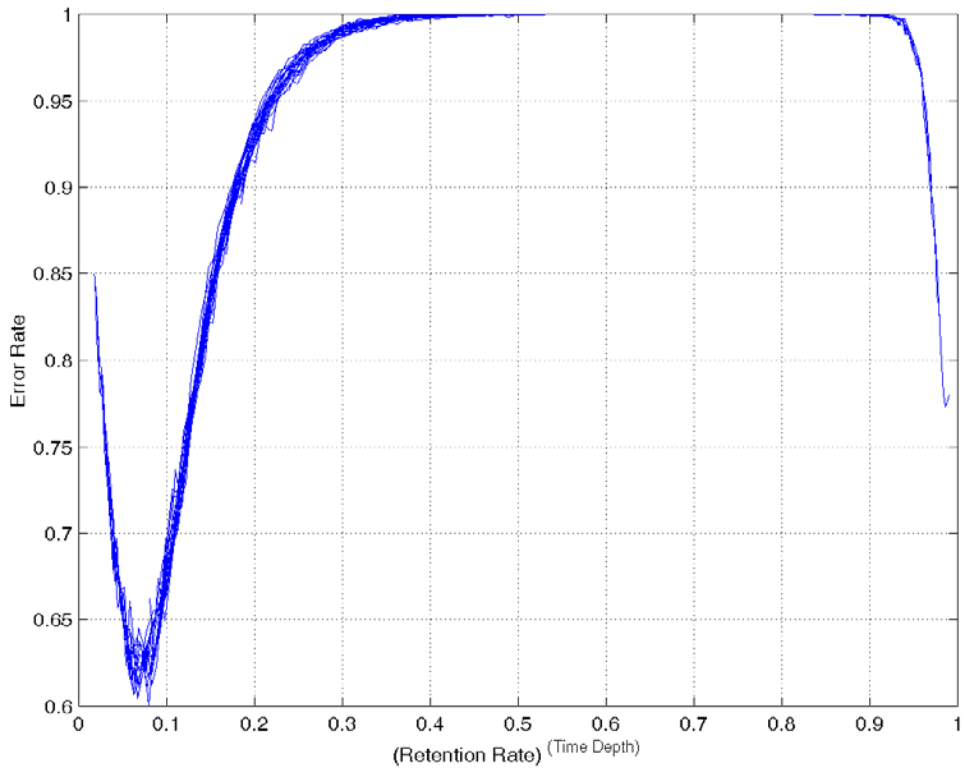


Fig. 10. Estimated miss rate for 100 meanings as a function of r^j — 20% borrowing.

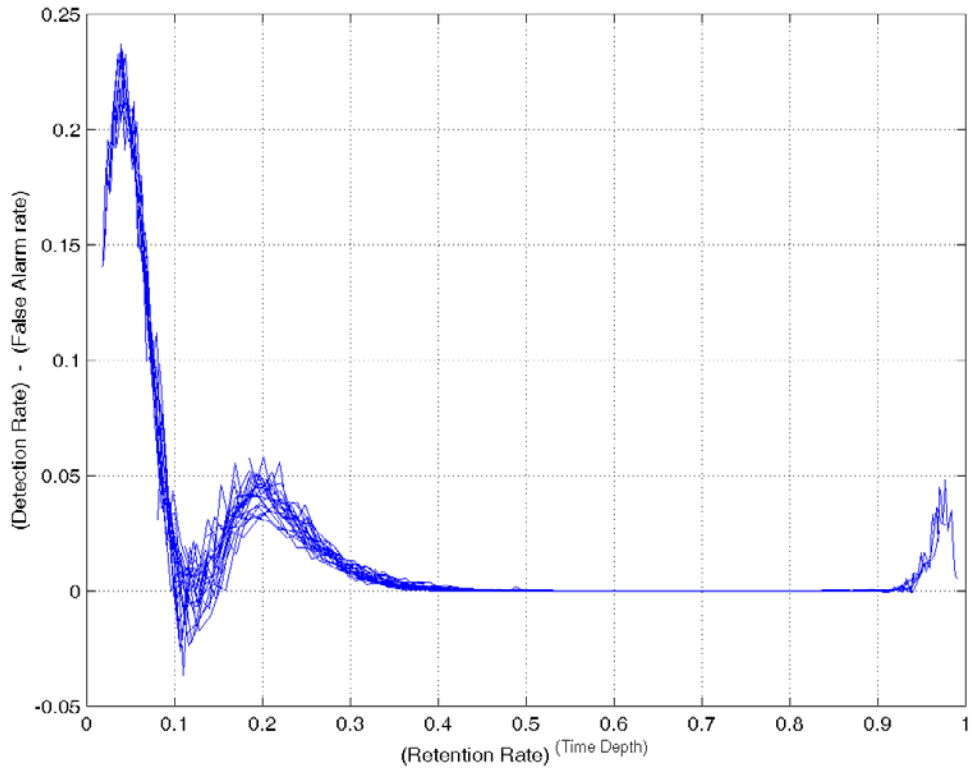


Fig. 11. Detection rate minus false alarm rate for 100 meanings as a function of r^t
— 20% borrowing.

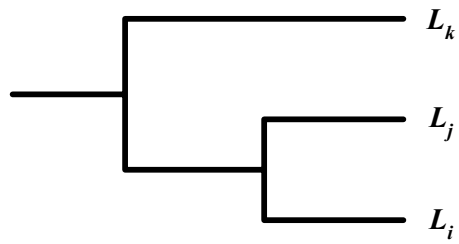


Fig. 12. The genetic classification of three languages, L_i , L_j and L_k .

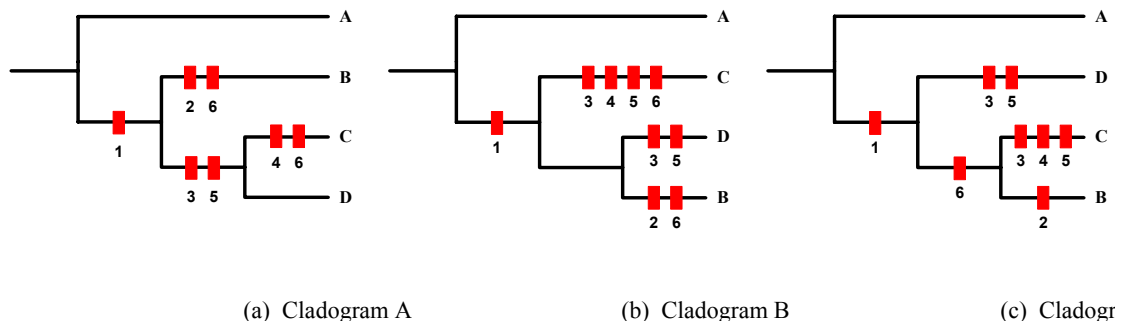


Fig. 13. Cladograms for three putative classifications of the taxa in Table 9, after Kitching et al. (1998).

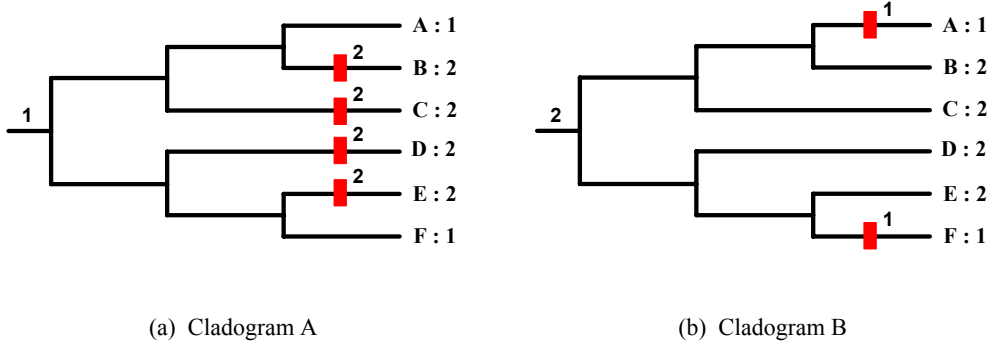


Fig. 14. Two cladograms with the same topology and terminal character states, but different lengths.

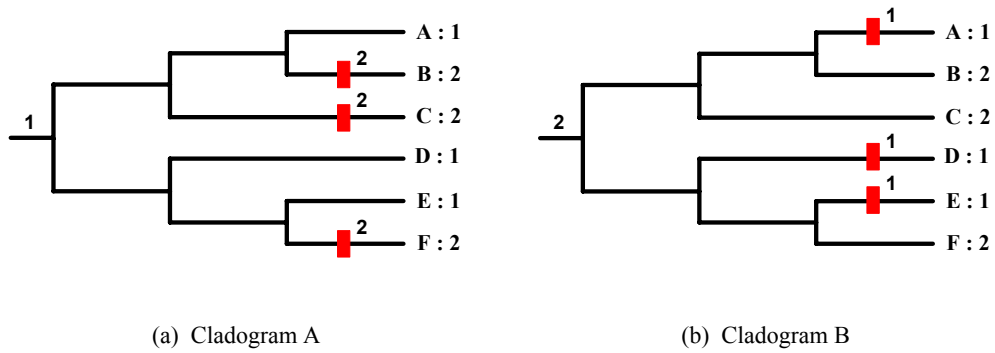


Fig. 15. Two cladograms with the same topology, terminal character states and length.

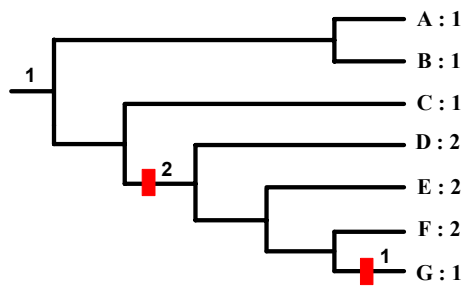


Fig. 16. A cladogram for which the most parsimonious assignment of direction of borrowing can be inferred.

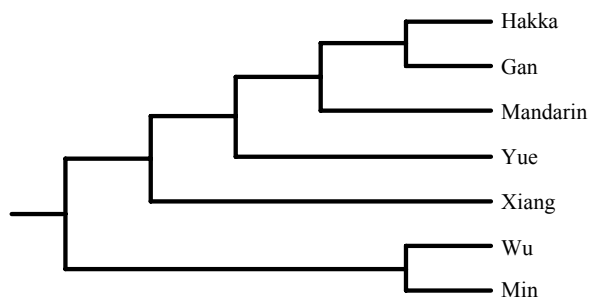


Fig. 17. A conventional classification of seven Chinese dialects, after You (2000).

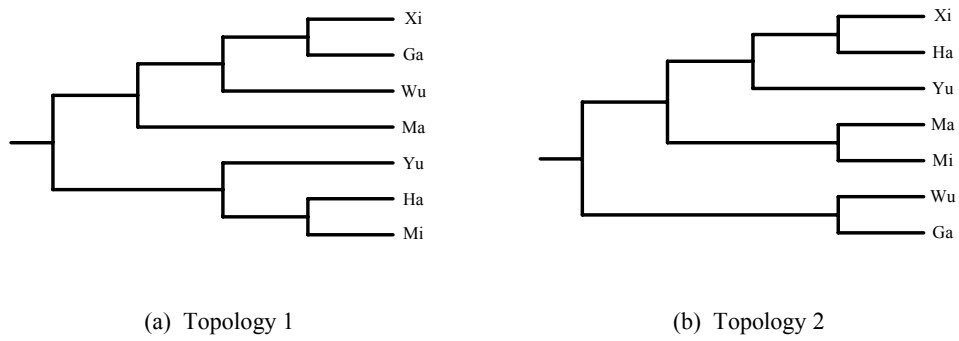


Fig. 18. Two possible topologies for seven Chinese dialects.

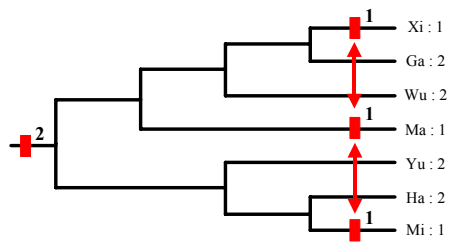


Fig. 19. The most parsimonious assignment of state changes to the character "feather" on Topology 1.

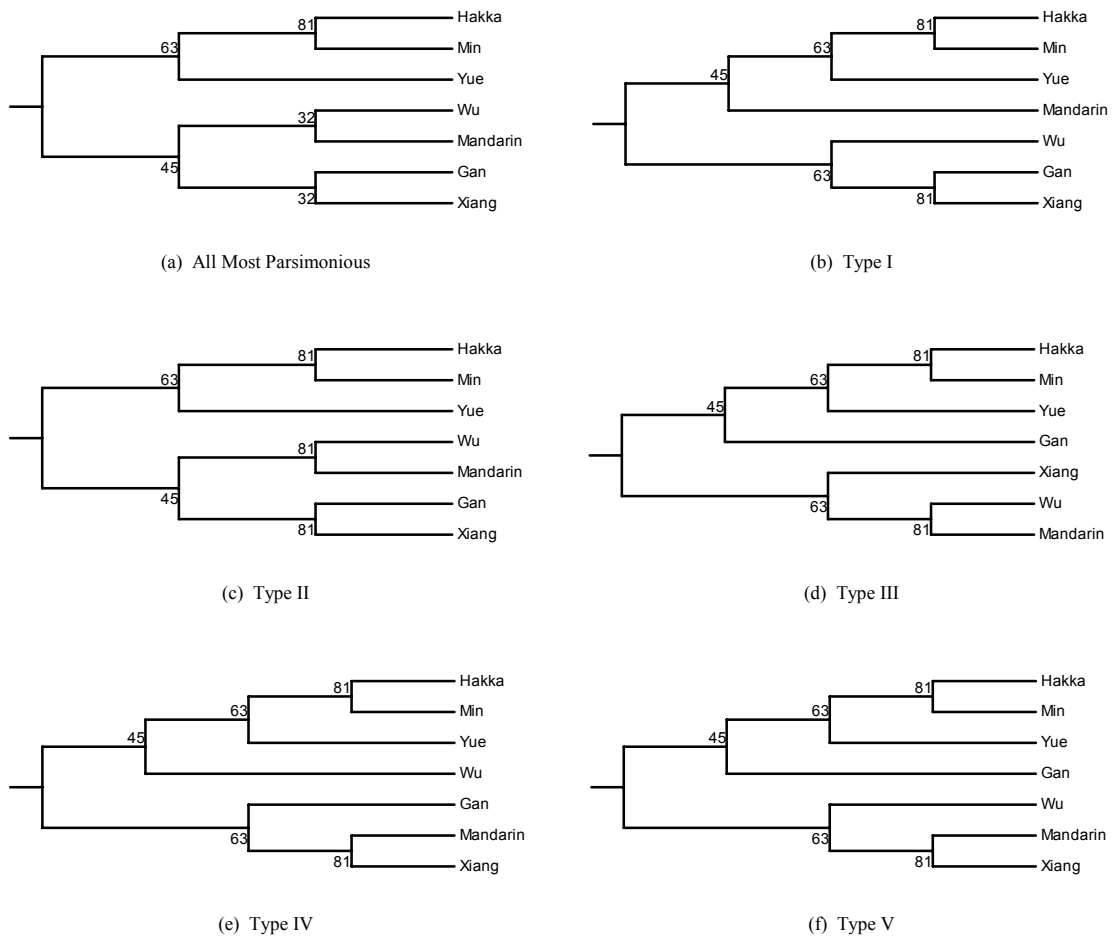


Fig.20. Consensus trees for (a) the 55 most parsimonious topologies, and (b–f) the topologies of Type I–V — each topology requires a minimum of seven character state changes. (At each node, the number indicates the percentage of topologies that are consistent with the consensus tree at that node.)

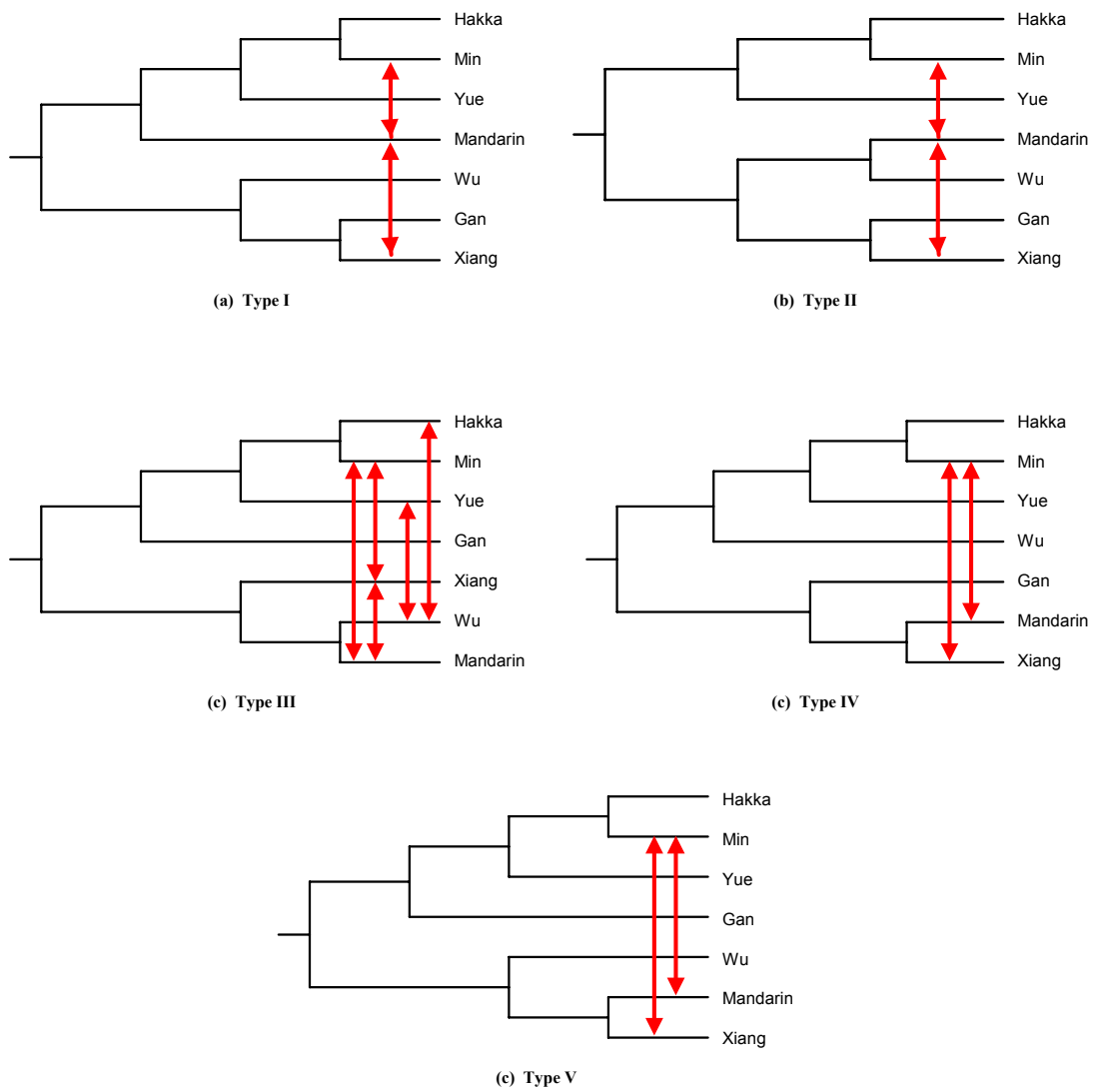


Figure 21. All most parsimonious assignments of borrowing of the character "feather" on the consensus trees for topologies of Type I-V.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
L_i	0	0	2	2	2	2	2	2	0	0
L_j	0	0	1	1	0	0	1	1	1	1
L_k	0	0	0	0	1	1	1	1	1	1

Table 1. The character states of ten meanings for languages L_i , L_j and L_k before contact between L_i and L_j .

	L_i	L_j	L_k
L_i	—	20%	20%
L_j	1.61	—	60%
L_k	1.61	0.51	—

Table 2. Pairwise lexical similarities (above leading diagonal) and lexical distances (below leading diagonal) for languages L_i , L_j and L_k *before* contact between L_i and L_j .

	L_i	L_j	L_k
L_i	—	40%	20%
L_j	0.92	—	60%
L_k	1.61	0.51	—

Table 3. Pairwise lexical similarities (above leading diagonal) and lexical distances (below leading diagonal) for languages L_i , L_j and L_k *after* contact between L_i and L_j .

	l_i	l_j	l_k
before	1.355	0.255	0.255
after	1.010	-0.090	0.600

Table 4. Branch lengths of the lexicostatistical tree for languages L_i , L_j and L_k both *before* and *after* contact between L_i and L_j .

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
L_i	0	0	0	2	2	2	2	2	0	0
L_j	1	1	0	0	0	0	0	1	1	1
L_k	0	0	0	0	0	1	1	1	1	1

Table 5. The character states of ten meanings for languages L_i , L_j and L_k — no contact.

	L_i	L_j	L_k
L_i	—	10%	30%
L_j	2.30	—	60%
L_k	1.20	0.51	—

Table 6. Pairwise lexical similarities (above leading diagonal) and lexical distances (below leading diagonal) for languages L_i , L_j and L_k — no contact.

l_i	l_j	l_k
1.495	0.805	-0.295

Table 7. Branch lengths of the lexicostatistical tree for languages L_i , L_j and L_k — no contact.

	M_1	M_2	M_{99}	M_{100}
L_i	0	0	...	0
L_j	0	0	...	1
L_k	1	0	...	1

Table 8. The character states of 100 meanings for three hypothetical, closely genetically related languages L_i , L_j and L_k .

Taxa	C_1	C_2	C_3	C_4	C_5	C_6
A	0	0	0	0	0	0
B	1	1	0	0	0	1
C	1	0	1	1	1	1
D	1	0	1	0	1	0

Table 9. The character states for four hypothetical taxa having 6 binary characters, $C_1 \dots C_6$, after Kitching et al. (1998).

Character	Xi	Ga	Wu	Ma	Ha	Mi	Yu
"eat"	1	1	1	1	2	2	2
"egg"	1	1	1	3	2	2	1
"eye"	1	1	1	1	2	2	1
"feather"	1	2	2	1	2	1	2
"give"	1	1	2	3	4	5	2
"grease"	1	2	1	3	2	2	2
"know"	1	1	1	2	2	2	2
"say"	1	3	2	2	1	1	1
"small"	1	2	2	2	1	1	1
"stand"	1	1	3	1	2	2	2
"sun"	1	2	1	1	2	2	3
"swim"	3	5	1	4	2	2	1
"walk"	1	1	1	1	2	2	1
"what"	3	1	4	1	2	1	2
"who"	1	1	2	3	5	2	4

Table 10. The character states of 15 informative characters for seven Chinese dialects: Xiang (Xi), Gan (Ga), Wu (Wu), Mandarin (Ma), Hakka (Ha), Min (Mi), and Yue (Yu).

Type:	Number of Topologies:	Borrowings:
I	11	"feather" (×2), "grease", "say", "small", "sun", "what"
II	11	"feather" (×2), "grease", "know", "small", "sun", "what"
III	11	"feather" (×2), "give", "know", "small", "what", "who"
IV	11	"feather", "grease", "know", "say", "small", "sun", "what"
V	11	"feather", "give", "know", "say", "small", "what", "who"
—	55	"feather" (55), "give" (22), "grease" (33), "know" (44), "say" (33), "small" (55), "sun" (33), "what" (55), "who" (22)

Table 11. Borrowings indicated among the 55 most parsimonious topologies classified into 5 types — seven instances of borrowing are indicated for each topology.

Character	Population	Parsimonious	Probability
"feather"	9900	55	0.932 *
"give"	6930	22	0.000
"grease"	8910	33	0.000
"know"	9900	44	0.000
"say"	8910	33	0.000
"small"	9900	55	0.932 *
"sun"	9504	33	0.000
"what"	8316	55	1.000 *
"who"	6930	22	0.000

Table 12. Number of topologies for which borrowing of a particular character is indicated:
a. among all possible topologies (Population); and
b. among the 55 most parsimonious topologies (Parsimonious).

Also listed is the probability that a random sample of 55 topologies would indicate *fewer than* the number of instances of borrowing observed among the 55 most parsimonious topologies.

(Asterisks (*) indicate statistically significant results.)

Character	Implied Borrowing	Types
"feather"	Mandarin → Xiang AND Mandarin → Min	I, II, III
	Hakka/Yue ↔ Wu AND Mandarin/Xiang → Min	III
	Xiang/Mandarin → Min	IV, V
"small"	Yue/Hakka/Min → Xiang	I, II, III, IV, V
	Wu/Mandarin ↔ Gan	II, III
"what"	Mandarin/Gan → Min	I, II, III, IV, V
	Hakka ↔ Yue	I, II, III, IV, V

Table 13. Summary of the most parsimonious assignments of borrowing of characters "feather", "small" and "what" among the seven Chinese dialects.

For each character, the multiple rows indicate alternative borrowing hypotheses.